# Bovine Genomics

# Bovine Genomics

Edited by

James E. Womack

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

This edition first published 2012 © 2012 by John Wiley & Sons, Inc.

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

*Editorial offices:* 2121 State Avenue, Ames, Iowa 50014-8300, USA
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK
9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

*Library of Congress Cataloging-in-Publication Data*

Bovine genomics / edited by James E. Womack.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-8138-2122-1 (hardcover : alk. paper
1. Cattle–Genome mapping. 2. Cattle–Genetics. 3. Veterinary genetics. I. Womack, James E.
SF756.65B68 2012
636.2′0821–dc23 2011049351

A catalogue record for this book is available from the British Library.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Set in 10/12 pt Dutch801BT by Aptara® Inc., New Delhi, India

**Disclaimer**
The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Website is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Website may provide or recommendations it may make. Further, readers should be aware that Internet Websites listed in this work may have changed or disappeared between when this work was written and when it is read.

1 2012

Cover photo: L1 Dominette 01449 with calf at USDA ARS, Miles City, Montana. Photograph courtesy of Dr. Mike MacNeil, Research Geneticist. Reprinted with permission.

# Contents

# List of Contributors

**David L. Adelson**
School of Molecular and Biomedical Science
The University of Adelaide
Adelaide, SA 5005, Australia

**Daniel G. Bradley**
Smurfit Institute of Genetics
Trinity College Dublin
Ireland

**Jared E. Decker**
Division of Animal Sciences
University of Missouri
Columbia, MO 65211-5300, USA

**Michel Georges**
Unit of Animal Genomics
GIGA-R & Faculty of Veterinary Medicine
University of Liège
Liège, Wallonia, Belgium

**Richard A. Gibbs**
Department of Molecular and Human Genetics
Baylor College of Medicine
Houston, TX 77030, USA

**M. E. Goddard**
Department of Agriculture and Food Systems
University of Melbourne
Melbourne, Victoria, Australia

**B. J. Hayes**
Biosciences Research Division, Department of Primary Industries Victoria
University of Melbourne
Melbourne, Victoria, Australia

**Denis M. Larkin**
Institute of Biological, Environmental and Rural Sciences
Aberystwyth University
Aberystwyth, SY23 3DA, UK

**Wansheng Liu**
Department of Dairy and Animal Science
Pennsylvania State University
University Park, PA 16802, USA

**Stephanie D. McKay**
Division of Animal Sciences
University of Missouri
Columbia, MO 65211-5300, USA

**Frank W. Nicholas**
Faculty of Veterinary Science
University of Sydney
NSW 2006
Australia

**F. Abel Ponce de León**
Department of Animal Science
University of Minnesota
St. Paul, MN 55108, USA

**Holly R. Ramey**
Division of Animal Sciences
University of Missouri
Columbia, MO 65211-5300, USA

**Megan M. Rolf**
Division of Animal Sciences
University of Missouri
Columbia, MO 65211-5300, USA

**Sheila M. Schmutz**
Department of Animal and Poultry Science
University of Saskatchewan
Saskatoon, SK S7N 5A8, Canada

**Robert D. Schnabel**
Division of Animal Sciences
University of Missouri
Columbia, MO 65211-5300, USA

**Morris Soller**
Department of Genetics
The Hebrew University of Jerusalem
91904 Jerusalem, Israel

**Jeremy F. Taylor**
Division of Animal Sciences
University of Missouri
Columbia, MO 65211-5300, USA

**Matthew D. Teasdale**
Smurfit Institute of Genetics
Trinity College Dublin
Ireland

**Xiuchun (Cindy) Tian**
Department of Animal Science, Center for Regenerative Biology
University of Connecticut
Storrs, CT 06269-4163, USA

**Joel I. Weller**
Institute of Animal Sciences
ARO, The Volcani Center
Bet Dagan 50250, Israel

**James E. Womack**
Department of Veterinary Pathobiology
Texas A&M University
College Station, TX 77843-4467, USA

**Kim C. Worley**
Department of Molecular and Human Genetics
Baylor College of Medicine
Houston, TX 77030, USA

# Foreword

Research in cattle genetics was profoundly changed in 2009 with public release of the cattle genome sequence and the publication of papers describing its content, function, and evolution. Since the development of mixed model equations by Henderson in the early 1950s, there has been no other event that has had similar impact on bovine biology and the science of dairy and beef cattle breeding. In ways foretold by the contemporary leaders in the field, of whom the editor and the authors are a part, genomics has been adopted as a foundational tool for the genetic improvement of cattle and other livestock species. One of the most gratifying stories to emerge from cattle genomics is the way that traditional animal breeding has been integrated with the new technologies and adopted by the industry, despite the doubts of many good friends along the way. Indeed, this was the vision of the earliest pioneers in animal genetics, such as Fred Hutt and Clyde Stormont, who grasped the verdant potential for genetic markers in animal breeding. We can only wonder what they would say now if they were alive to read this marvelous book!

*Bovine Genomics* begins with Matthew Teasdale's and Dan Bradley's updated review on the origins of domesticated cattle, providing current information on the timing of probable domestication events and an excellent summary of the archeological evidence as well as data from mitochondrial phylogeography and nuclear DNA that support the current consensus. Although the authors avoid the issue of taxonomic classification of indicine and taurine cattle (an ongoing source of confusion to students and practitioners alike!), they leave little doubt that modern cattle are the product of two or possibly more domestication events. At least for now, the Aurochsen appear as the forbearers of all cattle, but they have leapt into domesticated cattle lineages at several points in history.

Following two well-referenced reviews on Mendelian traits by Frank Nicholas and Sheila Schmutz, respectively, the reader is treated to historical perspectives from the "foundation sires" of cattle genetics, Morris Soller and Jim Womack. Coming from entirely different scientific and geographical worlds, these giants in the field provided the scientific rationale that was eventually used for sequencing the cattle genome. Soller gives a fascinating personal history of his transformation from a curious adolescent with a fascination for Morgan, to radical quantitative geneticist who envisioned and developed a theoretical framework for marker-assisted selection. This article is a must read for any serious student of animal genetics. The accounting is honest, detailed, accurate (from my perspective) and captures most of the important people and events leading up to modern, genomically driven animal breeding methods. It is incredible that Soller had it all figured out even before most of the current leaders in the field finished high school!

Next up is Womack's review of cattle gene mapping. It is hard for me to write dispassionately about Womack's contribution to the field, given that he has been a mentor and great collaborator for more nearly 20 years. Even though Womack has reviewed this subject in recent years, this time he has come up with some real gems! His quote of Frank Ruddle's response to the question "why map genes" brought a huge “LOL” (“gene mapping is good for you!”). Having mapped thousands of genes with Womack’s radiation hybrid panel, I often used that same line in convincing my own students and postdocs to persevere. As Womack details throughout the article, each gene becomes a landmark on which more detailed maps are built, similar to Jan Klein’s analogy between postage stamp collecting and MHC alleles (each very beautiful but with no immediately obvious value) that exploded in number after Peter Gorer’s famous discovery of the mouse H-2 complex. And, as the story turned out, the high-resolution maps that were produced using radiation hybrid analysis proved to be the critical scaffold reagent for a chromosome-based assembly of the draft cattle genome sequence. The reader of this review will come away with a true historical sense of how discoveries made in apparently disparate areas of science can have a transformative effect on other disciplines. Fortunately, Womack envisioned what was possible with the tools of somatic cell genetics, and this carried the field of animal genetics for an entire generation.

Womack’s article sets the stage nicely for the comprehensive review of linkage mapping by Stephanie McKay and Bob Schnabel, which is followed by a current view of the much maligned bovine sex chromosomes by Abel Ponce DeLeon and Wansheng Liu. I shall show my bias by commenting on the article by Denis Larkin, who summarized much of the work he conducted in my laboratory during the past 10 years on the subject of cattle comparative genomics and genome evolution. Larkin presents an expert technical review, and also gives us important insights into the relatively controversial interpretation that certain genome rearrangements in mammals may be adaptive. Although this idea has been floating around among evolutionary theorists and evolutionary biologists for more than half a century, there is now strong support for adaptive chromosome rearrangements gathering from work with yeast, plants, insects, and mammals. Larkin leaves us with strong anticipation that much will be learned from the multispecies comparisons of chromosome organization that will follow from the sequencing of thousands of species in the coming years.

The centerpiece of this volume is the review by Worley and Gibbs on the sequencing of “the” bovine genome. The community owes its gratitude to the Baylor group for providing the field with the critical resource on which the “new” cattle genetics is being built. Several excellent reviews follow on subjects ranging from genome architecture (Dave Adelson) to epigenetics (Cindy Tian), QTL mapping (Joel Weller), genome-wide association studies and linkage disequilibrium (Michael Goddard and Ben Hayes), and genomic selection in beef cattle (Jerry Taylor et al.). Brevity dictates that I restrain from commenting in detail on these articles, but readers will find that they match the stellar reputations of their authors.

This brings us to the final chapter by Michel Georges, one of the genetics community’s truly innovative scientists. The author critically reviews many areas of importance in cattle genetics, providing strong views on the candidate gene approach for single gene defects and QTL and on marker-assisted selection. With the hindsight gained from years of experience, and the foresight of a gifted scientist, Georges leaves no doubt concerning his enthusiasm for new genomic technologies for mutation sleuthing, and backs it up with several examples from his group’s work. There is much more in this excellent review for the reader to enjoy, but moreover, for the community to take as a bellwether of where cattle genetics and complex traits analysis will be going over the next few years.

While there are a few more topics that could have been covered, this edition of *Bovine Genomics* is a timestamp that marks the most dynamic period in the history of cattle genetics. The new resources for doing science with what was previously a very difficult animal to understand at the molecular and systems level has brought many talented young investigators to the field. I suspect that the next edition will show how much value the public investments in cattle research have brought to our understanding of biology in general, and to applications in animal agriculture at a critical time when demand for animal products is skyrocketing on a global scale. Keep your grill hot and your laptop warm!

Harris A. Lewin
Vice Chancellor for Research
Professor of Evolution and Ecology
Robert and Rosabel Osborne Endowed Chair
University of California, Davis
January, 2012

# Chapter 1
# The Origins of Cattle

*Matthew D. Teasdale and Daniel G. Bradley*

## Archeology and Domestication

The transformation of early human economies from nomadic hunter–gatherers to farmers is a pivotal moment in human evolution. Starting approximately 12,000 years ago, this process is entitled the Neolithic Revolution and encompassed the domestication of a variety of plants and animals (Bar-Yosef 1998). The archeological study of domestication requires a combination of classical and molecular approaches, which include the analysis of settlement patterns, food residues, and human, animal, and plant remains.

Settlement patterns provide an excellent source of evidence for the beginnings of domestication. Firstly, they provide direct evidence of a sedentary lifestyle, which is likely a prerequisite to Neolithisation. Secondly, the production of long-term housing requires specialist builders; these skills could likely only be supported if an agricultural economy was being practiced to offset the loss of labor from hunting. The evolution of particular building technologies within Neolithic core regions can also be informative, for example, the presence of grain stores and larger houses emerge as the Neolithic lifestyle develops (Cauvin 2000). Study of the surrounding areas can provide evidence for early attempts at domestication, for example, the manipulation of the landscape to control animal migration (Vigne 2011).

The analysis of organic residues found on cooking and storage artifacts is a relatively new technique in molecular archeology, which is providing exciting results especially in the field of domestication (Dudd et al. 1999; Copley et al. 2003; Copley et al. 2005; Outram et al. 2009). For example, lipid residues found on pottery can be used to deduce milk use and have allowed for the earliest date of specialized milking to be proposed as the seventh millennium BC (Evershed et al. 2008). Molecular archeology also allows the diet of early farmers to be inferred from the stable isotopes contained within their bones, analyses that have been fruitful in distinguishing the transition into farming (Richards et al. 2003; Liden et al. 2004; Eriksson et al. 2008).

The study of animal remains, however, is still the principal analysis for identifying domestication (Vigne 2011), with the differences in morphology of domesticates compared to their wild progenitor providing clues to this process. Cattle follow the general trend of domestic breeds being smaller than their wild relatives. However, the

*Bovine Genomics*, First Edition. Edited by James E. Womack.

usefulness of this factor alone to identify early signatures of domestication has recently been called into question (Zeder 2008, and references therein). More robust evidence for the beginnings of domestication may be found in the kill-off patterns of animals (age at which animals are killed) (Vigne and Helmer 2007). (Most hunters tend to target adult males to maximize the kill. In contrast, herders are thought to slaughter males young, apart from the few needed for herd propagation) (Zeder 2008). This leads to archeological remains dominated by young males and elderly females who are killed once they have passed their prime reproductive age (Vigne and Helmer 2007; Zeder 2008). The number of domesticate finds also increases through time at the proposed Neolithic sites, which allows for the time of domestication to be proposed (Bar-Yosef 1998; Vigne 2011).

## Bovine Mitochondrial DNA Diversity and Cattle Origins

The genetic description of a primary division within the genomes of domestic cattle, reflecting the difference between *Bos indicus* and *Bos taurus*, is almost 20 years old. However, the observation of morphological, behavioral, and physiological differences between the two taxa is an older one. In fact, Darwin (an ardent student of domestication), in *The Origin*, speculated that zebu had different domestic wild progenitors from observations on "the habits, voice and constitution etc of the humped Indian cattle," communicated to him by his correspondent from the subcontinent, Mr. Blyth.

Earlier studies of the bovine mitochondrial genome used both restriction fragment length polymorphism and limited control region sequencing and described two divergent clusters of sequences with limited diversity within each (Loftus et al. 1994). Any calibration of the difference between these clusters corresponded to hundreds of thousands of years and was clearly concordant with separate domestic origins for *B. indicus* and *B. taurus*. Phylogenies of these sequences had a simple structure, two groups separated by a single, long internal branch that suggested an interesting question, "Where was the missing phylogenetic history; were there other undescribed internal branches of bovine mitochondrial diversity?"

Two developments have filled out this internal region of the phylogeny. The first is the recovery of sequences from wild ox fossils. Bones discovered in Central, Northern, and Western Europe dating to periods before and sometimes during the Neolithic, yield a sequence type labeled P (for *Bos primigenius*) that is clearly divergent from the domestic family of sequences labeled T (for *B. taurus*) (Troy et al. 2001). A minority distinct European aurochs sequence, labeled E, has been described once from a German aurochs fossil (Edwards et al. 2007).

More recently, the study of bovine mitochondrial DNA (mtDNA) variation has matured into the examination of whole chromosome sequences; Figure 1.1 gives an unrooted phylogeny of a sample of T haplotype chromosomes plus other available complete chromosomes (Achilli et al. 2008, 2009). Interestingly, this has revealed the major *B. taurus* cluster to comprise two somewhat distinct types (T and Q) that were indistinguishable at the lower resolution analysis afforded by control region sequences. Also, two highly divergent lineages emerged in modern samples. After the analysis of several thousand modern sequences, a single P variant emerged

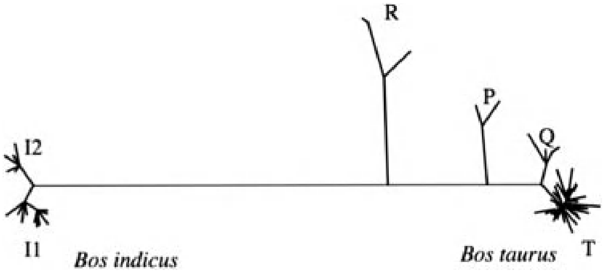


**Figure 1.1** Neighbor-joining tree of complete bovine mitochondrial DNA chromosomes. Cattle and wild ancestors segregate into discrete clusters, of which two indicine (I1, I2) and two taurine (T, Q) greatly predominate within modern samples.

from a Korean animal of ultimately European ancestry. The other P node is the first whole mtDNA from a bovine fossil—a 6700-year-old aurochs bone from Carsington Pasture Cave, England (Edwards et al. 2010). Second, a new lineage, R, was discovered in European cattle. Three clusters of *B. indicus* chromosomes are also clear.

This phylogeny invites several conclusions. Firstly, wild ox matrilines were diverse, with a branching complexity akin to patterns observed in wild bovines. The divergence between these implies geological (hundreds of thousands of years) rather than archeological (10,000 years) divergences. Secondly, the capture and subsequent thriving of these lineages in domestic populations was uneven and focused almost exclusively on three indicine, and two taurine lineages. Thirdly, this phylogenetic focus probably reflects a temporal and geographical concentration of domestication processes, and a more detailed examination of these key lineages will inform in more detail on this geography. Lastly, while unusual aurochs lineages do feature in modern samples, their extreme rarity implies limited secondary integration from the wild, rather than a major widening of the spatiotemporal focus of primary domestication.

## MtDNA Diversity Within *B. taurus*

The phylogeography of the major *B. taurus* lineage, T, has been extensively studied through sampling of short, informative sequences from the control region. T sublineages (labeled T1, T2, etc., and each diagnosed by one or a few substitutions) predominate in indigenous cattle from the Near East, Europe, Northern and Eastern Asia, and Africa. The Near East shows greatest diversity; Southern Europe displays a subset of this, and Northern Europe possesses least, with a single sublineage (T3) dominating. Using Vavilov's classic principal of diversity indicating domestic centers of origin, this pattern is consistent with the archeological evidence pointing toward domestication of *B. taurus* in the Near East (Troy et al. 2001). Some recent ancient DNA literature suggests that the wild oxen of Southern Europe may have possessed T haplotypes and that domestication of this lineage may have extended into that region. However, these sequences resemble modern T variation closely and are more difficult to interpret than, for example, a more distinct T type aurochs variant might be

(Beja-Pereira et al. 2006). This should be resolved soon by more complete surveying of Eurasian wild ox variation using next-generation sequencing and retrieval of whole mitochondrial chromosomes.

Two regionally distributed sublineage are worth considering. T4 equals T3 in frequency in far Eastern *B. taurus* populations but is undetected elsewhere, perhaps reflecting an input from wild oxen somewhere to the East of Anatolia or alternatively, a foundation bottleneck effect (Mannen et al. 2004). T1 is the African bovine lineage; here, other variants are secondary migrants from either Europe or the Near East and are restricted to Mediterranean regions. T1 is found only at low frequencies in the Near East and is an introgressor to Southern European populations in Italy and Iberia (Cymbron et al. 1999; Beja-Pereira et al. 2006). This points toward a relationship between the Near East and Africa, which is more distant than that with Europe, although it is currently unclear whether this reflects a more constricted migration of early domestics across the Sinai Peninsula or perhaps less likely, an input from the contemporary African wild ox.

## Archeology and Domestication in the Near East

The origin of cultivation in the Near East has been extensively studied and gives important insights into the domestication of a number of animal and plant species. The beginnings of the Neolithic lifestyle are thought to have emerged from the end of the Natufian culture, which occupied the Near East from approximately 12,500 to 11,500 cal years BP (Bar-Yosef 1998; Vigne 2011). These people are proposed as the first to have had a sedentary or at least semisedentary lifestyle in the Near East and possibly the world, likely supported by the high carrying capacity of the region at this time (Bar-Yosef 1998). Following this were two further important sedentary cultures: the Pre-Pottery Neolithic A and B (PPNA and PPNB). These are credited with the introduction of farming technologies that led ultimately to the domestication of grain, sheep, goat, cattle, and pig (Cauvin 2000; Vigne 2011).

The first archeological evidence for bovine domestication occurs in the Eastern slopes of the Taurus Mountains during the early PPNB (circa 10,500 BP) (Helmer et al. 2005). These early-domesticated animals are then thought to have spread from this core region through the whole of the Near East.

One of the most compelling pieces of evidence for this spread and a Near Eastern PPNB domestication of cattle is the early arrival of bovids in Cyprus. It had previously been thought that Cyprus was not colonized till about 8500 BP: however, the last 20 years have seen exciting discoveries that have pushed the earliest dates of human occupation to between 10,500 and 9000 years ago (Zeder 2008). With cattle thought to have been introduced to the island during the ninth millennium BC (Vigne et al. 2003; Peters et al. 2005; Zeder 2008), these early pastoralists would have to have traveled 60 km to the island by boat, taking not only cows but sheep, goats, and pigs with them (Vigne et al. 2003; Peters et al. 2005; Zeder 2008; Vigne et al. 2009). Although the remains of these animals do not display the morphological markers diagnostic of domestication, demographic profiles are consistent with domestication and their presence must have involved deliberate human transportation (Zeder 2008). That humans were willing to take the risks inherent in moving these animals by boat over such distances implies both that these animals were of great importance and that they had sufficient husbandry skills to enable their transport.

Archeological evidence suggests that domestic animals spread to Europe from the Near East during the early part of the seventh millennium BC, with the migrating farmers moving into Greece and the Balkan region (Pinhasi et al. 2005; Tresset and Vigne 2007; Pinhasi and von Cramon-Taubadel 2009). From this region, there are two proposed migratory routes of the Neolithic into the rest of Europe: (1) the Danubian and (2) the Mediterranean routes. The former is a proposed land migration following the Danube Valley; the latter route is suggested to have involved sea-based migration along the Mediterranean Coast into Europe (Price 2000).

Genetic evidence for these migrations is found in the genomes of both modern cattle and humans. Recent genetic data (Bramanti et al. 2009) suggest that the early farmers of Northern Europe were migrants and not descendent from the local hunter–gatherers. One fascinating human genetic variant intimately linked to cattle herding history is the mutation that confers lactase persistence. This trait is almost fixed in parts of Northern Europe, with markedly lower frequencies in Mediterranean regions—pointing toward it being a legacy of dairy-centered economies linked to the Danubian route. This is supported by recent fitting of data to simulated European genetic histories that indicates a Central European origin for the mutation (Ingram et al. 2009; Itan et al. 2009). Interestingly, this milk-related human trait variation mirrors milk protein genetic diversity differences among European cattle breeds, pointing toward culture–genetic coevolution in both species (Beja-Pereira et al. 2003). There are also sharp contrasts between Northern and Southern European cattle genomes in mtDNA, Y chromosomal marker, and autosomal marker diversity that are consistent with origins in separate migrations (Cymbron et al. 2005; Beja-Pereira et al. 2006; Negrini et al. 2007).

## The Origins of *B. indicus*

Recently, the phylogeography of *B. indicus* mtDNA has been comprehensively investigated. The two major lineages (labeled I1 and I2) are somewhat disjunct in distribution; both I1 and I2 are present at high frequency in South Asia. Both are also found in admixed and zebu populations further west but, notably, to the east of the subcontinent, I1 predominates almost absolutely in *B. indicus* from Southeast Asia and Southern China (Chen et al. 2010). This distribution gives temptation to conclude that whereas I1 is likely the product of a South Asian domestication center such as the Indus Valley region, I2 may have been initially captured in East Asia. However, an examination of genetic diversity within the lineage denies this latter possibility. Both I1 and I2 show significantly higher levels of diversity within the subcontinent than outside it. A transition from wild to domestic cattle is eminently plausible from archeological evidence from the Baluchistan region (in present day Pakistan), which is a well-documented key Neolithic center (Meadow 1993). I1 diversity is high in this region, which may well have been its site of domestication. The I2 diversity peak is less obvious, and this lineage may represent incorporation from the wild elsewhere in the subcontinent. There is some suggestion from the more limited I2 geographical dispersal of a different origin to I1; perhaps, migrations of animals carrying I1 to the east occurred from South Asian herds that had not yet incorporated I2 from Asian *B. primigenius.*

Thus, mtDNA clearly suggests a restriction of wild genetic diversity via the domestication process, with many divergent wild lineages being almost completely lost. It seems likely that this is due to a geographical limitation of cattle domestications primarily to the Near East and South Asia. Extremely rare exotic lineages in the modern population serve as exceptions to prove this rule. However, restriction of genetic diversity is not so clearly apparent with examination of autosomal polymorphism. The earliest indication of this came from the assaying of the 50 or so accessible proteins for electrophoretic variation that may be compared for levels of polymorphism across a wide species range. Here, it was clear that cattle showed heterozygosity typical of midsize mammals (Lenstra and Bradley 1999).

## Modeling Cattle Demographic History from Autosomal Sequence Variation

Vila et al. (2005) have argued from the magnitude of MHC diversities across domestic species that these are not consistent with a simple domestication model involving a single capture bottleneck. The most comprehensive analysis of bovine genetic variation to date, by the Bovine HapMap Consortium, finds higher levels of sequence diversity in all breeds surveyed; greater, for example, than those encountered in human and dog populations (Gibbs et al. 2009). However, examination of past population sizes on the basis of linkage disequilibrium decay with distance measured at medium density SNP (single nucleotide polymorphism) coverage estimates a 50-fold declaimed associated with domestication and further decline with the formation of modern breeds (MacEachern et al. 2009).

Recently, we modeled bovine population history based on site-frequency spectra of polymorphisms emerging from a survey of 37 kbp (17 genes), which have been sequenced in panels of African, European, and Indian cattle (Murray et al. 2010). Comparison of these spectra with those emerging from simulations using diffusion–approximation method (Gutenkunst et al. 2009) allowed the building of two best-fit models of past bovine demography (Figure 1.2).

These models (the analysis was limited to consideration in each of the three populations) involved (A) simply African, European, and Indian populations and (B) a better fit was achieved using a single combined African and European *B. taurus* population, a domestic *B. indicus* population, and a parallel South Asian wild ox population with secondary input into *B. indicus*. In each model, migration between populations and past population bottlenecks was allowed.

In model (A), a best fit involved an ancestral *B. taurus* population bottleneck but, notably, one with an onset that significantly predated the separation of African and European *taurus* ancestors by a factor of 2.75. The latter divergence within *B. taurus* is calibrated at 17 kyr ago, but with inherent uncertainty could plausibly overlap the domestic timeframe. The predomestic *B. taurus* bottleneck is also a feature of best-fit model (B) where it maps to between 46 and 36 kyr ago (SD = 11 kyr). This simple model may be forcing a complex population history (that may include a domestication

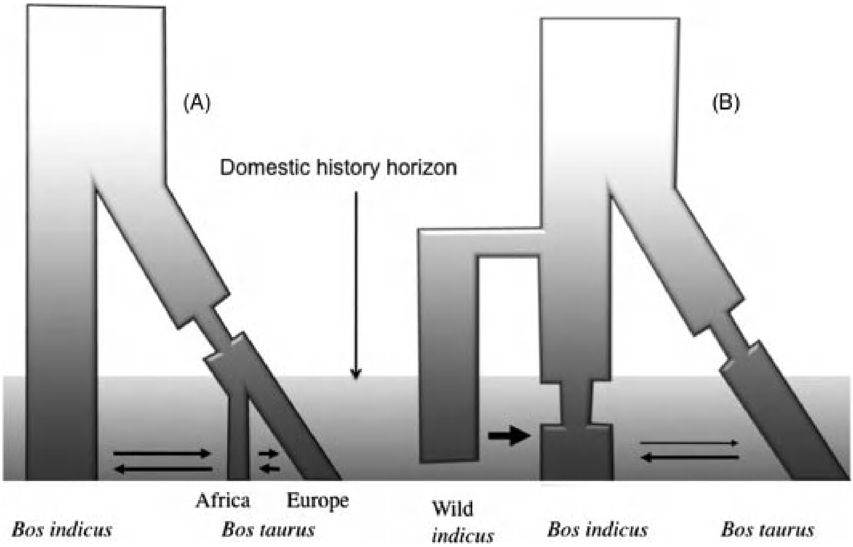


**Figure 1.2** Alternate best-fit models of bovine population history, each assuming three populations: (A) African and European *Bos taurus*, plus Indian *Bos indicus;* (B) combined *B. taurus, B. indicus*, plus a parallel contributing *B. indicus* wild ox population. Note that the domestic time horizon (denoted by shading) is highly approximate and the calibration of modeled events includes substantial uncertainty. However, some features include a separation of African and European that may be pre- or postdomestic (16.8 kyr), a *B. taurus* bottleneck that seems to predate this by a factor of 2.75 in onset (46.9–18.4 kyr), and an early *indicus* vs. *taurus* divergence (184.9 kyr).

population constriction) into a single episode but does seem to point toward an early bottleneck event that could reflect glaciation restriction of the West Asian aurochs. In contrast, the first model does not allow the fit of a bottleneck within *B. indicus* history. Model (B) does fit a *B. indicus* domestication bottleneck, but this is only with a remarkable 80% input from a parallel contemporaneous wild population. The separation of *B. indicus* and *B. taurus* ancestors concords with estimates from mtDNA and is of the order of hundreds of thousands of years.

Any modeling exercise, such as the aforementioned, should not be overinterpreted but it does point toward a complexity within South Asian domestication that is probably facilitated by substantial wild diversity that persists because of a relatively benign glacial period ecology. A contrast in history between the two taxa is mirrored by genetic diversity—nucleotide diversity is higher in *indicus* in a majority of the loci sampled and was observed to be twice as high in a single, extensively resequenced *B. indicus* breed compared to two *B. taurus* breeds by the Bovine HapMap Consortium (Gibbs et al. 2009).

Thus, our models, autosomal sequence diversity, and mtDNA phylogeography all seem to defy a unitary domestication narrative within South Asia. The first archeological evidence for domestic cattle occurs in Mehrgarh, in Baluchistan some 8000–7000 years BP, undoubtedly influenced by communication with the Fertile Crescent cultures and agricultural innovations. Bone morphology and artistic representations have allowed argument that these were *B. indicus* and transition in the nature of *Bos* bone collections suggests this is a center of zebu domestication (Meadow 1993).

Further cattle domestication centers have been suggested within the subcontinent but no archeology gives as secure a location as Baluchistan (Fuller 2006). However, large bovine bone finds in the eastern and southern parts of South Asia do suggest survival of the wild ox into the domestic period, and additional or continual wild incorporation is surely plausible. Also, wild bovines have certainly been domesticated at least four times to the east of Baluchistan, giving rise to yak, water buffalo, mithun, and domestic banteng. Caesar, in his early description of the aurochs painted a picture of a formidable animal, "Little below the elephant in size . . . Their strength and speed are extraordinary: they spare neither man nor wild beast they have espied." It might have been assumed that their capture and taming constituted singular and unlikely events in human history. Latest interpretations of genomic data point toward domestication processes that are more complex and repetitive in nature.

## References

Achilli, A., et al. (2008) Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Current Biology* **18**: R157–R158.

Achilli, A., et al. (2009) The multifaceted origin of taurine cattle reflected by the mitochondrial genome. *PLoS ONE* **4**: e5753.

Bar-Yosef, O. (1998) The Natufian culture in the Levant, threshold to the origins of agriculture. *Evolutionary Anthropology* **6**(5): 159–177.

Beja-Pereira, A., et al. (2003) Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nature Genetics* **35**: 311–313.

Beja-Pereira, A., et al. (2006) The origin of European cattle: evidence from modern and ancient DNA. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 8113–8118.

Bramanti, B., et al. (2009) Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326**: 137–140.

Cauvin, J. (2000) *The Birth of Gods and the Origins of Agriculture*. Cambridge: Cambridge University Press.

Caesar, J. (2005) *Caesar's Commentaries: On the Gallic War and on the Civil War*, edited by James H. Ford, translated by W.A. Macdevitt, p. 124. Special Edition Books.

Chen, S., et al. (2010) Zebu cattle are an exclusive legacy of the South Asia neolithic. *Molecular Biology and Evolution* **27**: 1–6.

Copley, M.S., et al. (2003) Direct chemical evidence for widespread dairying in prehistoric Britain. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 1524–1529.

Copley, M.S., Bland, H.A., Rose, P., Horton, M., Evershed, R.P. (2005) Gas chromatographic, mass spectrometric and stable carbon isotopic investigations of organic residues of plant oils and animal fats employed as illuminants in archaeological lamps from Egypt. *Analyst* **130**: 860–871.

Cymbron, T., Loftus, R.T., Malheiro, M.I., Bradley, D.G. (1999) Mitochondrial sequence variation suggests an African influence in Portuguese cattle. *Proceedings. Biological Sciences* **266**: 597–603.

Cymbron, T., Freeman, A.R., Isabel, M., Malheiro, Vigne, J.D., Bradley, D.G. (2005) Microsatellite diversity suggests different histories for Mediterranean and Northern European cattle populations. *Proceedings. Biological Sciences* **272**: 1837–1843.

Dudd, S.N., Evershed, R.P., Gibson, A.M. (1999) Evidence for varying patterns of exploitation of animal products in different prehistoric pottery traditions based on lipids preserved in surface and absorbed residues. *Journal of Archaeological Science* **26**: 1473–1482.

Edwards, C.J., et al. (2007) Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proceedings. Biological Sciences* **274**: 1377–1385.

Edwards, C.J., et al. (2010) A complete mitochondrial genome sequence from a mesolithic wild aurochs (Bos primigenius). *PLoS ONE* **5**: e9255.

Eriksson, G., et al. (2008) Same island, different diet: cultural evolution of food practice on Öland, Sweden, from the Mesolithic to the Roman Period. *Journal of Anthropological Archaeology* **27**: 520–543.

Evershed, R.P., et al. (2008) Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Nature* **455**: 528–531.

Fuller, D.Q. (2006) Agricultural origins and frontiers in South Asia: a working synthesis. *Journal of World Prehistory* **20**: 1–86.

Gibbs, R.A., et al. (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**: 528–532.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D. (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**: e1000695.

Helmer, D., Gourichon, H., Monchot, J., Peters, S.S.M. (2005) Identifying early domestic cattle from pre-pottery Neolithic sites on the midddle Euphrates using sexual dimorphism. In: *The First Steps of Animal Domestication: New Archaeozoological Approaches*, edited by J.-D. Vigne, J. Peters, and D. Helmer, pp. 86–95. Oxford: Oxbow Books.

Ingram, C.J., Mulcare, C.A., Itan, Y., Thomas, M.G., Swallow, D.M. (2009) Lactose digestion and the evolutionary genetics of lactase persistence. *Human Genetics* **124**: 579–591.

Itan, Y., Powell, A., Beaumont, M.A., Burger, J., Thomas, M.G. (2009) The origins of lactase persistence in Europe. *PLoS Computational Biology* **5**: e1000491.

Lenstra, J.A. and Bradley, D.G. (1999) Systematics and phylogeny of cattle. In: *The Genetics of Cattle*, edited by R. Fries and A. Ruvinsky, pp. 1–14. Wallingford: CAB International.

Liden, K., Eriksson, G., Nordqvist, B., Gotherstrom, A., Bendixen, E. (2004) "The wet and the wild followed by the dry and the tame" – or did they occur at the same time? Diet in Mesolithic Neolithic southern Sweden. *Antiquity* **78**: 23–33.

Loftus, R.T., MacHugh, D.E., Bradley, D.G., Sharp, P.M., Cunningham, P. (1994) Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 2757–2761.

MacEachern, S., Hayes, B., McEwan, J., Goddard, M. (2009) An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. *BMC Genomics* **10**: 181.

Mannen, H., et al. (2004) Independent mitochondrial origin and historical genetic differentiation in North Eastern Asian cattle. *Molecular Phylogenetics and Evolution* **32**: 539–544.

Meadow, R.H. (1993) Animal domestication in the Middle East: a revised view from the Eastern margin. In: *Harappan Civilization*, edited by G. Possehl, pp. 295–320. New Delhi: Oxford and IBH.

Murray, C., Huerta-Sanchez, E., Casey, F., Bradley, D.G. (2010) Cattle demographic history modelled from autosomal sequence variation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **365**: 2531–2539.

Negrini, R., et al. (2007) Differentiation of European cattle by AFLP fingerprinting. *Animal Genetics* **38**: 60–66.

Outram, A.K., et al. (2009) The earliest horse harnessing and milking. *Science* **323**: 1332–1335.

Peters, J., von den Driesch, A., Helmer, D. (2005) The uppper Euphrates-Tigris basin: cradel of agro-pastoralism? In: *The First Steps of Animal Domestication*, edited by J.-D. Vigne, J. Peters, and D. Helmer, pp. 94–124. Oxford: Oxbow Books.

Pinhasi, R., Fort, J., Ammerman, A.J. (2005) Tracing the origin and spread of agriculture in Europe. *PLoS Biology* **3**: e410.

Pinhasi, R. and von Cramon-Taubadel, N. (2009) Craniometric data supports demic diffusion model for the spread of agriculture into Europe. *PLoS ONE* **4**: e6747.

Price, T.D. (2000) *Europe's First Farmers*. Cambridge: Cambridge University Press.

Richards, M.P., Price, T.D., Koch, E. (2003) Mesolithic and Neolithic subsistence in Denmark: new stable isotope data. *Current Anthropology* **44**: 288–295.

Tresset, A. and Vigne, J.D. (2007) Substitution of species, techniques and symbols at the Mesolithic-Neolithic transition in Western Europe. *Proceeding of he British Academy* **144**: 189–210

Troy, C.S., et al. (2001) Genetic evidence for Near-Eastern origins of European cattle. *Nature* **410**: 1088–1091.

Vigne, J.-D. (2011) The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. *Comptes Rendus Biologies* **334**: 171–181.

Vigne, J.-D., Carrere, I., Guiliane, J. (2003) Unstable status of early domestic ungulates in the Near East. In: *The Neolithic of Cyprus*, edited by G.J. LeBrun, pp. 239–251. Athens: L'École française d'Athènes.

Vigne, J.D. and Helmer, D. (2007) Was milk a "secondary product" in the Old World Neolithisation process? Its role in the domestication of cattle, sheep and goats. *Anthropozoologica* **42**(2): 9–40.

Vigne, J.D., et al. (2009) Pre-Neolithic wild boar management and introduction to Cyprus more than 11,400 years ago. *Proceedings of the National Academy of Sciences of the United States of America*. **106**(38): 16135–16138.

Vila, C., Seddon, J., Ellegren, H. (2005) Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends in Genetics* **21**: 214–218.

Zeder, M.A. (2008) Domestication and early agriculture in the Mediterranean Basin: origins, diffusion, and impact. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 11597–11604.

# Chapter 2
# Mendelian Inheritance in Cattle

*Frank W. Nicholas*

## Introduction

One hundred years ago, there was much excitement among animal breeders and animal geneticists: Mendelism had only recently been rediscovered, and everyone was on the lookout for traits showing Mendelian inheritance in all living organisms, including domestic animals. The early results were summarized in two of the earliest books on genetics, namely those by Bateson (1909) and Castle (1916). In both books, the list of Mendelian cattle traits comprised various coat colors, the presence/absence of horns, and Dexter dwarfism. From those humble beginnings, in a little over 100 years, knowledge has progressed to the point where, at the time of submission of this chapter, at least 71 Mendelian cattle traits have been recorded as being characterized at the DNA level! Since much of this enormous advance in knowledge is the result of genomics research, it is appropriate to summarize the results for these 71 Mendelian traits in a book devoted to cattle genomics. Noting that Chapter 3 in this book deals with coat color, the present chapter concentrates on other Mendelian traits.

## A Classic Mendelian Cattle Trait: Presence/Absence of Horns

In cattle, one of the first Mendelian traits to attract attention was the presence/absence of horns. The inherited nature of this trait was well recognized (but not understood) long before the rediscovery of Mendelism (Darwin 1859; Darwin 1868). In 1906, the American agricultural polymath W.J. Spillman (who is not only regarded as a founding father of agricultural economics, but also independently rediscovered Mendelism while crossing strains of wheat) published a paper in *Science* (Spillman 1906a) and another in the newly-founded *Journal of Heredity* (Spillman 1906b), providing convincing evidence that the presence/absence of horns is a Mendelian trait, with polled being dominant to horned. This trait soon became a classic Mendelian trait, cited in many textbooks. Indeed, as delightfully recorded by Crow (1992), this trait even attracted the attention of the Nobel-prize winning physicist Erwin Schrödinger, who wrote two letters to J.B.S. Haldane in 1945, in relation to “the hornless cattle problem.” In these letters, Schrödinger derived an equation

*Bovine Genomics*, First Edition. Edited by James E. Womack.

that predicts the frequency of horned offspring in a closed herd after any number of generations of complete selection against horned bulls, but with no selection on cows.

Nothing much was added to our knowledge of this trait until the first wave of genomics tools provided sufficient microsatellite markers to enable Georges et al. (1993) to map the presence/absence of horns, to within a recombination fraction of 13% with two markers on chromosome BTA1. To current readers, such "loose" linkage might seem to be not worthy of much celebration. At the time, however, this result was sufficiently important and novel to warrant publication in *Nature Genetics*. By 2005, the region had been narrowed to 1 Mb (Drögemüller et al. 2005). Despite the chromosomal location of the gene having been known quite accurately for more than a decade, despite 71 other Mendelian cattle traits having been characterized at the DNA level, and despite an annotated bovine genome having been available for more than 2 years (Bovine Genome Sequencing and Analysis Consortium et al. 2009), at the time of submission of this chapter (October 2011) there is still no published paper reporting the actual gene for the presence/absence of horns. Interestingly (especially in a book devoted to genomics), the most recent paper on this trait, by Mariasegaram et al. (2010), reports a comparison of gene expression between polled and horned tissue, in the hope of gaining new insights into the biology of horn development.

## Bovine Mendelian Traits Characterized at the DNA Level

As mentioned previously, there are at least 71 bovine Mendelian traits that have been characterized at the DNA level. Publicly available details of all these traits, and of all other cattle traits that have ever been reported as being Mendelian, are continually updated in Online Mendelian Inheritance in Animals (OMIA), which is freely available at http://omia.angis.org.au. There is no point in reproducing the OMIA text in this chapter. Instead, highlights of some of the discoveries are mentioned. Whenever a trait is mentioned, its 10-digit OMIA ID (comprising a 6-digit trait ID and the 4-digit NCBI taxonomy ID for cattle, namely 9913) is also cited.

The history of the discoveries of the molecular basis of bovine Mendelian traits is summarized in Figure 2.1, which shows the first such discovery being published in 1985, and three others being published in the following 9 years. By 1995, following publication of the first genome-wide linkage maps (Bishop et al. 1994; Barendse et al. 1994), the pace had quickened, with mostly two or more traits being added each subsequent year. By 2010 (the most recent complete year available at the time of submission of this chapter), the record number of 12 new traits reflects the fruits of the genomics revolution, which by then was in full swing.

Tellingly, the discovery of the molecular lesion causing the first trait characterized at the DNA level (inherited goiter; OMIA 000424-9913) was made by a large transnational team of medical researchers (Ricketts et al. 1985) who had an obvious candidate gene, based on clinical signs alone. Even so, the effort required was quite substantial. The next trait, citrullinemia (OMIA 000194-9913; Dennis et al. 1989), also had an

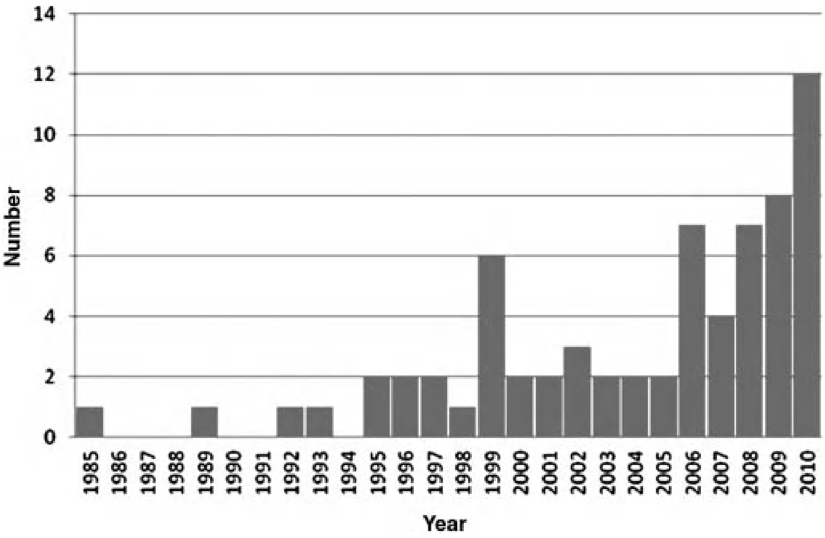


**Figure 2.1** The numbers of published discoveries of the molecular basis of Mendelian traits in cattle each year since the first such discovery in 1985.

obvious candidate gene that had been implicated in humans with the same enzyme deficiency. For the first time, animal scientists (Dennis and Healy) were involved, but in order to do the research, they crossed the Pacific Ocean to the Baylor College of Medicine in Texas, to work in the human genetics laboratory of Beaudet and O'Brien, who had pioneered the molecular analysis of human citrullinemia. Tellingly, the discovery of the DNA lesion in cattle was regarded as sufficiently groundbreaking to be published in *PNAS*. As linkage mapping became more powerful, the task became easier: even if there was no obvious candidate gene, the search could be narrowed down to "positional candidates," that is, genes located in the mapped region and which, when mutated, might give rise to the relevant phenotype. The task was made even easier with the advent of SNP "chips" in the late 2000s (e.g., Matukumalli et al. 2009). The power of this technology was illustrated in a landmark paper by Charlier et al. (2008), who showed how, with only a "handful" of affected and control animals (3–12 cases and 9–24 controls), it is possible to fine-map an autosomal recessive trait using a high-density SNP panel, to such an extent that it becomes relatively simple to choose among the small number of positional candidate genes residing in that small region. Most recently, the sequencing of positional candidate genes has been greatly facilitated by sequence capture followed by massively parallel (re)sequencing, which, in the first published example of the use of this strategy, enabled Drögemüller et al. (2010) to discover the molecular basis of arachnomelia (OMIA 000059-9913).

In the future, with the genomics revolution now in full swing and providing access to such powerful resources, we are certain to see an explosion of discoveries. These will all be documented in OMIA, which will continue to be freely available at http://omia.angis.org.au.

## A (Mostly) Morbid Map of the Bovine Genome

For several decades, human geneticists have had available updated versions of Victor McKusick's "morbid anatomy of the human genome" (McKusick 1986), more recently called the "morbid map," which lists disease-causing genes according to their chromosomal location. A regularly updated version of the human morbid map is obtainable from http://omim.org/downloads. With annotations of the bovine genome sequence assembly now becoming available, it is possible to assemble a similar map for cattle, as shown in Table 2.1. This map is "mostly morbid" because, for completeness, it includes all Mendelian traits that have been characterized at the DNA level, including traits such as coat color that are not necessarily disadvantageous.

Many of the genes listed in Table 2.1 were annotated by members of the Inherited Disorders team of the Bovine Genome Analysis Consortium. Most of their annotations are available at NCBI via the HomoloGene link from the relevant OMIA entry. Much of this annotation information is also viewable directly in the Bovine Genome Database at http://bovinegenome.org.

As a sign of the times, two genes have been included in the mostly morbid map on the basis of genome scans for multifactorial traits rather than for a Mendelian trait. In 2009, Liu et al. reported the results of a genome scan for the degree of white spotting in a Jersey/Holstein-Friesian F2. Two of the three significant quantitative trait loci (QTL) corresponded to two well-known coat-color genes, namely *KIT* (dominant white; OMIA 000209-9913) and *MITF* (white spotting; OMIA 000214-9913). These results were confirmed in a separate genome scan within Holstein-Friesians reported by Hayes et al. (2010).

In another sign of the times, the genes and locations for seven disorders are recorded with question marks. For each of these disorders, it is known that the causal mutation has been identified. Indeed, in all seven cases, a DNA test is commercially available. However, to the present author's knowledge, these discoveries have not appeared in a refereed scientific paper. The typical reason for this lack of traditional scientific reporting is that the discoverer's institution has insisted that some income be generated from the discovery, and the most cost-effective way to do this, especially since tests like this almost inevitably have only a limited lifespan, is to retain the results as a trade secret. The present author is not critical of this approach. Indeed, under instructions from his own institution, the present author and his colleagues were obliged to follow exactly the same strategy for Dexter dwarfism (OMIA 001271-9913): there was a delay of several years between the discovery of the mutation and its reporting. It would be a great benefit for the public good if a means could be developed for institutions to gain their pound of flesh and at the same time allow their employees to publish their results.

## Other Bovine Mendelian Traits

In the past decade, there have been several very useful reviews of Mendelian traits in cattle, concentrating mainly on inherited disorders. Readers requiring additional information are directed to reviews by Millar et al. (2000), Gentile and Testoni (2006),

**Table 2.1** A mostly morbid map of the bovine genome, incorporating all Mendelian traits that have been characterized at the DNA level, as on October 30, 2011.

| | | | Location in bovine genome assembly Btau_4.2[a] | | |
|---|---|---|---|---|---|
| Name of trait | MIA number | Gene | Chromosome | Band(s) or scaffold | Nucleotide range |
| Deficiency of uridine monophosphate synthase | 000262-9913 | *UMPS* | 1 | q31–q36 | 70331329. . .70381376 |
| Renal tubular dysplasia | 001135-9913 | *CLDN16* | 1 | q31–q33 | 78511695. . .78489435 |
| Leukocyte adhesion deficiency | 000595-9913 | *ITGB2* | 1 | | 146757642. . .146786802 |
| Muscular hypertrophy | 000683-9913 | *MSTN* | 2 | q14–q15 | 6532637. . .6539264 |
| Complex vertebral malformation | 001340-9913 | *SLC35A3* | 3 | | 46233205. . .46218305 |
| Scurs, type 2 | 001593-9913 | *TWIST1* | 4 | | 28978295. . .28976301 |
| Osteopetrosis | 000755-9913 | *SLC4A2* | 4 | | 117895305. . .117907857 |
| Coat color, roan | 001216-9913 | *KITLG* | 5 | | 20599086. . .20624323 |
| Ehlers–Danlos syndrome | 000327-9913 | *EPYC* | 5 | | 23601394. . .23560847 |
| Epidermolysis bullosa | 000340-9913 | *KRT5* | 5 | q14–q23 | 30274221. . .30280072 |
| Arachnomelia, BTA5 | 000059-9913 | *SUOX* | 5 | | 61761632. . .61757363 |
| Coat color, dilution | 001545-9913 | *PMEL* | 5 | | 6178763. . .61795738 |
| Hypotrichosis with coat-color dilution | 001544-9913 | *PMEL* | 5 | | 6178763. . .61795738 |
| Mannosidosis, beta | 000626-9913 | *MANBA* | 6 | q3 | 23676316. . .23803977 |
| Coat color, dominant white | 000209-9913 | *KIT* | 6 | | 72822200. . .72909476 |
| Coat color, color-sided | 001576-9913 | *KIT* | 6 | q3 | 72822200. . .72909476 |
| Dwarfism, Angus | 001485-9913 | *PRKG2* | 6 | | 99480894. . .99369084 |
| Chondrodysplasia | 000187-9913 | *EVC2* | 6 | | 107885450. . .107864143 |
| Ehlers–Danlos syndrome, type VII | 000328-9913 | *ADAMTS2* | 7 | | 1884060. . .2128274 |
| Mannosidosis, alpha | 000625-9913 | *MAN2B1* | 7 | | 11107375. . .11122713 |
| Myoclonus | 000689-9913 | *GLRA1* | 7 | | 62753526. . .62656324 |
| Coat color, brown | 001249-9913 | *TYRP1* | 8 | | 33490350. . .33474113 |

(*continued*)

**Table 2.1** (*Continued*)

| | | | Location in bovine genome assembly Btau_4.2[a] | | |
|---|---|---|---|---|---|
| Name of trait | MIA number | Gene | Chromosome | Band(s) or scaffold | Nucleotide range |
| Marfan syndrome | 000628-9913 | *FBN1* | 10 | | 63333012…63596903 |
| Spinal dysmyelination | 001247-9913 | *SPAST* | 11 | | 15331608…15386684 |
| Citrullinemia | 000194-9913 | *ASS1* | 11 | q28 | 104578816…104630861 |
| Beta-lactoglobulin, aberrant low expression | 001437-9913 | *LGB* | 11 | q28 | 107256533…107261250 |
| Neuronal ceroid lipofuscinosis, 5 | 001482-9913 | *CLN5* | 12 | | 52629673…52637493 |
| Coat color, agouti | 000201-9913 | *ASIP* | 13[b] | | 64234645…64239783[b] |
| Goitre, familial | 000424-9913 | *TG* | 14 | q12–q15 | 7894998…7658631 |
| Yellow fat | 001079-9913 | *BCO2* | 15 | | 20853556…20921259 |
| Syndactyly | 000963-9913 | *LRP4* | 15 | | 77343208…77305775 |
| Ichthyosis congenita | 000547-9913 | *ABCA12* | 16 | | 959283…1159171 |
| Trimethylaminuria | 001360-9913 | *FMO3* | 16 | | 1720314…1746378 |
| Axonopathy | 001106-9913 | *MFN2* | 16 | | 38452074…38425920 |
| Lethal trait A46 | 000593-9913 | *SLC39A4* | 17 | | 21439028…21443517 |
| Multiple ocular defects | 000733-9913 | *WFDC1* | 18 | | 9750651…9780814 |
| Coat color, extension | 001199-9913 | *MC1R* | 18 | | 13780542…13782293 |
| Maple syrup urine disease | 000627-9913 | *BCKDHA* | 18 | | 50228462…50247729 |
| Cardiomyopathy and woolly haircoat syndrome | 000161-9913 | *PPP1R13L* | 18 | | 52806645…52791856 |
| Cardiomyopathy, dilated | 000162-9913 | *OPA3* | 18 | | 52904087…52970557 |
| Abortion and stillbirth | 001565-9913 | *MIMT1* | 18 | | NA |
| Myasthenic syndrome, congenital | 000685-9913 | *CHRNE* | 19 | | 26859041…26863638 |
| Spherocytosis | 001228-9913 | *SLC4A1* | 19 | | 45500927…45484737 |
| Tail, crooked | 001452-9913 | *MRC2* | 19 | | 48588382…48647119 |
| Dwarfism, growth-hormone deficiency | 001473-9913 | *GH1* | 19 | q22 | 48768618…48772014 |
| Glycogen storage disease II | 000419-9913 | *GAA* | 19 | | 54015691…54003392 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| Dwarfism, Dexter | 001271-9913 | *ACAN* | 21 |  | 20147594. . .20216258 |
| Coat color, white spotting | 000214-9913 | *MITF* | 22 |  | 32387233. . .32353745 |
| Arachnomelia, BTA23 | 001541-9913 | *MOCS1* | 23 |  | 14431706. . .14397223 |
| Myopathy of the diaphragmatic muscles | 001319-9913 | *HSPA1B* | 23 | q22 | 27268676. . .27266578 |
| Protoporphyria | 000836-9913 | *FECH* | 24 |  | 59133675. . .59098774 |
| Spinal muscular atrophy | 000939-9913 | *KDSR* | 24 |  | 64284026. . .64222078 |
| Congenital muscular dystonia 1 | 001450-9913 | *ATP2A1* | 25 |  | 27738510. . .27721452 |
| Pseudomyotonia, congenital | 001464-9913 | *ATP2A1* | 25 |  | 27738510. . .27721452 |
| Mucopolysaccharidosis IIIB | 001342-9913 | *NAGLU* | 26 |  | 32493638. . .32500618 |
| Factor XI deficiency | 000363-9913 | *F11* | 27 |  | 17607726. . .17626870 |
| Chédiak–Higashi syndrome | 000185-9913 | *LYST* | 28 | q13–q14 | 7004429. . .6860482 |
| Coat color, albinism | 000202-9913 | *TYR* | 29 | qter | 6536886. . .6426495 |
| Congenital muscular dystonia 2 | 001451-9913 | *SLC6A5* | 29 |  | 25567977. . .25516144 |
| Glycogen storage disease V | 001139-9913 | *PYGM* | 29 |  | 44790056. . .44778256 |
| Thrombopathia | 001003-9913 | *RASGRP2* | 29 |  | 44877546. . .44865240 |
| Anhidrotic ectodermal dysplasia | 000543-9913 | *EDA* | X | Un.004.31 | 197273. . .451038 |
| Haemophilia A | 000437-9913 | *F8* | X | Un.004.16[b] | 344471. . .482199[b] |
| Sex reversal: XY female | 001230-9913 | *SRY* | Y | Un.004.2642.scaffold1 | 1. . .166 |
| Arthrogryposis multiplex congenita | 001465-9913 | ? | ? |  |  |
| Contractural arachnodactyly | 001511-9913 | ? | ? |  |  |
| Epilepsy | 000344-9913 | ? | ? |  |  |
| Hydrocephalus, neuropathic | 000487-9913 | ? | ? |  |  |
| Hypotrichosis | 000540-9913 | ? | ? |  |  |
| Pulmonary hypoplasia with anasarca | 001562-9913 | ? | ? |  |  |
| Tibial hemimelia | 001009-9913 | ? | ? |  |  |

[a] As gleaned from NCBI's HomoloGene; http://www.ncbi.nlm.nih.gov/homologene/.

[b] Information taken from Ensembl (http://www.ensembl.org) because this gene is not annotated in HomoloGene.

Agerholm (2007), Ibeagha-Awemu et al. (2008), Whitlock et al. (2008), Windsor and Agerholm (2009), and Windsor et al. (2011a, 2011b).

## Conclusion

The potential number of Mendelian traits is at least as great as the number of coding sequences, which is more than 22,000 (Bovine Genome Sequencing and Analysis Consortium et al. 2009). Add to this the well-documented examples of Mendelian traits due to mutations in noncoding regions (e.g., Callipyge muscular hypertrophy in sheep; OMIA 001354-9940), we are faced with an almost infinite potential number of Mendelian traits. Of course, many of these will be undetectable lethals. However, even if only a fraction of the potential number of Mendelian traits is actually identifiable, we are still a long way from knowing them all. The statistics for humans provide a clue to the future. At the time of submission of this chapter, there were 3288 Mendelian traits in humans with a known molecular basis, and a further 1776 whose molecular basis is still to be determined (http://omim.org/statistics/entry; accessed October 30, 2011). We can conclude that even though human geneticists are way ahead of cattle geneticists, both camps still have a long way to go before getting anywhere near a comprehensive catalog of Mendelian traits.

## References

Agerholm, J.S. (2007) Inherited disorders in Danish cattle. *Acta Pathologica, Microbiologica et Immunologica Scandinavica* **115**(Suppl 122): 1–76.

Barendse, W., et al. (1994) A genetic linkage map of the bovine genome. *Nature Genetics* **6**: 227–235.

Bateson, W. (1909) *Mendel's Principles of Heredity*. Cambridge: Cambridge University Press.

Bishop, M.D., et al. (1994) A genetic linkage map for cattle. *Genetics* **136**: 619–639.

Bovine Genome Sequencing and Analysis Consortium, Elsick, C.G., Tellam, R.L., Worley, K.C. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**(5926): 522–528.

Castle, W.E. (1916) *Genetics and Eugenics*. Cambridge, MA: Harvard University Press.

Charlier, C., et al. (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nature Genetics* **40**(4): 449–454.

Crow, J.F. (1992) Erwin Schrödinger and the hornless cattle problem. *Genetics* **130**: 237–239.

Darwin, C.R. (1859) *On the Origin of Species by Means of Natural Selection*. p. 14. London: John Murray.

Darwin, C.R. (1868) *The Variation of Animals and Plants under Domestication*. Vol. 2, p. 316. London: John Murray.

Dennis, J.A., Healy, P.J., Beaudet, A.L., Obrien, W.E. (1989) Molecular definition of bovine argininosuccinate synthetase deficiency. *Proceedings of the National Academy of Sciences of the United States of America* **86**: 7947–7951.

Drögemüller, C., Wöhlke, A., Mömke, S., Distl, O. (2005) Fine mapping of the polled locus to a 1-Mb region on bovine chromosome 1q12. *Mammalian Genome* **16**: 613–620.

Drögemüller, C., Tetens, J., Sigurdsson, S., Gentile, A., Testoni, S., Lindblad-Toh, K., Leeb, T. (2010) Identification of the bovine arachnomelia mutation by massively parallel sequencing implicates sulfite oxidase (SUOX) in bone development. *PLoS Genetics* **6**(8): e1001079.

Gentile, A. and Testoni, S. (2006) Inherited disorders of cattle: a selected review. *Slovenian Veterinary Research* **43**(Suppl): 15–16.

Georges, M., et al. (1993) Microsatellite mapping of a gene affecting horn development in *Bos taurus*. *Nature Genetics* **4**(2): 206–210.

Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J., Goddard, M.E. (2010) Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics* **6**(9): e1001139.

Ibeagha-Awemu, E.M., Kgwatalala, P., Zhao, X. (2008) A critical analysis of production-associated DNA polymorphisms in the genes of cattle, goat, sheep, and pig. *Mammalian Genome* **19**(9): 591–617.

Liu, L., Harris, B., Keehan, M., Zhang, Y. (2009) Genome scan for the degree of white spotting in dairy cattle. *Animal Genetics* **40**: 975–977.

Mariasegaram, M., Reverter, A., Barris, W., Lehnert, S.A., Dalrymple, B., Prayaga, K. (2010) Transcription profiling provides insights into gene pathways involved in horn and scurs development in cattle. *BMC Genomics* **11**: 370.

Matukumalli, L.K., et al. (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* **4**(4): e5350.

McKusick, V.A. (1986) The morbid anatomy of the human genome. *Medicine* **65**(1): 1–33.

Millar, P., Lauvergne, J.J., Dolling, C.H.S. (2000) *Mendelian Inheritance in Cattle*, EAAP Publication No 101. Wageningen: Wageningen Pers.

Ricketts, M.H., Pohl, V., Martynoff, G. de, Boyd, C.D., Bester, A.J., Jaarsveld, P.P. van, Vassart, G. (1985) Defective splicing of thyroglobulin gene transcripts in the congenital goitre of the Afrikander cattle. *EMBO Journal* **4**: 731–737.

Spillman, W.J. (1906a) A Mendelian character in cattle. *Science* **23**: 549–551.

Spillman, W.J. (1906b) Application of Mendel's law to a practical problem in breeding cattle. *Journal of Heredity* **2**: 173–177.

Whitlock, B.K., Kaiser, L., Maxwell, H.S. (2008) Heritable bovine fetal abnormalities. *Theriogenology* **70**: 535–549.

Windsor, P. and Agerholm, J. (2009) Inherited diseases of Australian Holstein-Friesian cattle. *Australian Veterinary Journal* **87**(5): 193–199.

Windsor, P., Kessell, A., Finnie, J. (2011a) Review of neurological diseases of ruminant livestock in Australia. V: congenital neurogenetic disorders of cattle. *Australian Veterinary Journal* **89**(10): 394–401.

Windsor, P., Kessell, A., Finnie, J. (2011b) Review of neurological diseases of ruminant livestock in Australia. VI: postnatal bovine, and ovine and caprine, neurogenetic disorders. *Australian Veterinary Journal* **89**(11): 432–438.

# Chapter 3
# Genetics of Coat Color in Cattle

*Sheila M. Schmutz*

## Introduction

Coat color and pattern have been a hallmark characteristic of many breeds of cattle since registries began, and therefore, undoubtedly before. Coat color is a very useful teaching tool for students at many levels because they can observe and relate to it easily. Coat color and pattern is not a single trait, but a complex set of traits that display gene interactions including epistasis.

Inheritance patterns for a single trait can be dominant or recessive, codominant as in case of roan, or quantitative as in the amount of white spotting on Holsteins (Table 3.1). Understanding the molecular and biochemical mechanisms involved in coat color and pattern leads us to studies in developmental biology and cellular differentiation. Genes involved in pigmentation include hormone receptors, signaling hormones, and transcription factors (Table 3.2). The mutations include base pair substitutions, insertions and deletions, copy number variation, and promoter mutations (Table 3.2).

There are several previous articles and book chapters that have focused on coat color inheritance in cattle (see Olson 1999). This information is not all reviewed here. This chapter focuses on the molecular genetics of coat colors and patterns. Because most DNA studies have been done using *Bos taurus* breeds, there is minimal coverage of *Bos indicus* cattle.

## Basic Coat Colors

### *MC1R*

As with many other mammals, the first cattle pigmentation locus that was studied was *melanocortin 1 receptor* (*MC1R)*, often known as the E locus. *MC1R* was mapped to BTA 18 (Klungland et al. 1995). Klungland et al. (1995) described three functional alleles in this gene: (1) $E^+$, (2) $E^D$, and (3) *e*. The $E^D$ allele is caused by a mutation that replaces the leucine at amino acid position 99 with a proline (Leu99Pro). The $E^D$ allele was so named because it is dominant to the wild-type $E^+$ allele. Cattle with the $E^D$ allele are black or a shade thereof, such as gray or brown. Klungland et al. (1995) suggested that the alpha melanocyte stimulating hormone ($\alpha$-MSH) binds exclusively

*Bovine Genomics*, First Edition. Edited by James E. Womack.

**Table 3.1** Suggested coat color loci and alleles in cattle.

| | |
|---|---|
| The alleles are listed in their "predicted" dominance hierarchy. Those in bold have been confirmed at the DNA level. | |
| **Basic colors** | |
| E (extension) = melanocortin 1 receptor (*MC1R*) | |
| $\mathbf{E^{D}}$ | Eumelanin is produced |
| $\mathbf{E^{+}}$ | Eumelanin or phaeomelanin can be produced |
| **E** | "Only" phaeomelanin produced (although a few black hairs possible) |
| A (agouti) = agouti signalling protein (*ASIP*) | |
| A | Shaded or solid, without stripes |
| $\mathbf{A^{br}}$ | Brindle |
| **Diluted colors** | |
| B (brown) = tyrosinase related protein 1 (*TYRP1*) | |
| **B** | Black eumelanin |
| **B** | Brown eumelanin |
| C (colored) = tyrosinase (*TYR*) | |
| C | Full pigmentation |
| $c^{P}$ | Colored points and white body (allele not found but maps to *TYR*) |
| $\mathbf{c^{a}}$ | Complete albinism (allele found for Braunvieh) |
| D (dilutes eumelanin and/or phaeomelanin) = *silver* gene (*SILV)* (codominant) | |
| D | Not diluted |
| $d^{C}$ | Diluted to white in the homozygote (i.e., Charolais) |
| $d^{H}$ | Diluted to cream in the homozygote (i.e., Highland) |
| K (black) = beta defensin 103 (*DEFB103*) | |
| $k^{+}$ | Black |
| $K^{VR}$ | Variant red |
| **White markings** | |
| R (roan) = KIT ligand (*KITLG)* (roan is codominant) | |
| R/R | White in homozygote |
| R/r | Roan |
| R/r | Colored in homozygote |

(*continued*)

**Table 3.1** (*Continued*)

| S (spotting) = *KIT*? gene (linkage studies suggest this is KIT, but no alleles identified yet) | |
|---|---|
| S | Solid colored |
| $s^H$ | Hereford pattern |
| $s^p$ | Piebald or random spotting (i.e., Holstein and Ayrshire) |
| $s^{cs}$ | Color sided (i.e., Pinzgauer) |
| ? (belted) = ? (trait maps to BTA3) | |
| $bt^+$ | No belt |
| Bt | White belt, complete or incomplete |

to *MC1R*, when the animal has an $E^D$ allele and that, in turn, only eumelanin pigment was then produced in the melanocyte. However, some data contradict this postulate (see Section "DEFB103").

The recessive *e* allele is caused by a premature stop codon at amino acid 155, which is a result of a frameshift caused by the deletion of base pair 310 (Klungland et al. 1995; Joerg et al. 1996). Only the agouti peptide binds to *MC1R* in cattle that have an *e*/*e* genotype and only phaeomelanin is produced. Such cattle are red. Although some authors have described such cattle as brown (Mohanty et al. 2008), the *e*/*e* genotype of *MC1R* only produces phaeomelanin, a red pigment, while true brown is eumelanin-based.

Cattle with the wild-type allele, $E^+$, are able to bind either $\alpha$-MSH or agouti peptide. Therefore, both eumelanin and phaeomelanin can be produced at different times of life, in different places on the body, or in distinct patterns such as brindle. The color and pattern in cattle that are homozygous $E^+/E^+$, therefore, depends on the genotype at other loci. Breeds such as Ayrshire, Brown Swiss, and Jersey are fixed for the $E^+/E^+$ genotype. All brindle cattle have at least one $E^+$ allele and no $E^D$ allele. The caped pattern of bison, and the coloration of some Highland cattle with a black cape and reddish body, are possible because of the $E^+$ allele.

Other variants have been reported in *MC1R* but these have not been shown to affect coat color. Rouzaud et al. (2000) reported a new variant, a four amino acid duplication beginning at nucleotide 699. This duplication was found in Gasconne and Aubrac cattle in both the heterozygous and homozygous state. Rouzaud et al. (2000) did not suggest that any coat color was associated with this duplication.

Graphodatskaya et al. (2002) discovered additional *MC1R* variants, which they also studied in vitro. The variant they designated $E^{d1}$ is 667C>T, changing an arginine to a tryptophan (Arg223Trp). They also described the same 12 bp duplication described previously by Rouzaud et al. (2000), which they designated as $E^{d2}$. Graphodatskaya et al. (2002) state that the duplication was not associated with coat color. Some Brown Swiss cattle have the duplication alone and others the Arg223Trp alone, so it would seem that neither is causing the taupe brown color of this breed (Dreger and Schmutz, unpublished data).

**Table 3.2** Genetic characterization of the coat color alleles discovered to date and example breeds harboring these alleles.

| Gene | BTA | Coat color | Allele | Genomic | cDNA | Example breed/s |
|---|---|---|---|---|---|---|
| *MC1R* | 18 | Variable | $E^+$ | Wild type | No change | Fixed in Brown Swiss, Jersey, etc. |
| | | Black | $E^D$ | g.296T>C | Leu99Pro | Angus, Galloway, and . . .? |
| | | Red | *e* | del 309 | Premature stop | Fixed in traditional Limousin, Hereford, Simmental, Charolais, etc. |
| *TYRP1* | 8 | Black | *B* | Wild type | No change | All common beef & dairy except Dexter |
| | | Brown | *b* | C>T | His434Tyr | Dexter (some) |
| *KITLG* | 5 | Colored | r | Wild type | No change | |
| | | Roan | R/r | C>A | Ala193Asp | Shorthorn, Belgian Blue |
| | | White | R/R | C>A | Ala193Asp | Shorthorn, Belgian Blue |
| *TYR* | 29 | Colored | C | Wild type | No change | |
| | | Albinism | c | insC | 316premature | Braunvieh |
| | | Color pointed | | Promoter? | stop? | Fixed in White Galloway, White Park |
| *DEFB103* | 27 | Variant red | K$^{VR}$ | CNV? | | Holstein |
| *PMEL* | 5 | Dark | D | Wild type | No change | |
| | | Medium | D/d$^c$ | | | |
| | | Pale | d$^c$ / d$^c$ | | c.64G>A | Charolais |
| | | Dun/yellow | D/d$^H$ | del of TTC | c.del50–52 | Highland |
| | | Silver dun/white | d$^H$ /d$^H$ | | | Highland |
| *KIT* | 6 | Solid | | Wild type | No change | |
| | | Hereford markings | ? | | Hereford | |
| | | Piebald spotting | | ? | | Holstein, Ayrshire, Simmental |
| | | Color sided | | ? | | Pinzgauer |
| ? | 3 | Belted | | ? | | Brown Swiss, Galloway |

Both $E^{d1}$ and $E^{d2}$ were found in Brown Swiss cattle. An Ile297Trp variant, that Graphodatskaya et al. (2002) designated as $e^{f}$, was reported in a red Holstein bull that was heterozygous for *e*.

## *POMC*

$\alpha$-MSH is one of the products of the *pro-opiomelanocortin* (*POMC)* gene. $\alpha$-MSH is essential in many pathways and hence, no functional mutation is typically found in this gene in a living mammal. A 12-bp deletion in *POMC* was reported by Deobald (2009) (c.293_304delTTGGGGGCGCGG) that would result in the loss of four amino acids of cattle $\gamma$-MSH. This mutation was rare in the population she studied, with a minor allele frequency of 0.05 or less depending upon the breed. Hence, it is unclear if no homozygotes were discovered because the allele was simply rare or because the homozygous animal would not have survived. Cattle heterozygous for this deletion exhibited no difference in coat color compared to those with no copy of the deletion.

## *ASIP*

The *agouti signal protein* (*ASIP*) is the gene that produces the agouti peptide. This gene is polymorphic in several species. In cattle, there is potentially one functional mutation found to date. No polymorphisms were detected in 20 cattle of nine breeds (Royo et al. 2005). Girardot et al. (2006) found one copy of a LINE insertion in the 5′ UTR in four Normande cattle that were brindle. They also examined 20 cattle of four other breeds for this LINE insertion: Holstein, Limousin, Parthenaise, and Montbéliarde. The LINE insertion was only present in some of the Montbéliarde cattle. Girardot et al. (2006) suggest that both a copy of the LINE insertion and an $E^{+}$ allele at *MC1R* are needed for the brindle pattern and all the Montbéliarde cattle were *e*/*e*. This LINE insertion has also been observed in some Highland cattle that were $E^{+}/E^{+}$ or $E^{+}/e$, and not brindle (Schmutz and Dreger, in press).

Brindle cattle have a distinctive striped pattern of alternating stripes of phaeomelanin and eumelanin. These can be red and black, yellow and gray, or even cream and gray. Texas Longhorn cattle and Highland cattle can both be brindle, as can the Korean Hanwoo (Mohanty et al. 2008), and the dairy breeds, Jersey and Normande. The pattern is much more difficult to recognize as stripes in the longhaired Highlands than in shorthaired breeds.

## *TYRP1*

The *tyrosinase related protein one* gene (*TYRP1*) is the gene reported to cause brown or chocolate coat color in several species. A His434Tyr mutation has been reported in Dexter cattle (Berryere et al. 2003). Cattle that are homozygous for this allele, *b*/*b*, and also have at least one $E^{D}$ allele are dun brown. In Dexters, these cattle have traditionally been called dun and occur in shades from caramel to chocolate. The term "dun" is also used to describe a different coat color in Highland and Galloway cattle that is not caused by a mutation in *TYRP1*. One hundred and twenty-one *B. taurus* cattle of 19 breeds were studied and this mutation was not found in any of them. It appears to be unique to Dexter cattle (Berryere et al. 2003).

## *TYR*

The *tyrosinase* gene (*TYR*) is the primary gene causing complete albinism in many species of animals. Albino cattle have been reported in many breeds including Shorthorn (Greene et al. 1973), Braunvieh (Winzenried and Lauvergne 1970), Guernsey (Leipold et al. 1968), Holstein (Petersen et al. 1944), and Austrian Murboden (Schleger 1959).

An insertion of a cytosine at position 926 of the *TYR* mRNA sequence causes a frameshift mutation leading to a premature stop codon at residue 316 in Brown Swiss or Braunvieh cattle (GenBank AY162287) (Schmutz et al. 2004). Albino cattle were homozygous for this mutation. These cattle were sun sensitive and had visual and hearing deficits. Concerted efforts to DNA test Braunvieh cattle in the United States in recent years have reduced this originally rare allele to almost zero.

Albinism is also periodically reported in Holstein cattle. However, the mutation found in Brown Swiss was not found in albino Holstein cattle (Schmutz et al. 2004, unpublished data).

A pattern often called "colored points" exists in many species. In cats, it is known as Siamese. In mice and rabbits, it is known as Himalayan. These species have mutations in *TYR* that are temperature sensitive leading to high levels of expression of the gene and pigmentation in the cooler points or extremities: ears, feet, and tail tip. The color of the points is determined by the genotype at other loci. Both White Galloway cattle (Figure 3.1) and White Park cattle exhibit this pattern. Family studies of White Galloway cattle show that this trait mapped to the *TYR* gene with a significant LOD score and complete concordance, but no mutation in the coding sequence was found (Schmidtz 2002). RNA from skin biopsies of the darkly pigmented ear showed expression of *TYR*, whereas skin biopsies of the white skin from the shoulder area did not (Schmidtz and Schmutz, unpublished data). This suggests that a mutation in the promoter sequence may be responsible for this trait.

Whether this "colored points" phenotype is the same phenotype referred to as "black-ear" by Chen et al. (1994) is not clear. They say that this gene was found in four local Chinese breeds and that they considered it common in cattle of tropical origin.

## *DEFB103*

*Beta-defensin 103*, formerly known as *DEFB300*, on cattle chromosome 27 is a gene that has only recently been shown to be involved in coat color in dogs (Candille et al. 2007) and cattle (Dreger and Schmutz 2010). Variant red is the name designated by the Holstein Associations of Canada and the United States for Holstein cattle that are red even though they have an $E^D$ allele at *MC1R*. This suggests that there is an epistatic relationship of a particular genotype of *DEFB103* and the *MC1R* genotype.

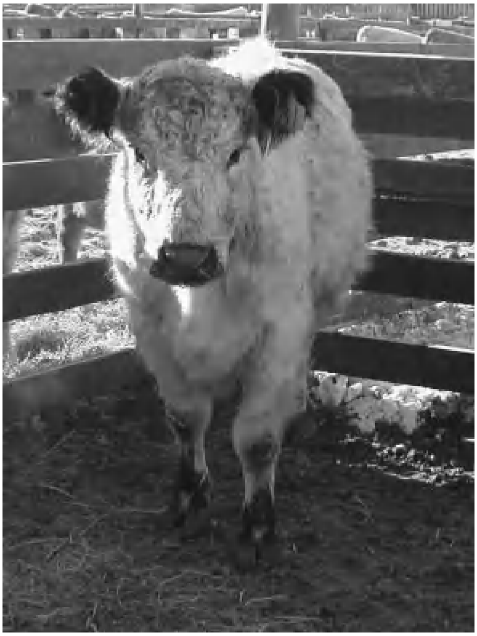


**Figure 3.1** A White Galloway illustrating the colored point pattern.

A multigeneration family of Holstein cattle with this rare phenotype cosegregated with a specific haplotype of 5 SNPs in the 5′ UTR of *DEFB103* (GenBank EU715240) (Dreger and Schmutz 2010). However, this haplotype also occurs in black cattle and hence, none of these variants is the causative mutation. *DEFB103* exists in at least five copies in some cattle (Dreger and Schmutz 2010) and therefore, this coat color could also be affected by copy number variation.

## Shades

There are many shades of the two basic coat colors, red and black. Gelbvieh were named "yellow cattle" even though they appear to be as red as a Limousin, although not as dark red as Maine Anjou. There is no known gene or allele that causes the subtle differences in shade of red in these cattle.

It has been suggested that bulls are often a darker red than cows, and that cattle may darken with age. Gilmore et al. (1961) examined the extent and location of black hairs in Jersey males, before and after castration. Although black hairs were not present at birth, they did develop and then were absent within 1 year of castration in some parts of the body such as the head, but not in others such as the legs.

Some genetics textbooks use the shade of the red on Ayrshire cattle as an example of a sex-influenced trait (Elrod and Stansfield 2002), meaning that the heterozygote was dark red in males and lighter red in females. No gene has been reported that would explain this phenomenon to date, but it has not been studied using molecular genetics either.

However, some genes have been studied in relation to the shade of color in cattle. The *silver* gene or *PMEL* has been shown to affect the shade of coat color in some breeds. This gene has been called *PMEL17* and *SILV* in the past.

## *PMEL*

Studies have been done on the *silver* gene (*PMEL*), which is composed of 11 exons, with the coding sequence contained in exons 2–11. Some of these studies suggest that specific *PMEL* mutations may act as alleles involved in the dilution of one or both pigments.

Kühn and Weikard (2007a) used crossbred calves with either a black-and-white or red-and-white Holstein parent and a Charolais parent to create 133 F2 offspring. They mapped diluted coat color to microsatellite ETH10 on BTA5 near the *PMEL* gene. They studied a c.64G>A mutation that has previously been reported in Charolais cattle (Oulmouden et al. 2005) and found a correlation with the shade of coat color. However, this mutation did not explain the total variation in shade observed. Further study indicated that additional splice variants of *PMEL* occurred in skin and other tissues (Kühn and Weikard 2007b).

A similar F2 and Backcross study was conducted by Gutiérrez-Gil et al. (2007), which also concluded that the *PMEL* c.64G>A mutation, acting in a codominant fashion, fit with the coat color in 93% of the cattle. Their study also suggested that a gene on BTA28, *lysosomal trafficking regulator* (*LYST*), might be involved in some subtle variation in the gray or intermediate shade of cattle.

A H2015R mutation in *LYST* was reported in Japanese black cattle with Chédiak–Higashi syndrome 1 (Kunieda et al. 1999). Such cattle typically have increased bleeding time and were a paler shade of black.

A 3-bp deletion in exon 1 of *PMEL* eliminates a leucine residue in the signal peptide region (GenBank EF363684). This deletion has been observed in all Highland cattle that are not a dark red or black (Schmutz and Dreger, in press). There are basically six solid coat colors in Highland cattle, in addition to brindle. The interaction of the alleles of *MC1R* and *PMEL* account for these six coat colors in Highland cattle (Table 3.3).

A limited number of Galloway cattle tested suggest that this same gene interaction will explain the solid coat colors in this breed. Although dun and silver dun are not uncommon in Galloway, any shade of red is rare.

**Table 3.3** Coat colors in Highland cattle resulting from the genotype at both *PMEL* and *MC1R*.

| *PEML* genotype | *MC1R* genotype | |
|---|---|---|
| | $E^D$/-(any second allele) | e/e or e/$E^+$ or $E^+$/$E^+$ |
| +/+ | Black | Red (dark red) |
| +/del | Dun | Yellow (light red) |
| del/del | Silver dun | Cream |

Some diluted black cattle have been described to have a thinner coat than usual. Because the tail coat is often very sparse, they have occasionally been dubbed "rat-tail." Schalles and Cundiff (1999) reported that all affected cattle had at least one allele for black coat color (i.e., $E^D$ of *MC1R*) and were heterozygous for another allele at another unidentified locus. Although the coat quality was affected, they found no significant difference in growth traits. Jolly et al. (2008) suggest that the second gene is *PMEL*, formerly known as *PMEL17* or *SILV*.

Olson (1999) reported observations of dilute black Simmentals, which still retain white spotting, with normal length and texture of hairs in the white areas and sparse or short hairs only in the gray areas. Such a condition in dogs is known as black hair follicular dysplasia and has been shown to be associated with mutations in the *MLPH* gene (Philipp et al. 2005). To the best of my knowledge, this gene has not been studied in such cattle.

## White Markings

### *KITLG*

Roan is a pattern that consists of intermingled pigmented and unpigmented hairs. Roan occurs in Shorthorn as red roan and Belgian Blue cattle as blue roan. The roan pattern is inherited as a codominant trait (Barrington and Peterson 1906). Heterozygous cattle are roan, and cattle homozygous for the mutation are almost entirely white. Roan was mapped to BTA5 in Belgian Blue cattle by Charlier et al. (1996). An Ala193Asp mutation in the *KIT ligand* gene (*KITLG*), formerly called the *mast cell growth factor* gene (*MGF*), was identified as the causative mutation in both Shorthorn and Belgian Blue cattle (Seitz et al. 1999). The amount of roan on heterozygous cattle varies among individuals. How this mutation functions to eradicate pigment in only some hairs remains a mystery.

Considerable earlier research suggested that some white cattle in both breeds had fertility problems, commonly known as white heifer disease (Charlier et al. 1996). Bulls were not reported to have fertility problems and not all white heifers were sterile.

### *KIT and White Spotting Patterns*

White spotting or white markings occur in many breeds of cattle such as Holsteins, Simmental, Hereford, etc. There are white markings such as a white belt around the middle section of the animal that is called belted. There are also patterns of white markings such as white undersides and a white area along the spine with pigmented areas along the sides of the torso, which is a phenotype called "color sided."

In pigs, several types of white markings, including complete white, are caused by various mutations in the gene *KIT* (reviewed in Andersson 2009). Several researchers have suggested that *KIT* is a good candidate gene for at least some types of white spotting. Grosz and MacNeil (1999) mapped the pattern of Hereford cattle, which is a white face, legs, and underside (Figure 3.2A) to region on BTA6 that includes *KIT*. In the same year, Reinsch et al. (1999) published that there was a strong correlation between the amount of white on Holsteins (Figure 3.2B) and markers near *KIT*. They

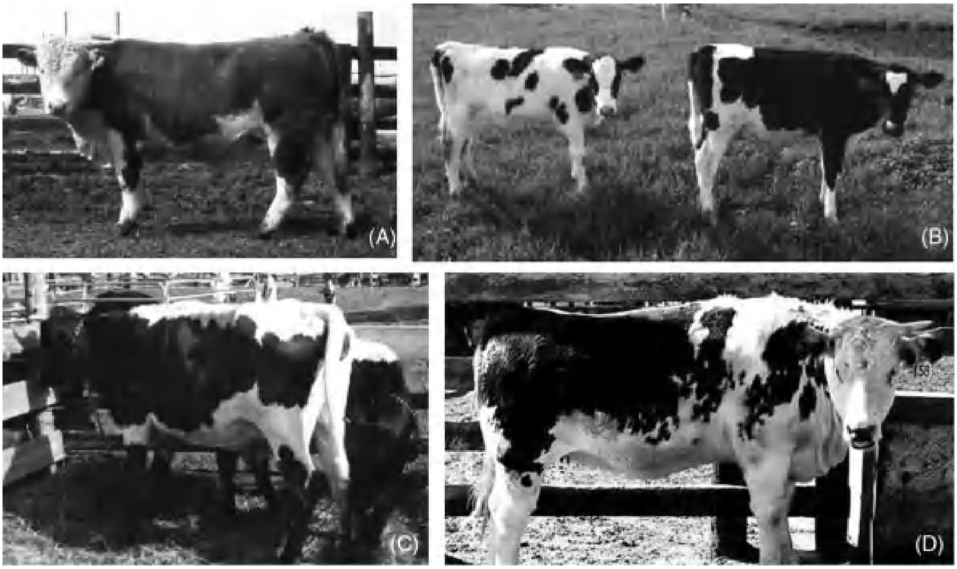


**Figure 3.2** (A) A Hereford bull illustrating the Hereford pattern. (B) Two Holstein calves illustrating the variability in the amount of white spotting. (C) A Pinzgauer illustrating the color-sided pattern with solid red. (D) A crossbred heifer illustrating the color-sided pattern with some speckling in the pigmented areas.

suggested that *KIT* was a quantitative trait loci (QTL) for the amount of white. Liu et al. (2009) used a genome scan approach in F2 Holstein × Jersey cattle, with SNPs and microsatellites to follow the quantity of white. They also found a significant peak on BTA6 in the region including *KIT*. They suggested that another significant peak was found on BTA22 in the region including the *microphthalmia related transcription factor* gene (*MITF*). Other studies have used certain coat color traits or patterns to verify approaches such as genome-wide analysis or HapMap strategies (Stella et al. 2010).

The color-sided pattern occurs occasionally in Brown Swiss. It is apparently considered lucky to have one in a dairy herd. Leeb (2008) presented data that this pattern was linked to *KIT*. This is the pattern of Pinzgauer cattle (Figure 3.2C). In some breeds, the color portion of the torso is speckled (Figure 3.2D) rather than solid colored (Olson 1999).

Interestingly, no researcher has published a mutation in the *KIT* gene or promoter region that could be a causative mutation for any white marking pattern in cattle yet, even though several years have passed. Is this because some white marking patterns are essentially fixed in a breed? Would there ever be a Hereford without the characteristic white markings? Or a Holstein that is solid black or red? There has been a movement in the United States away from white spotting in Simmental, but grading up is allowed in North America so this loss of white spotting is likely through the integration of alleles from other breeds such as Angus.

## *Belted*

The belted pattern occurs in Dutch Belted or Lakenvelder, Belted Galloway, and occasionally in Brown Swiss cattle. The white belt around the midsection of the cattle

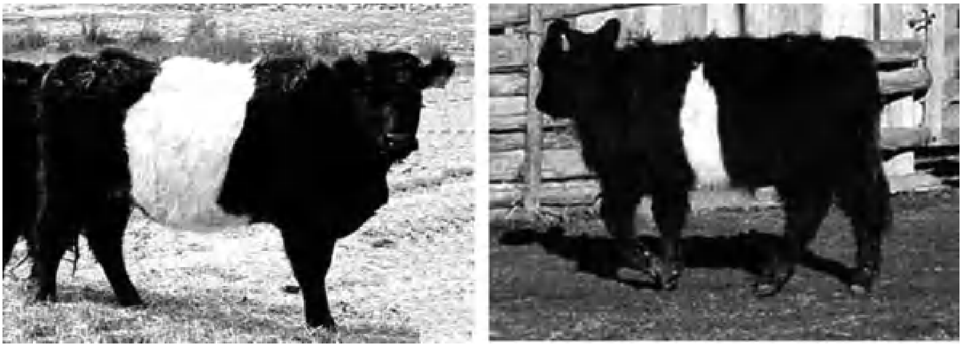


**Figure 3.3** Two belted Galloway cattle illustrating a complete belt on the left, and an incomplete belt on the right.

varies in size and shape (Figure 3.3). If it does not completely encircle the midsection, some breed associations do not allow the animal to be registered as a belted animal. It is inherited as a dominant trait (Schmutz et al. 2001; Drögemüller et al. 2009).

Drögemüller et al. (2009, 2010) have mapped the belted pattern to a 336-kb region near the telomeric end of BTA3, using a genome scan. This region corresponds to a region on human chromosome 2 and mouse chromosome 1. The candidate gene in this region that appeared the most promising was a transcription factor, *HES6* for *hairy enhancer of split 6* from *Drosophila*. Although the complete coding sequence of *HES6* was studied, and they identified ten polymorphisms, no polymorphism cosegregated with the belted pattern.

## Unsolved Colors and Patterns

Some colors or patterns discussed in the preceding pages have been mapped to a chromosomal region and perhaps even to a candidate gene in that region, but the specific mutation has not yet been identified. These include color pointed, color sided, belted, and variant red.

Although Brown Swiss or Braunvieh cattle are called brown, there appears to be no mutation in the gene causing brown pigmentation in cattle and other species, *TYRP1*. Therefore, what gene is causing the taupe brown color in this breed?

The white to gray colors common in many *B. indicus* breeds may also be due to mutations in *PMEL*, but to the best of my knowledge, have not been studied. However, the shading on some Brahma (Figure 3.4) would not be explained by *PMEL*.

White facial markings such as a blaze (Figure 3.5A) have not been studied. Likewise, the dark goggles that occur in some cattle with white faces, such as Herefords and some Simmentals (Figure 3.5B), have not been studied. Brockle face is the name given to irregular pigmented patches on a white-faced animal (Figure 3.5C). The gene that causes this pattern has not been identified.

Oculocutaneous albinism, known as OCA, has been reported in Angus cattle (Cole et al. 1984). The main feature of this phenotype is a loss of pigmentation in the eye. However, the black coat color is typically also tinged a rust color. It is considered to be inherited as an autosomal recessive trait. A mutation has not been identified

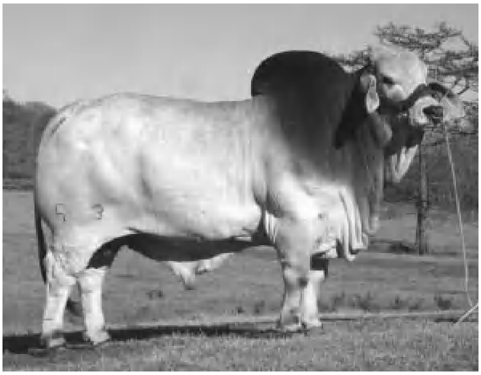


**Figure 3.4** A Brahma bull with dark-shaded gray over the hump and very pale gray on the rest of the body. Courtesy of J.D. Hudgins Inc.

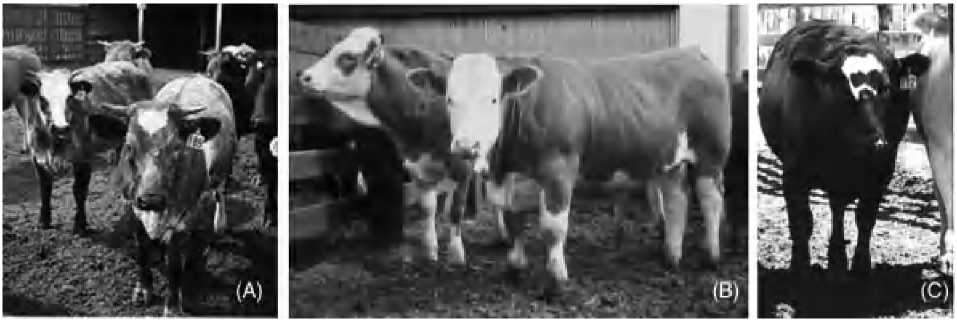


**Figure 3.5** White facial markings of various types. (A) The crossbred roan steer on the left has a large white blaze and the one on the right, a small marking sometimes called a star. (B) Two Simmental cows with white faces, although the one on the left also has goggles. (C) A brockle-faced crossbred steer.

during gene sequencing of obvious candidate genes in affected cattle, such as *TYR* and *TYRP1* (Genbank AF400250) (Schmutz et al., unpublished data). Another form of ocular albinism, accompanied by hearing loss, was found to occur in a family of Fleckvieh cattle caused by a R210I mutation in *MITF* (Philipp et al. 2011).

## References

Andersson, L. (2009) Studying phenotypic evolution in domestic animals: a walk in the footsteps of Charles Darwin. *Cold Spring Harbor Symposium Quantitative Biology* **74**: 319–325.

Barrington, A. and Peterson, K. (1906) On the inheritance of coat color in cattle. I. Shorthorn crosses and pure Shorthorn. *Biometrica* **4**: 427–437.

Berryere, T.G., Schmutz, S.M., Schimpf, R.J., Cowan, C.M., Potter, J. (2003) *TYRP1* is associated with dun brown coat colour in Dexter cattle or how now brown cow? *Animal Genetics* **34**: 169–175.

Candille, S.J., Kaelin, C.B, Cattanach, B.M., Yu, B., Thompson, D.A., Nix, M.A., Kerns, J.A., Schmutz, S.M., Millhauser, G.L., Barsh, G.S. (2007) A beta-defensin mutation causes black coat color in domestic dogs. *Science* **318**: 1418–1423.

Charlier, C., Denys, B., Belanche, J.I., Coppieters W., Grobet, L., Mni, M., Womack, E.J., Hanset, R., Georges, M. (1996) Microsatellite mapping of the bovine roan locus: a major determinant of White Heifer disease. *Mammalian Genome* **7**: 138–142.

Chen, Y., Wang, Y., Cao, H, Pang, Z., Yang, G. (1994) Black-ear gene and blood polymorphism in four southern Chinese cattle groups. *Animal Genetics* **25**(Suppl 1): 89–90.

Cole, D., Leipold, H., Schalles, R. (1984) Oculocutaneous hypopigmentation of Angus cattle. *Bovine Practitioner* **19**: 92–99.

Deobald, H.M. (2009) Characterization of pro-opiomelanocortin gene variants and their effect on carcass traits in beef cattle. M.Sc. Thesis. University of Saskatchewan, Saskatoon, SK, Canada.

Dreger, D.L. and Schmutz, S.M. (2010) The variant red coat colour phenotype of Holstein Cattle Maps to BTA27. *Animal Genetics* **41**: 109–112.

Drögemüller, C., Engensteiner, M., Moser, S., Rieder, S., Leeb, T. (2009) Genetic mapping of the belt pattern in Brown Swiss cattle to BTA03. *Animal Genetics* **40**: 225–229.

Drögemüller, C., Demmel, S., Engensteiner, M., Rieder, S., Leeb, T. (2010) A shared 336 kb haplotype associated with the belt pattern in three divergent cattle breeds. *Animal Genetics* **41**: 304–307.

Elrod, S.L. and Stansfield, W.D. (2002) *Schaum's Outline of Theory and Problems in Genetics*. 4th edition, New York: McGraw-Hill Professional.

Gilmore, L.O., Fechheimer, N.S., Baldwin, C.S. (1961) Inheritance of black hair patterns in cattle lacking the extension factor for black (E). IV. Partitioning phenotypes by castration. *Ohio Journal of Science* **61**: 273–277.

Girardot, M., Guibert, S., Laforet, M.P., Gallard, Y., Larroque, H., Oulmouden, A. (2006) The insertion of a full-length *Bos taurus* LINE element is responsible for a transcriptional deregulation of the Normande Agouti gene. *Pigment Cell Research* **19**: 346–355.

Graphodatskaya, D., Joerg, H., Stranzinger, G. (2002) Molecular and pharmacological characterisation of the MSH-R alleles in Swiss cattle breeds. *Journal of Receptor Signal Transduction Research* **22**: 421–430.

Greene, H.J., Leipold, H.W., Gelatt, K.M., Huston, K. (1973) Complete albinism in beef Shorthorn calves. *Journal of Heredity* **64**: 189–192.

Grosz, M.D. and MacNeil, M.D. (1999) The "spotted" locus maps to bovine chromosome 6 in a Hereford-Cross population. *Journal of Heredity* **90**: 233–236.

Gutiérrez-Gil, B., Wiener, P., Williams, J.L. (2007). Genetic effects on coat colour in cattle, dilution of eumelanin and phaeomelanin pigments in an F2-Backcross Charolais×Holstein population. *BMC Genetics* **16**: 56–67.

Joerg, H., Fries, H.R., Meijerink, E., Stransinger, G.F. (1996) Red coat color in Holstein cattle is associated with a deletion in the *MSHR* gene. *Mammalian Genome* **7**: 317–318.

Jolly, R.D., Wills, J.L., Kenny, J.E., Cahill, J.I., Howe, L. (2008) Coat-colour dilution and hypotrichosis in Hereford crossbred calves. *New Zealand Veterinary Journal* **56**: 74–77.

Klungland, H., Vage, D.I., Gomez-Raya, L., Adalsteinsson, S., Lien, S. (1995) The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination. *Mammalian Genome* **6**: 636–639.

Kühn, C. and Weikard, R. (2007a) An investigation into the genetic background of coat colour dilution in a Charolais × German Holstein F2 resource population. *Animal Genetics* **38**: 109–113.

Kühn, C. and Weikard, R. (2007b) Multiple splice variants within the bovine silver homologue (*SILV*) gene affecting coat color in cattle indicate a function additional to fibril formation in melanophores. *BMC Genomics* **24**: 335–348.

Kunieda, T., Nakagiri, M., Takami, M., Ide, H., Ogawa, H. (1999) Cloning of bovine LYST gene and identification of a missense mutation associated with Chediak-Higashi syndrome of cattle. *Mammalian Genome* **10**: 1146–1149.

Leeb, T. (2008) White patterns in Horse and Cattle. Comparative Genetics of Coat Color Workshop, International Society of Animal Genetics Conference, Amsterdam. June 26, 2008.

Leipold, H.W., Huston, K., Gelatt, K.N. (1968) Complete albinism in a Guernsey calf. *Journal of Heredity* **59**: 218–220.

Liu, L., Harris, B., Keehan, M., Zhang, Y. (2009) Genome scan for the degree of white spotting in dairy cattle. *Animal Genetics* **40**: 975–977.

Mohanty, T.R., Seo, K.S., Park, K.M, Choi, T.J., Choe, H.S., Baik, D.H., Hwang, I.H. (2008) Molecular variation in pigmentation genes contributing to coat colour in native Korean Hanwoo cattle. *Animal Genetics* **39**: 550–553.

Olson, T.A. (1999) *The Genetics of Cattle*, edited by R. Fries and A. Ruvinsky. New York: CABI Publishing.

Oulmouden, A., Julien, R., Laforet, J.M., Leveziel, H. (2005) Use of silver gene for authentication of the racial origin of animal populations, and of the derivative products thereof. Patent WO2005/019473.

Petersen, W.E., Gilmore, LO., Fitch, J.B. (1944) Albinism in cattle. *Journal of Heredity* **35**: 135–144.

Philipp, U., Hamann, H., Mecklenburg, L., Nishino, S., Mignot, E., Schmutz, S.M., Leeb, T. (2005) Polymorphisms within the canine MLPH gene are associated with dilute coat colour in dogs. *BMC Genetics* **6**: 34–49.

Philipp, U., Lupp, B., Mömke, S., Stein, V., Tipold, A., Eule, J.C., Rehage, J., Distl, O. (2011) A *MITF* mutation associated with a dominant white phenotype and bilateral deafness in German Fleckvieh cattle. *PLoS ONE* **6**(12): e28857.

Reinsch, N., et al. (1999) A QTL for the degree of spotting in cattle shows synteny with the KIT locus on chromosome 6. *Journal of Heredity* **90**: 629–634.

Rouzaud, F., Martin, J., Gallet, P.F., Delourme, D., Goulemot-Leger, V., Amigues, Y., Ménissier, F., Levéziel, H., Julien, R., Oulmouden, A. (2000) A first genotyping assay of French cattle breeds based on a new allele of the extension gene encoding the melanocortin-1 receptor (*MC1R*). *Genetics Selection and Evolution* **32**: 511–520.

Royo, L.J., Alvarez, I., Fernández, I., Arranz, J.J, Gómez, E., Goyache, F. (2005) The coding sequence of the ASIP gene is identical in nine wild-type coloured cattle breeds. *Journal of Animal Breeding and Genetics* **122**: 357–360.

Schalles, R.R. and Cundiff, L.V. (1999) Inheritance of the "rat-tail" syndrome and its effect on calf performance. *Journal of Animal Science* **77**: 1144–1147.

Schleger, W. (1959) Auftreten eines Albinokalbes bei der Murbodnerrasse. *Wiener Tierärztliche Monatsschrift* **46**: 196–199.

Schmidtz, B.H. (2002) Characterization and mapping of cattle tyrosinase. M.Sc. Thesis, University of Saskatchewan, Saskatoon, SK, Canada.

Schmutz, S.M., Berryere, T.G., Moker, J.S., Bradley, D.J. (2001) Inheritance of the belt pattern in Belted Galloway cattle. Plant & Animal Genome IX Conference, San Diego, CA.

Schmutz, S.M., Berryere, T.G., Ciobanu, D.C., Mileham, A.J., Schmidtz, B.H., Fredholm, M. (2004) A form of albinism in cattle is caused by a tyrosinase frameshift mutation. *Mammalian Genome* **15**: 62–67.

Schmutz, S.M. and Dreger, D.L. (2012) Interaction of *MC1R* and *PMEL* alleles on solid coat colors in Highland Cattle. *Animal Genetics* **in press**.

Seitz, J.J., Schmutz, S.M, Thue, T.D., Buchanan, F.C. (1999) A missense mutation in the bovine MGF gene is associated with the roan phenotype in Belgian Blue and Shorthorn cattle. *Mammalian Genome* **10**: 710–712.

Stella, A., Ajmone-Marsan, P., Lazzari, B., Boettcher, P. (2010) Identification of selection signatures in cattle breeds selected for dairy production. *Genetics* **185**: 1451–1461.

Winzenried, H.U. and Lauvergne, J.J. (1970) Spontanes Auftreten von Albinos in der Schweizerischen Braunviehrasse. *Schweizer Archiv fur Tierheilkunde* **112**: 581–587.

# Chapter 4
# From Quantitative Genetics to Quantitative Genomics: A Personal Odyssey

*Morris Soller*

> Quantitative Genetics: A science of heredity of quantitative traits based on inferences from observed phenotypes.
>
> Quantitative Genomics: A science of heredity of quantitative traits based on inferences from observed phenotypes and DNA structure.

I had always had a liking for genetics—ever since I came across a copy of T.H. Morgan's book *The Theory of the Gene* (Morgan 1926) at age 12, which had somehow gotten into the children's section of our neighborhood Carnegie library, and became enamored of the full page facing color photos of a normal "red" *Drosophila* eye and of the "white eye" mutant, the first *Drosophila* mutant to be discovered. In my senior year in high school, three classmates and I decided to do a genetics project for the Westinghouse science competition—we thought a microorganism would be a good model system. Our school provided us with a spare room to use as a lab. We constructed an incubator out of an old refrigerator and pooled our savings to buy a used binocular microscope. A search of Kudo's classic *Protozoology* text (Kudo 1946) turned up *Paramecia aurelia* as a protozoan that could be crossed. We wrote to Tracy Sonnenborne at Indiana State University who had worked out the *P. aurelia* mating types, asking for cultures. He replied generously, remarking that he was also planning on doing genetic studies in *P. aurelia* and hoped we would not mind his "treading on our toes." All was going along swimmingly, when one of our group did some extracurricular chemistry in our lab that blew up in his face (no lasting damage, but over 100 glass splinters had to be removed), and our lab privileges were revoked.

Motivated by a desire to participate in the rebuilding of the Jewish nation in its ancestral homeland, I decided to pursue a BS degree in agriculture, and ended up a Dairy Science major at Rutgers University in New Jersey. In 1951, my last year of undergraduate studies, one of my instructors gave me a copy of J.L. Lush's book *Animal Breeding Plans* (Lush 1943)—I had been complaining to him about the lack of intellectual challenge in some of my dairy science courses, and he said "Read this Morris—it will be an intellectual challenge." Indeed it was. More than that, here was a way to combine dairy science and genetics! I went on to do a senior thesis and later a

*Bovine Genomics*, First Edition. Edited by James E. Womack.

PhD thesis (also at Rutgers) on dairy sire progeny testing, which at the time was based on dam–daughter comparisons. The puzzle I dealt with in my PhD thesis was the fact that in spite of intense pedigree selection of candidate sires for progeny testing, on the average, the mean production of the dams was distinctly better than their daughters. In retrospect, the solution was obvious—the dams were a selected group, their daughters were unselected and hence, regressed toward the mean. Identifying and documenting this took a couple of years! In any event, in my senior thesis I was so impressed by the beauty of the dam–daughter comparison that I concluded that it could not be improved upon, and suggested that the next step would be to search for associations between Mendelian genetic traits and production traits, giving "wry neck" as an example of a Mendelian trait—at the time I was not aware of the just published studies of Clyde Stormont and the Madison Laboratory on cattle blood group genetics (Stormont et al. 1951).

In 1957, with PhD in hand, my family and I made the move to Israel, where I worked at the Department of Animal Science at the Israel Agricultural Research Station, the Volcani Institute (now the Agricultural Research Organization, ARO) at Bet Dagan, and also taught genetics and general biology at Bar-Ilan University—a new university that had just opened its doors the year before. Although there were a number of highly competent animal breeders in Israel, I was the first to have professional training in the new biometrical methodologies. I was warmly welcomed and rapidly integrated into all aspects of Israel animal breeding. Those were exciting times. In addition to *Animal Breeding Plans*, I.M. Lerner proposed selection of layer chicken males on the performance of their half-sisters (Lerner 1950), and Alan Robertson defined the four avenues of genetic improvement enabling a systematic analysis and comparison of breeding plans (Robertson and Rendel 1950). In dairy cattle, Alan further proposed the contemporary comparison for progeny tests in place of the dam–daughter comparison (Johansson and Robertson 1952), and the "waiting period" design for sire selection (Robertson and Rendel 1950). In this design, a large number of young sires are progeny tested on a limited number of daughters each, and then kept in "waiting" until the progeny tests become available to choose the very best for widespread service. Remarkably, when I came to Israel, I found this scheme already being implemented by Reuven Bar-Anan and Moshe Heiman at the Hasherut AI Center. This was 4 years before the introduction of large-scale progeny testing of young bulls by the English Milk Marketing Board in 1961, and may have been the first actual implementation of this scheme anywhere in the world! At the same time, broiler breeding in Israel was taking off, and in between these two main pillars of activity, there were layer chickens, sheep and goats, bees, and fish. I was busy gathering data, calculating heritabilities and genetic correlations, and then plugging them into animal breeding plans à la Lush, Lerner, and Robertson, searching for optimal four-path selection combinations to maximize genetic progress. Our main contribution in this period was to introduce the concept of "present value" into animal breeding (Soller et al. 1966), which led in other hands to the ability to evaluate the expected summed future benefits of a breeding program.

However, by 1966, I came to feel that there must be more to life than heritabilities and genetic correlations—once again it seemed to me that animal breeding methodology had come as far as it could go. With BLUP (Best Linear Unbiased Predictor) and computers just a few years down the road, how mistaken that was! At the same time, students of mine at Bar-Ilan University got me interested in behavior genetics. We did an experiment on brain size and maze running in mice, and DNA was just shaking the scientific tree, with all sorts of fascinating fruits dropping off. This led me to a postdoctorate in brain biochemistry with focus on DNA in Henry Mahler's laboratory at Indiana State University, Bloomington.

We came to Chicago (my wife's birthplace) in summer of 1966, just in time for the meeting of American Society for Human Genetics, where I gave an oral presentation of our Bar-Ilan mouse work, which was eventually published in *Behavior Genetics* (Padeh and Soller 1976). At the meeting, J.M. Thoday gave a plenary session talk on his work on polygene mapping in *Drosophila* and made some suggestions for implementing this in human genetics (Thoday 1966). That struck a responsive chord and remained in my mind. Meanwhile, after the year at Bloomington I realized that I needed much more training to do work in biochemistry, and spent the next 6 years as a visiting scholar in Biochemistry at Northwestern University in Chicago, working in Harold Koenig's Neurochemistry Laboratory at Veterans Administration (VA) Research Hospital in downtown Chicago; at the same time I was teaching genetics and biology at Roosevelt University, a mile or two down Michigan Avenue from the VA Hospital.

In 1970, The Department of Genetics at The Hebrew University of Jerusalem initiated a teaching group on applied genetics, aimed at training students for work in plant and animal genetic improvement, and I was invited to join. This brought the family back to Israel, and myself to animal breeding in the spring of 1972. It turned out to be an auspicious move. The discovery of widespread isozyme variation in *Drosophila pseudoobscura* and other natural populations by Richard Lewontin was revolutionizing population genetics (Lewontin and Hubby 1966); and Daniel Zohary of the Hebrew University and Eviatar Nevo of University of Haifa were pioneering the application of this methodology in wild populations of barley (Nevo et al. 1979) and wheat (Nevo et al. 1982). To further explore the potential of isozyme markers for genetic improvement, Zohary and his colleague Rom Moav established a laboratory for molecular markers as part of the applied genetics group. Thus, I had joined a department that was actively pursuing applications of molecular markers in agricultural plants and animals.

As part of my teaching responsibilities at the Hebrew University, I taught the course in quantitative genetics. Traditionally, a course in quantitative genetics focused on the *quantitative*-half of the title (analysis of variance, heritability, prediction equations, and so on) based on the assumption that a very large number of loci, individually of very small effects, were the source of genetic variation in quantitative traits (the "infinitesimal model"). To make things a bit more interesting, I decided to focus the course on the *genetics*-half of the title, on the basis of the idea that at least some fraction of quantitative genetic variation might be due to individual loci of detectable effect. At the time, this was known as the "polygene" model, which attributed genetic variation in quantitative traits to the joint action of a large (but not "infinite") number of nonallelic loci, of small but individually appreciable effect.

To emphasize this shift, I renamed the course "Genetics of Quantitative Traits." A literature search uncovered a surprisingly large number of studies supporting the polygene notion. Among the most impressive were the "inbred backcross" lines of Wehrhahn and Allard (1965) that separated out the polygenes contributing to the difference between two pure lines of wheat in a quantitative trait (heading date) into a series of inbred lines; and the "chromosome substitution line" studies of E.R.

Sears (Sears 1969), in which individual chromosomes of the Hope variety of wheat were substituted for the corresponding chromosome of the Chinese Spring variety, while retaining all other Chinese Spring chromosomes. Each substituted chromosome had its own spectrum of quantitative effects, indicating that the genetic content of the individual chromosomes differed widely with respect to quantitative effects. Subsequent studies by Law (1966) of the chromosome 7 substitution line used genetic markers to localize specific quantitative effects to specific regions within the substituted wheat chromosome. Even more supportive of the polygene model were the *Drosophila* studies of Thoday and colleagues on genetics of sternopleural bristle number in *Drosophila*. These studies were able to localize genetic effects on bristle number to defined chromosomal regions, using *Drosophila* mapping lines marked by morphological traits, and specialized "chromosome-dissection" methodologies (Spickett and Thoday 1966). Thoday (1961) took this a step further showing how the quantitative gene effects could be localized to a point location by means of progeny testing, using a "moment" method to implement what is today termed "interval mapping," that is, using information from a marker bracket to locate a quantitative trait locus (QTL) more precisely within the bracket. This was "QTL mapping" in the fullest sense of the word, but still discussed in terms of "polygenes."

Both the wheat chromosome substitution and the *Drosophila* bristle number experimental designs were based on species-specific genetic and reproductive infrastructures, while the backcross inbred design was limited to selfers and inbred lines and required many generations to implement. Consequently, I did not view these experiments as being directly applicable to polygene mapping in agricultural plants or animals. Amazingly, I was unaware at the time of the pioneering experiments by Sax (1923), now so widely cited, showing how standard linkage experiments could associate quantitative effects with genetic markers. In truth, that is not so surprising. The paper is not cited in *Animal Breeding Plans*, or in the early editions of Falconer's *Introduction to Quantitative Genetics* (Falconer 1960), the much loved standard text for quantitative genetics in outcrossing animal populations. Sax's paper was well known to the plant biometrical groups, and is cited in Mather and Jinks (1949) text *Biometrical Genetics* (Mather and Jinks 1949). But there was little contact between the plant and animal groups. The animal breeding groups (centered in Ames and Edinburgh) were focused on quantitative genetics of outcrossing populations. The plant breeding groups (centered in North Carolina and Birmingham) were focused on quantitative genetics of selfers. The two groups even used different notations for the same statistical entities. I was only made aware of the Sax paper some time later by my colleague Giora Simchen at the Hebrew University, who had received his PhD in quantitative genetics at Birmingham.

Instead, the turning point in my thinking came from Spickett and Thoday (1966), in which they used the usual specialized *Drosophila* chromosome-dissection methods to identify interactions between mapped polygenes affecting bristle number. Along the way, however, they did a simple F2 experiment using *Drosophila* morphological markers, and noted that the results of the simple experiment were very similar to those obtained by the much more complex chromosome-dissection methods. Reading this, I looked at the experiment and said to myself "I can do that!" In other words, given the genetic markers, this is something we could do in cattle or tomato. At this point, my quantitative genetics and biometrical training kicked in, and I began to ask the basic experimental design questions: “What would be the nature of the crosses to uncover polygenes in plants (selfers) and in animals (outcrossers)?” “What will be the statistical power of these experiments as a function of number of individuals sampled, and effect of the quantitative locus—can such experiments be implemented with reasonable likelihood of success in populations of feasible size?” “How could the information be used in actual breeding practice?” These questions led to the two basic QTL-mapping design papers with my colleague Avraham Genizi: the F2 and BC design for mapping in pure lines (Soller et al. 1976), and the full-sib and half-sib design for mapping in outcrossing populations (Soller and Genizi 1978); and to a pair of application papers for marker-assisted introgression (Soller and Plotkin-Hazan 1977) and marker-assisted selection (MAS) (Soller 1978). It should be noted that Hermann Geldermann had even earlier explored in some detail, various aspects of estimating quantitative effects of marked chromosome segments and utilization of this information in breeding (Geldermann 1972; 1975). In these papers, he also introduced the term “quantitative trait locus” to refer to the specific element at the DNA level that was being mapped. The acronym “QTL” became widespread shortly thereafter, but I am not clear on its provenance. It was only in 1978 that I became aware of Geldermann’s work, when a referee of my 1978 paper on MAS in dairy cattle brought it to my attention, and for this reason it was not cited in my early papers.

With the theory in place, the challenge was to test these ideas in practice. Together with Tom Brody, who directed the marker laboratory, we considered using an animal model (chicken) as we had facilities for working with chickens at the Hebrew University. However, the only markers readily available for chicken were blood group markers (Briles 1984), and although we purchased a set of reagents from W.E. Briles, the logistics of a major chicken experiment were too daunting. Instead, at Tom’s suggestion we turned to a tomato model, with fewer chromosomes, availability of morphologically marked tester lines, and many known isozyme markers pioneered by Charles Rick (Tanksley and Khush 2004). Tom idolized Rick for his work with tomato, and we actually wrote and asked Rick if he would join us in a BARD (Binational Agricultural Research and Development Foundation) proposal for QTL mapping in tomato, but he was unable to do so. In any event, we received invaluable guidance in design and implementation of the experiment from my former colleagues Raphael Frankel and Dvorah Lapushner of Department of Field Crops at ARO. At their suggestion, to maximize exposure of QTL, we used the F2 of a cross between a morphologically marked line of domesticated tomato (*Lycopersicon esculentum*) and a close wild relative (*Lycopersicon pimpinellifolium*) bearing very small berry-like fruit. The cross between the two produces fully fertile hybrids. The parental strains differed in six morphological and four isozyme markers, giving us a total of ten markers for analysis. In all, we reared a total of 1684 valid plants, and measured 18 different quantitative traits in each plant. The morphological markers were scored on the F2 plants. The electrophoretic markers were scored on three F3 seedlings from each F2 plant, to determine whether the F2 parent was homozygous or heterozygous. The ARO group carried out the initial crosses, generating the F1 and F2 seedlings, and all fieldwork needed to plant and rear the F2 seedlings to maturity. They also suggested and defined for us the traits to be measured. The fieldwork of phenotyping and genotyping was done by myself, Tom, and Joel Weller, as his PhD research. The experiment was completed in 1981, and Joel’s PhD awarded by 1984. A large number of highly significant effects were uncovered, many more than would be expected by chance alone, and we were even able to explore epistatic relations among the identified QTL. Writing up this complex experiment and getting it through the review process at *Genetics* took ages, so the publication was not until 1988 (Weller et al. 1988). Indeed, the paper was first rejected by *Genetics* on the grounds that it was too specialized, and we had to take out Joel's maximum likelihood estimate of QTL location and publish this separately in *Biometrics* (Weller 1986).

In part, however, the delay was on my side, due to a seminal event that took place in 1980 and marked the true beginning of the transformation of quantitative genetics into quantitative genomics. This was the realization by Solomon and Bodmer (1979) and independently by Botstein et al. (1980) that restriction length polymorphisms first identified at the hemoglobin gene by Kan and Dozy (1978) using Southern blot methodology (Southern 1975) could serve as a new class of genetic marker, identified at the DNA level. These markers, promised to be orders of magnitude more plentiful than the morphological, biochemical, electrophoretic, and immunological markers available until then.

I first became aware of RFLP (restriction fragment length polymorphism) markers in 1980, through a chance meeting with my future colleague, Jacques (Jacqui) Beckmann. Jacqui had recently joined ARO, after his PhD (Weizmann Institute) and postdoc training in molecular biology at Edinburgh and San Diego. Inspired by the successes of the newly discovered *Agrobacterium tumefaciens* system, his original intention at ARO was to work on transforming plant cells, and he had a BARD grant for this with Ron Davis. After attending a seminar at the ARO on use of isozyme markers to identifying paternal and maternal contributions to avocado pollination, it struck Jacqui that there might be a better way at the DNA level, with possibly unlimited number of easily accessible markers. This was stimulated by a landmark paper in *Cell* by Alec Jeffreys, showing that when he digested human DNA from some of his lab colleagues, ran it on a gel, and performed a Southern blot with Hb cDNA as a probe, the size of the bands differed from one individual to another (Jeffreys 1979). Although Jeffreys conceived of this as providing a means to assess DNA sequence variants in man, Jacqui realized that this in itself constituted a genetic marker; in fact, it was an RFLP, but that was before the term was coined. Based on this, Jacqui conceived the idea of searching for such markers for parental identification in plants. He shared this with Giora Simchen, who encouraged him and connected him with a brilliant MS student, Moshe Rom, and together they embarked on a quest for this new class of genetic marker in tomato. In 1980, Jacqui met Ron Davis, his BARD co-PI and one of the coauthors on the Botstein et al. (1980) RFLP paper, at an EMBO (European Molecular Biology Organization) meeting in Heidelberg, and shared these thoughts with him. Ron mentioned that they had just published a paper on exactly this class of marker (which had meanwhile been termed "restriction fragment length polymorphism"), but had not thought of its applications in any species but man. This was the first Jacqui had heard of RFLPs. At that same meeting, Jacqui met and befriended Frances and Ben Burr, and shared his idea on the use of RFLPs in plant breeding with them. They liked the idea, and asked Jacqui if it was acceptable on his side if they implemented it in corn, which was their experimental species. Jacqui agreed, and the Burrs went ahead to demonstrate RFLPs in corn with S. Evola as postdoc. They graciously acknowledged Jacqui's contribution by including him as a coauthor on their 1983 paper (Burr et al. 1983).

At the time, Jacqui was coming up weekly to Jerusalem to teach a course on yeast genetics, and on this particular day I was walking the corridor passing my colleague Giora Simchen talking to Jacqui. Giora who was aware of my interest in genetic markers stopped me and said "Moshe, here is someone you should meet," introducing me to Jacqui. On the spot, Jacqui made me aware of the existence of RFLPs and shared with me his thought that they would be useful in plant parentage identification. I recall replying "We will do better than that—with RFLPs we will revolutionize plant and animal breeding." Jacqui and I then wrote our mega opus on RFLPs and their potential use for parentage determination and genetic improvement. We sent the paper to *Theoretical and Applied Genetics* (*TAG*) in 1981 with Alan Robertson as the editor. He recommended that it be rewritten as two separate papers, which we did and resubmitted in 1982. Sadly, Alan was already in the initial stages of Alzheimer's disease, and the papers sat on his desk for a year, before finally being published in 1983 (Soller and Beckmann 1983; Beckmann and Soller 1983).

In the meantime, in spring of 1982, I gave a round of seminars in the United States and Great Britain, introducing our tomato QTL mapping experiment and RFLPs to groups at North Carolina State University, USDA (United States Department of Agriculture) Beltsville, Cornell University, and Birmingham University in England, ending up with a presentation at the British Poultry Breeders' Roundtable in Edinburgh. Charles Stuber was at my seminar at North Carolina and he liked the idea and implemented a similar experiment in corn using isozyme markers, after first writing to ask if I would mind! (Edwards et al. 1987). Amazingly, both of us had been preceded by almost 10 years by Abraham Korol and his colleagues in Moldava (Zhuchenko et al. 1975, 1978, 1979), who had followed a very similar line of thought. Korol later repatriated to Israel and continues as a leading figure in QTL mapping and analysis.

The seminar series was followed in the same summer by a self-invited workshop presentation at the 2nd World Congress Genetics Applied to Livestock Production, Madrid, 1982 (Soller and Beckmann 1982). At this meeting, I introduced the potential of RFLP markers for QTL mapping and MAS, and the talk aroused considerable interest. All of this public relations activity and our two 1983 papers in TAG led to invitations in 1985 to speak at the Armidale Conference on Biotechnology Applied to Livestock Production in Australia; to the Symposium on Molecular Genetics at the meeting of the Poultry Science Association of America at Ames; and at the Royal Veterinary and Agricultural University, Copenhagen.

The Armidale Conference was of special importance because it led to a contact from Peter Brumby at ILCA (International Livestock Center for Africa, Addis Ababa) asking whether RFLPs might be used to characterize trypanotolerant cattle breeds in Africa. In my usual enthusiastic way I replied, "We can do better than that—we will map the trypanotolerance loci." Peter was on his way to a position at World Bank, so communication with ILCA lapsed. However, John Hodges, a close friend and colleague had just begun to work at FAO (Food and Agriculture Organization), Rome, and he arranged a consultantship for me with the International Trypanotolerance Center (Banjul, The Gambia) that was in process of being established by Ian McIntyre and centered on the trypanotolerant N'Dama cattle breed, a longhorn *Bos taurus* breed. Ian was enthusiastic about the possibility of mapping trypanotolerance loci based on crosses of N'Dama to susceptible cattle, and together with Tony Davies organized an International Seminar on Genetic Aspects of Trypanotolerance at Banjul in 1987, to celebrate the official opening of the ITC (International Trypanotolerance Center). At the seminar, Jacqui and I presented the trypanotolerance mapping proposal (Soller and Beckmann 1987). For various reasons, the project did not move forward at ITC, although USAID (United States Agency for International Development) provided funding to Jim Womack, myself, and ITC for the initial steps, including developing the RFLP markers for the mapping. However, Jack Doyle, Director of Research at ILRAD (International Laboratory for Research on Animal Diseases) was present at the seminar and he encouraged Alan Teale to organize a livestock genetics meeting at ILRAD. I was unable to attend, but Jacqui came and was enormously impressed by the facilities at ILRAD and convinced Alan of the potential of QTL mapping. Jack Doyle backed this and helped convince Ross Gray, Director of ILRI (International Livestock Research Institute), and then the Board that this was something to which ILRI should make a major commitment. Peter Doherty and Ole Nielsen also provided strong support. Independently of this, a small N'Dama herd (five males and four females) had already been produced at ILRI by implantation of N'Dama embryos collected at ITC into surrogate Boran (*Bos indicus*) females, so the biological infrastructure was already in place. In 1989, construction of large full-sib F2 families was initiated under Alan's direction. The F2 animals were born between November 1992 and September 1996, reared, challenged, phenotyped, and genotyped, and results were published in 2003 (Hanotte et al. 2003). This was one of the first dedicated QTL mapping experiments in livestock, though it may have been preceded by a similar design for mapping tick resistance in Australia. While this experiment was in progress, Alan began to wonder about use of a mouse model. He knew of the immunology done by Ivan Morrisson and Sam Black at ILRAD that had shown differences in innate resistance among mouse strains (Morrison et al. 1978; Mahan et al. 1986). Detailed examination of their data convinced him that they should also begin mapping in mice, while they awaited the development of the F2 cattle population. Steve Kemp seconded this enthusiastically, and took a leading role in turning the initial thoughts into meaningful experiments (Kemp et al. 1997).

Meanwhile, the immediate goal in our home laboratory was to demonstrate that RFLP markers could actually be found. In 1982, Jacqui and I submitted a BARD proposal with Benn Burr as co-PI, and we were funded in 1983—surely among the first funded proposals for RFLP search in agricultural species. The Department of Genetics provided me a superb technician, Alona Naveh, with experience in DNA work and Alona, graduate student Yechezkel Kashi, and postdoc Eric Hallerman, under overall supervision of Jacqui Beckmann, embarked on a successful search for bovine RFLPs using bovine Growth Hormone and Prolactin as probes (Beckmann et al. 1986; Hallerman et al. 1987). In 1985, Jim approached us on submitting a BARD proposal for mapping the bovine genome using markers and somatic cell hybrids. This was funded in 1985 and renewed in 1989. On our side, we used the funds primarily for marker development, while Jim went on to develop the comparative bovine/human/mouse synteny map (Dietz et al. 1992a, 1992b), showing extensive correspondence between the three genomes and becoming an essential tool for comparative genomics and positional gene cloning.

Analyzing the trypanotolerance-mapping cross required working out a theory for mapping in crosses between outcrossing populations that shared marker alleles (Beckmann and Soller 1988). On completing the theory, I realized to my horror that diallelic RFLP markers would be very inefficient for mapping in a cross of this type. Fortunately, in the next year, through use of the newly available PCR (polymerase chain reaction) reaction and sequencing gel, a number of groups (Weber and May 1989; Litt and Luty 1989; Tautz 1989) independently discovered high repeat number variability of short tandem sequences (SSR or microsatellites). This provided a large group of relatively low cost, highly informative polyallelic markers. Jacqui immediately realized the importance of this new class of marker for our mapping studies, and already in 1990 proposed that the basic livestock and poultry marker maps be based on these markers (Beckmann and Soller 1990), as indeed happened. While all this was going on, Adam Friedmann, a close colleague in the Genetics Department, was engaged in exploring RFLPs at the human major histocompatibility complex (MHC) locus, looking for associations of markers at the human MHC locus and the human autoimmune disease pemphigus vulgaris. In a breakthrough series of papers, he and his colleagues were able to show that people sharing identical human leukocyte antigen (HLA) alleles as identified by antibodies or cellular reactions, differed at the DNA level, enabling disease associations and relative risk to be assessed more precisely (Szafer et al. 1987). In these studies, Adam was one of the very first to utilize the PCR for genomic analysis (Scharf et al. 1989) and thus, was able to provide us with invaluable guidance in applying this technique in our laboratory. In 1985, Ehud (Uddi) Lipkin, a member of the Ganigar Collective Settlement where he was in charge of the dairy herd, joined us as an MSc student. Adam took him under his wing and directed him in an RFLP analysis of the bovine MHC locus. After completing his MSc degree, Uddi went back to the farm and the dairy herd, but returned to us as a PhD student in 1991, again with Adam as his direct mentor on DNA pooling, and has been with us ever since. Uddi's PhD study established the possibility of doing selective DNA pooling using shadow-band corrected microsatellite markers, and was the first to identify the extremely powerful effect of QTL located on BTA6 on milk production traits (Lipkin et al. 1998). As time went on, Adam developed a greater interest in our work. We jointly directed a number of MS and PhD students, and when I eventually reached retirement age, Adam provided Uddi, myself, and our students with office and laboratory space, and we continue to work together to this day.

In January 1986, Michael Grossman and Gene Eisen organized the first Gordon Conference on Quantitative Genetics and Biotechnology, which was attended by over 100 scientists from universities, government, and industry, and at which I gave a talk on QTL mapping and MAS. In 1987, the first EC meeting on Mapping the Bovine Genome was held in Brussels. Michel Georges was there—he had just completed his MS in which he also demonstrated RFLPs at the growth hormone (GH) and thyroglobulin loci. In 1989, Rudy Fries developed the first DNA-marker-based map of the bovine genome (Fries et al. 1989). In 1991, ILRAD organized a workshop on Trypanotolerance and Genome Mapping, and at this meeting, Jim Womack, Jay Hetzel, and Alan Teale set up the first bovine mapping resource consisting of a panel of full- and half-sib families, which they each contributed to the general pool (Hetzel 1991). This provided the first set of framework families (known as the CSIRO reference families) for bovine map development (Barendse et al. 1994). At the same time, USDA MARC (Meat Animal Research Center) set up an independent set of reference families (Bishop et al. 1994). In 1984, I became interested in chicken endogenous viruses as a possible source of genetic markers and for direct effects on production traits (Iraqi et al. 1991). We received support for this and other work related to the development of the chicken marker map (Khatib et al. 1993) in a series of three BARD grants over the period 1985–1994 with Lyman Crittenden, who did the original work on chicken endogenous viruses, and also with his colleague Larry Bacon as co-PIs. This got Lyman involved in genome mapping and, around 1988, Lyman and his colleague Jerry Dodgson created a chicken reference family population for marker mapping, based on backcross of a partially inbred red jungle fowl to a highly inbred White Leghorn line, and went on to develop a chicken genome map consisting of RFLP, RAPD (random amplification of polymorphic DNA), and Chicken Repeat Element 1 markers (Crittenden et al. 1993). At about the same time, Nat Bumstead created an independent resource population and genetic map based on a backcross of two highly inbred White Leghorn lines differing in disease resistance, and consisting entirely of RFLP markers (Bumstead and Palyga 1992). Considerably later (Groenen et al. 1998), a third set of reference families was developed by Martien Groenen at Wageningen Agricultural University as the F2 products of a cross between two broiler dam lines, and a consensus map was constructed in 2000 (Groenen et al. 2000). Doctoral students in our lab, Anne Shalom, Hasan Khatib, and Mathias Mosig, developed microsatellite markers for both the chicken and cattle maps (Khatib et al. 1993; Shalom et al. 1995).

In 1987, we proposed the "trait-based" mapping design that showed how individuals in the tails of a distribution could be used for highly informative mapping (Lebowitz et al. 1987). All this time, QTL mapping remained a specialized, if highly active enclave within the agricultural plant and animal community. In 1989, however, Lander and Botstein (1989) published an influential paper in *Genetics*, introducing QTL mapping to the general genetics public. This paper also independently reintroduced the concept of interval mapping, this time proposing the use of maximum likelihood methods for its implementation; and independently proposed a variation of the "trait-based design" under the felicitous term "selective genotyping," which is now the accepted designation for this design. In an important methodological advance during the same period, Haley and Knott (1992) showed how interval mapping could be implemented by least squares using standard statistical packages instead of ad hoc maximum likelihood procedures, which greatly simplified QTL mapping statistics.

In the fall of 1987, I spent a 3-month sabbatical at University of Illinois with support from George Miller endowment, and gave the first credit course ever on "molecular markers in plant and animal genetic improvement." About 30 students and staff participated, among them were Rohan Fernando and Michael Grossman. In 1989, Rohan and Michael were working on a paper for genetic evaluation with autosomal and X-chromosomal loci under equilibrium, when Rohan realized that the same principles could be applied more importantly to the challenge of incorporating marker information (MQTL) with phenotypic information in BLUP evaluation. This led to their seminal paper on incorporating marker information in BLUP prediction (Fernando and Grossman 1989). Harris Lewin also participated, and it turned out that he had a set of data that were suitable for QTL mapping, and used them to map some of the first QTL in cattle (Beever et al. 1990). Pride of place must go to Hermann Geldermann, I believe, who published the first QTL mapping study in dairy cattle in 1985, using a panel of blood group, protein, and biochemical markers (Geldermann et al. 1985). Although in his usual prescient way, Alan Robertson preceded everyone in his 1961 paper with Neimann Sorensen using blood group markers (Neimann-Sorenson and Robertson 1961). Joel Weller was on sabbatical at University of Illinois at the time, and at the suggestion of my graduate student Yechezkel Kashi and with an assist from Daniel Gianola, we explored the use of progeny test information for QTL mapping, resulting in the well-known paper introducing the widely used granddaughter design for QTL mapping in dairy cattle (Weller et al. 1990). This was followed by the founding by Harris Lewin of the Dairy Bull Repository for QTL mapping based on the granddaughter design (Da et al. 1994). Some time later, USDA at Beltsville began a similar effort and the two repositories were eventually combined (Ashwell et al. 2000).

In winter of 1990, there was a Banbury Conference on Mapping the Genome of Agricultural Animals at Cold Spring Harbor, followed in the spring of 1990 by the Allerton Conference on Mapping Domestic Animal Genomes. At the Allerton meeting, I gave a talk on use of population-wide linkage disequilibrium for QTL mapping. I followed this up a bit with my colleague Avraham Genizi, but did not pursue it further when our first paper was rejected by *Genetical Research*, basically on grounds of "triviality." As we all know, Mike Goddard and his students did develop the topic in a series of outstanding papers, turning it into a tremendous practical success when the advent of the dense SNP chips enabled whole-genome selection (WGS) based on genomic estimates of EBV (Meuwissen et al. 2001). In the summer of 1990, I chaired a workshop on Molecular Mapping of Quantitative Genes at the 4th World Congress of Genetics Applied to Livestock Production, held in Edinburgh. Hermann Geldermann and Michel Georges led off, and were followed by Joel Weller, Harris Lewin, Max Rothschild, H.C. Hines, Jay Hetzel, and Jossi Hillel (Soller 1990a). At this meeting, quantitative genomics can be said to have come of age. The venue for the workshop had to be moved from one of the smaller meeting rooms to one of the large lecture halls, with loudspeakers in the halls for the overflow, and the excitement was palpable. D.S. Falconer was heard to remark after the workshop "I will have to rewrite my book," and indeed he did so not long after with Trudy Mackay in the fourth edition of his classical text (Falconer and Mackay 1996). In 1990, we also published the first detailed study of MAS based on QTL mapping (Kashi et al. 1990) and of use of replicated progenies for QTL mapping (Soller and Beckmann 1990). That year, I also wrote a review for *Journal of Dairy Science*, setting forth the challenge of achieving a complete QTL map of the bovine genome by the year 2000 (Soller 1990b). In the spring of 1992, I gave an intensive 1-week course on QTL mapping and MAS at Wageningen Agricultural University in the Netherlands.

In 1990, Ariel Darvasi joined my laboratory as MS and then PhD student. Together, we analyzed in depth selective genotyping (Darvasi and Soller 1992) and selective DNA pooling (Darvasi and Soller 1994). Ariel with Joel Weller showed that even with a highly saturated marker map, confidence interval of QTL map location was a function of allele substitution effect at the QTL and sample size raising the problem of QTL map resolution (Darvasi et al. 1993). This led to the development of the advanced intercross line (AIL) design for high-resolution QTL mapping (Darvasi and Soller 1995). An early application of this methodology was for high-resolution mapping of QTL affecting trypanotolerance in crosses between inbred lines of mice by Fuad Iraqi (who did his PhD in our laboratory on endogenous viruses). Fuad developed an F6 AIL between resistant and susceptible strains (Iraqi et al. 2000) and used it to obtain refined estimates of QTL map location, and even more strikingly, to show that one of the original QTL mapped by Kemp et al. (1997) could be resolved into three separate loci. Additional studies with more advanced generations, further sharpened QTL map location (Nganga et al. 2010).

In 1989, GenMark, the first commercial company aimed at implementing MAS for livestock genetic improvement, was founded with Michel Georges as Director of Research. Based on the work done in GenMark, the first complete dairy cattle genome scan of dairy cattle for QTL affecting production and functional traits was published in 1995 by Michel and his colleagues (Georges et al. 1995). In contrast to the QTL mapping approach described in detail in this memoir, Max Rothschild has consistently and successfully championed the candidate gene approach to QTL identification (Rothschild and Soller 1997). First, by establishing associations of swine Swine Leukocyte Antigen (SLA) antigens and SLA RFLPs on production traits Rothschild et al. 1986), and in a landmark paper showing association of the swine estrogen receptor locus with litter size (Rothschild et al. 1996).

With these influential and widely cited studies by Michel and Max, the new quantitative genomic paradigm was firmly established, with many laboratories taking part in development of theory and application. Our own laboratory is now one of many, toiling in the vineyard. Our particular interest has been in the theory and application of selective DNA pooling for mapping of production and functional traits in dairy cattle and poultry, which we have pursued in our own laboratory, and in cooperation with colleagues in Israel, Italy, and the United States (Mosig et al. 2001; Heifetz et al. 2007; Korol et al. 2007; Bagnato et al. 2008). Together with the rest of the bovine community, we look forward with anticipation to the new genomic worlds that are being revealed through the lens of dense SNP chips and next-generation sequencing.

## Acknowledgment

I acknowledge with thanks the generous assistance of my colleagues Jacques Beckmann, Jerry Dodgson, Rohan Fernando, Adam Friedmann, Hermann Geldermann, Michel Georges, Michael Grossman, Olivier Hanotte, Harris Lewin, Ehud Lipkin, Eviatar Nevo, Max Rothschild, Alan Teale, and James E. Womack, all of whom were kind enough to share with me their memories and documents of the events that unfolded in the period 1980–1995. As a personal memoir reflecting events as I experienced them, this is subject to all of the vagaries of personal involvement and memory. Any inaccuracies are mine alone and I apologize for any sins of omission or commission.

## References

Ashwell, M.S., Van Tassell, C.P., Sonstegard, T.S. (2000) The cooperative dairy DNA repository: a new resource for quantitative trait loci detection and verification. Proc. 8th Plant Animal Genome Conference, San Diego, CA, p. 360.

Bagnato, A., Schiavini, F., Rossoni, A., Maltecca, C., Dolezal, M., Medugorac, I., Sölkner, J., Russo, V., Friedman, A., Soller, M., Lipkin, E. (2008) Quantitative trait loci affecting milk yield and protein percent in a three countries Brown Swiss population. *Journal of Dairy Science* **91**: 767–783.

Barendse, W., et al. (1994) A genetic linkage map of the bovine genome. *Nature Genetics* **6**: 227–235.

Beckmann, J.S., Kashi, Y., Hallerman, E.M., Nave, A., Soller, M. (1986) Restriction fragment length polymorphism among Israeli Holstein-Friesian dairy bulls. *Animal Genetics* **17**: 25–38.

Beckmann, J.S. and Soller, M. (1983) Restriction fragment length polymorphisms in genetic improvement: methodologies, mapping and costs. *Theoretical and Applied Genetics* **67**: 35–43.

Beckmann, J.S. and Soller, M. (1988) Detection of linkage between marker loci and loci affecting quantitative traits in crosses between segregating populations. *Theoretical and Applied Genetics* **76**: 228–236.

Beckmann, J.S. and Soller, M. (1990) Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. *Biotechnology* **8**: 930–933.

Beever, J.E., George, P.D., Fernando, R.L., Stormont, C.J., Lewin, H.A. (1990) Associations between genetic markers and growth and carcass traits in a paternal half-sib family of Angus cattle. *Journal of Animal Science* **68**: 337–344.

Bishop, M.D., et al. (1994) A Genetic Linkage Map for Cattle. *Genetics* **136**: 619–639.

Botstein, D., White, R.I., Skolnick, M., Davis, R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**: 314–331.

Briles, W.E. (1984) Early chicken blood group investigations. *Immunogenetics* **20**: 217–226.

Bumstead, N. and Palyga, J. (1992) A preliminary linkage map of the chicken genome. *Genomics* **13**: 690–697.

Burr, B., Evola, S.V., Burr, F.A., Beckmann, J.S. (1983) The application of restriction fragment length polymorphism to plant breeding. In: *Genetic Engineering: Principles and Methods*, edited by J.K. Setlow and A. Hollaender, Vol. 5, pp. 45–59. New York: Plenum Press.

Crittenden, L.B., Provencher, L., Santangelo, L., Levin, I., Abplanalp, H., Briles, R.W., Briles, W.E., Dodgson, J.B. (1993) Characterization of a Red Jungle Fowl by White Leghorn backcross reference population for molecular mapping of the chicken genome. *Poultry Science* **72**: 334–348.

Da, Y., Ron, M., Yanai, A., Band, M., Everts, R.E., Heyen, D.W., Weller, J.I., Wiggans, G.R., Lewin, H.A. (1994) The dairy bull DNA repository: A resource for mapping quantitative trait loci. Proc. 5th World Congress on Genetics Applied to Livestock Production, Guelph, ON, Canada. Vol. 21, pp. 229–232.

Darvasi, A. and Soller, M. (1992) Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and Applied Genetics* **85**: 353–359.

Darvasi, A. and Soller, M. (1994) Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics* **138**: 1365–1373.

Darvasi, A. and Soller, M. (1995) Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* **141**: 1199–1207.

Darvasi, A., Weinreb, A., Minke, V., Weller, J.I., Soller, M. (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* **134**: 943–951.

Dietz, A.B., Georges, M., Threadgill, D.W., Womack, J.E., Schuler, L.A. (1992a) Somatic cell mapping, polymorphism, and linkage analysis of bovine prolactin-related proteins and placental lactogen. *Genomics* **14**: 137–143.

Dietz, A.B., Neibergs, H.L., Womack, J.E. (1992b) Assignment of eight loci to bovine syntenic groups by use of PCR: extension of a comparative gene map. *Mammalian Genome* **3**: 106–111.

Edwards, M.D., Stuber, C.W., Wendel, J.F. (1987) Molecular-marker facilitated investigations of quantitative trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics* **116**: 113–125.

Falconer, D.S. (1960) *Introduction to Quantitative Genetics*. Edinburgh: Oliver and Boyd.

Falconer, D.S. and Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*. 4th edition. Harlow: Longman Group Ltd.

Fernando, R. and Grossman, M. (1989) Marker assisted selection using best linear unbiased prediction. *Genetics Selection Evolution* **21**: 467–477.

Fries, T., Beckmann, J.S., Georges, M., Soller, M., Womack, J. (1989) The bovine gene map. *Animal Genetics* **20**: 3–29.

Geldermann, H. (1972) Molekulargenetische Probleme in der tierzuchterischen Forschung und Praxis. B1. Die Bedeutung der monogener Vererbung bei molecularen Substanzen fur die Zuchtung auf polygene Leistungsmerkmale. *Tierzuchter* **21**: 15–16.

Geldermann, H. (1975) Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. *Theoretical and Applied Genetics* **46**: 319–330.

Geldermann, H., Pieper, U., Roth, B. (1985) Effects of marked chromosome sections on milk performance in cattle. *Theoretical and Applied Genetics* **70**: 138–146.

Georges, M., et al. (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**: 907–920.

Groenen, M.A.M., Crooijmans, R.P.M.A., Veenendaal, A., Cheng, H.H., Siwek, M., Van der Poel, J.J. (1998) A comprehensive microsatellite linkage map of the chicken genome. *Genomics* **49**: 265–274.

Groenen, M.A.M., et al. (2000) A consensus linkage map of the chicken genome. *Genome Research* **10**: 137–147.

Haley, C.S. and Knott, S.A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315–324.

Hallerman, E.M., Nave, A., Kashi, Y., Soller, M., Beckmann, J.S. (1987) Restriction fragment length polymorphisms in dairy and beef cattle at the growth hormone and prolactin loci. *Animal Genetics* **18**: 213–222.

Hanotte, O., et al. (2003) Mapping of QTL controlling resistance to trypanosomosis in an experimental cross of trypanotolerant West African N'Dama cattle (*Bos taurus*) and trypanosusceptible East African Boran cattle (*Bos indicus*). *Proceedings of the National Academy of Sciences of the United States of America* **100**: 7443–7448.

Heifetz, E.M., Fulton, J.E., O'Sullivan, N.P., Arthur, J.A., Wang, J., Dekkers, J.C.M., Soller, M. (2007) Mapping QTL affecting susceptibility to Marek's disease in a backcross population of layer chickens. *Genetics* **177**: 2417–2431.

Hetzel, D.J.S. (1991) The use of reference families for genome mapping in domestic livestock. In: *Gene-Mapping Techniques and Applications*, edited by L.B. Schook, H.A. Lewin, and D.G. McLaren, pp. 51–64. Oxford: Dekker.

Iraqi, F., Clapcott, S.J., Kumari, P., Haley, C.S., Kemp, S.J., Teale, A.J. (2000) Fine mapping of trypanosomiasis resistance loci in murine advanced intercross lines. *Mammalian Genome* **11**: 645–648.

Iraqi, F., Soller, M., Beckmann, J.S. (1991) Endogenous viruses in commercial laying populations. *Poultry Science* **70**: 665–679.

Jeffreys, A.J. (1979) DNA sequence variants in the G gamma-, A gamma-, delta- and beta-globin genes of man. *Cell* **18**: 1–10.

Johansson, I. and Robertson, A. (1952) Progeny testing in the breeding of farm animals. Proc. of the British Society of Animal Production pp. 79–105.

Kan, Y.S. and Dozy, A.M. (1978) Antenatal diagnosis of sickle cell anemia by DNA analysis of amniotic fluid cells. *Lancet* **2**: 910–912.

Kashi, Y., Hallerman, E.M., Soller, M. (1990) Marker-assisted selection of candidate bulls for progeny testing programs. *Animal Production* **51**: 63–74.

Kemp, S.J., Iraqi, F., Darvasi, A., Soller, M., Teale, A.J. (1997) Localization of genes controlling resistance to trypanosomiasis in mice. *Nature Genetics* **16**: 194–196.

Khatib, H., Genislav, E., Crittenden, L.B., Bumstead, N., Soller, M. (1993) Sequence tagged microsatellite sites as markers in chicken reference and resource populations. *Animal Genetics* **24**: 355–362.

Korol, A., Frenkel, Z., Cohen, L., Lipkin, U., Soller, M. (2007) Fractioned DNA pooling: A new cost effective strategy for fine QTL mapping. *Genetics* **176**: 2611–2625.

Kudo, R.R. (1946) *Protozoology*. 3rd edition. Springfield: Charles C. Thomas Publisher.

Lander, ES. and Botstein D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.

Law, C.N. (1966) The location of genetic factors affecting a quantitative character in wheat. *Genetics* **53**: 487–498.

Lebowitz, R., Soller, M., Beckmann, J.H.S. (1987) Trait based designs for determination of linkage between marker loci and quantitative trait loci. *Theoretical and Applied Genetics* **72**: 556–562.

Lerner, I.M. (1950) *Population Genetics and Animal Improvement*. Cambridge: University Press.

Lewontin, R.C. and Hubby, J.L. (1966) A molecular approach to the study of genic heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics* **54**: 595–609.

Lipkin, E., Mosig, M.O., Darvasi, A., Ezra, E., Shalom, A., Friedmann, A., Soller, M. (1998) Mapping loci controlling milk protein percentage in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers. *Genetics* **149**: 1557–1567.

Litt, M. and Luty, J.A. (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics* **44**: 397–401.

Lush, J.L. (1943) *Animal Breeding Plans*. Ames: Iowa State College Press.

Mahan, S.M., Hendershot, L, Black, S.J. (1986) Control of trypanodestructive antibody responses and parasitemia in mice infected with Trypanosoma (Duttonella) vivax. *Infection and Immunity.* **54**: 213–221.

Mather, K. and Jinks, J.L. (1949) *Biometrical Genetics*. London: Chapman and Hall Ltd.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Morgan, T.H. (1926) *The Theory of the Gene*. New Haven: Yale University Press.

Morrison, W.I., Roelants, G.E., Mayor-Withey, K.S., Murray, M. (1978) Susceptibility of inbred strains of mice to Trypanosoma congolense: correlation with changes in spleen lymphocyte populations. *Clinical & Experimental Immunology* **32**: 25–40

Mosig, M.O., Lipkin, E., Khutoreskaya, G., Tchouryzna, E., Ezra, E., Soller, M., Friedmann, A. (2001) A whole genome scan for QTL affecting milk protein percent in Israel-Holstein cattle by means of selective milk DNA pooling in a daughter design. *Genetics* **157**: 1683–1698.

Neimann-Sorenson, A. and Robertson, A. (1961) The association between blood groups and several production characters in three Danish cattle breeds. *Acta Agriculturae Scandinavica* **11**: 163–196.

Nevo, E., Golenberg, E., Beiles, A., Brown, A.H.D., Zohary, D. (1982) Genetic diversity and environmental associations of wild wheat. Triticum discoccoides in Israel. *Theoretical and Applied Genetics* **62**: 241–254.

Nevo, E., Zohary, D., Brown, A.H.D., Haber, M. (1979) Genetic diversity and environmental associations of wild barley, Hordeum spontaneum in Israel. *Evolution* **33**: L815–L833.

Nganga, J.K., Soller, M., Iraqi, F.A. (2010) High resolution mapping of trypanosomosis resistance loci *Tir*2 and *Tir*3 using F12 advanced intercross lines with major locus *Tir*1 fixed for susceptible allele. *BMC Genomics* **11**: 394–416.

Padeh, B. and Soller, M. (1976) Genetic and environmental correlations between brain weight and maze learning in inbred strains of mice and their F-1 hybrids. *Behavior Genetics* **6**: 31–41.

Robertson, A. and Rendel, J.M. (1950) The use of progeny testing with artificial insemination in dairy cattle. *Journal of Genetics* **50**: 21–31.

Rothschild, M. and Soller, M. (1997) Candidate gene analysis to detect genes controlling traits of economic importance in domestic livestock. *Probe* **8**: 13–20.

Rothschild, M.F., Renard, C., Bolet, G., Dando, P., Vaiman, M. (1986) Effect of SLA haplotypes on birth and weaning weights in pigs. *Animal Genetics* **17**: 267–272.

Rothschild, M., et al. (1996) The estrogen receptor locus is associated with a major gene influencing litter size in pigs. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 201–205.

Sax, K. (1923) The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. *Genetics* **8**: 552–560.

Scharf, S.J., Friedmann, A., Steinman, L., Brautbar, C., Erlich, H.A. (1989) Specific HLA-DQB and HLA-DRBl alleles confer susceptibility to pemphigus vulgaris. *Proceedings of the National Academy of Sciences of the United States of America* **86**: 6215–6219.

Sears, E.R. (1969) Wheat cytogenetics. *Annual Review Genetics* **3**: 451–468.

Shalom, A., Mosig, M.O., Barendse, W., Soller, M., Friedmann, A. (1995) Dinucleotide repeat polymorphism at the bovine HUJ673, HUJ121, HUJ174, HUJ225, HUJI13, and HUJI29 loci. *Animal Genetics* **26**: 202.

Soller, M. (1978) The use of loci associated with quantitative effects in dairy cattle improvement. *Animal Production* **27**: 133–139.

Soller, M. (1990a) Molecular Mapping of quantitative genes. Workshop 8.3, 4th World Congress Genetics Applied to Livestock Production, Edinburgh, July 1990. Vol. 3, p. 91.

Soller, M. (1990b) Genetic mapping of the bovine genome using DNA-level markers with particular attention to loci affecting quantitative traits of economic importance. *Journal of Dairy Science* **73**: 2628–2646.

Soller, M., Bar-Anan, R., Pasternak, H. (1966) Selection of dairy sires for growth rate and milk production. *Animal Production* **8**: 109–119.

Soller, M. and Beckmann J.S. (1982) Restriction fragment length polymorphisms and genetic improvement. Proc. 2nd World Congress on Genetics Applied to Livestock Prod., Madrid, Oct. 1982. Vol. 6, pp. 396–404.

Soller, M. and Beckmann, J.S. (1983) Genetic polymorphisms in varietal identification and genetic improvement. *Theoretical and Applied Genetics* **67**: 25–33.

Soller, M. and Beckmann, J.S. (1987) Toward an understanding of the genetic basis of trypanotolerance in the N'Dama cattle of West Africa. Consultation Report submitted to FAO, Rome (March, 1987).

Soller, M. and Beckmann, J.S. (1990) Marker-based mapping of quantitative trait loci using replicated progenies. *Theoretical and Applied Genetics* **80**: 205–208.

Soller, M. and Genizi, A. (1978) The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. *Biometrics* **34**: 47–55.

Soller, M. and Plotkin-Hazan, J. (1977) The use of marker alleles for the introgression of linked quantitative alleles. *Theoretical and Applied Genetics* **51**: 133–137.

Soller, M., Brody, T., Genizi, A. (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47**: 35–39.

Solomon, E.T. and Bodmer, W.F. (1979) Evolution of sickle variant gene. *Lancet* **1**: 923.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology* **98**: 503–517.

Spickett, S.P. and Thoday J.M. (1966) Regular responses to selection; Interaction between located polygenes. *Genetic Research* **7**: 96–121.

Stormont, C., Owen, R.D., Irwin, M.R. (1951) The B and C systems of bovine blood groups. *Genetics* **36**: 134–161.

Szafer, F., et al. (1987) Detection of disease-specific restriction fragment length polymorphisms in pemphigus vulgaris linked to the DQw1 and DQw3 alleles of the HLA-D region. *Proceedings of the National Academy of Sciences of the United States of America* **84**: 6542–6545.

Tanksley, S.D. and Khush G.S. (2004) Charles Madera Rick. *Biographical Memoirs*. Vol 84, pp. 307–320. Washington D.C.: National Academy of Sciences (NAS), National Academies Press.

Tautz, D. (1989) Hypervariability of simple sequences as a general source of polymorphic DNA markers. *Nucleic Acids Research* **17**: 6463–6471.

Thoday, J.M. (1961) Location of polygenes. *Nature* **191**: 368–370.

Thoday, J.M. (1966) New Insights into continuous variation. Proc. 3rd International Congress of Human Genetics, University of Chicago, September 5–10 1966, edited by J.F. Crow and J.V. Neel. Baltimore: The Johns Hopkins Press.

Weber, J.L. and May, P.E. (1989) Abundant class of human polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* **44**: 388–396.

Wehrhahn, C. and Allard, R.W. (1965) The detection and measurement of the effects of individual genes involved in the inheritance of a quantitative character in wheat. *Genetics* **51**: 109–119.

Weller, J.I. (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**: 627–640.

Weller, J.I., Kashi, Y., Soller, M. (1990) Daughter and granddaughter designs for mapping of quantitative trait loci in dairy cattle. *Journal of Dairy Science* **73**: 2525–2537.

Weller, J.I., Soller, M., Brody, T. (1988) Linkage analysis of quantitative traits in an interspecific cross of tomato (Lycopersicon esculentum x Lycopersicon pimpinellifolium) by means of genetic markers. *Genetics* **118**: 329–339.

Zhuchenko, A.A., Andryschenko, V.K., Korol, A.B., Korochkina, S. (1975) Mapping of genetic factors of quantitative characters in tomato. *Cytologia i Genetika* (USSR) **9**(2): 101–105 (in Russian). English translation: *Cytology and Genetics* (1978), **9**(2): 4–7.

Zhuchenko, A.A., Korol, A.B., Andryuschenko, V. (1978) Linkage between loci of quantitative characters and marker loci. 1. The model. *Genetika* (USSR) **14**: 771–778 (in Russian).

Zhuchenko, A.A., Samovol, A.P., Korol, A.B., Andryuschenko, V.K. (1979) Linkage between quantitative trait loci and marker loci. The influence of three tomato chromosomes on the variation for five quantitative traits in backcross generation. *Genetika* (USSR) **15**: 672–683 (in Russian).

# Chapter 5
# Cartography of the Bovine Genome

*James E. Womack*

## Introduction

Why map genes? In the midst of his pioneering efforts to develop comprehensive gene maps of human chromosomes in the 1970s and 1980s, Frank Ruddle was often asked this question. While predicting that good maps would enhance the pace of discovery in human genetics and would provide focal points for posing and testing hypotheses, he frequently added, “Gene mapping is good for you!” (Ruddle 1984). Exploring and recording landmarks in the genomes of target species continues to be a fascinating endeavor, providing gratification in overcoming technical challenges as well as esthetic satisfaction in seeing a graphic representation of a genome at an enhanced level of resolution. Satellite imagery of the earth’s surface has not eliminated the need for maps of lesser resolution. In fact, the historical maps of lesser resolution are important to the interpretation and annotation of high-resolution photography. In this chapter, I review mapping of the bovine genome in a historical perspective. Applications of these maps to comparative mapping and contributions of the maps to assembly of the whole genome sequence are addressed in Chapters 8 and 9, respectively. The bovine linkage map is also the subject of Chapter 6.

## Somatic Cell Mapping

The fusion of cultured somatic cells (Barski et al. 1961) coupled with the observation of preferential loss of human chromosomes in human/mouse hybrids (Weiss and Ephrussi 1966) paved the way for the development of the first comprehensive maps of the human genome (Ruddle 1972; Ruddle and Creagan 1975). These prerecombinant DNA maps were largely dependent on isoenzymes (Markert 1968) as markers, although the parasexual approach flourished when DNA markers were subsequently developed (Ruddle 1981). Both isoenzyme and DNA markers proved extremely valuable for comparative gene mapping in mammals, which was initially restricted to human–mouse comparisons (Minna et al. 1976; Nadeau and Tayler 1984). The somatic cell method, however, proved effective for rapid and efficient mapping of the genomes of other mammals, including cattle, and facilitated the early comparisons of animal genome organization beyond primates and rodents (O’Brien et al. 1988).

*Bovine Genomics*, First Edition. Edited by James E. Womack.

Heuertz and Hors-Cayla (1978) fused bovine and hamster cells specifically for the purpose of bovine gene mapping and followed the segregation of enzyme markers to define the first three autosomal synteny groups in cattle (Heuertz and Hors-Cayla 1981). In an independent study, Shimizu et al. (1981) described a group of syntenic markers on the bovine X chromosome. The three autosomal groups were confirmed by Echard et al. (1984). Utilizing hybridoma cells that had been initially fused to produce bovine monoclonal antibodies, Dain et al. (1984) made the first assignment of a bovine syntenic group to an autosome. Meanwhile, my laboratory had generated yet another panel of hybrid cells that we used to produce a gene map of the cow, consisting of 35 markers defining 23 syntenic groups, including singlets, with assignments of markers to five chromosomes (Womack and Moll 1986). This map revealed greater conservation of synteny between cow and human than between mouse and human and was subsequently used in the development of higher-resolution whole genome maps of cattle with comprehensive chromosomal assignments (Fries et al. 1993).

## *Synteny and Syntenic Groups*

Hybrid somatic cells constructed for gene mapping normally segregate the chromosomes of one species as intact bodies. Thus, in the absence of karyotypic analysis of individual hybrid cell lines, the "map" produced consists of groups of markers that segregate together. These markers are then assumed to be on the same chromosome, although the specific chromosome may or may not initially be identified. These markers on a common chromosome have been erroneously described as "linked" by a number of investigators, including myself, although most of us knew that "linkage" is a genetic term describing the relationship of markers in the analysis of meiotic events. Although it was not immediately recognized and implemented, Renwick (1971) provided a solution to the nomenclature dilemma regarding markers cosegregating in hybrid somatic cells. He coined, or at least introduced into the genetic literature, the term "synteny" to describe markers on the same chromosome, regardless of whether linkage can be demonstrated in meiotic products. The term literally means "on the same ribbon" (Gk: *syn* = together; *taenia* = ribbon). Thus, two markers may be sufficiently far apart on a chromosome that linkage cannot be demonstrated, yet they can still be described as syntenic. "Asynteny" is conversely the state of being on different chromosomes (Renwick 1971).

Unfortunately, the terms synteny and syntenic are commonly abused in today's genetic literature. Terminology for describing chromosomal conservation between species at the gene level was initially recommended in a report of the Comparative Mapping Committee (Lalley et al. 1987) and subsequently adopted by the mammalian genetics community as reviewed by Nadeau (1989). "Synteny conservation" defines two or more pairs of homologous genes on the same chromosome in two or more species. Whether from laziness or simple ignorance of the derivation of the word, "synteny" is often used in place of "synteny conservation" or "conservation of synteny." It is not uncommon to read that a portion of chromosome A in species X is syntenic with chromosome B in species Y, which is a literal impossibility outside an interspecific translocation that creates a common ribbon from two chromosomes of different species. To say that synteny is conserved between portions of chromosome

A in species X and chromosome B in species Y reflects an accurate understanding of the very useful word Renwick introduced to our genetic vocabulary.

## *In Situ Hybridization and Chromosome Identification*

Somatic cell maps produce synteny (or syntenic) groups, markers that are on the same chromosome. The linear order of the markers, their position on the chromosome, or even which chromosome they are on, are not necessarily known. Synteny groups in the early human maps were almost immediately assigned to specific chromosomes because the hybrid clones were karyotyped and the segregation of individual chromosomes was correlated with the segregation of markers, thanks to the development of banding techniques that made individual human chromosomes readily distinguishable (Caspersson et al. 1968, 1970). Cattle chromosomes, on the other hand, have been (and remain) difficult to distinguish. The 29 autosomal pairs are all acrocentric and banding patterns of many, especially the smaller ones, are similar, making the autosomes recalcitrant to identification except when lined up side by side in a karyogram. The banding technologies that advanced human cytogenetics have been applied to many mammals, including cattle. These include G-banding, Q-banding, R-banding, and C-banding with multiple variants of each. A systematic nomenclature for cattle chromosomes was first proposed at Reading University in 1976 (Ford et al. 1980). Although the Reading Conference did not produce a model ideogram, it served as a standard among the small community of livestock cytogeneticists. The Reading standard was improved in 1989 at a conference organized to produce an International System for Cytogenetic Nomenclature of Domestic Animals (ISCNDA) in Jouy-en-Josas (Di Berardino et al. 1990). Improved technology using prometaphase chromosomes and sequential G- and R-banding resolved more than 400 bands across the cattle chromosomes. Standard karyotypes and accompanying band-numbered ideograms provided useful tools for the growing number of scientists exploring the cattle genome at the chromosome level. However, inconsistencies between the Reading and ISCNDA standards were problematic. These issues were largely resolved at the 9th North American Colloquium on Domestic Animal Cytogenetics and Gene Nomenclature at Texas A&M University in 1995 (Popescu et al. 1996). By 1995, marker genes were available for each chromosome, thanks to somatic cell genetics and in situ hybridization (Fries et al. 1993). The publication includes the marker genes, along with the previously identified synteny groups, the Reading and ISCNDA designations, conserved synteny with human and sheep chromosomes, and the relative lengths of each chromosome, all presented in tabular form. Fries and Popescu (1999) revised the ISCNDA ideograms according to the Texas Standard and published state-of-the-art Q- and R-banded karyotypes. Although cattle chromosomes can still be distinguished in only a small number of laboratories, and not totally without disagreements, the inclusion of marker genes has made the Texas Standard the chromosomal anchor for subsequent radiation hybrid (RH) mapping, physical mapping, and whole genome sequencing.

In situ hybridization of molecular probes directly to cattle chromosomes was essential to the development of gene maps in cattle. Initial studies with radioactive probes were labor and time intensive but, nonetheless, successful in placing a handful of genes onto cattle chromosomes. Representative studies include those by Fries et al. (1986), Fries et al. (1988), Hediger et al. (1990), and Threadgill et al. (1990). These studies not only localized genes to chromosomes but in most cases were coupled with somatic cell genetics to assign synteny groups to chromosomes. A major breakthrough in chromosome cartography came with the development of fluorescence in situ hybridization (FISH) in the late 1980s (Pinkel et al. 1986). This technology not only eliminated the need for radioactive isotopes, it replaced the statistical evaluation of silver grains over chromosomes with direct observation of fluorescent signal. Application to bovine chromosomes resulted in the rapid assignment of all synteny groups to chromosomes (e.g., Gallagher et al. 1992; Hayes and Petit 1993; Friedl and Rottmann 1994). Because FISH can accommodate large probes, it can be used with large insert cloning vectors such as bacterial artificial chromosomes (BACs), yeast artificial chromosomes (YACs), and cosmids, integrating chromosomal maps with physical or linkage maps that employ the insert. For example, chromosome maps and linkage maps were integrated by FISH mapping of cosmids carrying microsatellites used to develop the bovine linkage map (Solinas Toldo et al. 1993; Ferretti et al. 1997). Multiple FISH assignments on the same chromosome provide the orientation of RH maps, physical maps, and the sequence of bovine chromosomes.

## Radiation Hybrid Mapping

The incredible revelation by Goss and Harris (1975), that irradiation of the donor cell line prior to the construction of hybrid somatic cells could order genes on a chromosome at a high level of resolution, was largely ignored by the human genetics community for 15 years. The rediscovery of RH mapping by Cox et al. (1990) produced an immediate quantum leap in the number of ordered markers on human chromosomes and the technology was subsequently adopted by those of us working with other mammalian species. When donor cells are irradiated prior to fusion with a recipient cell line, fragments of the donor chromosomes rather than whole chromosomes are randomly retained. The average fragment size is inversely proportional to the radiation dose, providing flexibility to the protocol depending on the level of resolution desired in the resultant maps. Loci in close proximity to each other are concordantly retained at a higher frequency than loci separated by a greater physical distance on the same chromosome, thus syntenic markers can be ordered on a map by the same rationale behind linkage mapping. In fact, the methods used in RH mapping are similar to those used in linkage mapping, employing two-point and multipoint analysis, maximum likelihood, map function, and lod scores. A variety of algorithms have been developed for RH mapping as reviewed by Matise et al. (1999). RH mapping is a variant of somatic cell mapping, both of which have the advantage over linkage mapping of not requiring polymorphic markers. Thus, RH mapping provided for the first time a robust method for ordering markers on chromosomes without the requirement of polymorphic markers and large numbers of meiotic products.

We developed a 5000 rad panel of RH clones for bovine genome mapping (Womack et al. 1997) that have been dispersed internationally and used for mapping several thousand markers. This panel was used initially to produce maps of individual bovine chromosomes (Band et al. 1998; Yang and Womack 1998; Rexroad and Womack 1999;

Amarante et al. 2000; Ozawa et al. 2000; Amaral et al. 2002; Antoniou et al. 2002; Ashwell et al. 2002; Goldammer et al. 2002; Schläpfer et al. 2002; Kurar et al. 2003) and also provided the platform for whole genome maps at increasingly higher levels of resolution (Band et al. 2000; Larkin et al. 2003; Everts-van der Wind et al. 2004, 2006). The panel has been instrumental in the search for functional elements underlying quantitative trait loci (QTL) (Takeda et al. 2002; Winter et al. 2002; Brunner et al. 2003; Schwerin et al. 2003; Goldammer et al. 2004; Sugimoto et al. 2006; Weikard et al. 2006), and has been employed to facilitate scaffold assembly in the bovine physical map (Marques et al. 2007) and in the final assembly of the whole genome sequence (Bovine Genome Sequencing and Analysis Consortium 2009). Rexroad et al. (2000) constructed a 12,000 rad panel, Williams et al. (2002) constructed a 3000 rad panel, and Itoh et al. (2005) constructed a 7000 rad panel, resulting in a rich resource of tools for bovine genome analysis as well as a variety of comprehensive bovine RH maps with varying levels of resolution and different types of markers (Williams et al. 2002; Itoh et al. 2005; McKay et al. 2006).

## *Comparative Mapping*

The cattle genome has been an integral component of the study of comparative genomics in mammals and is the subject of Chapter 8. However, the roots of comparative genomics are in comparative mapping, so a brief review of comparative mapping is in order here. The first comparative map including bovine genes was a somatic cell map that included only 32 loci on 21 cattle chromosomes (Womack and Moll 1986). Although the cattle genes were not ordered and therefore, intrachromosomal disruptions could not be observed, a comparison of cattle and human homologs revealed fewer disruptions in synteny (three) than did the comparison of the same homologs in humans and mice (nine). The prediction, that the big picture of mammalian chromosome evolution would eventually prove to be more conservative than was apparent from human–mouse comparisons, was on target (Murphy et al. 2005). The hypothesis that cattle–human genomes were highly conserved relative to mouse–human received substantial support from Zoo-FISH painting experiments. Hybridization of each of the single human chromosome painting probes to cattle chromosomes (Hayes 1995; Solinas Toldo et al. 1995; Chowdhary et al. 1996) revealed extensive homology compared to similar experiments with human probes on mouse chromosomes (Scherthan et al. 1994; Wienberg and Stanyon 1997). While somatic cell mapping of homologs of mapped human genes defined borders of conserved synteny on human chromosomes, and hybridization of human chromosome paints on cattle chromosomes defined borders of conserved synteny on cattle chromosomes, neither addressed conservation of order within conserved syntenic segments.

In situ hybridization of gene probes provided early insight into the order of bovine genes in cattle comparative maps (Hayes et al. 1993; Eggen and Fries 1995; Lòpez-Corrales et al. 1998). A comprehensive comparative map, based largely on in situ hybridization of gene markers was published by Hayes (1995). A high-resolution comparative map integrating FISH mapping at 598 loci with physical mapping (Schibler et al. 2006) highlighted the value of FISH for comparative mapping in cattle. RH mapping has proved to be a robust method for producing ordered comparative maps. The earliest RH maps of single bovine chromosomes often included comparison to their human counterparts (Yang et al. 1998; Gu et al. 1999; Gautier et al. 2002), and the comparison of cattle to human RH maps, which we termed "parallel radiation hybrid mapping" (Yang and Womack 1998; Rexroad and Womack 1999), provided the first insights into the extent of internal rearrangements accompanying the divergence of cattle and human chromosomes. A series of papers from Harris Lewin's laboratory (Band et al. 2000; Larkin et al. 2003; Everts-van der Wind et al. 2004, 2006; Itoh et al. 2005) present whole genome comparative RH maps that described internal rearrangements within conserved synteny blocks. These are described in detail in Chapter 8.

## Linkage Mapping

While somatic cell and RH maps provided extremely informative comparative maps and the first opportunities to exploit human and mouse genome data to interpret cattle genetics, the search for economically important traits in cattle awaited the development of linkage maps. Linkage mapping in cattle is discussed in detail in Chapter 6 and is only briefly outlined here. The first whole genome linkage maps for cattle were published in 1994. A set of resource families assembled by a consortium of cattle geneticists, were genotyped by a variety of laboratories around the world to produce a map of 202 markers comprising 36 linkage groups (Barendse et al. 1994). The mapping markers used were primarily microsatellites, although some restriction fragment length polymorphisms (RFLPs) in coding genes were included in order to integrate linkage maps with synteny maps and to permit comparisons of linkage between species. Linkage groups were assigned to 28 of the bovine autosomes as well as to the sex chromosomes, and the orientation of linkage groups was based on somatic cell and in situ hybridization maps as previously described. A linkage map produced independently at USDA-MARC (United States Department of Agriculture-Meat Animal Research Center) was published in the same year. The MARC map (Bishop et al. 1994) contained 313 markers in 30 linkage groups with 24 assignments to chromosomes or synteny groups. Both of these maps were expanded to second-generation maps published in 1997. Barendse et al. (1997) produced a map of 746 markers in 31 linkage groups, one for each chromosome, while Kappes et al. (1997) increased the MARC map to more than 1200 markers with an average marker interval of 2.5 cM. Other linkage maps were also being generated (Georges et al. 1995; Ma et al. 1996), and in combination with the development of markers for fine mapping, these maps opened the door to QTL mapping in cattle. Subsequent linkage mapping has progressed through a third-generation map (Ihara et al. 2004) with more than 3000 markers, mostly microsatellites, to the current generation of maps made possible by the whole genome sequencing initiative and single nucleotide polymorphism (SNP) discovery. High-throughput genotyping produced a 3000 SNP map for McKay et al. (2006) and contributed to the 17,000 marker composite map of Snelling et al. (2007). Other maps are being generated as investigators apply SNP chips of increasing marker density to populations segregating QTL and Mendelian traits of economic or biological interest.

## Physical Mapping

While FISH, somatic cell, and RH mapping employ physical methods (as opposed to genetic), the term "physical map" generally refers to a clone-based map of ordered contigs derived from large insert libraries. Physical maps provide enhanced resolution to comparative mapping, a resource for mining genes underlying traits of interest, and of course, a foundation for whole genome sequencing. A number of BAC and YAC libraries have been constructed from cattle nuclear genomes (Libert et al. 1993; Cai et al. 1995; Hills et al. 1999; Zhu et al. 1999; Buitkamp et al. 2000; Warren et al. 2000; Eggen et al. 2001). These libraries were extremely useful for building scaffolds under QTL and ultimately in gene mining, and a first-generation whole-genome physical map was constructed using the INRA (Eggen et al. 2001) and CHORI-240 (http://bacpac.chori.org/bovine240.htm) libraries (Schibler et al. 2004). This map contained 6615 contigs assembled from 100,923 clones, and was anchored with 747 clones at 1303 loci defined by microsatellites, genes, expressed sequence tags (ESTs), and BAC ends. This map was updated by Schibler et al. (2006) for high-resolution comparative mapping as described previously. The updated map contained 5081 contigs, 860 of which were anchored to the bovine genome and a sizable number anchored to the human genome.

## Cattle Genome Sequencing

As described in Chapter 9, the sequencing and annotation of the bovine genome was completed and reported in 2009 (Bovine Genome Sequencing and Analysis Consortium 2009). The project was initiated with funding from the NIH (National Institutes of Health) in response to a white paper proposal submitted in 2002. The success of the proposal was undoubtedly based largely on the body of knowledge produced by an active community of "cartographers" in the preceding quarter century.

## References

Amaral, M.E.J., Kata, S.R., Womack, J.E. (2002) A radiation hybrid map of bovine X chromosome (BTAX). *Mammalian Genome* **13**: 268–271.

Amarante, M.R.V., Yang, Y.P., Kata, S.R., Lopes, C.R., Womack, J.E. (2000) RH maps of bovine Chromosomes 15 and 29: conservation of human Chromosomes 11 and 5. *Mammalian Genome* **11**: 364–368.

Antoniou, E., Gallagher, D. Jr., Taylor, J., Davis, S., Womack, J., Grosz, M. (2002) A comparative map of bovine chromosome 25 with human chromosomes 7 and 16. *Cytogenetic and Genome Research* **97**: 128–132.

Ashwell, M.S., Sonstegard, T.S., Kata, S., Womack, J.E. (2002) A radiation hybrid map of bovine chromosome 27. *Animal Genetics* 75–76.

Band, M., Larson, J.H., Womack, J.E., Lewin, H.A. (1998) A radiation hybrid map of BTA23: Identification of a chromosomal rearrangement leading to separation of the cattle MHC class II subregions. *Genomics* **53**: 269–275.

Band, M.R., et al. (2000) An ordered comparative map of the cattle and human genomes. *Genome Research* **10**: 1359–1368.

Barendse, W., et al. (1994) A genetic linkage map of the bovine genome. *Nature Genetics* **6**: 227–235.

Barendse, W., et al. (1997) A medium-density genetic linkage map of the bovine genome. *Mammalian Genome* **8**: 21–28.

Barski, G., Sorieul, S., Cornefert, F. (1961) "Hybrid" type cells in combined cultures of two different mammalian cell strains. *Journal of the National Cancer Institute* **26**: 1269–1291.

Bishop, M.D., et al. (1994) A genetic linkage map for cattle. *Genetics* **136**: 619–639.

Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522–528.

Brunner, R.M., Sanftleben, H., Goldammer, T., Kühn, C., Weikard, R., Kata, S.R., Womack, J.E., Schwerin, M. (2003) The telomeric region of BTA18 containing a potential QTL region for health in cattle exhibits high similarity to the HSA19q region in humans. *Genomics* **81**: 270–278.

Buitkamp, J., Kollers, S., Durstewitz, G., Fries, R., Welzel, K., Schafer, K., Kellermann, A., Lehrach, H. (2000) Construction and characterization of a gridded cattle BAC library. *Animal Genetics* **31**: 347–351.

Cai, L., Taylor, J.F., Wing, R.A., Gallgher, D.S., Woo, S.S., Davis, S.K. (1995) Construction and characterization of bovine bacterial artificial chromosome library. *Genomics* **29**: 413–425.

Caspersson, T., Farber, S., Foley, G.E., Kudynowski, J., Modest, E.J., Simonsson, E., Wagh, U., Zech, L. (1968) Chemical differentiation along metaphase chromosomes. *Experimental Cell Research* **49**: 219–222.

Caspersson, T., Zech, L., Johansson, C. (1970) Differential binding of alkylating fluorochromes in human chromosomes. *Experimental Cell Research* **60**: 315–319.

Chowdhary, B.P., Fronicke, L., Gustavsson, I., Scherthan, H. (1996) Comparative analysis of the cattle and human genomes: detection of ZOO-FISH and gene mapping-based chromosomal homologies. *Mammalian Genome* **7**: 297–302.

Cox, D.R., Burmeister, M., Price, E.R., Kim, S., Myers, R.M. (1990) Radiation hybrid mapping: A somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250**: 245–250.

Dain, A.R., Tucker, E.M., Donker, R.A., Clarke, S.W. (1984) Chromosome mapping in cattle using mouse myeloma/calf lymph node hybridomas. *Biochemical Genetics* **22**: 249–439.

Di Berardino, D., Hayes, H., Fries, R., Long, S.E (1990) International System for Cytogenetic Nomenclature of Domestic Animals (ISCNDA 1989). *Cytogenetics and Cell Genetics* **53**: 65–79.

Echard, G., Gellin, J., Benne, F., Gillois, M. (1984) Progress in gene mapping in cattle and pigs using somatic cell hybridization. Human-gene mapping 7. *Cytogenetics and Cell Genetics* **37**: 458–459.

Eggen, A. and Fries, R. (1995) An integrated cytogenetic and meiotic map of the bovine genome. *Animal Genetics* **26**: 215–236.

Eggen, A., Gautier, M., Billaut, A., Petit, E., Hayes, H., Laurent, P., Urban, C., Pfister-Genskow, M., Eilertsen, K., Bishop, M.D. (2001) Construction and characterization of a bovine BAC library with four genome-equivalent coverage. *Genetics Selection Evolution* **33**: 543–548.

Everts-van der Wind, A., Larkin, D.M., Green, C.A., Elliott, J.S., Olmstead, C., Chiu, R., Schein, J.E., Marra, M.A., Womack, J.E., Lewin, H.A. (2006) A high-resolution whole-genome cattle-human comparative map reveals details of mammalian chromosome evolution. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 18526–18531.

Everts-van der Wind, A., et al. (2004) A 1463 gene cattle-human comparative map with anchor points defined by human genome sequence coordinates. *Genome Research* **14**: 1424–1437.

Ferretti, L., et al. (1997) Cosmid-derived markers anchoring the bovine genetic map to the physical map. *Mammalian Genome* **8**: 29–36.

Ford, C.E., Pollok, D.L., Gustavsson, I. (1980) Proc. 1st International Conference for the Standardization of Banded Karyotypes of Domestic Animals, Reading, England, 1976. *Hereditas* **92**: 145–162.

Friedl, R. and Rottmann, O. (1994) Assignment of the cation independent mannose 6-phosphate/insulin-like growth factor II receptor to bovine chromosome 9q27–28 by fluorescent *in situ* hybridization. *Animal Genetics* **25**: 191–193.

Fries, R. and Popescu, P. (1999) Cytogenetics and physical chromosome maps. In: *The Genetics of Cattle*, edited by R. Fries and A. Ruvinski, pp. 247–327. Oxon: CAB International.

Fries, R., Hediger, R., Stranzinger, G. (1986) Tentative chromosomal localization of the bovine major histocompatibility complex by in situ hybridization. *Animal Genetics* **17**: 287–294.

Fries, R., Hediger, R., Stranzinger, G. (1988) The loci for parathyroid hormone and beta-globin are closely linked and map to chromosome 15 in cattle. *Genomics* **3**: 302–307.

Fries, R., Eggen, A., Womack, J.E. (1993) The bovine genome map. *Mammalian Genome* **4**: 405–428.

Gallagher, D.S., Grosz, M., Basrur, P.K. Skow, L., Womack, J.E. (1992) Chromosomal localization in cattle of BoLA and HSP70 genes by fluorescent in situ hybridization and confirmation of the identity of an autosome to X chromosome translocation. *Animal Genetics* **23**: 81.

Gautier, M., Hayes, H., Eggen, A. (2002) An extensive and comprehensive radiation hybrid map of bovine Chromosome 15: comparison with human Chromosome 11. *Mammalian Genome* **13**: 316–319.

Georges, M., et al. (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**: 907–920.

Goldammer, T., Kata, S.R., Brunner, R.M., Dorroch, U., Sanftleben, H., Schwerin, M., Womack, J.E. (2002) A comparative radiation hybrid map of bovine chromosome 18 and homologous chromosomes in human and mice. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 2106–2111.

Goldammer, T., Kata, S.R., Brunner, R.M., Kühn, C., Weikard, R., Womack, J.E., Schwerin, M. (2004) High-resolution comparative mapping between human chromosomes 4 and 8 and bovine chromosome 27 provides genes and segments serving as positional candidates for udder health in cattle. *Genomics* **84**: 696–706.

Goss, S.J. and Harris, H. (1975) New method for mapping genes in human chromosomes. *Nature* **255**: 680.

Gu, Z., Womack, J.E., Kirkpatrick, B.W. (1999) A radiation hybrid map of bovine chromosome 7 and comparative mapping with human chromosome 19 p arm. *Mammalian Genome* **10**: 1112–1114.

Hayes, H. (1995) Chromosome painting with human chromosome-specific DNA libraries reveals the extent and distribution of conserved segments in bovine chromosomes. *Cytogenetic and Cell Genetics* **71**: 168–174.

Hayes, H. and Petit, E. (1993) Mapping of the beta-lactoglobulin gene and of an immunoglobulin M heavy chain-like sequence to homoeologous cattle, sheep, and goat chromosomes. *Mammalian Genome* **4**: 207–210.

Hayes, H.C., Popescu, P., Dutrillaux, B. (1993) Comparative gene mapping of lactoperoxidase, retinoblastoma, and alpha-lactalbumin genes in cattle, sheep, and goats. *Mammalian Genome* **4**: 593–597.

Hediger, R., Johnson, S.E., Barendse, W., Drinkwater, R.D., Moore, S.S., Hetzel, J. (1990) Assignment of the growth hormone gene locus to 19q26-qter in cattle and to 11q25-qter in sheep by *in situ* hybridization. *Genomics* **8**: 171–174.

Heuertz, S. and Hors-Cayla, M.-C. (1978) Carte génétique des bovins par la technique d'hybridation cellulaire. Localisation sur le chromosome X de la glucose-6-phosphate déshydrogénase, la phosphoglycérate kinase, l'$\alpha$-galactosidase a et l'hypoxanthine guanine phosphoribosyl transférase. *Annales de Génétique* **21**: 197–202.

Heuertz, S. and Hors-Cayla, M.-C. (1981) Cattle gene mapping by somatic cell hybridization study of 17 enzyme markers. *Cytogenetics and Cell Genetics* **30**: 137–145.

Hills, D., Tracey, S., Masabanda, J., Fries, R., Schalkwyk, L.C., Lehrach, H., Miller, J.R., Williams, J.L. (1999) A bovine YAC library containing four- to five-fold genome equivalents. *Mammalian Genome* **10**: 837–838.

Ihara, N., et al. (2004) A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Research* **14**: 1987–1998.

Itoh, T., Watanabe, T., Ihara, N., Mariani, P., Beattie, C.W., Sugimoto, Y., Takasuga, A. (2005) A comprehensive radiation hybrid map of the bovine genome comprising 5593 loci. *Genomics* **85**: 413–424.

Kappes, S.M., Keele, J.W., Stone, R.T., McGraw, R.A., Sonstegard, T.S., Smith, T.P., Lopez-Corrales, N.L., Beattie, C.W.. (1997) A second-generation linkage map of the bovine genome. *Genome Research* **7**: 235–249.

Kurar, E., Womack, J.E., Kirkpatrick, B.W. (2003) A radiation hybrid map of bovine chromosome 24 and comparative mapping with human chromosome 18. *Animal Genetics* **34**: 198–204.

Lalley, P.A., O'Brien, S.J, Créau-Goldberg, N., Davisson, M.T., Roderick, T.H., Echard, G., Womack, J.E., Graves, J.M., Doolittle, D.P., Guidi, J.N. (1987) Report on the committee on comparative mapping. *Cytogenetics and Cell Genetics* **46**: 367–389.

Larkin, D.M., et al. (2003) A cattle-human comparative map built with cattle BAC-ends and human genome sequence. *Genome Research* **13**: 1966–1973.

Libert, F., Lefort, A., Okimoto, R., Womack, J., Georges, M. (1993) Construction of a bovine genomic library of large yeast artificial chromosome clones. *Genomics* **18**: 270–276.

Lòpez-Corrales, N.L., Sonstegard, T.S., Smith, T.P.L. (1998) Comparative gene mapping: cytogenetic localization of PROC, EN1, ALPI, TNP1, and IL1B in cattle and sheep reveals a conserved rearrangement relative to human genome. *Cytogenetics and Cell Genetics* **83**: 35–38.

Ma, R.Z., et al. (1996) A male linkage map of the cattle (*Bos taurus*) genome. *Journal of Heredity* **87**: 261–271.

Markert, C.L. (1968) The molecular basis for isozymes. *Annals of the New York Academy of Sciences* **151**: 14–40.

Marques, E., de Givry, S., Stothard, P., Murdoch, B., Wang, Z., Womack, J., and Moore, S.M. (2007) A high resolution radiation hybrid map of bovine chromosome 14 identifies scaffold rearrangement in the latest bovine assembly. *BMC Genomics* **8**: 254.

Matise, T.C., Wasmuth, J.J., Myers, R.M., McPherson, J.D. (1999) Somatic cell genetics and radiation hybrid mapping. In: *Genome Analysis: a Laboratory Manual, Mapping Genomes*, edited by B. Birren, E.D. Green, P. Hieter, S. Klapholz, R.M. Myers, H. Riethman, and J. Roskams, pp. 259–267. New York: Cold Spring Harbor Laboratory Press.

McKay, S.D., et al. (2006) Construction of bovine whole-genome radiation hybrid and linkage maps using high-throughput genotyping. *Animal Genetics* **38**: 120–125.

Minna, J.D., Lalley, P.A., Francke, U. (1976) Comparative mapping using somatic cell hybrids. *In vitro* **12**: 726–733.

Murphy, W.J., et al. (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**: 613–617.

Nadeau, J.H. (1989) Maps of linkage and synteny homologies between mouse and man. *Trends in Genetics* **5**: 82–86.

Nadeau, J.H. and Tayler, B.A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of America* **81**: 814–818.

O'Brien, S.J., Seuanez, N.H., Womack, J.E. (1988) Mammalian genome organization: an evolutionary view. *Anuual Review of Genetics* **22**: 323–351.

Ozawa, A., Band, M.R., Larson, J.H., Donovan, J., Green, C.A., Womack, J.E., Lewin, H.A. (2000) Comparative organization of cattle chromosome 5 revealed by comparative mapping by annotation and sequence similarity and radiation hybrid mapping. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 4150–4155.

Pinkel, D., Straume, T., Gray, J.W. (1986) Cytogenetic analysis using quantitative, high-sensitivity fluorescence hybridization. *Proceedings of the National Academy of Sciences of the United States of America* **83**: 2934–2938.

Popescu, C.P., Long, S., Riggs, P., Womack, J., Schmutz, S., Fries, R., Gallagher, D.S. (1996) Standardization of cattle karyotype nomenclature: report of the committee for the standardization of the cattle karyotype. *Cytogenetics and Cell Genetics* **74**: 259–261.

Renwick, J.H. (1971) The mapping of human chromosomes. *Annual Reviews of Genetics* **5**: 81–120.

Rexroad, C.E. III and Womack, J.E. (1999) Parallel RH mapping of BTA1 with HSA3 and HSA21. *Mammalian Genome* **10**: 1095–1097.

Rexroad, C.E. III, Owens, E.K., Johnson, J.S., Womack, J.E. (2000) A 12000 rad whole genome radiation hybrid panel for high resolution mapping in cattle: characterization of the centromeric end of chromosome 1. *Animal Genetics* **31**: 262–265.

Ruddle, F.H. (1972) Linkage analysis using somatic cell hybrids. *Advancement in Human Genetics* **3**: 173–235.

Ruddle, F.H. (1981) A new era in mammalian gene mapping: somatic cell genetics and recombinant DNA methodologies. *Nature* **294**: 115–120.

Ruddle, F.H. (1984) The William Allan Memorial Award Address: Reverse genetics and beyond. *American Journal of Human Genetics* **36**: 944–954.

Ruddle, F.H. and Creagan, R.P. (1975) Parasexual approaches to the genetics of man. *Annual Review of Genetics* **9**: 407–486.

Scherthan, H., Cremer, T., Arnason, U., Weier, H.U., Lima-de-Faria, A., Froenicke, L. (1994) Comparative chromosome painting discloses homologous segments in distantly related mammals. *Nature Genetics* **6**: 342–347.

Schibler, L., Roig, A., Mahé, M.-F., Save, J.-C., Gautier, M., Taourit, S., Boichard, D., Eggen, A., Cribiu, E.P. (2004) A first generation bovine BAC-based physical map. *Genetics Selection Evolution* **36**: 105–122.

Schibler, L., Vaiman, D., Oustry, A., Giraud-Delville, C., Cribiu, E.P. (2006) Comparative gene mapping: a fine-scale survey of chromosome rearrangements between ruminants and humans. *Genome Research* **8**: 901–915.

Schläpfer, J., Stahlberger-Saitbekova, N., Comincini, S., Gaillard, C., Hills, D., Meyer, R.K., William, J.L., Zurbriggen, A., Dolf, G. (2002) A higher resolution radiation hybrid map of the bovine chromosome 13. *Genetics Selection Evolution* **34**: 255–267.

Schwerin, M., Czernek-Schafer, D., Goldammer, T., Kata, S.R., Womack, J.E., Pareek, R., Pareek, C., Krzysztof, W., Brunner, R.M. (2003) Application of disease-associated differentially expressed genes - Mining for functional candidate genes for mastitis resistance in cattle. *Genetics Selection Evolution* **35**: S19–S34.

Shimizu, N., Shimizu, Y., Kondo, I., Woods, C., Wegner, T. (1981) The bovine genes for phosphoglycerate kinase, glucose-6-phosphate dehydrogenase, alpha-galactosidase, and hypoxanthine phosphoribosyltransferase are linked to the X chromosome in cattle-mouse cell hybrids. *Cytogenetics and Cell Genetics* **29**: 26–31.

Snelling, W.M., et al. (2007) A physical map of the bovine genome. *Genome Biology* **8**: R165.

Solinas Toldo, S., Fries, R., Steffen, P., Neibergs, H.L., Barendse, W., Womack, J.E., Hetzel, D.J.S., Stranzinger, G. (1993) Physically mapped, cosmid-derived microsatellite markers as anchor loci on bovine chromosomes. *Mammalian Genome* **4**: 720–727.

Solinas Toldo, S., Lengauer, C., Fries, R. (1995) Comparative genome map of human and cattle. *Genomics* **27**: 489–496.

Sugimoto, M., Fujikawa, A., Womack, J.E., Sugimoto, Y. (2006) Evidence that bovine forebrain embryonic zinc finger-like gene influences immune response associated with mastitis resistance. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 6454–6459,

Takeda, H., et al. (2002) Positional cloning of the gene LIMBIN responsible for bovine chondrodysplastic dwarfism. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 10549–10554.

Threadgill, D.W., Fries, R., Faber, L.K, Vassart, G., Gunawardana, A., Stranzinger, G., Womack, J.E. (1990) The thyroglobulin gene is syntenic with the MYC and MOS protooncogenes and carbonic anhydrase II and maps to chromosome 14 in cattle. *Cytogenetics and Cell Genetics* **53**: 32–36.

Warren, W., Smith, T.P., Rexroad, C.E III, Fahrenkrug, S.C., Allison, T., Shu, C.L., Catanese, J., de Jong, J. (2000) Construction and characterization of a new bovine bacterial artificial chromosome library with 10 genome-equivalent coverage. *Mammalian Genome* **11**: 662–663.

Weikard, R., Goldammer, T., Laurent, P., Womack, J.E., Kuehn, C. (2006) A gene-based high-resolution comparative radiation hybrid map as a framework for genome sequence assembly of a bovine chromosome 6 region associated with QTL for growth, body composition, and mild performance traits. *BMC Genomics* **7**: 53 (1–15).

Weiss, M.C. and Ephrussi, B. (1966) Studies of interspecific (rat x mouse) somatic hybrids. II. Lactic dehydrogenase and β-glucuronidase. *Genetics* **54**: 1111–1122.

Wienberg, J. and Stanyon, R. (1997) Comparative painting of mammalian chromosomes. *Current Opinion in Genetics Development* **7**: 784–791.

Williams, J.L., et al. (2002) A bovine whole-genome radiation hybrid panel and outline map. *Mammalian Genome* **13**: 469–474.

Winter, A., Krämer, W., Werner, F.A., Kollers, S., Kata, S., Durstewitz, G., Buitkamp, J., Womack, J.E., Thaller, G., Fries, R. (2002) Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (*DGAT1*) with variation at a quantitative trait locus for milk fat content. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 9300–9305.

Womack, J.E. and Moll, Y.D. (1986) Gene map of the cow: conservation of linkage with mouse and man. *Journal of Heredity* **77**: 2–7.

Womack, J.E., Johnson, J.S., Owen, E.K., Rexroad, C.E. III, Schläpfer, J., Yang, Y.P. (1997) A whole-genome radiation hybrid panel for bovine gene mapping. *Mammalian Genome* **8**: 854–856.

Yang, Y.-P. and Womack, J.E. (1998) Parallel radiation hybrid mapping: A powerful tool for high-resolution genomic comparison. *Genome Research* **8**: 731–736.

Yang, Y.-P., Rexroad, C.E. III, Schlapfer, J., Womack, J.E. (1998) An integrated radiation hybrid map of bovine chromosome 19 and ordered comparative mapping with human chromosome 17. *Genomics* **48**: 93–99.

Zhu, B., et al. (1999) A 5x genome coverage bovine BAC library: production, characterization, and distribution. *Mammalian Genome* **10**: 706–709.

Chapter 6

# History of Linkage Mapping the Bovine Genome

*Stephanie D. McKay and Robert D. Schnabel*

## Introduction

The driving force behind the construction of genetic linkage maps in cattle has been the hunt for quantitative trait loci (QTL) and economically important traits. Bovine genetic linkage mapping closely followed the evolution of molecular markers from restriction fragment length polymorphisms (RFLP) through the various classes of variable number tandem repeats (VNTR) and then finally single nucleotide polymorphisms (SNP). This led to the bovine genome being the fourth most densely mapped mammalian genome by 1997 (Barendse et al. 1997), after human, mouse, and rat. An initial surge of genomewide low-density linkage maps were produced in the mid 1990s that had sufficient map density for the identification of QTL regions. However, these QTL regions were quite large and necessitated refocusing from generating whole genome maps to chromosome-specific higher density maps in order to fine map QTL identified from genome scans. Eventually, the bovine genome sequencing initiative (The Bovine Genome Sequencing and Analysis Consortium 2009) prompted a resurrection of genomewide linkage maps with increased marker density and increased resolution that aided in the assembly of the genome. Even now, after the bovine genome sequence has been assembled and annotated, there are linkage maps being generated that have been utilized for identifying and correcting discrepancies in the sequence assembly (Arias et al. 2009). The following discussion is a brief history of the development of the various genomewide linkage maps and two examples of how chromosome-specific maps were used to localize economically important traits.

## The First Genome-wide Maps

In 1994, two manuscripts were published that independently reported genetic linkage maps in cattle (Barendse et al. 1994; Bishop et al. 1994). Prior to these publications, the state of the bovine genome map primarily consisted of cytogenetic maps with a few microsatellites assigned (see Chapter 5 for review of cytologic and radiation hybrid (RH) mapping). The map presented by Bishop et al. (1994) contained 313 polymorphic

*Bovine Genomics*, First Edition. Edited by James E. Womack.

markers of which 280 were microsatellites and were ordered in 30 linkage groups that were anchored to 24 autosomal chromosomes and four synteny groups. Linkage groups were unable to be assigned to five bovine autosomes: (1) BTA2, (2) BTA9, (3) BTA12, (4) BTA22, and (5) BTA27. The linkage groups ranged in size from 28.2 to 152.5 cM with an average length of 74.2 cM. Comparatively, Barendse et al. (1994) published a genetic linkage map of the bovine genome the same year and reported that their map covered about 90% of the expected length of the bovine genome. Of the 202 markers mapped, 144 were microsatellites and formed 36 linkage groups. The sex-averaged map totaled 2513 cM, representing 90% of the estimated 2800 cM bovine genome.

In addition to adding markers to the bovine map, Barendse et al. (1994) presented novel results with respect to sex-specific maps and conservation of synteny. First, the authors reported that there was less than a 50-cM difference between sex-specific total map lengths, which is cited as evidence of very little difference between the recombination frequency of males and females. Female linkage maps are expected to be longer than male linkage maps. However, female maps in humans have been documented to be almost double the length of the male maps (Matise et al. 1994). Second, 56 of the 202 mapped markers originated in humans, mice, and sheep, allowing comparative mapping using the genetic maps of humans and mice. Examination of the comparative maps indicated that the greatest conservation of synteny was between humans and cattle. However, as the authors pointed out, conservation of synteny does not guarantee conservation of gene order.

Georges et al. (1995) in an effort to advance QTL discovery, constructed a linkage map for the specific purpose of mapping QTL controlling milk production traits in Holstein cattle. Of the 159 microsatellite markers genotyped, 138 were assigned to 27 linkage groups. These 27 linkage groups were assigned to 24 autosomal synteny groups producing a total map length of 1645 cM flanked by linked markers. The authors assumed the male cattle genome measured 2500 cM and their map would then account for approximately 66% of the genome. Interestingly, the simulation study performed by the authors indicated that a map generated with 150 randomly selected microsatellite markers would be expected to cover approximately 1343 cM. The discrepancy between the obtained map distance and the simulated map distance was attributed to typing errors that produced inflated map distances. The 104,523 genotypes produced in the project underwent additional analysis that resulted in identification of QTL controlling milk production on five autosomal chromosomes. This represented the first successful use of the same animals to produce both linkage maps and identify QTL regions in cattle. One of the interesting comments from Georges et al. concerned the production capacity of the genotyping process. Microsatellite markers were so convenient and abundant that they were able to semiautomate the genotyping process and an individual was reportedly able to produce approximately 10,000 genotypes per month. Today, using the BovineSNP50 BeadChip (Matukumalli et al. 2009), one individual is capable of producing in excess of 270 million genotypes per month, and in the very near future it will be possible and cost effective to simply resequence the entire genome of an individual.

In an attempt to identify discrepancies in marker order and map distances, Ma et al. (1996) constructed a linkage map comprising 269 markers. Of the 249 microsatellites in this data set, 140 were selected from previously described maps. The resulting male linkage map contained 35 linkage groups assigned to all 29 bovine autosomes and 8

**Table 6.1** Comparison and progression of genome-wide linkage maps in cattle.

| Author year | Number of markers | Number of SNPs | Total map length (cM) | Autosomal map length (cM) |
|---|---|---|---|---|
| Bishop et al. 1994 | 313 | | $2464^A$ | |
| Barendse et al. 1994 | 202 | | | $2513^A$ |
| Georges et al. 1995 | 159 | | | $1645^M$ |
| Ma et al. 1996 | 269 | | | $1975^M$ |
| Barendse et al. 1997 | 746 | | $3567^M$ | $3532^A$ |
| | | | $3765^F$ | $3528^M$ |
| | | | | $3587^F$ |
| Kappes et al. 1997 | 1250 | | $2990^A$ | $2839^A$ |
| Ihara et al. 2004 | 3960 | 79 | $3160^A$ | $3013^A$ |
| Snelling et al. 2005 | 4585 | 918 | $3058^A$ | |
| McKay et al. 2007 | 2701 | 2701 | | $2890^A$ |
| Arias et al.2009 | 7066 | 6769 | $3249^A$ | $3097^A$ |

A, M, F indicate sex-averaged, male-, or female-specific map lengths, respectively.

of the autosomes included multiple linkage groups. The average intermarker distance was reported to be 9.73 cM with a total map length of 1975 cM. The addition of previously mapped markers into the current male linkage map enabled the authors to compare chromosome assignments, marker order, and map intervals between their map and previously described maps (Barendse et al. 1994; Bishop et al. 1994; Georges et al. 1995). Generally, the marker order was consistent between maps with a few disagreements present. In fact, discrepancies in chromosome assignment were found for only four microsatellites. However, large discrepancies were found on several chromosomes for 68 common marker intervals, with some marker intervals differing by more than 10 cM. Differences within and between sex recombination rates may explain the differences when comparing male and sex-averaged linkage maps (Table 6.1).

## Second-Generation Maps

In 1997, Barendse et al. (1997) published a second bovine genetic map with an increased density that represented a 95% coverage of the bovine genome. This map contained 746 polymorphic markers that resulted in 31 linkage groups, one linkage group per chromosome. While the total autosomal map length was reported to be 3532 cM, the difference between sex-specific autosomal maps was reported to be 58 cM in favor of the female, which was consistent with the author's previously published findings (Barendse et al. 1994). The authors revisited the question of gene order conservation in conserved synteny between human–cattle comparative maps and concluded these regions did not demonstrate conserved gene order. Coincidently, this medium-density map was published at approximately the same time when scientist from the human genome project revealed a map of more than 16,000 human genes (Schuler et al. 1996), thus providing a more thorough human map for comparative purposes. These findings reaffirm the observation that karyotypic evolution, namely

Robertsonian translocations and inversion within a chromosome tend to keep genes syntenic, but exchange between chromosomes disrupt gene order within the syntenic groups. Subsequently, many more markers would need to be mapped in cattle in order to determine the boundaries for blocks of conserved synteny.

At this point, bovine linkage maps had a marker density sufficient for detecting QTL regions but lacked the resolution needed for efficient use of marker-assisted selection (Kappes et al. 1997). Therefore, a second-generation linkage map was assembled. This map included 623 previously mapped markers originating from the studies previously described, and an additional 627 new markers were incorporated producing a linkage map covering 2990 cM with an average interval length of 2.5 cM. Kappes et al. (1997) examined error detection by comparing linkage group distances between maps with common markers. The resulting comparisons indicated that the maps produced by Barendse et al. (1994) and Bishop et al. (1994) were considerably larger than the map presented by Kappes et al. (1997), with the difference in lengths attributed to the greater ability for error detection with increased marker density. However, the male maps presented by Ma et al. (1996) and Georges et al. (1995) are more similar in length to the Kappes map than the previous Barendse et al. (1994) or Bishop et al. (1994) maps. Additionally, the female map presented by Kappes et al. (1997) was 71 cM longer than the corresponding male map, which is in agreement with the Barendse et al. (1994) finding of a less than 50 cM difference between sex-specific maps. The increased marker density in the map presented by Kappes et al. (1997) was four times greater than previously published maps and clearly improved the map resolution needed for QTL detection and marker-assisted selection. However, as marker density increased and the ability to detect genotyping errors improved, the effect of an undetected genotyping error also increased.

The availability of linkage maps with sufficient density to identify QTL regions of approximately 20 cM transformed the perspective in which linkage mapping was approached. Despite a sufficient resolution to identify 20 cM QTL regions, the need still persisted for smaller intermarker distances and for markers associated with known genes. Addressing these needs would facilitate the gain of additional information from markers in QTL regions and improve the identification of positional candidate genes. Therefore, instead of generating genome-wide linkage maps, additional markers were developed specifically to fine map regions of interest on chromosomes detected to harbor QTLs (Ponce de Leon et al. 1996; Yeh et al. 1996; Sonstegard et al. 1997b, 2001; Sun et al. 1997; Taylor et al. 1997; Casas et al. 1999; Larsen et al. 1999; Gu et al. 2000; Smith et al. 2000; Kurar et al. 2002; Snelling et al. 2004). It was not until early the next century that the need for whole genome linkage maps resurfaced.

## Third-Generation Maps

In the early 2000s, advancements in microsatellite scoring again aided the development of higher density maps. Ihara et al. (2004) integrated markers from the second-generation map of Kappes et al. (1997) with 2293 new microsatellite markers, resulting in a 3160-cM genetic map encompassing 3960 polymorphic markers covering all autosomes and the X chromosome with an average intermarker distance of 1.4 cM. This dramatic increase in the number of microsatellite markers mapped on the Ihara map was the result of a concentrated effort to develop additional markers and subsequently narrow QTL critical regions. Despite the drastic increase in the number of markers mapped, comparisons between the total map length of the Ihara et al. (2004) and Kappes et al. (1997) maps indicate that the Ihara map was only 170 cM longer than the Kappes map. This minor difference in total map length is the result of increasing the number of markers without increasing the potential number of informative meiosis needed to ensure accurate marker order. Of the 3960 markers on the Ihara map, only 2389 have distinct map locations, meaning that multiple markers were binned and share the same map location. Nevertheless, the authors estimate that the resulting map had a potential genetic resolution of approximately 0.8 cM compared with 2.5 cM of the Kappes map.

By the early-to-mid 2000s, it was becoming possible to discover and genotype SNPs relatively easily and they were, therefore, being integrated into genetic maps. Ihara et al. (2004) initially incorporated SNPs; however, Snelling et al. (2005) took advantage of expressed sequence tags (ESTs) and bacterial artificial chromosome end sequences (BES) in order to identify 918 SNPs. These new markers were integrated into existing maps and were thus able to add, by proxy, a large number of genes into the bovine linkage map. The resulting map was 3058 cM and contained 4585 total genetic markers including 3612 microsatellites and 918 SNPs. While this map contained 16% more markers, it was 102 cM shorter than Ihara et al. (2004). The authors were concerned about the accuracy of placing SNPs onto the linkage map because of the less informative nature of SNPs compared to microsatellites. However, 80% of the developed SNPs were positioned on the linkage map. The increased length of the Snelling et al. (2005) map compared to the autosomal maps of Ihara et al. (2004) and Kappes et al. (1997) is attributed in part to incorrect ordering of markers. Correlations between marker position between the Snelling et al. (2005) and Ihara et al. (2004) maps was $r>0.99$. Adding SNP markers resulted in a shift in position of only 8% of the markers on the autosomal map. However, the mapping population utilized in this study allows for a maximum of 412 informative meiosis, thus limiting the resolution of the genetic map. If markers were not separated by recombination events in the mapping population, then the markers cannot be correctly ordered.

The first use of high-throughput genotyping for linkage mapping purposes was seen in 2007 (McKay et al. 2007). Two custom oligo pooled assay (OPA) totaling 3072 SNP markers were generated based on the available sequence from the genome sequencing effort and genotyped on the Illumina BeadStation (Illumina, San Diego, CA, USA). Using 80 registered Angus sires, a linkage map was constructed with genotypes produced from these custom OPAs. Construction of this linkage map was somewhat unconventional and involved multiple steps that utilized homologous bovine sequence coordinates, orthologous human sequence coordinates, human–bovine comparative maps (Itoh et al. 2005), RH map locations (McKay et al. 2007), as well as CRIMAP (Green et al. 1990). Of the 3072 SNPs contained in the OPAs, 2701 were successfully mapped, resulting in a 2890 cM linkage map. When this map was compared to the bovine genome sequence (Btau_2.0), disagreement in marker order were found on seven chromosomes, inversions were identified on an additional two chromosomes, and 133 discordant SNP markers were identified.

In 2007, Snelling et al. created a composite map of the bovine genome comprising two genetic maps (Williams et al. 2002; Ihara et al. 2004; McKay et al. 2007; Snelling et al. 2007) and vectors from three RH panels (Everts-van der Wind et al. 2004; Itoh et al. 2005; Jann et al. 2006; McKay et al. 2007). These data were consolidated to produce a single composite map totaling 17,254 unique markers. Consolidating RH and linkage data involved addressing the limitations of both genetic and RH mapping. Specifically, the available genetic maps lack the recombination events necessary to separate closely linked markers and RH maps may have unreliable whole chromosomes marker order. Exploiting the long-range resolution of linkage maps and the short-range resolution of RH maps, a composite map has the potential to overcome the shortcomings of each type of map. Indeed, this composite map proved to be a valuable resource, which was used to aid in the production of an alternative assembly (UMD_3.1) of the bovine genome (Zimin et al. 2009).

One would think that after a whole genome sequence was publicly available for an organism that mapping efforts would cease. Historically, this has not been the case. In fact, after the human genome sequence assembly was released, mapping efforts continued in an attempt to identify discrepancies and improve the latest version of the assembly, as well as build recombination maps in order to study genomic variability (Kong et al. 2002). Similar events took place after the bovine genome sequence was publicly available. The first linkage map released post bovine assembly was that of Arias et al. (2009), which used the Affymetrix GeneChip Bovine Mapping 10K SNP kit to genotype 1679 animals resulting in 7066 markers mapped. Based upon this manuscript, there appears to be an indirectly proportional relationship between the cost of genotyping and the complexity of map construction. In order to generate high-density linkage maps, the authors utilized a complex mapping scheme that involved five rounds of mapping initiated by construction of low density microsatellite-based linkage maps with subsequent mapping rounds that incorporated SNPs from the Affymetrix 10K SNP chip. The resulting maps contained a total of 7066 markers mapped, including 6769 SNPs, 294 microsatellites, and 3 haplotypes. The average correlation between the Arias et al. (2009) linkage map and version Btau_4.0 of the bovine genome sequence was reported as 0.985. While the overall correlation was high, discrepancies were noted on BTA3 and BTA27. The large number of informative meiosis resulted in a higher resolution map than that presented by Snelling et al. (2005), while the length of the autosomal map presented was only 83.9 cM larger than the Ihara et al. (2004) genetic map. Additional high-density linkage maps are likely to be produced based on the large number of cattle that are being genotyped using SNP chips. As in other model organisms, these linkage maps will be a resource for further refinement of the bovine reference genome assembly.

## Single Chromosome Maps

As previously stated, once the marker density of genome-wide linkage maps was sufficient to support QTL analysis, the focus on linkage mapping switched to fine mapping chromosomes of interest where economically important QTL regions had been detected. Numerous fine mapping projects were undertaken and several led to narrowing a QTL region to less than 5 Mb or even aided in determining the causative mutation for an economically important trait. In order to demonstrate the role that linkage mapping has played in QTL discovery, we shall address two prominent examples; the *POLL* locus on BTA1 and milk production QTL on BTA6.

## BTA1

The hunt for one of the most infamous economically important traits in bovine began in 1993 when the *POLL* locus was mapped to BTA1 (Georges et al. 1993). Nine paternal half-sib families, representing three *Bos taurus* breeds, and their 138 offspring were genotyped with 38 minisatellite markers and 233 microsatellites (Georges et al. 1991). Pairwise linkage analysis was performed between each marker and the *polled* locus. Subsequently, linkage was identified between the *POLL* locus and microsatellite markers GMPOLL-1 and GMPOLL-2. While the authors were able to determine that the *polled* locus did not lie between the two anonymous microsatellites, they were not able to definitively determine the position of the *polled* locus relative to the two microsatellites. These markers were located on bovine synteny group 10, which the authors were able to assign to BTA1 using a somatic cell hybrid panel (Dietz et al. 1992). However, the authors were unable to show evidence for linkage between these markers and any other marker mapped to BTA1 and, consequently, were unable to determine gene order. Mapping the *polled* locus to BTA1 confirmed the hypothesis that the mode of inheritance for the *polled* locus was autosomal dominant. Brenneman et al. (1996) independently confirmed the localization of the *POLL* locus to BTA1 and refined the marker order for the proximal region of BTA1 utilizing a *Bos indicus* × *B. taurus* cross generating 209 reciprocal backcross and F2 progeny along with their 60 parents and grandparents. Known as the Angleton Project (Kim et al. 2003), these animals were utilized to generate a genetic map of BTA1 comprising 14 microsatellites that spanned 124.6 cM with an average interval size of 9.6 cM. The authors were able to determine that *POLL* was mapped to the proximal end of BTA 1 within 4.9 cM of TGLA49.

Over the next few years, additional markers were mapped to BTA1. Some mapping efforts included only a handful of markers (Schmutz et al. 1995; Band et al. 1997; Harlizius et al. 1997), while others integrated existing informative meiosis from multiple laboratories to generate a consensus framework map of BTA1 (Taylor et al. 1998). Among these efforts to fine map BTA1, Sonstegard et al. (1997a) constructed a BTA1-specific λ library that yielded 44 additional microsatellites that were linked to BTA1. The genetic map of BTA1 included 84 microsatellites spanning 153.8 cM, which improved the map resolution twofold and resulted in the highest resolution linkage map of any bovine chromosome at that time.

Additional strategies employed for marker development included a comparative genomics approach that exploited available bovine ESTs to construct a sequence-ready ~4-Mb single bacterial artificial chromosome (BAC) contig of the polled region of BTA1 (Drogemuller et al. 2005). In an effort to fine map the bovine *polled* locus, the newly constructed BAC contig was integrated with the existing linkage map, and random sequences of 13 BAC clones facilitated the development of 20 new microsatellite markers in the *polled* region of BTA1. Fine mapping the polled region guided the generation of 19 recombinant haplotypes that were used to narrow the *polled* locus to a 1-Mb segment. While 13 genes had been physically mapped in the *polled* critical region in cattle and an additional 18 genes were known in the orthologous human chromosomal region, no obvious functional candidate gene was identified. The following year a similar strategy was undertaken as an intermediate step toward identifying the causative mutation (Wunderlich et al. 2006). A 2.5-Mb BAC contig was constructed that spanned the *polled* locus. The BAC clones utilized by Wunderlich et al. (2006) were likewise utilized in the bovine genome sequencing initiative, directly tying this contig to the bovine genome sequence. Even with a 1-Mb critical region and 15 years of community effort, the causative mutation underlying the *POLL* locus has yet to be identified. Linkage mapping was successful in determining the chromosome harboring the *polled* locus and further fine mapping the critical region. Furthermore, the recent assembly of the bovine genome has catalyzed the development of additional resources that have invigorated the hunt for the causative mutation associated with the poll phenotype.

## *BTA6*

The first genome scan for milk production QTL was published by Georges et al. (1995). As previously discussed, a linkage map was constructed that spanned 1645 cM. QTL analysis was undertaken for five milk production traits and QTLs were identified on bovine chromosomes 1, 6, 9, 10, and 20. A milk yield (MY) QTL on BTA6 was among the QTLs identified, and additional QTLs for fat percentage (FP) and protein percentage (PP) were identified in the same region of BTA6. Interestingly, the reported increase in MY did not increase the fat yield (FY) or protein yield (PY), which resulted in drastically reduced FP and PP. Of the five chromosomes where QTL were identified, only BTA6 contained a known candidate gene, the casein locus. However, the QTL position and the effect observed did not necessarily support the casein locus as an ideal candidate gene. Dissection of the BTA6 QTL region originally identified by Georges et al. (1995) would progress for a decade, until two competing causative mutations were proposed.

Because of the large number of milk production QTLs segregating on BTA6, this chromosome became the focus of several fine mapping efforts (Kuhn et al. 1999; Wiener et al. 2000; Ron et al. 2001; Freyer et al. 2002). Among these, Ron et al. (2001) genotyped 12 microsatellite markers in nine Israeli Holstein sire families containing 2978 daughters. Two chromosomal regions were found to be associated with milk production traits, one near the center of BTA6 close to microsatellite marker BM143, and the other near marker BM415 at the telomeric end of BTA6 at approximately 80 cM. In this instance, even though only 12 microsatellite markers were used, the number of individuals genotyped was almost double that of Georges et al. (1995), enabling the QTL near the center of BTA6 to be localized to a 4-cM region around marker BM143. Building upon the previous QTL work performed on BTA6, Freyer et al. (2002) genotyped the sons of five German Holstein-Friesian sires with 16 microsatellites, four of which were common to the Ron et al. (2001) study. Five QTLs were reported, including two MY QTLs at approximately 47 cM and 91 cM, two PP QTL located at 44 and 67 cM, as well as a QTL affecting both FY and PP at 70 cM, thus reaffirming the presence of multiple milk production QTLs on BTA6.

The QTL near marker BM143 became the target of investigation by several laboratories due to its large effect and the fact that it had been validated in several populations. Olsen et al. (2004) targeted a 31-cM region of BTA6 harboring the PP and FP QTL with ten publicly available microsatellite markers, using 35 elite sire families containing 1098 sons and 680,000 daughters. However, a lack of recombination events between closely spaced markers prevented accurate ordering of this region.

Ultimately, the marker order was determined using the map reported by Weikard et al. (2002). Despite this, Olsen was able to localize the FP and PP QTL to a 7.5-cM interval between markers BMS2508 and FBN12. This 7.5-cM interval included the 4-cM confidence interval proposed by Ron et al. (2001). In an attempt to accurately order the markers on BTA6, Olsen et al. (2005) mapped 20 newly discovered SNPs located within ten genes, thus producing a linkage map of BTA6 that spanned approximately 90 cM with an average intermarker distance of 2.43 cM. Unfortunately, issues regarding the number of informative meiosis needed to accurately fine map these chromosomal regions persisted. Again, the authors utilized previously published maps (Weikard et al. 2002) for ordering markers on BTA6. However, in order to confirm the marker order on BTA6 and refine the human–cattle comparative map for HSA4/BTA6, the authors RH mapped ten genes using the Roslin–Cambridge 3000 rad RH panel (Williams et al. 2002). Subsequently, the human–cattle comparative map indicated the QTL region of interest on BTA6 corresponded with two blocks of conserved synteny on HSA4, one of which contained the genes *ABCG2*, *IBSP*, and *SPP1*. In an effort to further refine the critical region, the authors constructed a BAC clone-based physical map. Utilizing the genetic, RH, and BAC clone-based physical map, the authors were able to refine the QTL region to an estimated 420-kb region on BTA6 between markers *ABCG2* and *LAP3*.

Simultaneous to the studies discussed previously, Schnabel et al. (2005) had fine mapped regions of BTA6 with 38 microsatellite markers specifically targeting the region near marker BM143 using 3147 Holstein bulls from 45 half-sib families. Numerous QTL were identified further confirming the presence of multiple milk related QTL on BTA6. Schnabel et al. (2005) probed the conserved syntenic region on HSA4 corresponding to the 420-kb critical region for candidate genes and identified *OPN* (SPP1) among the four known human genes located in this region as a functional candidate gene. Subsequently, a 12.3-kb region of BTA6 harboring the *OPN* gene was sequenced and a candidate causal mutation (OPN3907) was identified. Cohen-Zinder et al. (2005) also fine mapped this QTL but chose the *ABCG2* gene, which is 150 kb downstream of *OPN*, as a candidate. They identified a Y581S polymorphism in *ABCG2* that was concordant with the QTL segregation status of their sires and concluded that this was the causal polymorphism rather than that identified by Schnabel et al. (2005). Subsequent analysis by Schnabel et al. (unpublished data) indicated that the *ABCG2* Y581S and OPN3907 mutations were in complete linkage disequilibrium in the US Holstein sires tested. Ultimately, the group that originally identified the *ABCG2* Y581S mutation was able to demonstrate its causality because the OPN3907 and *ABCG2* Y581S mutations were not in complete linkage disequilibrium in the Norwegian Red population (Olsen et al. 2007). The history of the milk production QTL near marker BM143 on BTA6 illustrates the progression of mapping resolution from chromosomal localization to finally distinguishing between two competing candidate causal mutations.

## Conclusion

The contributions of linkage mapping toward the advancement of bovine genomics is undisputed. The progression of linkage maps coincided with marker development and technological ability to genotype ever increasing numbers of markers. Enhancements in linkage mapping have facilitated progress toward identification and fine mapping of QTL while simultaneously contributing toward our knowledge of human/cattle synteny and chromosome evolution. Even though we have global positioning satellites today, sometimes an old-fashioned map is still the best tool for locating something of interest. Likewise, even though we have a draft bovine genome assembly, it seems fitting that we are relying on the "old" technology of linkage maps to resolve ambiguities in the reference assembly.

## References

Arias, J.A., Keehan, M., Fisher, P., Coppieters, W., Spelman, R. (2009) A high density linkage map of the bovine genome. *BMC Genetics* **10**: 18.

Band, M., Eggen, A., Bishop, M.D., Ron, M. (1997) Isolation of microsatellites from a bovine YAC clone harbouring the SOD1 gene. *Animal Genetics* **28**: 363–366.

Barendse, W., et al. (1994) A genetic linkage map of the bovine genome. *Nature Genetics* **6**: 227–235.

Barendse, W., et al. (1997) A medium-density genetic linkage map of the bovine genome. *Mammalian Genome* **8**: 21–28.

Bishop, M.D., et al. (1994) A genetic linkage map for cattle. *Genetics* **136**: 619–639.

Brenneman, R.A., Davis, S.K., Sanders, J.O., Burns, B.M., Wheeler, T.C., Turner, J.W., Taylor, J.F. (1996) The polled locus maps to BTA1 in a *Bos indicus* × *Bos taurus* cross. *Journal of Heredity* **87**: 156–161.

Casas, E., et al. (1999) Bovine chromosome 4 workshop: consensus and comprehensive linkage maps. *Animal Genetics* **30**: 375–377.

Cohen-Zinder, M., et al. (2005) Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research* **15**: 936–944.

The Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522–528.

Dietz, A.B., Neibergs, H.L., Womack, J.E. (1992) Assignment of eight loci to bovine syntenic groups by use of PCR: extension of a comparative gene map. *Mammalian Genome* **3**: 106–111.

Drogemuller, C., Wohlke, A., Momke, S., Distl, O. (2005) Fine mapping of the polled locus to a 1-Mb region on bovine chromosome 1q12. *Mammalian Genome* **16**: 613–620.

Everts-van der Wind, A., et al. (2004) A 1463 gene cattle-human comparative map with anchor points defined by human genome sequence coordinates. *Genome Research* **14**: 1424–1437.

Freyer, G., Kuhn, C., Weikard, R., Zhang, Q., Mayer, M., Hoeschele, I. (2002) Multiple QTL on chromosome six in dairy cattle affecting yield and content traits. *Journal of Animal Breeding and Genetics* **119**: 69–82.

Georges, M., et al. (1991) Characterization of a set of variable number of tandem repeat markers conserved in bovidae. *Genomics* **11**: 24–32.

Georges, M., et al. (1993) Microsatellite mapping of a gene affecting horn development in *Bos taurus*. *Nature Genetics* **4**: 206–210.

Georges, M., et al. (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**: 907–920.

Green, P., Falls, K., Crooks, S. (1990) Documentation for CRI-MAP version 2.4. Washington University School of Medicine, St. Louis.

Gu, Z., et al. (2000) Consensus and comprehensive linkage maps of bovine chromosome 7. *Animal Genetics* **31**: 206–209.

Harlizius, B., Tammen, I., Eichler, K., Eggen, A., Hetzel, D.J. (1997) New markers on bovine chromosome 1 are closely linked to the polled gene in Simmental and Pinzgauer cattle. *Mammalian Genome* **8**: 255–257.

Ihara, N., et al. (2004) A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Research* **14**: 1987–1998.

Itoh, T., Watanabe, T., Ihara, N., Mariani, P., Beattie, C.W., Sugimoto, Y., Takasuga, A. (2005) A comprehensive radiation hybrid map of the bovine genome comprising 5593 loci. *Genomics* **85**: 413–424.

Jann, O.C., et al. (2006) A second generation radiation hybrid map to aid the assembly of the bovine genome sequence. *BMC Genomics* **7**: 283.

Kappes, S.M., Keele, J.W., Stone, R.T., McGraw, R.A., Sonstegard, T.S., Smith, T.P., Lopez-Corrales, N.L., Beattie, C.W. (1997) A second-generation linkage map of the bovine genome. *Genome Research* **7**: 235–249.

Kim, J.J., Farnir, F., Savell, J., Taylor, J.F. (2003) Detection of quantitative trait loci for growth and beef carcass fatness traits in a cross between *Bos taurus* (Angus) and *Bos indicus* (Brahman) cattle. *Journal of Animal Science* **81**: 1933–1942.

Kong, A., et al. (2002) A high-resolution recombination map of the human genome. *Nature Genetics* **31**: 241–247.

Kuhn, C., Freyer, G., Weikard, R., Goldammer, T., Schwerin, M. (1999) Detection of QTL for milk production traits in cattle by application of a specifically developed marker map of BTA6. *Animal Genetics* **30**: 333–340.

Kurar, E., et al. (2002) Consensus and comprehensive linkage maps of bovine chromosome 24. *Animal Genetics* **33**: 460–463.

Larsen, N.J., Hayes, H., Bishop, M., Davis, S.K., Taylor, J.F., Kirkpatrick, B.W. (1999) A comparative linkage and physical map of bovine chromosome 24 with human chromosome 18. *Mammalian Genome* **10**: 482–487.

Ma, R.Z., et al. (1996) A male linkage map of the cattle (*Bos taurus*) genome. *Journal of Heredity* **87**: 261–271.

Matise, T.C., Perlin, M., Chakravarti, A. (1994) Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nature Genetics* **6**: 384–390.

Matukumalli, L.K., et al. (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* **4**: e5350.

McKay, S.D., et al. (2007) Construction of bovine whole-genome radiation hybrid and linkage maps using high-throughput genotyping. *Animal Genetics* **38**: 120–125.

Olsen, H.G., Lien, S., Svendsen, M., Nilsen, H., Roseth, A., Aasland Opsal, M., Meuwissen, T.H. (2004) Fine mapping of milk production QTL on BTA6 by combined linkage and linkage disequilibrium analysis. *Journal of Dairy Science* **87**: 690–698.

Olsen, H.G., Lien, S., Gautier, M., Nilsen, H., Roseth, A., Berg, P.R., Sundsaasen, K.K., Svendsen, M., Meuwissen, T.H. (2005) Mapping of a milk production quantitative trait locus to a 420-kb region on bovine chromosome 6. *Genetics* **169**: 275–283.

Olsen, H.G., Nilsen, H., Hayes, B., Berg, P.R., Svendsen, M., Lien, S., Meuwissen, T. (2007) Genetic support for a quantitative trait nucleotide in the ABCG2 gene affecting milk composition of dairy cattle. *BMC Genetics* **8**: 32.

Ponce de Leon, F.A., Ambady, S., Hawkins, G.A., Kappes, S.M., Bishop, M.D., Robl, J.M., Beattie, C.W. (1996) Development of a bovine X chromosome linkage group and painting probes to assess cattle, sheep, and goat X chromosome segment homologies. *Proceedings of National Academy of Sciences of the United States of America* **93**: 3450–3454.

Ron, M., Kliger, D., Feldmesser, E., Seroussi, E., Ezra, E., Weller, J.I. (2001) Multiple quantitative trait locus analysis of bovine chromosome 6 in the Israeli Holstein population by a daughter design. *Genetics* **159**: 727–735.

Schmutz, S.M., Marquess, F.L., Berryere, T.G., Moker, J.S. (1995) DNA marker-assisted selection of the polled condition in Charolais cattle. *Mammalian Genome* **6**: 710–713.

Schnabel, R.D., Kim, J.J., Ashwell, M.S., Sonstegard, T.S., Van Tassell, C.P., Connor, E.E, Taylor, J.F. (2005) Fine-mapping milk production quantitative trait loci on BTA6: analysis of the bovine osteopontin gene. *Proceedings of National Academy of Sciences of the United States of America* **102**: 6896–6901.

Schuler, G.D., et al. (1996) A gene map of the human genome. *Science* **274**: 540-546.

Smith, T.P., Casas, E., Rexroad, C.E., 3rd, Kappes, S.M., Keele, J.W. (2000) Bovine CAPN1 maps to a region of BTA29 containing a quantitative trait locus for meat tenderness. *Journal of Animal Science* **78**: 2589-2594.

Snelling, W.M., Casas, E., Stone, R.T., Keele, J.W., Harhay, G.P., Bennett, G.L., Smith, T.P. (2005) Linkage mapping bovine EST-based SNP. *BMC Genomics* **6**: 74.

Snelling, W.M., et al. (2004) Integrating linkage and radiation hybrid mapping data for bovine chromosome 15. *BMC Genomics* **5**: 77.

Snelling, W.M., et al. (2007) A physical map of the bovine genome. *Genome Biology* **8**: R165.

Sonstegard, T.S., Abel Ponce de Leon, F., Beattie, C.W., Kappes, S.M. (1997a) A chromosome-specific microdissected library increases marker density on bovine chromosome 1. *Genome Research* **7**: 76–80.

Sonstegard, T.S., Lopez-Corrales, N.L., Kappes, S.M., Stone, R.T., Ambady, S., Ponce de Leon, F.A., Beattie, C.W. (1997b) An integrated genetic and physical map of the bovine X chromosome. *Mammalian Genome* **8**: 16–20.

Sonstegard, T.S., et al. (2001) Consensus and comprehensive linkage maps of the bovine sex chromosomes. *Animal Genetics* **32**: 115–117.

Sun, H.S., et al. (1997) Comparative linkage mapping of human chromosome 13 and bovine chromosome 12. *Genomics* **39**: 47–54.

Taylor, J.F., Lutaaya, E., Sanders, J.O., Turner, J.W., Davis, S.K. (1997) A medium density microsatellite map of BTA10: reassignment of INRA69. *Animal Genetics* **28**: 360–362.

Taylor, J.F., et al. (1998) Report of the first workshop on the genetic map of bovine chromosome 1. *Animal Genetics* **29**: 228–235.

Weikard, R., Kuhn, C., Goldammer, T., Laurent, P., Womack, J.E., Schwerin, M. (2002) Targeted construction of a high-resolution, integrated, comprehensive, and comparative map for a region specific to bovine chromosome 6 based on radiation hybrid mapping. *Genomics* **79**: 768–776.

Wiener, P., Maclean, I., Williams, J.L., Woolliams, J.A. (2000) Testing for the presence of previously identified QTL for milk production traits in new populations. *Animal Genetics* **31**: 385–395.

Williams, J.L., et al. (2002) A bovine whole-genome radiation hybrid panel and outline map. *Mammalian Genome* **13**: 469–474.

Wunderlich, K.R., et al. (2006) A 2.5-Mb contig constructed from Angus, Longhorn and horned Hereford DNA spanning the polled interval on bovine chromosome 1. *Animal Genetics* **37**: 592–594.

Yeh, C.C., Taylor, J.F., Gallagher, D.S., Sanders, J.O., Turner, J.W., Davis, S.K. (1996) Genetic and physical mapping of the bovine X chromosome. *Genomics* **32**: 245–252.

Zimin, A.V., et al. (2009) A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biology* **10**: R42.

# Chapter 7
# Bovine X and Y Chromosomes

*F. Abel Ponce de León and Wansheng Liu*

Sex chromosomes evolved from a pair of autosomes (Muller 1914) and are believed to be the result of genetic sex determination that originated when a sex-determining gene was acquired by one member of the pair to become the sex specific determining chromosome. This gave origin to the male heterogamety, XX female:XY male, and female heterogamety, ZW female:ZZ male, systems. The former is observed in mammals, some species of turtles, insects, lizards, and even some plants and the latter in birds, amphibians, snakes, and some species of fish, turtles, insects, and lizards (Modi and Crews 2005). Sex chromosomes show a relative gradient of morphological and size differentiation moving from undifferentiated sex chromosomes in fishes and amphibians (Ohno 1967) to those of mammals and birds. The X and Z chromosomes are large and gene rich, while, in comparison, the Y and W chromosomes are significantly smaller, gene poor, and contain large heterochromatic blocks. However, the gene content of the XY and ZW systems is different (Nanda et al. 1999).

In the broad sense, the X and Y chromosomes have two regions: (1) the pseudoautosomal region (PAR), which is the recombining region, and (2) the X-specific and Y-specific regions that do not pair and therefore do not recombine during meiosis.

Our current understanding of X chromosome evolution in mammals is that it is formed by four evolutionary strata and the PAR (Graves 2006). The first evolutionary layer known as the X conserved region (XCR) was identified by the comparison of human X orthologous genes across mammals. This XCR, a conserved block of euchromatin, represents the original autosome pair from which sex chromosomes evolved (Glas et al. 1999) about 166 million years ago (MYA) (Veyrunes et al. 2008). The second X chromosome evolutionary strata is defined by genes that are orthologous to autosomal genes in marsupials and monotremes; therefore, this region was only added about 90–50 MYA and is known as the X added region (XAR). However, comparisons of chicken homologs to the human X chromosome subdivided XCR into two strata and the XAR (Nanda et al. 1999; Kohn et al. 2004). A further refinement of our understanding of these evolutionary strata was achieved by comparing gene sequences between the human Y and X chromosomes. The oldest group of genes (more divergent) corresponds to the XCR stratum I, and the second oldest to XCR stratum II. Similarly, the XAR contains two clusters of genes (evolutionary strata III and IV) differentiated on the basis of their homology/divergence with copies found on the Y chromosome and the PAR (Lahn and Page 1999b).

*Bovine Genomics*, First Edition. Edited by James E. Womack.

Comparisons of human to marsupial Y chromosome genes and human Y chromosome to X chromosome genes have provided information to delineate our current understanding of the evolutionary regions of the Y chromosome. The Y chromosome essentially contains a small Y conserved region (YCR), a large Y added region (YAR), and genes transposed from other autosomes.

Based on comparative genomic studies in insects and vertebrates, it has been postulated that the Y chromosome has accumulated male advantage genes, suppressed recombination by accumulating mutations and deletions in the nonrecombining region, and therefore is degrading (Aitken and Graves 2002). This concept can also be extended to the W chromosome. Wilson and Makova (2009) provide a thorough review on the evolution of XY and ZW systems on the basis of genomic analysis.

## Cytogenetic Analysis of Bovine Sex Chromosomes

The bovine chromosome complement includes 29 pairs of autosomes, all acrocentrics, and the submetacentric X and Y chromosomes, which can be readily distinguished from the autosomes in metaphase preparations. The X chromosome is a large chromosome and the Y chromosome is one of the smallest. Cytogenetic banding techniques (Evans et al. 1973) have been used to identify each of the autosomes and sex chromosomes. Banding techniques Giemsa (G-banding) and Reverse to Giemsa (R- banding) have also been used for band pattern comparisons among cattle, sheep, and goats, and results have supported the hypothesis of a common origin of all bovids as proposed by Wurster and Benirschke (1968). This was later corroborated by more detailed analysis of band homologies in many other bovid species (Buckland and Evans 1978; Bunch and Nadler 1980; Di Berardino et al. 1981; Mensher et al. 1989; Iannuzzi et al. 1990; Hayes et al. 1991; Gallagher and Womack 1992). However, the X chromosome among bovids varies in morphology from submetacentric, as in cattle, to acrocentric in sheep, goat, and suni, and in size due to the acquisition of heterochromatic blocks, as in kudu (Robinson et al. 1998).

The advent of fluorescent in situ hybridization technology (FISH) coupled to the development of chromosome-specific painting probes allowed the identification of interspecies chromosome homologies. Homologies between the bovine X chromosome and goat and sheep X chromosomes were demonstrated by FISH analysis using short arm ($BTAX_p$) and long arm ($BTAX_q$) painting probes. The $BTAX_p$ probe showed homology with the goat and sheep Xq34–q41 region and demonstrated that $BTAX_p$ moved as a conserved euchromatic block among these species (Ponce de León et al. 1996). Likewise, Robinson et al. (1998) used a combination of the bovine X chromosome arm-specific painting probes and one bovine bacterial artificial chromosome (BAC) probe to assess X-chromosome repatterning and euchromatic block orientation among 22 bovid species representing 22 tribes and eight out of the nine bovid subfamilies. These researchers found that $BTAX_p$ had been moved as a euchromatic block during bovid X-chromosome evolution. Further FISH analysis of BAC 101 located at the proximal region of $BTAX_p$ permitted the orientation of this euchromatic block and allowed these authors to describe three bovid X-chromosome types. One type is represented by the cattle (Subfamily Bovinae, Tribe Bovini) submetacentric chromosome. A second type is represented by the eland (Subfamily Bovinae, Tribe Tragelaphini) acrocentric X chromosome that shows proximal region homology and same orientation of the euchromatic block as the cattle $X_{p.}$ The third type is found in all other subfamilies including the Caprinae. In this latter type, the $BTAX_p$ euchromatic block is inverted and is distally and interstitially located, or better described, is flanked by $BTAX_q$ blocks of DNA. Assuming the "suni acrocentric type" as the ancestral chromosome type (Hayes et al. 1991; Robinson et al. 1998), it is possible to infer that the generation of the "eland acrocentric type" arose by a paracentric inversion of the $BTAX_p$ block that relocated this block close to the centromere in the eland. On the other hand, the generation of the cattle submetacentric X chromosome is more complex and has been proposed to have been originated from the "eland acrocentric type" by a single transposition (Robinson et al. 1998). According to these authors, this rearrangement requires three breakpoints and a shift of a chromosomal segment to another region of the same chromosome while maintaining the same orientation of the homologous $BTAX_p$ translocated segment as in the eland acrocentric chromosome.

Another characteristic of the bovine sex chromosomes is that they do not have prominent centromeres as their autosomes and as a consequence show negative centromeric C-band staining.

Cytogenetic analysis of the bovine Y chromosome is, in comparison to the X chromosome, limited. Y-chromosome morphology and size differ among bovis; it is submetacentric in cattle (BTAY), sheep (OARY), and goats (CHIY) and, acrocentric in zebu (BINY) and in river buffalo (BBUY), to mention a few. Some laboratories refer to the sheep and goat Y chromosomes as being metacentric as well (Di Meo et al. 2005). This might be due to Y-chromosome size polymorphisms among breeds and/or deletions in some male lineages.

Because of their small size and the fact that they are largely heterochromatic, Y-chromosome banding techniques do not offer enough resolution for chromosome rearrangement and evolutionary studies among bovids. This coupled to the significant paucity in the identification of Y-chromosome molecular markers led our laboratory to develop a bovine Y chromosome-specific DNA library and chromosome painting probe that allowed the localization of the PAR at Xq42–43 (Figure 7.1). To confirm this finding, an Xqter (Figure 7.1)-specific painting probe was also developed to allow the identification of the PAR at Yp13 (Ponce de León and Carpio 1995).

The availability of Y chromosome-specific molecular probes permitted Di Meo et al. (2005) to synergistically use chromosome banding and gene/marker localizations by FISH to infer Y-chromosome similarities and possible evolutionary patterns within and between Bovinae (BTAY, BINY, BBUY), and Caprinae (OARY, CHIY). Their work describes the existence of a C-band located distally and in close proximity to the PAR in all species. This C-band appears to have the same location as the always observable positive R-band. Based on the alignment of the prominent positive R-band among Y chromosomes of these five species and FISH localization of eight gene/markers (*DXYS3*, *SLC25A6l SRY*, *ZFY*, *DYZ10* described in Bovmap; and UMN0504, UMN0301, and UMN0304 described in Liu et al. 2002), Di Meo et al. (2005) hypothesized that Y chromosomal rearrangements between these species are the result of a pericentric inversion or a centromeric transposition between BTAY and BINY, pericentric inversion between BTAY and BBUY, pericentric inversion with a major loss of heterochromatin between BBUY and OARY/CHIY, and a centromere transposition with loss of heterochromatin between BTAY and OARY/CHIY. Marker order comparisons between BTAY and BBUY radiation hybrid ($RH_{5000}$) maps have

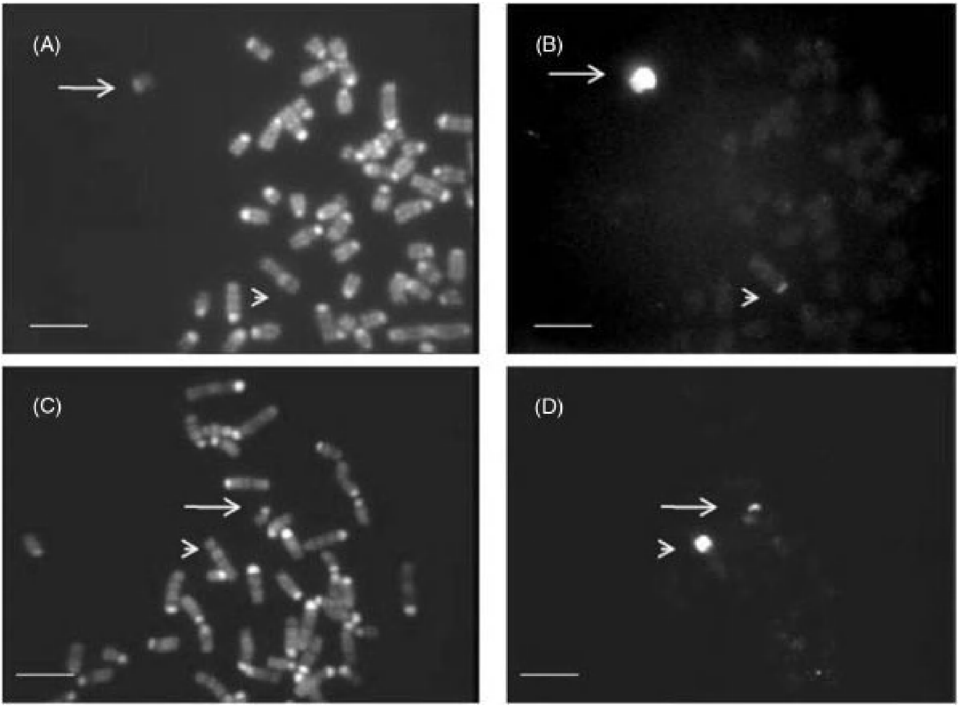


**Figure 7.1** Localization of the pseudoautosomal region (PAR) with BTAY and BTAXq42-43 chromosome-painting probes. (A and B) Same partial bovine male metaphase showing chromosome R-Banding patterns (A) and FITC hybridization signals (B) obtained with the whole BTAY chromosome painting probe. (C and D) Same partial bovine male metaphase showing chromosome R-Banding patterns (C) and FITC hybridization signals (D) obtained with the BTAXq42-43 painting probe. BTAY is identified by arrows and BTAX is identified by arrowheads. The pseudoautosomal region is clearly delineated on BTAX (B, arrowhead) when the BTAY chromosome painting probe is used. Similarly, the pseudoautosomal region is clearly delineated on BTAY (D, arrow) when the BTAXq42-43 chromosome painting probe is used. (Bars = 10 $\mu$m.)

confirmed Di Meo et al.'s (2005) hypothesis that proposed the morphological difference between BTAY and BBUY to be the result of a pericentric inversion with addition or loss of heterochromatin (Stafuzza et al. 2009). Similar future studies in other bovid species are necessary to illustrate the evolutionary changes of the Y chromosome in Bovidae.

## X and Y Chromosome Genetic and Physical Maps

The third-generation comprehensive genetic map (Ihara et al. 2004) was constructed on the basis of >880,000 genotypes across cattle reference families (United States Department of Agriculture, Meat Animal Research Center (USDA-MARC)) incorporating 2325 microsatellites into the second-generation genetic map developed by Kappes et al. (1997). The third-generation genetic map spans 29 sex-averaged autosomal linkage groups and the X-specific linkage group. This map has a total length of 3160 cM and includes 3960 marker loci localized in 2389 sites. The X linkage group has a length of 146.5 cM and includes 189 markers at 83 positions with an average interval of 1.8 cM and a maximum interval of 10.2 cM. The PAR boundary was localized within 1 cM interval between the XBM31 and IOBT1489-DXS23 markers in the second-generation map (Sonstegard et al. 2001) and between TGLA325 and BM861 in the third generation map (Ihara et al. 2004).

RH maps have also been developed for BTAX and BTAY. The first whole genome bovine RH (WG-RH) panel was developed by Womack et al. (1997). This RH panel was derived from a culture of normal diploid fibroblast obtained from an Angus bull. Fibroblasts were subject to a total radiation dose of 5000 rads, chemically fused to the recipient thymidine kinase negative A23 Chinese hamster cell line and selected in hypoxanthine–aminopterin–thymidine (HAT) medium in the presence of Ouabain. One hundred and one cloned cells lines were derived and constitute the $RH_{5000}$ cell panel. At present, three different hamster-cattle WG-RH panels have been constructed at 5000, 7000, and 12,000 REF rads, respectively (Womack et al. 1997; Liu et al. 2002). The third-generation RH bovine map (Everts-van der Wind et al. 2005) comprises the localization of 3484 markers of which 163 have been ordered along BTAX and of these 144 were found to have orthologs in the human X chromosome (HSAX). BTAX was found to have complete homology to HSAX sharing seven homologous synteny blocks. There are now over 5307 genetic markers mapped on the bovine genome. Of these, 1507 markers are type I (INRA, bovine genome databases, http://locus.jouy.inra.fr/cgi-bin/bovmap/intro.pl), and 3800 are type II (Ihara et al. 2004). However, none of these maps include the male-specific Y (MSY) region and are only limited to the RH map of the PAR.

As indicated before, the BTAY-specific region does not undergo recombination during meiosis. This MSY region represents about 95% of the chromosome length and essentially comprises repetitive sequences making physical mapping and chromosome sequencing difficult. This nonrecombining region also makes genetic mapping impossible. Because of these limitations, a first-generation BTAY $RH_{7000}$ map (Figure 7.2) was generated (Liu et al. 2002). Thirteen markers were localized in the PAR region and 46 markers in the MSY. The *AMELY* gene was localized in the MSY close to the pseudoautosomal boundary (PAB) region and both the *SRY* and *TSPY* genes in the MSY region. Although the level of resolution of this cell hybrid panel did not allow precise localization of the SRY gene, the latter was mapped to the distal region of BTAYq by FISH (Liu and Ponce de León 2004). Retention frequencies of Y-chromosome markers ranged from 18.5% to 76.5%. Retention frequencies higher than 55% were indicative of multiple marker copies making the map order of these markers difficult to achieve. The multiple copy *TSPY* gene was among the genetic markers that have a retention frequency higher than 55%.

The bovine genome sequencing project (http://www.hgsc.bcm.tmc.edu/) has now generated a 7× sequence genome (Btau_4.0). Data is accessible at the Bovine Genome Database (http://genomes.arc.georgetown.edu/drupal/bovine/) and at the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/).

The Btau_4.0 statistics indicates that BTAX has a length of 89 Mbp and 150.5 cM made up by 107 contigs containing 1168 expressed sequence tag (EST) transcripts and 793 genes.

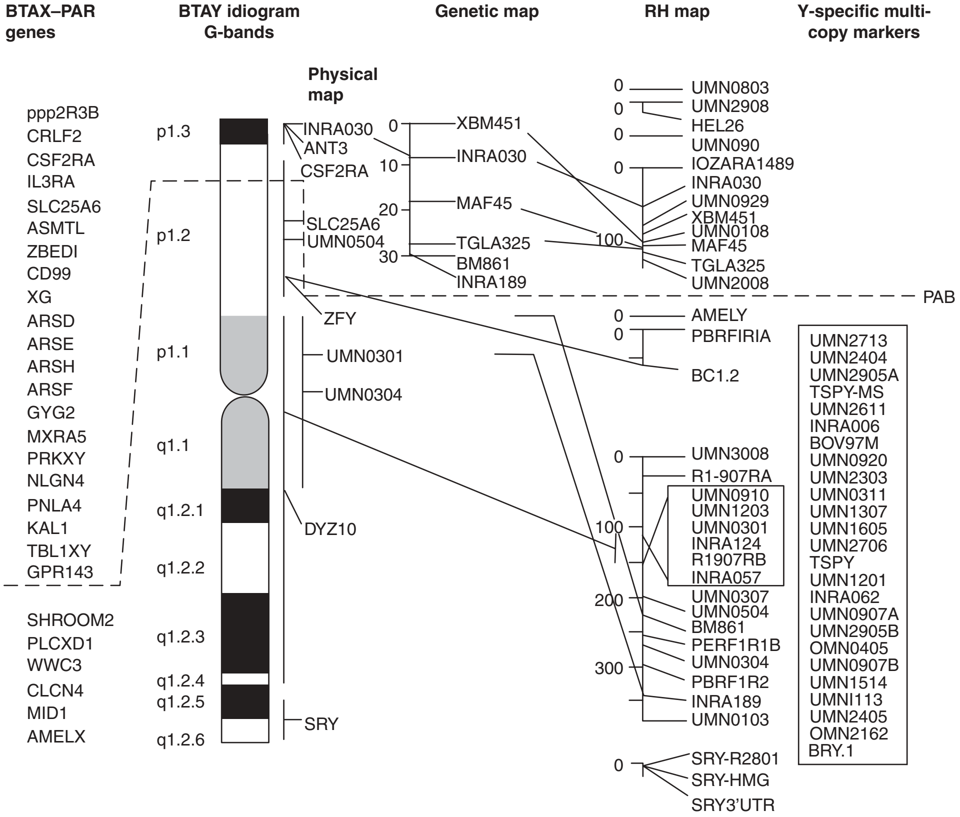
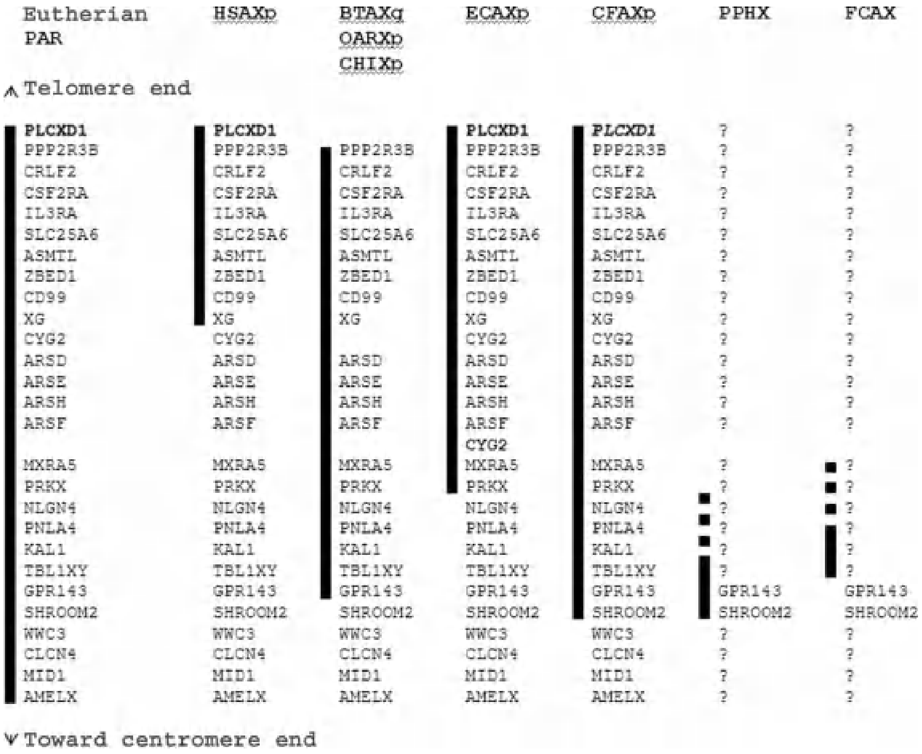


**Figure 7.2** BTAY genetic, physical and RH maps (Modified from Liu and Ponce de León, 2007). From left to right: A list of identified BTAX pseudoautosomal region (PAR) genes (Das et al., 2009); a physical map of genes assigned by fluorescent in situ hybridization (FISH); an idiogram of the G-banded BTAY; a genetic linkage map for the PAR (Kappes et al., 1997); and the RH$_{7000}$ map of BTAY (Liu et al., 2002). The PAB dotted line indicates the pseudoautosomal boundary. Markers in the box of the RH map are centromeric, and in the far right box are Y specific multicopy markers, which could not be mapped on the RH map.

## The Pseudoautosomal Region

The PAR is the region with highest homology between the X and Y chromosomes. Only one PAR region has been observed in Bovinae by FISH analysis (Ponce de León and Carpio 1995; Robinson et al. 1998) and synaptonemal complex analysis (Switonski and Stranzinger 1998). The size of the bovine PAR has been estimated to span 5.9 Mb and like in human and mouse its GC and CpG island content decreases from Xq ter toward the PAB in BTAX (Das et al. 2009).

Van Laere et al. (2008) have compared PAR, X-specific, and autosomal sequences of the bovine genome (Btau_4.0 build) to the available Y-chromosome sequences and comparable sequences obtained from the human genome NCBI 36 build. Their results indicate that there is a good correlation between GC and CpG island content with recombinational activity being higher in the PAR, next in the autosomes, followed by X-specific, and lower on the Y-specific sequences, in that order. There is also



| Eutherian PAR | HSAXp | BTAXq<br>OARXp<br>CHIXp | ECAXp | CFAXp | PPHX | FCAX |
|---|---|---|---|---|---|---|
| ^ Telomere end | | | | | | |
| PLCXD1 | PLCXD1 | | PLCXD1 | PLCXD1 | ? | ? |
| PPP2R3B | PPP2R3B | PPP2R3B | PPP2R3B | PPP2R3B | ? | ? |
| CRLF2 | CRLF2 | CRLF2 | CRLF2 | CRLF2 | ? | ? |
| CSF2RA | CSF2RA | CSF2RA | CSF2RA | CSF2RA | ? | ? |
| IL3RA | IL3RA | IL3RA | IL3RA | IL3RA | ? | ? |
| SLC25A6 | SLC25A6 | SLC25A6 | SLC25A6 | SLC25A6 | ? | ? |
| ASMTL | ASMTL | ASMTL | ASMTL | ASMTL | ? | ? |
| ZBED1 | ZBED1 | ZBED1 | ZBED1 | ZBED1 | ? | ? |
| CD99 | CD99 | CD99 | CD99 | CD99 | ? | ? |
| XG | XG | XG | XG | XG | ? | ? |
| CYG2 | CYG2 | | CYG2 | CYG2 | ? | ? |
| ARSD | ARSD | ARSD | ARSD | ARSD | ? | ? |
| ARSE | ARSE | ARSE | ARSE | ARSE | ? | ? |
| ARSH | ARSH | ARSH | ARSH | ARSH | ? | ? |
| ARSF | ARSF | ARSF | ARSF | ARSF | ? | ? |
| | | | CYG2 | | | |
| MXRA5 | MXRA5 | MXRA5 | MXRA5 | MXRA5 | ? | ? |
| PRKX | PRKX | PRKX | PRKX | PRKX | ? | ? |
| NLGN4 | NLGN4 | NLGN4 | NLGN4 | NLGN4 | ? | ? |
| PNLA4 | PNLA4 | PNLA4 | PNLA4 | PNLA4 | ? | ? |
| KAL1 | KAL1 | KAL1 | KAL1 | KAL1 | ? | ? |
| TBL1XY | TBL1XY | TBL1XY | TBL1XY | TBL1XY | ? | ? |
| GPR143 | GPR143 | GPR143 | GPR143 | GPR143 | GPR143 | GPR143 |
| SHROOM2 | SHROOM2 | SHROOM2 | SHROOM2 | SHROOM2 | SHROOM2 | SHROOM2 |
| WWC3 | WWC3 | WWC3 | WWC3 | WWC3 | ? | ? |
| CLCN4 | CLCN4 | CLCN4 | CLCN4 | CLCN4 | ? | ? |
| MID1 | MID1 | MID1 | MID1 | MID1 | ? | ? |
| AMELX | AMELX | AMELX | AMELX | AMELX | ? | ? |
| v Toward centromere end | | | | | | |

**Figure 7.3** Comparative X chromosome pseudoautosomal regions (PARs) and pseudoautosomal boundaries (PABs) of ruminant and non ruminant species (Modified from Das et al., 2009). Alignment of genes found in the PAR in sequential order from the telomeric end towards to centromere for the ancestral eutherian sex X, human (HSAXp), bovine, ovine, caprine (BTAXq, OARp, CHIp), horse (ECAXp), dog (CFAXp), porpoise (PPHX) and cat (FCAX). Black bars at left of each column represent the set of genes found in the PAR for each of the represented species. The centromeric end of the black bars represents the approximate location of the PAB. The PAR linkage order of genes for PPHX and FCAX is not yet determined. The PAB location for PPHX and FCAX is based on the work of Van Laere et al., (2008).

higher recombination rate in the human PAR than in the bovine. Another important observation is that the higher GC and CpG island content rate observed for the PAR is not significantly higher than in the autosomal regions close to the telomeres. As in humans, the density of CpG islands in bovine is highest in autosomes, followed by the X chromosome, Y chromosome, and PAR. It was also found that the bovine PAR is enriched with repeat sequences. The PAR density of short interspersed nuclear element and/or short interspersed repeats (SINEs) is more than twice the density of the X-specific region and that of the autosomal average. Also, the bovine PAR has a higher long interspersed nuclear element and/or long interspersed repeat (LINE) density than the rest of the X-specific region.

PAR gene linkage conservation among bovine, ovine, caprine, human, horse, and dog has been described (Figure 7.3; Das et al. 2009). However, the *CYG2* gene is located between *ARSF* and *MXRA5* in ruminant species and between *XG* and *ARSD* in nonruminant species. Also, the gene *PLCXD1* is X-chromosome specific in ruminants and therefore not found in the ruminant PAR. Since the human and equine PAR linkage groups include the *PLCXD1* gene and both are considered ancestral to the ruminant, the BTAX-specific localization of *PLCXD1* is a de novo event in the evolution of the ruminant sex chromosomes (Das et al. 2009).

## Pseudoautosomal Boundary

Van Laere et al. (2008) identified the bovine PAB to be located between the *SHROOM2* and *GPR143* genes on the X chromosome. Sequence homology between X and Y PAR sites was found to be near perfect at 99.97%. Fine alignment of sequences and comparison between X- and Y-specific regions adjacent to the PAR allowed the identification of a 413 bp fragment with reduced homology at 86.20% separating the PAR from the nonhomologous gonosome-specific regions. Also, at the boundary site between the gonosome-specific sequences and the 413 bp reduced homology segment there are sequences that represent the tRNA portion of the Bov-tA1 SINE element on the X chromosome and a Bov-tA2 SINE element on the Y chromosome. This finding is indicative that the PAB was created by intrachromatid recombination between these SINE elements that are ruminant specific (Shimamura et al. 1999). Therefore, Van Laere et al. (2008) concluded that the bovine PAB occurred after ruminants diverged from other mammals and further proved that the bovine PAB is ruminant specific by comparing PAB sequences of other ruminants (bison, yak, banteng, zebu, and sheep) as well as by comparing the female to male gene copy ratio of the *SHROOM2* and *GPR143* genes in ruminants (cattle), nonruminants (horse, cats, dogs, mice, and humans) and cetacean (porpoises) assumed to be a close relative of ruminants that diverged more than 50 MYA. This latter analysis confirmed the already known X-specific location of these genes in human and mouse and the X-specific location of *SHROOM2* and PAR location of *GPR143* in cattle. Further, it allowed these researchers to conclude that both genes were located on the X-specific region of the horse and therefore implying that the PAB in the horse as well as in human and mouse is located more distally than in cattle. They also found that both genes were located in the PAR region of porpoise and dog and therefore it implied that the PAB in these two species was located more proximal. In cats, however, results indicated that both genes are also located on the Y chromosome and closely related to the X-specific gametolog sequences indicative of a possible recent transposition. However, the precise location of the PAB in cats cannot be defined with current information.

## The MSY Region

The male-specific region, comprising 95% of the DNA content of the Y chromosome, can be divided into two regions, (1) euchromatic and (2) heterochromatic. According to the human Y-chromosome sequence, the euchromatic region contains at least four different types of sequences: (1) X-transposed (99% similarity to the Xq21), (2) X-degenerate (60%–96% to the X), (3) ampliconic, and (4) centromere repetitive sequences (Skaletsky et al. 2003). This euchromatic region also harbors all genes of the MSY, whereas the heterochromatic region contains Y-specific repetitive sequences. The absence of recombination at meiosis, the abundance of Y-specific repetitive sequences, the tendency of its genes to degenerate during evolution, and the functional coherence of its gene content in male growth, spermatogenesis, and fertility (Lahn and Page 1997) are some of the characteristics that make the MSY unique among all other nuclear chromosomes. The absence of recombination makes genetic mapping of the MSY virtually impossible, and the depth, breadth, and complexity of the repetitive sequences make sequencing extremely difficult. Therefore, mapping and sequencing strategies applied successfully elsewhere in the genome have faltered in the MSY, making the mammalian Y chromosome a difficult target for linkage mapping and sequencing (Tilford et al. 2001; Liu and Ponce de León 2007). These difficulties led all mammalian genome sequencing projects including the Bovine Genome Project chose to sequence DNA from females (Lander et al. 2001; Waterston et al. 2002; Krzywinski et al. 2004). To date, only the human, chimpanzee, and mouse Y chromosomes have been sequenced (Skaletsky et al. 2003; Kuroki et al. 2006; Aflöldi 2008), and the bovine Y chromosome is currently being sequenced by a joint effort between Baylor College of Medicine and the Massachusetts Institute of Technology (Ding et al. 2009).

Unlike the X-specific region that is highly conserved among mammalian species, the Y-specific region (MSY) is poorly conserved. The variation observed in MSY sequences and gene content is believed to be generated through two different mechanisms. One mechanism is the differential retention of genes from the proto-X/Y chromosomes during the process of Y-chromosome degeneration in different lineages (Graves 2006). For example, the *NLGN4Y* gene is present on the human and horse MSY region (Skaletsky et al. 2003; Raudsepp et al. 2004), but not on the mouse, cat, pig, and cattle MSY. Degeneration of the Y chromosome was driven by several synergistic evolutionary forces including recombination suppression, Muller's ratchet, background selection, the Hill Robertson effect with weak selection, and hitchhiking of deleterious alleles by favorable mutations (Charlesworth and Charlesworth 2000; Roze and Borton 2006). Independent Y-chromosome decay during evolution (Graves 2006; Pearks Wilkerson et al. 2008) led to different lineages retaining different subsets of Y genes and a diverse and lineage-specific Y-chromosome gene content.

The second mechanism that led to diverse Y gene content is the autosome-to-Y transposition. It is believed that the autosome-to-Y transposition of male fertility genes is a recurrent theme in mammalian Y-chromosome evolution (Hurst 1994; Saxena et al. 1996; Graves 2000). As a result, the content of male-beneficial genes in MSY has increased in spite of a 95% loss of the ancestral Y-chromosome genes due to absence of recombination. Autosome-to-Y transposition events apparently occurred separately in different lineages with newly acquired Y-chromosome genes from diverse genomic locations (Murphy et al. 2006). This resulted in lineage-specific Y-chromosome genes (families) that account for a significant portion of the gene (and sequence) variation among mammalian Y chromosomes. The human *DAZ* gene family was derived from the transposition of the autosomal *DAZL* that maps to the subtelomeric region on HSA3p24.3 (Saxena et al. 1996), while *CDY* arrived on the human Y chromosome through retrotransposition of *CDYL* on HSA6 (Lahn and Page 1999b; Skaletsky et al. 2003) during primate evolution. The mouse *Ssty1* was derived from a retroposition of an autosomal gene *Spin1* on chromosome 13 (Church et al. 2009). The feline *TETY1* and *FLJ36031* gene families originated through autosome-to-Y transposition before (*FLJ36031*) and after (*TETY1*) the divergence of cat and dogs, respectively (Murphy et al. 2006). We have recently reported the bovid lineage-specific Y-chromosome genes, *ZNF280BY*, *ZNF280AY*, and *PRAMEY*, which were derived from a transposition of a gene block (*ZNF280B-ZNF280A-PRAME*) on BTA17 (Liu et al. 2009; Yang 2009; Chang et al. 2010).

Sequencing of the human MSY revealed a total of 156 transcripts, including 78 protein-coding genes that collectively encode only 27 distinct proteins (Skaletsky et al. 2003) and 78 noncoding RNAs. These protein-coding genes are classified into three categories: (1) X-degenerate genes (*SRY*, *RPS4Y1*, *ZFY*, *TBL1Y*, *PRKY*, *USP9Y*, *DDX3Y*, *UTY*, *TMSB4Y*, *NLGN4Y*, *CYorf15A* and *CYorf15B*, *JARID1D*, *EIF1AY*, and *RPS4Y2*), which are all single copy, have an X-chromosome counterpart, and are largely housekeeping genes with broad expression profiles, or in some cases have acquired more specific functions, such as *SRY*, which regulates male sex determination (Murphy et al. 2006); (2) X-transposed genes (*TGIF2LY* and *PCDH11Y*), which have recently moved from the X to the Y; and (3) Y-specific ampliconic genes (*RBMY*, *DAZ*, *TSPY*, *CDY*, *BPY2*, *XKRY*, *PRY*, *HSFY*, and *VCY*), which are multicopy located in the palindromes of the ampliconic region, and are expressed exclusively in testes. These genes presumably enhance male spermatogenesis, and have been acquired from many genomic sources. A proposed growth control Y (*GCY*) gene that is associated to the control of embryonic growth, stature, and development of teeth was assigned to a region near the centromere of the human Y (Ogata and Matsuo 1993; Kirsch et al. 2000, 2002a, 2002b, 2004). But *GCY* has not yet been confirmed at a transcriptional level. If it is confirmed, it may have a potential value for growth selection in animal breeding.

The gene content of the bovine MSY is, however, still unknown. Earlier investigations were focused on the development of Y chromosome-specific markers and the identification of bovine MSY genes by a comparative mapping approach. To accelerate this process, Ponce de León and Carpio (1995) generated a BTAY-specific DNA phage library with an average insert size of 675 bp, representing approximately 3.8× coverage of BTAY (Ponce de León 1996). This library has proven to be a very important resource not only for generating Y-specific markers and building the first-generation BTAY RH map (Figure 7.2; Liu et al. 2002), but also for targeting the BTAY gene content by direct testis cDNA selection (Liu et al. 2010). A list of BTAY markers is summarized in Table 7.1 to reflect worldwide efforts in this regard (Bondioli et al. 1989; Miller and Koopman 1990; Matthews and Reed 1991; Vaiman et al. 1994; Cui et al. 1995; Vogel et al. 1997a, 1997b; Xiao et al. 1998). There are about 100 markers available so far (Table 7.1; also see in INRA BOVMAP Database (http://dga.jouy.inra.fr/cgi-bin/lgbc/loci_part.operl?MAPYN=Mapping&BASE=cattle&PARTIE=BTAY).

To date, a total of 15 orthologs ((1) *UBE1AY*, (2) *AMELY*, (3) *DDX3Y*, (4) *USP9Y*, (5) *UTY*, (6) *EIF1AY*, (7) *EIF2S3Y*, (8) *OFD1Y*, (9) *RBMY*, (10) *ZFY*, (11) *TSPY*, (12) *HSFY*, (13) *SRY*, (14) *DAZ*, and (15) *CDY*) of human or mouse Y chromosome-related genes have been identified in cattle (Table 7.1; Liu 2010). These genes except for *DAZ* and *CDY* are physically mapped on BTAY either by a comparative approach (*RBM1A1*, *ZFY*, *DDX3Y*) (Liu et al. 2009), or by restriction mapping (*AMELY*, *TSPY*), or by FISH and/or RH mapping (*SRY*, *DDX3Y*, and *UTY*) (Liu et al. 2002, 2009), or by testis direct cDNA selection and male-specific polymerase chain reaction (PCR) (*UBE1AY*, *AMELY*, *DDX3Y*, *USP9Y*, *UTY*, *EIF1AY*, *EIF2S3Y*, *OFD1Y*, *RBMY*, *ZFY*, *TSPY*, *HSFY*) (Liu et al. unpublished data). The *DAZ* and *CDY* gene families are not present on BTAY, while their autosomal copies, *DAZL* and *CDYL*, do exist in the bovine genome (Liu et al. 2007; Wang et al. 2008).

To identify the gene content of the bovine MSY, experiments for direct testis-cDNA selection (Del Mastro and Lovett 1997) with the BTAY-specific DNA library as probes were carried out (Liu 2010; Chang et al. 2011). A magnetic beads system

**Table 7.1** List of mapped loci on the bovine Y chromosome.[a]

| Locus name | Map position | Gene name or type of marker | Reference |
|---|---|---|---|
| AF275611 | Y | Bovine STS AF275611 | Laurent et al. 2000[b] |
| AF275612 | Y | Bovine STS AF275612 | Laurent et al. 2000[b] |
| AF275618 | Y | Bovine STS AF275618 | Laurent et al. 2000[b] |
| AMELY | Yp | Amelogenin, Y-linked | Ennis and Gallagher 1994 |
| ANT3 | Yp, Xq | ADP/ATP translocase 3 | Liu and Ponce de León 2004 |
| ASMT | Y | Acetylserotonin *O*-methyltransferase | Donohue et al. 1992 |
| BL22 | Y | DNA segment (BL22) (XBM31) | Sonstegard et al. 1997 |
| BL22A | Yp | Microsatellite | Sonstegard et al. 1997; Liu et al. 2002 |
| BL22B | Yp | Microsatellite | Sonstegard et al. 1997; Liu et al. 2002 |
| BYM-1 | Y | Microsatellite BYM-1 | Ward et al. 2001 |
| CSF2RA | Yp, Xq | Colony-stimulating factor 2 receptor alpha | Liu and Ponce de León 2004 |
| DDX3Y | Y | DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked | Liu et al. 2009 |
| DYS001 | Y | DNA segment (OPA.06.3100) | Antoniou and Skidmore 1995 |
| DYS1 | Y | DNA segment (bov35m) | Miller and Koopman 1990 |
| DYS2 | Y | DNA segment (bov97m) | Miller and Koopman 1990 |
| DYS23 | Y | Microsatellite (IOBT1489) | Sonstegard et al. 2001 |
| DYS3 | Y | Microsatellite (INRA008) | Vaiman et al. 1994 |
| DYS4 | Y | Microsatellite (INRA057) | Vaiman et al. 1994 |
| DYS5 | Y | Microsatellite (INRA062) | Vaiman et al. 1994 |
| DYS6 | Y | Microsatellite (INRA124) | Vaiman et al. 1994 |
| DYS7 | Y | Microsatellite (INRA126) | Vaiman et al. 1994 |
| DYS8 | Y | Microsatellite (BM861) | Kappes et al. 1997 |
| DYZ1 | Yp12 | DNA segment (DYZ-1) | Perret et al. 1990 |
| DYZ10 | Yq | Microsatellite (IDVGA50) | Mezzelani et al. 1995 |
| DYZ3 | Y | DNA segment (ES5(2)) | Bondioli et al. 1989 |
| DYZ4 | Y | DNA segment (ES8) | Bondioli et al. 1989; Schwerin et al. 1992 |
| DYZ5 | Yp12 | DNA segment (ES6.0) | Schwerin et al. 1992 |
| DYZ6 | Yp12 | DNA segment (BC1.2) | Schwerin et al. 1992 |
| DYZ7 | Y | DNA segment (BRY.1) | Schwerin et al. 1992 |
| DYZ8 | Y | DNA segment (BRY.2) | Matthews and Reed 1992 |
| DYZ9 | Y | DNA segment (BRY.3) | Matthews and Reed 1992 |
| DZY10 | Yq | Microsatellite (IDVGA50) | Mezzelani et al. 1995 |
| EIF1AY | Yp | Eukaryotic translation initiation factor 1A, Y-linked | Liu et al. (unpublished); Van Laere et al. 2008 |

(*continued*)

**Table 7.1** (*Continued*)

| Locus name | Map position | Gene name or type of marker | Reference |
|---|---|---|---|
| EIF2S3Y | Yp | Eukaryotic translation initiation factor 2, subunit 3, structural gene Y-linked | Liu et al. (unpublished) |
| FBNY | Y | DNA fragment FBNY | Weikard et al. 2001 |
| HEL26 (DXS29) | Yp | Microsatellite | Vilkki et al. 1995 |
| HSFY | Y | Heat shock transcription factor, Y-linked; >100 copies | Liu et al. (unpublished) |
| INRA189 | Yq | Microsatellite (INRA189) | Kappes et al. 1997 |
| INRA30 | Yp13 | Microsatellite (INRA030) | Kappes et al. 1997 |
| MAF45 | Yp | Microsatellite (MAF45, ovine) | Kappes et al. 1997 |
| MCM74 | Y | Microsatellite (MCM74) | Sonstegard et al. 2001 |
| OFD1Y | Yp | Oral-facial-digital syndrome 1, Y-linked | Liu et al. (unpublished) |
| PBRF1R1A | Y | DNA segment | Liu et al. 2002 |
| PBRF1R1B | Y | DNA segment | Liu et al. 2002 |
| PBRF1R2 | Y | DNA segment | Liu et al. 2002 |
| PRAMEY | Yq | Preferentially expressed antigen in melanoma, Y-linked; >10 copies | Liu et al. (unpublished) |
| R1-0907RA | Y | DNA segment | Liu et al. 2002 |
| R1-0907RB | Y | DNA segment | Liu et al. 2002 |
| RBMY | Y | RNA binding motif protein, Y-linked | Liu et al. (unpublished); Skaletsky et al. 2003 |
| SRY | Yq | Sex-determining region Y | Liu and Ponce de León 2004 |
| SRY-HMG | Y | Sex-determining region Y | Cui et al. 1995; Liu et al. 2002 |
| TGLA325 | Yp | Microsatellite (TGLA325) | Kappes et al. 1997 |
| TSPY | Y | Testis-specific protein, Y-linked; multicopy | Jakubiczka et al. 1993; Verkaar et al. 2004 |
| UBE1Y | Yq | Ubiquitin-activating enzyme E1, Chr Y | Liu et al. (unpublished) |
| UBE2D3Y | Yq | Ubiquitin-conjugating enzyme E2D 3, Y-linked; multicopy, pseudogenes | Liu et al. (unpublished) |
| UMN0103 | Y | Microsatellite | Liu et al. 2002 |
| UMN0108 | Yp | Microsatellite | Liu et al. 2002 |
| UMN0301 | Y | Microsatellite | Liu et al. 2002 |
| UMN0304 | Y | Microsatellite | Liu et al. 2002 |
| UMN0307 | Y | Microsatellite | Liu et al. 2002 |
| UMN0311 | Y | Microsatellite | Liu et al. 2002 |
| UMN0406 | Y | Microsatellite | Liu et al. 2002 |
| UMN0504 | Y | Microsatellite | Liu et al. 2002 |
| UMN0705 | Y | TSPY-microsatellite | Liu et al. 2002 |

**Table 7.1** (*Continued*)

| Locus name | Map position | Gene name or type of marker | Reference |
|---|---|---|---|
| UMN0803 | Yp | Microsatellite | Liu et al. 2002 |
| UMN0905 | Yp | Microsatellite | Liu et al. 2002 |
| UMN0907A | Y | Microsatellite | Liu et al. 2002 |
| UMN0907B | Y | Microsatellite | Liu et al. 2002 |
| UMN0910 | Y | Microsatellite | Liu et al. 2002 |
| UMN0920 | Y | Microsatellite | Liu et al. 2002 |
| UMN0929 | Yp | Microsatellite | Liu et al. 2002 |
| UMN1113 | Y | Microsatellite | Liu et al. 2002 |
| UMN1201 | Y | Microsatellite | Liu et al. 2002 |
| UMN1203 | Y | Microsatellite | Liu et al. 2002 |
| UMN1307 | Y | Microsatellite | Liu et al. 2002 |
| UMN1514 | Y | Microsatellite | Liu et al. 2002 |
| UMN1605 | Y | Microsatellite | Liu et al. 2002 |
| UMN2008 | Yp | Microsatellite | Liu et al. 2002 |
| UMN2102 (BTIGA50) | Y | Microsatellite | Liu et al. 2002 |
| UMN2303 | Y | Microsatellite | Liu et al. 2002 |
| UMN2404 | Y | Microsatellite | Liu et al. 2002 |
| UMN2405 (BTMS2437) | Y | Microsatellite | Liu et al. 2002 |
| UMN2611 | Y | Microsatellite | Liu et al. 2002 |
| UMN2706 | Y | Microsatellite | Liu et al. 2002 |
| UMN2713 | Y | Microsatellite | Liu et al. 2002 |
| UMN2905F | Y | Microsatellite | Liu et al. 2002 |
| UMN2905M | Y | Microsatellite | Liu et al. 2002 |
| UMN2908 | Yp | Microsatellite | Liu et al. 2002 |
| UMN3008 | Y | Microsatellite | Liu et al. 2002 |
| USP9Y | Yp | Ubiquitin-specific peptidase 9, Y-linked | Liu et al. (unpublished) |
| UTY | Yp | Ubiquitously transcribed tetratricopeptide repeat gene, Y-linked | Liu et al. (unpublished) |
| XBM31 | Yp | Microsatellite (XBM31) | Ponce de León et al. 1996 |
| XBM451 | Yp | Microsatellite (XBM451) | Kappes et al. 1997 |
| ZFY | Yp12 | Zinc finger protein, Y-linked | Xiao et al. 1998 |
| ZNF280AY | Yq | Zinc finger protein 280A, Y-linked; >100 copies, pseudogenes | Liu et al. (unpublished) |
| ZNF280BY | Yq | Zinc finger protein 280B, Y-linked; >100 copies | Liu et al. (unpublished) |

[a] Modified from Liu and Ponce de León 2007.

[b] Information can be found at http://dga.jouy.inra.fr/cgi-bin/lgbc/loci_part.operl?MAPYN=Mapping&BASE=cattle&PARTIE=BTAY.

for Y-chromosome gene enrichment and a TA-cloning system to clone the selected cDNAs were applied. Sequencing of about 750 selected cDNA clones resulted in approximately 270 clear sequences, which were categorized into three groups on the basis of blast analysis results and annotations. The first group contained bovine homologs of known Y-chromosome genes reported previously for human and other mammalian species. These include ten single-copy X-degenerated genes ((1) *UBE1AY*, (2) *AMELY*, (3) *OFD1Y*, (4) *DDX3Y*, (5) *USP9Y*, (6) *UTY*, (7) *EIF1AY*, (8) *EIF2S3Y*, (9) *RBMY*, and (10) *ZFY*), and two multiple copy X-degenerated genes, (1) *HSFY* and (2) *TSPY* (Liu et al. Unpublished data). The second group was formed by BTAY-specific transcripts with high similarity to predicted genes from the Bovine Genome Sequence (build 4) on BTAX and autosomes (or unmapped). Four genes/transcripts (1) *ZNF280BY*, (2) *ZNF280AY*, (3) *PRAMEY*, and (4) *UBE2D3Y*, which have a copy on an autosome, are multiple-copy gene families on BTAY. Our preliminary analysis indicated that the first three genes are clustered together in BTA17, which transposed to BTAY and amplified thereafter during bovine evolution (Chang et al. 2010). Although the *ZNF280AY* and *UBE2D3Y* transcripts have multicopies on BTAY and are still active at the transcriptional level, they all are pseudogenes on the Y. However, their counterparts in BTA17 are functional. The third group included over 100 BTAY novel transcripts that do not match any genes in GenBank. At least ten different transcripts (or families of transcripts) have been proved to be multiple copy genes all located in the MSY region. Preliminary analysis indicated that these novel transcripts are noncoding RNAs, very much like the noncoding RNAs identified on the human Y chromosome (Liu et al. 2009).

According to the available literature information, we placed the MSY genes in the following order starting from the PAB: *EIF1AY*-*AMELY*-*OFD1Y*-*USP9Y*-*UTY*-*DDX3Y*-*ZFY*-*EIF2S3Y*-*TSPY*-(amplified multicopy genes including *ZNF280BY*, *ZNF280AY*, *PRAMEY*, *UBE2D3Y*, *HSFY*, and the novel noncoding transcripts)-*UBE1AY*-*SRY* (Liu and Ponce de León 2004; Ding et al. 2009; Liu 2010). We do not know yet where *RBMY* is located on BTAY. The localization of *SRY* in the distal region of BTAYq (Liu et al. 2002; Liu and Ponce de León 2004) is unusual as *SRY* is usually located near the PAB on most mammalian Y chromosomes (Graves et al. 1998).

To our knowledge, only four BTAY genes ((1) *AMELY*, (2) *SRY*, (3) *DDX3Y*, and (4) *TSPY*) were previously characterized in detail. The bovine *AMEL* genes reside on both the X and Y chromosomes and are expressed only in tooth buds. Alternative mRNA splicing generates at least seven messages, five from the *AMELX* primary transcript and two from the *AMELY* (Gibson et al. 1991; Yuan et al. 1996). Similar to the bovine *AMELX/Y* genes, the bovine DEAD box protein gene also has a Y-copy (*DDX3Y*) and an X-copy (*DDX3X*). Two transcripts of the bovine *DDX3Y* (*DDX3Y-L* and *DDX3Y-S*) were isolated, corresponding to the long and short transcripts of the human *DDX3Y* and mouse *Ddx3y* gene (Foresta et al. 2000; Vong et al. 2006). The two transcripts are identical except for a 3-bp (AGT) insertion and an expanded 3′UTR in *bDDX3Y-L*. The *bDDX3Y-S* encodes a peptide of 660 amino acids (aa), while the *bDDX3Y-L* encodes a peptide of 661 aa due to an additional serine (S) insertion. Both *DDX3Y* isoforms contain the conserved DEAD-box motif. The bovine *DDX3Y* is composed of 17 exons. The homologous gene on the X chromosome, *bDDX3X*, is highly conserved and shows similar genomic structure as well as 83% and 88% similarity to the Y mRNA copy and protein, respectively (Liu et al. 2009).

An autosomal paralog of the bovine *DDX3X/Y*, named *PL10*, was also identified and mapped to BTA15 by RH mapping. *PL10* is a processed pseudogene with a similarity of 88.1% to *DDX3Y* and 93.7% to *DDX3X* mRNA, suggesting that *bPL10* is a retroposon of *bDDX3X*. RT-PCR analyses showed that *DDX3Y-L, DDX3Y-S, DDX3X*, and *PL10* were all widely expressed with predominant expression in testis and brain. In situ hybridization analysis on testicular sections revealed that sense and antisense RNAs of *DDX3Y-L, DDX3Y-S*, and *DDX3X* are expressed in interstitial cells (Liu et al. 2009). A further phylogenetic analysis (Chang and Liu 2010) of the bovine *DDX3* gene families (*DDX3X/DDX3Y/PL10*) and their orthologs in a variety of species from yeast, plants, to animals, including humans (Rosner and Rinkevich 2007), revealed that the evolution of *DDX3Y* homologs was under positive selection and the elevated Ka/Ks ratios observed in eutherian lineages for *DDX3Y*, but not for *PL10* and *DDX3X*, suggest relaxed selective constraints on *DDX3Y* (Chang and Liu 2010). All other bovine Y-chromosome genes mentioned previously are currently being sequenced and characterized.

As stated before, two human Y-chromosome gene families, (1) *DAZ* and (2) *CDY*, are autosomal in cattle (Liu et al. 2007; Wang et al. 2008). Because the two gene families were derived from an autosome in the primate-lineage through autosome-to-Y transposition, and are the most important candidates for Azoospermia Factor (AZF) and male infertility (Saxena et al. 1996; Lahn and Page 1997) in humans, we believe it is necessary to give a brief introduction about the bovine orthologs, *DAZL* and *CDYL*, of the human *DAZ* and *CDY* even though these two genes are not localized in BTAY.

In human, the *DAZ* gene family has four copies (*DAZ1-4*) on the Y chromosome and one copy (*DAZ*-like, or *DAZL*) on chromosome 3 (Cauffman et al. 2005). Deletions and single nucleotide polymorphisms (SNPs) identified in *DAZ* and/or *DAZL* have been linked to subfertility and infertility in several species including human (Teng et al. 2002), mouse (Ruggiu et al. 1997), fly (Eberhart et al. 1996), and frog (Houston and King 2000). The bovine *DAZL* contains 11 exons, encodes a protein of 295 aa, and is highly (96%) conserved when compared to human and mouse *DAZL*. Two transcript variants were found for the bovine *DAZL*, which are expressed in bovine testis only, while the human and mouse *DAZL* are expressed in both male and female gonads (Ruggiu et al. 1997; Teng et al. 2002). Sixteen SNPs for the bovine *DAZL* gene have been reported. A preliminary association study indicated that these SNPs are associated with bull fertility (Liu 2008). Similarly, there are four copies of the *CDY* gene on the human Y chromosome, and two copies, *CDYL* (*CDY*-like) and *CDYL2*, on human autosomes (Lahn and Page 1999a). It is believed that the progenitor of this gene family was duplicated to generate *CDYL* and *CDYL2*, and that *CDY* arose by retroposition of *CDYL* to the Y chromosome and was retained only in simian mammals (Lahn and Page 1999a; Dorus et al. 2003). This explains why in bovine (Wang et al. 2008) and in mice (Dorus et al. 2003) only autosomal *CDYL* and *CDYL2* genes were found. The bovine *CDYL* and *CDYL2* are highly similar to the human orthologs at both mRNA (81% and 82%) and protein (89% and 94%) levels. However, the similarity between the bovine *CDYL* and *CDYL2* proteins is low (41%). The bovine *CDYL* and *CDYL2* genes were assigned to BTA23 and BTA18 by RH mapping, respectively. Sequence analyses indicated that there are at least four transcript variants that yield three protein isoforms for the bovine *CDYL* gene. Expression analysis in different bovine tissues showed the bovine *CDYL* variant 2 to be expressed only in testis and variants 1, 3, and 4 expressed predominantly in testis and at very low or undetectable levels in other tissues, whereas the *CDYL2* was expressed ubiquitously. Examination of bovine testis by in situ hybridization revealed that the *CDYL* and *CDYL2* transcripts were found mainly in spermatids, though the amounts of transcripts varied among genes and isoform variants. In addition, antisense transcripts were detected in the bovine *CDYL* variants 2/3, and 4, and the *CDYL2* gene (Wang et al. 2008). These results indicated that the bovine *DAZL* and *CDYL* genes, just like the human *DAZ* and *CDY*, play essential roles in spermatogenesis and fertility.

## Bovine Y-Chromosome Phylogeny

Bovine phylogenetic studies based on mitochondrial DNA sequences indicate that domestication of taurine (*Bos taurus*) in the Near East and zebuine (*Bos indicus*) in the Indus Valley required at least two genetically distinct auroch (*Bos primigenius*) species. The mitochondrial macrohaplogroup T (*B. taurus*) has six haplogroups (T, T1, T2, T3, T4 and T5) and the Q haplogroup of European auroch origin identified in the Near East cattle populations (Troy et al. 2001; Mannen et al. 2004; Achilli et al. 2008). Two separate haplogroups P and R of European auroch ancestry were identified in an animal from Korea and animals from the Italian peninsula, respectively. Haplogroups Q, P, and R are found at low frequencies while the T3 haplogroup predominates in European cattle, which originated from the expansion of a small cattle population domesticated in the Near East. T4 was found in Japanese cattle and is a derived clade within T3 suggesting its origin from either the same genetic source as the T3 founder sequence(s), or at most from a genetically (and geographically) closely related population of aurochs. The T1 haplogroup was found mainly in Northern Africa while the T2 haplogroup was found mainly in Continental Europe, Anatolia, and the Middle East. The T haplogroup represents the mitochondrial sequence more similar to the original and was found predominantly in Anatolia and the Middle East. The introgression of haplogroup P most likely took place either in Northern or Central Europe, while haplogroup Q was possibly acquired from a different population of aurochs that might have ranged only south of the Alps (Beja-Pereira et al. 2006). The pre-Neolithic macrohaplogroup T, also found in some European aurochs samples, has led some authors to argue that this haplogroup was not restricted to the Near East, and that wild haplogroup T females may have been incorporated locally into the European domestic pool (Beja-Pereira et al. 2006).

Zebu (*B. indicus*) origin of domestication has been determined on the basis of the two major haplogroups I1 and I2, which are well represented in India suggesting either a single domestication event followed by introgression of wild auroch females into protodomesticated herds (Baig et al. 2005), or more probably that domestication included two different wild female populations (Chen et al. 2010).

Y-chromosome phylogenetic studies are rare (Verkaar et al. 2004), and most have been focusing on taurine and zebuine crosses (Hanotte et al. 2000; Anderung et al. 2007; Edwards et al. 2007). It is only relatively recently that Götherström et al. (2005) identified five SNPs that permitted the identification of contemporary breeds into three Y-chromosome haplogroups (Y1, Y2, Y3). The Y1 haplogroup was found to be prevalent among cattle in Northwestern Europe, Y2 was prevalent in Southern

Europe, and Anatolian cattle and the Y3 haplogroup was identified only in zebu. These findings indicated that the Y2 haplotypes represent cattle domesticated in the Near East while the Y1 haplotype represents European aurochs demonstrating the male lineage genetic influence of European aurochs in the formation of contemporary European cattle.

Even though Götherström et al. (2005) confirmed the mitochondrial phylogenetic studies described previously, the very few markers being used do not confer the level of robustness that is necessary, especially when ancient DNA studies do not support crossbreeding of wild European auroch and domesticated cattle (Edwards et al. 2007; Bollongino et al. 2008), and medieval Scandinavian samples are found to belong to the Y2 haplogroup (Svensson and Götherström 2008).

In an effort to increase the marker coverage of the Y chromosome and the robustness of bovine Y-chromosome phylogenetic studies, Perez-Pardal et al. (2010a, 2010b) used a set of interspersed multilocus microsatellites (IMMs) described by Liu et al. (2003). These MSY microsatellites yield several amplified bands of different sizes using a single primer pair in a single PCR reaction from a single male DNA sample. Two out of five IMMs tested for amplification only in male DNA, polymorphism (presence or absence of an amplified band), paternal compatibility, and correct and repeatable scoring were used for this study. IMM UMN2405 yielded a total of 30 amplified bands of which 25 were polymorphic and UMN2303 yielded 23 polymorphic bands out of a total of 38. Essentially, this study assessed 48 polymorphic loci. The use of these markers (IMMs) not only confirmed the findings already described by Götherström et al. (2005), but it increased the resolution of detection by identifying a subhaplogroup within the Y2 haplogroup in cattle Y chromosomes sampled from Northern Italy, Northern Atlantic Europe, Mongolia, and Japan, which might correspond to the mitochondrial haplogroup Q identified by Achilli et al. (2008).

The synergistic use of Y-specific microsatellites and the SNPs described by Götherström et al. (2005) increased analytical resolution and allowed at least two different Y2-haplotypic subfamilies to be distinguished, one of them in Northern Italy and the other restricted to the African continent (Perez-Pardal et al. 2010b).

Taken together these studies have further suggested that there has been introgression of wild sire European auroch genetics into domesticated herds, that cattle domestication in Africa most probably included local Y2 wild auroch sires, and that the high genetic similarity found in Asian zebu supports a single domestication event. Overall, there is a need to develop and characterize more Y-chromosome markers to refine the few phylogenetic studies based on the male lineage.

## Sex Chromosome Abnormalities

Numerical (aneuploidy) and structural (translocations, deletions, etc.) chromosome abnormalities generally lead to reproductive failure and/or reduced fertility. Information on bovine sex chromosome abnormalities is scanty as most subfertile and/or infertile animals are culled before any cytogenetic analysis is carried out. Our knowledge of sex chromosome abnormalities mostly derives from observations and studies in other domestic animals and humans. However, Mikaye and Kaneda (1988) reported two cases, one mosaic and one chimeric with 60,XY/61,XYY and XX/XY karyotypes, respectively, among unilateral cryptorchid bulls. In bovine, mosaics XX/XY have been commonly reported for freemartin animals (Marcum 1974) where the male chimeric twin animal is most commonly sterile (Dunn et al. 1979; Schmutz et al. 1996). It is important to state here that freemartin chimeras are only somatic chimeras and not germ line. Also, Swartz and Vogt (1983) reported that out of 71 nulliparous heifers belonging to 11 breeds and various crossbreeds, 18.3% showed chromosomal abnormalities. These abnormalities included one tetraploid/diploid mosaic, five 1/29 translocations, two trisomy X, two 59,XO/60,XX mosaics, one 60,XX/60,XY chimeric mosaics, and two mixoploid mosaics with karyotypes 59,XO/60,XX/61,XXX and 59,XO/60,XX/61XO, respectively. Unfortunately, these studies, as in most cases, did not include germ line synaptonemal complex analysis to determine if these chromosomal abnormalities were also observed in the germ line.

Chromosome rearrangements involving sex chromosomes and autosomes are rare in mammals because most result in a high rate of nonviable conceptuses. In cattle, Basrur et al. (1992) reported the identification of X-autosome translocation (X-AT) cow carriers in a Limousin–Jersey crossbred. This study and its subsequent studies represent to date the most comprehensive analysis of X-AT in cattle. This translocation was demonstrated to have involved the X chromosome and chromosome 23 (t(Xp+;23q−)) (Gallagher et al. 1992; Basrur et al. 2001a, 2001b). Cow carriers of this translocation showed higher rates of fertilization failure, abnormal embryos, and return to estrus. Carriers also showed a relatively high rate of abortion by the second trimester of gestation and only 13 live births were obtained among which four translocation carriers were identified. In vitro cell cultures from different tissues of these X-AT carriers allowed the assessment of their X inactivation pattern. These cell cultures showed that X-AT carriers preferentially late-replicated the normal X chromosome, therefore, assuring normality of expression for genes located in the X chromosome and autosome 23. Authors suggested that this selective process favoring cells in which the genes of the normal X chromosome are inactivated in the translocation carrier females may be the mechanism that helps these embryos escape the adverse effects of their aneuploidy. Also, X-AT infertile carrier bulls were studied. Their semen contained few and malformed spermatozoa, although testicular histological studies of seminiferous tubules indicated that all stages of spermatogenesis were present. Synaptonemal complex analysis of spermatocytes of these X-AT carriers showed a large proportion of spermatocytes with trivalent configuration consisting of $X_p+$, the normal chromosome 23 in partial synapsis with $X_p$, and the Y chromosome broken away from the PAR. Implying a loss of $Y_q$ that led the authors to suggest an association of this chromatin loss to the sperm head malformation and oligospermia observed in these X-AT carrier bulls (Basrur et al. 2001b).

These few reports demonstrate that chromosomal abnormalities in general and sex chromosomal abnormalities in particular are as common in cattle as they are in other farm animals but that most of these animals are culled before any cytogenetic assessment is done.

## Sex Chromosomes QTLs

Detection of quantitative trait loci (QTLs), that is those loci that control genetic additive effects, have been possible because of the availability of abundant genetic molecular markers, phenotype records, and appropriate resource populations. However, relatively few QTLs have been reported to be located on the bovine sex chromosomes. The following information is available at http://www.genome.iastate.edu/cgi-bin/QTLdb/BT/index.

Detection of the QTL for bovine spongiform encephalopathy (BSE) resistance and/or susceptibility was based on a genome scan of 360 daughters from four half-sib families with 268 BSE-affected and 92 unaffected animals of the Holstein breed. This study analyzed 173 microsatellite spanning 29 autosomes and the PAR. Interval mapping analysis by linear regression extended to a multiple-QTL analysis that identified QTL on other chromosomes as cofactors was used. Significant QTL effects were identified to be located on the X/Y PAR, between the TGLA325 and INRA30 markers, and on BTA17. Four other suggestive QTL autosomal sites were also identified (Zhang et al. 2004)

QTLs for dystocia or calving difficulty across parities as a direct effect of the calf (DYSTd) was investigated in German Holstein using a granddaughter design. Genotypes of 263 markers covering all autosomes and the PAR were generated for 473 sons when the traits were assessed for first parity and on 1237 sons when traits were evaluated for second or later parities. Of the eight calving traits that were analyzed, only the DYSTd QTL was assigned to the PAR between Marker MAF45 and INRA30 (Seidenspinner et al. 2009).

In a separate half-sib family design study to identify maternal dystocia (DYSTm), stillbirths (SB), nonreturn rate at 90 days (NONR), functional herd life, and somatic cell count QTLs, researchers included 246 microsatellite markers, eight single strand conformation polymorphisms, four protein polymorphisms, and five erythrocyte antigen loci. The population sample included 16 German Holstein paternal half-sib families and 872 bulls. DYSTm, NONR, and SB QTLs were found in the PAR between markers MAF45 and INRA030 (Kuhn et al. 2003). However, 15 other significant and/or suggestive autosomal QTLs have been reported for DYSTm, one autosomal QTL site for NONR, and 21 autosomal sites for SB.

Three more QTLs, (1) milk energy yield (EY), (2) milk yield (MY), and (3) milk protein yield (PY), have been assigned to the bovine PAR. The analysis was based on a granddaughter design and breeding values of the first three lactations estimated with a random regression animal model. Genetic markers used in this study were the same as described previously (Kuhn et al. 2003). These three QTLs were all located between the INRA030 and MAF45 microsatellite markers in the PAR. MY also has about 158 significant and/or suggestive autosomal QTL sites identified. Likewise, EY and PY have nine and 88 autosomal QTL sites identified (Harder et al. 2006).

Sandor et al. (2006) using a granddaughter design and 22 X chromosome specific and three PAR microsatellite markers on Holstein-Friesian animals identified five significant QTLs of which four ((1) fat yield, (2) direct durability, (3) milking speed, and (4) rear leg set) were localized in the X-specific region and one (durable prestation) in the PAR. There are also 31 autosomal significant and/or suggestive QTLs reported for fat yield and, to our knowledge, no autosomal location for the other described traits. These researchers found also a higher level of linkage disequilibrium among X-chromosome markers than autosomal markers. This finding indicates a higher gonosomal than autosomal effective population size, which is not compatible with the small male-to-female ratio found in dairy breeding populations. It is important to recognize that when the X chromosome-specific region is treated as an autosomal chromosome, a sex difference in the phenotype can lead to the identification of a false linkage; in this case, a QTL on the X chromosome-specific region (Broman et al. 2006). X chromosome region-specific QTL associations when using the daughter and granddaughter designs should be revised as these designs do not measure linkage associations based on the grandsire X chromosome-specific region recombinations between markers and QTLs.

## References

Achilli, A., et al. (2008) Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Current Biology* **18**(4): 157–158.

Aflöldi, J.E. (2008) Ph.D. Thesis. Sequence of the Mouse Y Chromosome.

Aitken, R.J. and Graves, J.A.M. (2002) Human spermatozoa: the future of sex. *Nature* **415**(6875): 963–963.

Anderung, C., Hellborg, L., Seddon, J., Hanotte, O., Götherström, A. (2007) Investigation of X- and Y-specific single nucleotide polymorphisms in taurine (Bos taurus) and indicine (Bos indicus) cattle. *Animal Genetics* **38**: 595–600.

Antoniou, E. and Skidmore, C.J. (1995) A bovine Y-specific marker amplified by RAPD. *Animal Genetics* **26**(6): 444–445.

Baig, M., Beja-Pereira, A., Mohammad, R., Kulkarni, K., Farah, S., Luikart, G. (2005) Phylogeography and origin of Indian domestic cattle. *Current Science* **89**: 38–40.

Basrur, P.K., Pinheiro, L.E., Berepubo, N.A., Reyes, E.R., Popescu, P.C. (1992) X chromosome inactivation in X autosome translocation carrier cows. *Genome/National Research Council Canada* **35**(4): 667–675.

Basrur, P.K., Koykul, W., Baguma-Nibasheka, M., King, W.A., Ambady, S., Ponce de León, F.A. (2001a) Synaptic pattern of sex complements and sperm head malformation in X-autosome translocation carrier bulls. *Molecular Reproduction and Development* **59**(1): 67–77.

Basrur, P.K., Reyes, E.R., Farazmand, A., King, W.A., Popescu, P.C. (2001b) X-autosome translocation and low fertility in a family of crossbred cattle. *Animal Reproduction Science* **67**(1–2): 1–16.

Beja-Pereira, A., et al. (2006) The origin of European cattle: evidence from modern and ancient DNA. *Proceedings of the National Academy of Sciences of the United States of America* **103**(21): 8113–8118.

Bollongino, R., Elsner, J., Vigne, J.-D, Burger, J. (2008) Y-SNPs do not indicate hybridisation between European aurochs and domestic cattle. *PLoS ONE* 3, http://dx.doi.org/10.1371%2Fjournal.pone.0003418.

Bondioli, K.R., Ellis, S.B., Pryor, J.H., Williams, M.W., Harpold, M.M. (1989) The use of male-specific chromosomal DNA fragments to determine the sex of bovine preimplantation embryos. *Theriogenology* **31**: 95–104.

Broman, K.W., Sen, S., Owens, S.E., Manichaikul, A., Southard-Smith, E.M., Churchill, G.A. (2006) The X chromosome in quantitative trait locus mapping. *Genetics* **174**(4): 2151–2158.

Buckland, R.A. and Evans, H.J. (1978) Cytogenetic aspects of phylogeny in the Bovidae. I. G-banding. *Cytogenetics and Cell Genetics* **21**: 42–63.

Bunch, T.D. and Nadler, C.F. (1980) Giemsa-band patterns of the Tahr and chromosomal evolution of the tribe Caprini. *Journal of Heredity* **71**: 110–116.

Cauffman, G., Van de Velde, H., Liebaers, I., Van Steirteghem, A. (2005) DAZL expression in human oocytes, preimplantation embryos and embryonic stem cells. *Molecular Human Reproduction* **11**: 405–411.

Chang, T.-C. and Liu, W.-S (2010) The molecular evolution of PL10 homologs. *BMC Evolutionary Biology* **10**: 127.

Chang, T.-C., Yang, Y., Yasue, H., Liu, W.-S (2011) ZNF280BY and PRAMEY: autosome derived Y chromosome gene families in cattle. *BMC Genomics*. **12**: 13.

Charlesworth, B. and Charlesworth, D. (2000) The degeneration of Y chromosomes. *Philosophical transactions of the Royal Society of London: Series B, Biological sciences* **355**(1403): 1563–1572.

Chen, S., et al. (2010) Zebu cattle are an exclusive legacy of the South Asia neolithic. *Molecular Biology and Evolution* **27**(1): 1–6.

Church, D.M., et al. (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biology* **7**(5): e1000112.

Cui, X., Kato, Y., Sato, S., Sutou, S. (1995) Mapping of bovine Sry gene on the distal tip of the long arm and murine Sry on the short arm of the Y Chr by the method of fluorescence in situ hybridization (FISH). *Animal Science and Technology (Japan)* **66**: 441–444.

Das, P.J., Chowdhary, B.P., Raudsepp, T. (2009) Characterization of the bovine pseudoautosomal region and comparison with sheep, goat, and other mammalian pseudoautosomal regions. *Cytogenetic and Genome Research* **126**(1–2): 139–147.

Del Mastro, R.G. and Lovett, M. (1997) Isolation of coding sequences from genomic regions using direct selection. *Methods in Molecular Biology* **68**: 183–199.

Di Berardino, D., Iannuzzi, L., Bettini, T.M., Matassino, D. (1981) Ag-NORs variation and banding homologies in two species of Bovidae: Bubalus bubalis and Bos taurus. *Canadian Journal of Genetics and Cytology* **23**(1): 89–99.

Di Meo, G.P., et al. (2005) Chromosome evolution and improved cytogenetic maps of the Y chromosome in cattle, zebu, river buffalo, sheep and goat. *Chromosome Research* **13**(4): 349–355.

Ding, Y., Worley, K.C., Page, D.C. (2009) Upgrading the bovine draft genome- Y chromosome finishing. Bovine Genome Consortium, May 9–11, 2009.

Donohue, S.J., Roseboom, P.H., Klein, D.C. (1992) Bovine hydroxyindole-O-methyltransferase. Significant sequence revision. *Journal of Biological Chemistry* **267**(8): 5184–5185.

Dorus, S., Gilbert, S.L., Forster, M.L., Barndt, R.J., Lahn, B.T. (2003) The CDY-related gene family: coordinated evolution in copy number, expression profile and protein sequence. *Human Molecular Genetics* **12**(14): 1643–1650.

Dunn, H.O., McEntee, K., Hall, C.E., Johnson, R.H., Jr , Stone, W.H. (1979) Cytogenetic and reproductive studies of bulls born co-twin with freemartins. *Reproduction* **57**(1): 21–30.

Eberhart, C.G., Maines, J.Z., Wasserman, S.A. (1996) Meiotic cell cycle requirement for a fly homologue of human Deleted in Azoospermia. *Nature* **381**: 783–785.

Edwards, C.J., et al. (2007) Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proceedings of Biological Sciences/The Royal Society* **274**(1616): 1377–1385.

Ennis, S. and Gallagher, T.F. (1994) A PCR-based sex-determination assay in cattle based on the bovine amelogenin locus. *Animal Genetics* **25**(6): 425–427.

Evans, H.J., Buckland, R.A., Sumner, A.T. (1973) Chromosome homology and heterochromatin in goat, sheep and ox studied by banding techniques. *Chromosoma* **42**: 42–63.

Everts-van der Wind, A., Larkin, D.M., Green, C.A., Elliott, J.S., Olmstead, C.A., Chiu, R., Schein, J.E., Marra, M.A., Womack, J.E., Lewin, H.A. (2005) A high-resolution whole-genome cattle-human comparative map reveals details of mammalian chromosome evolution. *Proceedings of the National Academy of Sciences of the United States of America* **102**(51): 18526–18531.

Foresta, C., Ferlin, A., Moro, E. (2000) Deletion and expression analysis of azfa genes on the human Y chromosome revealed a major role for dby in male infertility. *Human Molecular Genetics* **9**: 1161–1169.

Gallagher, D.S. and Womack, J.E. (1992) Chromosome conservation in the Bovidae. *Journal of Heredity* **83**: 287–298.

Gallagher, D.S. Jr, Basrur, P.K. and Womack, J.E. (1992) Identification of an autosome to X chromosome translocation in the domestic cow. *Journal of Heredity* **83**: 451–453.

Gibson, C., Golub, E., Herold, R., Risser, M., Ding, W., Shimokawa, H., Young, M., Termine, J., Rosenbloom, J. (1991) Structure and expression of the bovine amelogenin gene. *Biochemistry* **30**: 1075–1079.

Glas, R., Marshall Graves, J.A., Toder, R., Ferguson-Smith, M., O'Brien, P.C. (1999) Cross-species chromosome painting between human and marsupial directly demonstrates the ancient region of the mammalian X. *Mammalian Genome* **10**(11): 1115–1116.

Götherström, A., Anderung, C., Hellborg, L., Elburg, R., Smith, C., Bradley, D.G., Ellegren, H. (2005) Cattle domestication in the Near East was followed by hybridization with aurochs bulls in Europe. *Proceedings Biological Sciences/The Royal Society* **272**(1579): 2345–2350.

Graves, J.A.M. (2000) Human Y chromosome, sex determination, and spermatogenesis—a feminist view. *Biology of Reproduction* **63**(3): 667–676.

Graves, J.A.M. (2006) Sex chromosome specialization and degeneration in mammals. *Cell* **124**(5): 901–914.

Graves, J.A.M., Wakefield, M.J., Toder, R. (1998) The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Human Molecular Genetics* **7**(13): 1991–1996.

Hanotte, O., Tawah, C.L., Bradley, D.G., Okomo, K., Verjee, Y., Ochieng, J., Rege, J.E.O. (2000) Geographic distribution and frequency of a taurine Bos taurine and an indicine Bos indicus Y-specific allele amongst sub-Saharan African cattle breeds. *Molecular Ecology* **9**: 387–396.

Harder, B., et al. (2006) Mapping of quantitative trait loci for lactation persistency traits in German Holstein dairy cattle. *Journal of Animal Breeding and Genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie* **123**(2): 89–96.

Hayes, H., Petit, E., Dutrillaux, B. (1991) Comparison of RBG-banded karyotypes of cattle, sheep, and goats. *Cytogenetics and Cell Genetics* **57**(1): 51–55.

Houston, D.W. and King, M.L. (2000) A critical role for Xdazl, a germ plasm-localized RNA, in the differentiation of primordial germ cells in Xenopus. *Development* **127**: 447–456.

Hurst, L.D. (1994) Embryonic growth and the evolution of the mammalian Y chromosome. I. The Y as an attractor for selfish growth factors. *Heredity* **73**(3): 223–232.

Iannuzzi, L., Di Meo, G.P., Perucatti, A., Ferrara, L. (1990) A comparison of G- and R-banding patterns in cattle and river buffalo prometaphase chromosomes. *Caryologia* **43**: 283–290.

Ihara, N., et al. (2004) A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Research* **14**(10A): 1987–1998.

Jakubiczka, S., Schnieders, F., Schmidtke, J. (1993) A bovine homologue of the human TSPY gene. *Genomics* **17**(3): 732–735.

Kappes, S.M., Keele, J.W., Stone, R.T., McGraw, R.A., Sonstegard, T.S., Smith, T.P., Lopez-Corrales, N.L., Beattie, C.W. (1997) A second-generation linkage map of the bovine genome. *Genome Research* **7**(3): 235–249.

Kirsch, S., Weiss, B., De Rosa, M., Ogata, T., Lombardi, G., Rappold, G.A. (2000) FISH deletion mapping defines a single location for the Y chromosome stature gene, GCY. *Journal of Medical Genetics* **37**(8): 593–599.

Kirsch, S., Weiss, B., Schon, K., Rappold, G.A. (2002a) The definition of the Y chromosome growth-control gene (GCY) critical region: relevance of terminal and interstitial deletions. *Journal of Pediatric Endocrinology & Metabolism* **15**(Suppl 5): 1295–1300.

Kirsch, S., Weiss, B., Kleiman, S., Roberts, K., Pryor, J., Milunsky, A., Ferlin, A., Foresta, C., Matthijs, G., Rappold, G.A. (2002b) Localisation of the Y chromosome stature gene to a 700 kb interval in close proximity to the centromere. *Journal of Medical Genetics* **39**(7): 507–513.

Kirsch, S., Weiss, B., Zumbach, K., Rappold, G. (2004) Molecular and evolutionary analysis of the growth-controlling region on the human Y chromosome. *Human Genetics* **114**(2): 173–181.

Kohn, M., Kehrer-Sawatzki, H., Vogel, W., Graves, J.A., Hameister, H. (2004) Wide genome comparisons reveal the origins of the human X chromosome. *Trends in Genetics* **20**(12): 598–603.

Krzywinski, M., et al. (2004) Integrated and sequence-ordered BAC- and YAC- based physical maps for the rat genome. *Genome Research* **14**(4): 766–779.

Kuhn, Ch., et al. (2003) Quantitative trait loci mapping of functional traits in the German Holstein cattle population. *Journal of Dairy Science* **86**(1): 360–368.

Kuroki, Y., et al. (2006) Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nature Genetics* **38**(2): 158–167.

Lahn, B.T. and Page, D.C. (1997) Functional coherence of the human Y-chromosome. *Science* **278**: 675–680.

Lahn, B.T. and Page, D.C. (1999a) Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nature Genetics* **21**(4): 429–433.

Lahn, B.T. and Page, D.C. (1999b) Four evolutionary strata on the human X chromosome. *Science* **286**(5441): 964–967.

Lander, E.S., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860–921.

Liu, W.-S. (2008) Polymorphisms of the bovine DAZL gene are associated with male fertility. *PAG-XVI San Diego, CA.* P174.

Liu W.-S. (2010) Comparative genomics of the Y chromosome and male fertility. In: *Reproductive Genomics in Domestic Animals*, edited by Zhihua Jiang and Troy L. Ott, pp. 129–155. Ames: Wiley-Blackwell.

Liu, W.-S. and Ponce de León, F.A. (2004) Assignment of *SRY, ANT3* and *CSF2RA* to the Bovine Y Chromosome by FISH and RH mapping. *Animal Biotechnology* **15**: 103–109.

Liu, W.-S and Ponce de León, F.A. (2007) Mapping of the Bovine Y chromosome. *Electronic Journal of Biology* **3**(1): 5–12.

Liu, W.S., Mariani, P., Beattie, C.W., Alexander, L.J., Ponce de León, F.A. (2002) A radiation hybrid map for the bovine Y chromosome. *Mammalian Genome* **13**(6): 320–326.

Liu, W-S., Beattie, C.W., Ponce de León, F.A. (2003) Bovine Y chromosome microsatellite polymorphism. *Cytogenet Genome Res* **10**: 53–58.

Liu, W.-S., Wang, A., Uno, Y., Galtz, D., Beattie, C.W., Ponce de León, F.A. (2007) Genomic structure and transcript variants of the bovine DAZL gene. *Cytogenetic and Genome Research* **116**(1–2): 65–71.

Liu, W.-S., Wang, A.-H; Yang, Y., Chang, T.-C; Landrito, E., Yasue, H. (2009) Molecular characterization of the DDX3Y gene and its homologs in cattle. *Cytogenetic and Genome Research* **126**(4): 318–328.

Liu, W.-S., Chang, T.-C; Yang, Y., Yasue, H., Crow, J.A., Retzel, E. (2010) Functional genomics of the bovine Y-chromosome. PAG XVIII *San Diego, CA Abstract W548*.

Mannen, H., et al. (2004) Independent mitochondrial origin and historical genetic differentiation in North Eastern Asian cattle. *Molecular Phylogenetics and Evolution* **32**(2): 539–544.

Marcum, J.B. (1974) The freemartin syndrome. *Animal Breeding Abstracts* **42**: 227–242.

Matthews, M.E. and Reed, K.C. (1991) A DNA sequence that is present in both sexes of Artiodactyla is repeated on the Y chromosome of cattle, sheep, and goats. *Cytogenetic and Genome Research* **56**(1): 40–44.

Matthews, M.E. and Reed, K.C. (1992) Sequences from a family of bovine Y-chromosomal repeats. *Genomics* **13**(4): 1267–1273.

Mensher, S.H., Bunch, T.D. and Maciulis, A. (1989) High resolution banded karyotype and idiogram of the goat: a sheep-goat G-banded comparison. *Journal of Heredity* **80**: 150–155.

Mezzelani, A., Zhang, Y., Redaelli, L., Castiglioni, B., Leone, P., Williams, J.L., Toldo, S.S., Wigger, G., Fries, R., Ferretti, L. (1995) Chromosomal localization and molecular characterization of 53 cosmid-derived bovine microsatellites. *Mammalian Genome* **95**(9): 629–635.

Mikaye, Y.-I. and Kaneda, Y. (1988) Chromosomal abnormalities in bulls with unilateral cryptorchidism. *Journal. Faculty of Agriculture (Iwate University)* **19**: 11–19.

Miller, J.R. and Koopman, M. (1990) Isolation and characterization of two male-specific DNA fragments from the bovine genome. *Animal Genetics* **21**: 77–82.

Modi, W.S. and Crews, D. (2005) Sex chromosomes and sex determination in reptiles. *Current Opinion in Genetics and Development* **15**(6): 660–665.

Muller, H.J. (1914) A gene for the fourth chromosome of *Drosophila*. *Journal of Experimental Zoology* **17**: 325–336.

Murphy, W.J., Pearks Wilkerson, A.J., Raudsepp, T., Agarwala, R., Schaffer, A.A., Stanyon, R., Chowdhary, B.P. (2006) Novel gene acquisition on carnivore Y chromosomes. *PLoS Genetics* **2**(3): e43.

Nanda, I., et al. (1999) 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nature Genetics* **21**(3): 258–259.

Ogata, T. and Matsuo, N. (1993) Sex chromosome aberrations and stature: deduction of the principal factors involved in the determination of adult height. *Human Genetics* **91**(6): 551–562.

Ohno, S. (1967) *Sex Chromosomes and Sex-Linked Genes*. Berlin: Springer.

Pearks Wilkerson, A.J., Raudsepp, T., Graves, T., Albracht, D., Warren, W., Chowdhary, B.P., Skow, L.C., Murphy, W.J. (2008) Gene discovery and comparative analysis of X-degenerate genes from the domestic cat Y chromosome. *Genomics* **92**(5): 329–338.

Perez-Pardal, L., et al. (2010a) Multiple paternal origins of domestic cattle revealed by Y-specific interspersed multilocus microsatellites. *Heredity*, www.nature.com/hdy/journal/vaop/ncurrent/full/hdy201030a.html

Perez-Pardal, L., et al. (2010b) Y-specific microsatellites reveal an African subfamily in taurine (Bos taurus) cattle. *Animal Genetics* **41**(232–241).

Perret, J., Shia, Y.C., Fries, R., Vassart, G., Georges, M. (1990) A polymorphic satellite sequence maps to the pericentric region of the bovine Y chromosome. *Genomics* **6**(3): 482–490.

Ponce de León, F.A. (1996) Microdissected chromosome libraries for livestock species. *Archivos de Zootecnia* **45**: 165–174.

Ponce de León, F.A. and Carpio, C. (1995) Identification of the bovine X chromosome pseudoautosomal region. (Abstract). In: *9th North American Colloquium on Domestic Animal Cytogenetics and Gene Mapping.* Texas A&M University. p. 8.

Ponce de León, F.A., Ambady, S., Hawkins, G.A., Kappes, S.M., Bishop, M.D., Robl, J.M., Beattie, C.W. (1996) Development of a bovine X chromosome linkage group and painting probes to assess cattle, sheep, and goat X chromosome segment homologies. *Proceedings of the National Academy of Sciences of the United States of America* **93**(8): 3450–3454.

Raudsepp, T., Santani, A., Wallner, B., Kata, S.R., Ren, C., Zhang, H.B., Womack, J.E., Skow, L.C., Chowdhary, B.P. (2004) A detailed physical map of the horse Y chromosome. *Proceedings of the National Academy of Sciences of the United States of America* **101**(25): 9321–9326.

Robinson, T.J., Harrison, W.R., Ponce de Leon, F.A., Davis, S.K., Elder, F.F. (1998) A molecular cytogenetic analysis of X chromosome repatterning in the Bovidae: transpositions, inversions, and phylogenetic inference. *Cytogenetics and Cell Genetics* **80**(1–4): 179–184.

Rosner, A. and Rinkevich, B. (2007) The DDX3 subfamily of the DEAD box helicases: divergent roles as unveiled by studying different organisms and in vitro assays. *Current Medicinal Chemistry* **14**: 2517–2525.

Roze, D. and Barton, N.H. (2006) The Hill-Robertson effect and the evolution of recombination. *Genetics* **173**(3): 1793–1811.

Ruggiu, M., Speed, R., Taggart, M., McKay, S.J., Kilanowski, F., Saunders, P., Dorin, J., Cooke, H.J. (1997) The mouse Dazla gene encodes a cytoplasmic protein essential for gametogenesis. *Nature* **389**: 73–77.

Sandor, C., Farnir, F., Hansoul, S., Coppieters, W., Meuwissen, T., Georges, M. (2006) Linkage disequilibrium on the bovine X chromosome: characterization and use in quantitative trait locus mapping. *Genetics* **173**(3): 1777–1786.

Saxena, R., et al. (1996) The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nature Genetics* **14**(3): 292–299.

Schmutz, S.M., Moker, J.S., Clark, G., Orr, J.P. (1996) Chromosomal aneuploidy associated with spontaneous abortions and neonatal losses in cattle. *Journal of Veterinary Diagnostic Investigation* **8**: 91–95.

Schwerin, M., Gallagher, D.S., Jr; Miller, J.R., Thomsen, P.D. (1992) Mapping of repetitive bovine DNA sequences on cattle Y chromosomes. *Cytogenetics and Cell Genetics* **61**(3): 189–194.

Seidenspinner, T., Bennewitz, J., Reinhardt, F., Thaller, G. (2009) Need for sharp phenotypes in QTL detection for calving traits in dairy cattle. *Journal of Animal Breeding and Genetics* **126**(6): 455–462.

Shimamura, M., Abe, H., Nikaido, M., Ohshima, K., Okada, N. (1999) Genealogy of families of SINEs in cetaceans and artiodactyls: The presence of a huge superfamily of tRNA(Glu)-derived families of SINEs. *Molecular Biology and Evolution* **16**: 1046–1060.

Skaletsky, H., et al. (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**(6942): 825–837.

Sonstegard, T., Lopez-Corrales, N., Kappes, S., Stone, R., Ambady, S., Ponce de León, F.A., Beattie, C. (1997) An integrated genetic and physical map of the bovine X Chromosome. *Mammalian Genome* **8**(1): 16–20.

Sonstegard, T.S., et al. (2001) Consensus and comprehensive linkage maps of the bovine sex chromosomes. *Animal Genetics* **32**: 115–117.

Stafuzza, N.B., et al. (2009) Comparative RH maps of the river buffalo and bovine Y chromosomes. *Cytogenetics and Genome Research* **126**(1–2): 132–138.

Svensson, E. and Götherström, A. (2008) Temporal fluctuations of Y-chromosomal variation in Bos taurus. *Biology Letters* **4**: 752–754.

Swartz, H.A. and Vogt, D.W. (1983) Chromosome abnormalities as a cause of reproductive inefficiency in heifers. *Journal of Heredity* **74**: 320–324.

Switonski, M. and Stranzinger, G. (1998) Studies of synaptonemal complexes in farm animals: A review. *Journal of Heredity* **89**(6): 473–480.

Teng, Y.-N., Lin, Y.-M., Lin, Y.-H., Tsao, S.-Y., Hsu, C.-C., Lin, S.-J., Tsai, W.-C., Kuo, P.-L. (2002) Association of a single-nucleotide polymorphism of the deleted-in-azoospermia-like gene with susceptibility to spermatogenic failure. *Journal of Clinical Endocrinology and Metabolism* **87**: 5258–5264.

Tilford, C.A., et al. (2001) A physical map of the human Y chromosome. *Nature* **409**(6822): 943–945.

Troy, C.S., MacHugh, D.E., Bailey, J.F., Magee, D.A., Loftus, R.T., Cunningham, P., Chamberlain, A.T., Sykes, B.C., Bradley, D.G. (2001) Genetic evidence for Near-Eastern origins of European cattle. *Nature* **410**(6832): 1088–1091.

Vaiman, D., et al. (1994) A set of 99 cattle microsatellites: characterization, synteny mapping, and polymorphism. *Mammalian Genome* **5**: 288–297.

Van Laere, A.-S., Coppieters, W., Georges, M. (2008) Characterization of the bovine pseudoautosomal boundary: Documenting the evolutionary history of mammalian sex chromosomes. *Genome Research* **18**(12): 1884–1895.

Verkaar, E.L., Zijlstra, C., van ‘t Veld, E.M., Boutaga, K., van Boxtel, D.C., Lenstra, J.A. (2004) Organization and concerted evolution of the ampliconic Y-chromosomal TSPY genes from cattle. *Genomics* **84**(3): 468–474.

Veyrunes, F., et al. (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Research* **18**(6): 965–973.

Vilkki, J., Sandholm, J., Kostia, S., Varvio, S.L. (1995) Four SINE-associated polymorphic bovine microsatellites (HEL23-HEL26). *Animal Genetics* **26**(3): 206.

Vogel, T., Dechend, F., Manz, E., Jung, C., Jakubiczka, S., Fehr, S., Schmidtke, J., Schnieders, F. (1997a). Organization and expression of bovine TSPY. *Mammalian Genome* **8**(7): 491–496.

Vogel, T., Borgmann, S., Dechend, F., Hecht, W., Schmidtke, J. (1997b) Conserved Y-chromosomal location of TSPY in Bovidae. *Chromosome Research* **5**(3): 182–185.

Vong, Q.P., Li, Y., Lau, Y.F., Dym, M., Rennert, O.M., Chan, W.Y. (2006) Structural characterization and expression studies of dby and its homologs in the mouse. *Journal of Andrology* **27**: 653–661.

Wang, A.-H., Yasue, H., Li, L., Tagashima, M., Ponce de León, F.A., Liu, W.-S (2008) Molecular characterization of the bovine chromodomain Y-like (CDYL) genes. *Animal Genetics* **39**: 207–216.

Ward, T.J., Skow, L.C., Gallagher, D.S., Schnabel, R.D., Nall, C.A., Kolenda, C.E., Davis, S.K., Taylor, J.F., Derr, J.N. (2001) Differential introgression of uniparentally inherited markers in bison populations with hybrid ancestries. *Animal Genetics* **32**(2): 89–91.

Waterston, R.H., et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915): 520–562.

Weikard, R., Kuhn, C., Brunner, R.M., Roschlau, D., Pitra, C., Laurent, P., Schwerin, M. (2001) Sex determination in cattle based on simultaneous amplification of a new male-specific DNA sequence and an autosomal locus using the same primers. *Molecular Reproduction and Development* **60**(1): 13–19

Wilson, M.A. and Makova, K.D. (2009) Genomic analyses of sex chromosome evolution. *Annual Review of Genomics and Human Genetics* **10**: 333–354.

Womack, J.E., Johnson, J.S., Owens, E.K., Rexroad III, C.E., Schlipfer, J., Yang, Y. (1997) A whole-genome radiation hybrid panel for bovine gene mapping. *Mammalian Genome* **8**: 854–856.

Wurster, D.H. and Benirschke, K. (1968) Chromosome studies in superfamily Bovidea. *Chromosoma* **25**: 152–171.

Xiao, C., Tsuchiya, K., Sutou, S. (1998) Cloning and mapping of bovine ZFX gene to the long arm of the X-Chr (Xq34) and homologous mapping of ZFY gene to the distal region of the short arm of the bovine (Yp13), ovine (Yp12-p13), and caprine (Yp12-p13) Y chromosome. *Mammalian Genome* **9**(2): 125–130.

Yang, Y. (2009) Isolation and characterization of male fertility-related genes on the bovine Y chromosome. Society for the Study of Reproduction, Annual Meeting, Pittsburg, PA, 18–22.

Yuan, Z.A., Collier, P.M., Rosenbloom, J., Gibson, C.W. (1996) Analysis of amelogenin mRNA during bovine tooth development. *Archives of Oral Biology* **41**: 205–213.

Zhang, C., De Koning, D.J., Hernandez-Sanchez, J., Haley, C.S., Williams, J.L., Wiener, P. (2004) Mapping of multiple quantitative trait loci affecting bovine spongiformencephalopathy. *Genetics* **167**(4): 1863–1872.

# Chapter 8
# Cattle Comparative Genomics and Chromosomal Evolution

*Denis M. Larkin*

## Introduction

The cattle genome sequencing and assembly was completed in 2009. This was the first whole genome assembly of a species from the order *Cetartiodactyla,* an order distinct from the human and mouse lineage. Cetartiodactyls appeared about 60 million years ago (Springer et al. 2003) and show a unique variety of adaptive features. For example, cetartiodactyls are the only mammals adapted to live in the ocean (whales and dolphins). In addition, the Tibetan antelope lives in very high altitudes and is adapted to deal with high hypoxia. Other cetartiodactyls demonstrate distinct features related to genome organization, for example, Indian Muntjac has the lowest number of chromosomes among all karyotyped mammals (Tsipouri et al. 2008). In addition, cetartiodactyls have the highest number of economically important and domesticated species due to their unique ability to produce energy-dense products, such as fat, milk, and meat (Bovine Genome Sequencing and Analysis Consortium 2009).

Among the livestock species, cattle have one of the best and most detailed set of comparative maps available, mostly due to its economical importance. Whereas somatic cell hybrid maps and cross-species chromosome painting with human and other species DNA probes have provided an important but patched correspondence between the cattle, human, mouse, and pig genomes (Womack and Moll 1986; Hayes 1995; Chowdhary et al. 1998; Schmitz et al. 1998), the real breakthrough in cattle comparative studies started with the introduction of high-resolution ordered radiation hybrid (RH) maps (Womack et al. 1997; Band et al. 2000; Everts-van der Wind et al. 2004, 2005) and the COMPASS-based approach of marker selection for mapping (Rebeiz and Lewin 2000; Larkin et al. 2003). High-quality physical maps are required to order the whole-genome sequence scaffolds on chromosomes (Lewin et al. 2009). The cattle genome was assembled to chromosomes using the IL-TX (Illinois-Texas) RH map at the Baylor College of Medicine (Btau_4.0) and using the British Columbia Cancer Research Centre (BCCRC) integrated fingerprint map at the University of Maryland (UMD_3.1) (Bovine Genome Sequencing and Analysis Consortium 2009; Zimin et al. 2009). These assemblies contain the same raw sequences but differ in N50 contig size, number of sequence reads placed on chromosomes, and also contain a

*Bovine Genomics*, First Edition. Edited by James E. Womack.

lot of small and several large-scale chromosomal structural differences. These regions represent intervals that have to be explored and fixed during further polishing of the assemblies.

Sequencing of the cattle genome has been achieved as the result of collaborative effort between at least six countries. An assembly of the genome became possible due to numerous projects started in the 1970s to understand the organization of cattle chromosomes (Heuertz and Hors-Cayla 1978; Womack and Moll 1986), to generate microsatellites (Barendse et al. 1997) for gene mapping (Itoh et al. 2003), and to construct high-resolution physical (Everts-van der Wind et al. 2005; Snelling et al. 2007, 2010) and linkage maps (Ihara et al. 2004). With the availability of the genome sequence and accurate assembly, it became feasible to perform the analysis of the genome at a level that no one could imagine until very recently (Bovine Genome Sequencing and Analysis Consortium 2009).

The comparative analysis shows that the cattle genome is an invaluable resource for studying mammalian genome evolution. Unique genome features in cattle developed in course of speciation and adaptation are reflected by gene mutations, sequence losses, duplications, and repositions due to multiple chromosomal rearrangements that distinguish the cattle genome from other mammals and a putative mammalian ancestor (Murphy et al. 2005; Bovine Genome Sequencing and Analysis Consortium 2009; Larkin et al. 2009). On the other hand, when compared to other mapped mammalian genomes, the cattle genome in some chromosome regions still maintain the ancestral organization, allowing for the detection of the evolutionary events that occurred in the course of genome evolution in other species (Murphy et al. 2005).

## Chromosomal Rearrangements and Genome Evolution

The whole-genome ordered comparative maps identify approximately 201–211 large blocks of homologous synteny (HSBs) between the human and cattle chromosomes (Everts-van der Wind et al. 2005; Bovine Genome Sequencing and Analysis Consortium 2009). Comparable numbers are reported for the comparison of completely sequenced cattle and human genomes with the highest number of 268 HSBs being reported by Zimin et al. (2009). Comparison of the cattle genome with the genomes of other ferungulate species (dog and pig) led to the identification of 124 chromosomal evolutionary breakpoint regions (EBRs) in the cattle lineage, of which 100 are putatively cattle/ruminant specific and 24 are shared by pig and cattle, and could be Cetartiodactyla/Artiodactyla specific. There are nine additional breakpoint regions that are shared by the cattle, pig, and dog, and may represent the evolutionary events in the ancestral Ferungulate lineage (Table 8.1). Interestingly, cattle chromosome 16 (BTA16) is populated with four ferungulate-specific rearrangements, suggesting that those originated in the common ancestor of Carnivora and Artiodactyla and the ancestral organization of this genomic interval is still preserved in human and other euarchontoglires (Bovine Genome Sequencing and Analysis Consortium 2009). Such a low number of superordinal chromosomal rearrangements is in agreement with the previous observation of a low rate of chromosomal rearrangements at the early stages of chromosomal evolution in eutherian mammals. This rate was estimated as ~0.1–0.2 large-scale rearrangement per million years (Murphy et al.

**Table 8.1** Classification of evolutionary breakpoint regions in ferungulate genomes.

| EBR classification | Number |
|---|---|
| Superordinal | |
| Cetartiodactyla-Carnivora | 9 |
| Order-specific | |
| Cetartiodactyla | 24 |
| Lineage-specific | |
| Cattle | 100 |
| Pig | 77 |
| Dog | 82 |

*Source:* Bovine Genome Sequencing and Analysis Consortium 2009.

2005) for both Ferungulate and Euarchontoglires lineages. This rate has significantly increased within different mammalian orders after the C-T boundary about 65 million years ago with the highest rate observed in murid rodents (~141 large-scale order-specific rearrangements).

It was demonstrated that the EBRs are associated with the positions of segmentally duplicated sequences (SDs) in the human genome (Bailey et al. 2004; Murphy et al. 2005) and that most likely, SDs promote EBRs causing nonallelic homologous recombination (NAHR) between the chromosomal intervals containing similar SDs. The cattle genome comparative analysis confirms this observation, showing that 10-kb sequence intervals overlapping with the cattle/ruminant EBRs contain approximately seven times more segmentally duplicated nucleotides than the other intervals of the cattle genome. Strikingly, artiodactyl-specific EBRs (shared by the cattle and pig genomes) contained ~14 times more segmental duplications than other genomic regions suggesting that there are hotspots for the insertion of SDs in artiodactyl genomes (Table 8.2) (Bovine Genome Sequencing and Analysis Consortium 2009).

Another possible source of sequence elements that could cause NAHR and lead to the formation of chromosomal rearrangements are recent repetitive elements in the genome that still hold a high sequence similarity among different copies and are present in the genome in hundreds or thousands of copies. Indeed, when density of

**Table 8.2** Distribution of segmental duplications in cattle chromosomes.

| Region | SD in 10-kbp intervals[a] |
|---|---|
| Cattle-specific EBRs | 11.7% |
| Cetartiodactyl-specific EBRs | 23.0% |
| Other intervals | 1.7% |

SD, segmentally duplicated sequences; EBR, evolutionary breakpoint region.
[a] Number of bases from SDs was calculated for all 10-kb intervals in the cattle genome and the percentage of SDs bases was calculated for the EBRs and other intervals (Bovine Genome Sequencing and Analysis Consortium 2009).

the lineage-specific retrotransposable elements was compared in the EBRs and in the rest of the cattle genome, a strong positive correlation was observed between some lineage-specific LINE-L1 and LINE-RTE elements and cattle EBRs. In the dog and mouse genomes, EBRs were found enriched for LTR-ERV1 elements that were active in each of these lineages. Another group of repeats, tRNA$^{Glu}$-derived SINEs originating in the common ancestor of all artiodactyls had a higher-than-expected density in the artiodactyl-specific breakpoint regions, but not in the cattle-specific breakpoints. This suggests that in mammals, chromosomal rearrangements tend to occur in the regions with a high density of repetitive elements that are still active and, therefore, have a high sequence similarity between different copies required for an NAHR. In confirmation to this conclusion, a negative correlation between the density of old retrotransposable elements (such as LINE-L2 and some SINEs) and EBRs in all mammalian genomes was noticed suggesting that an active insertion of new mobile elements either destroys copies of old repetitive elements or forms new regions in the genome (Figure 8.1). Later these new intervals could be used as templates for NAHR and form material for chromosome structural changes. During evolution different

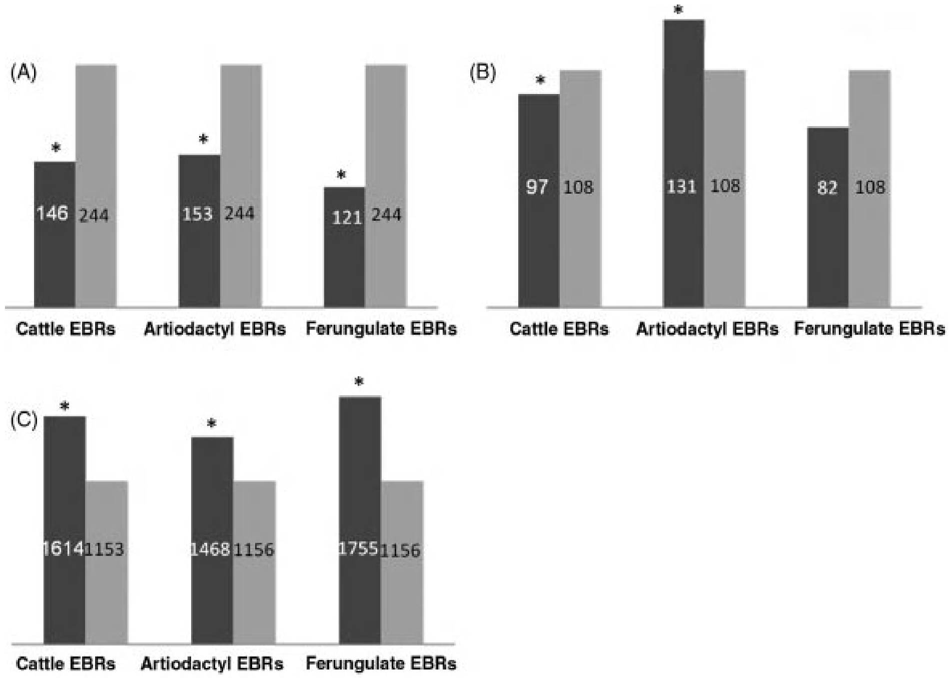


**Figure 8.1** Average densities of bases from (A) LINE-L2, (B) SINE-tRNA-GLU and (C) LINE-L1 retrotransposable elements in 10-kb intervals of the cattle genome overlapping with cattle, artiodactyl, and ferungulate evolutionary breakpoint regions (dark gray) compared with densities of the elements in all other 10-kb intervals of the cattle genome (light gray). Asterisks (∗) indicate statistically significant differences (FDR $< 0.05$). The data suggest that the evolutionary breakpoint regions tend to occur in the regions with low density of "old" (e.g., LINE-L2) and high density of "new" (e.g., LINE-L1) elements (Bovine Genome Sequencing and Analysis Consortium 2009).

copies of the same mobile element will accumulate different mutations making the sequences not suitable for NAHR anymore. These observations are in agreement with a recently proposed theory that chromosomal rearrangements in mammalian genomes are occurring in the fragile regions that are subject to birth and death processes in different genomes (Alekseyev and Pevzner 2010).

## Chromosomal Rearrangements and Adaptation

EBRs may be connected to speciation not only due to the reproductive isolation they may be causing in populations (Brown and O'Neill 2010), but also due to the changes in gene regulation and networks they may cause by moving genes to a new regulatory environment (De et al. 2009), or causing gene duplications and deletions. Analysis of the cattle and other amniote genomes provides a support for the hypothesis of adaptive value of EBRs. For example, Everts-van der Wind et al. (2004) reported that the evolutionary breakpoints between the cattle and human genomes are enriched for genes, at least in the human genome. Lately, this observation was confirmed by the multispecies genome comparisons (Murphy et al. 2005; Larkin et al. 2009). Larkin et al. (2009) demonstrated that the amniote-specific EBRs are significantly enriched for genes involved in the *organism's response to external stimuli* (Larkin et al. 2009).

A cattle-specific EBR was found responsible for the formation of a new bidirectional promoter that controls the expression of the *CYB5R4* gene (Piontkivska et al. 2009). This gene is involved in diabetes in humans and is a good candidate to be involved in the evolution of energy flow in cattle (Bovine Genome Sequencing and Analysis Consortium 2009). Another striking connection between the positions of the EBRs and gene family expansions in the cattle genome is an expansion and reorganization of a $\beta$-defensin gene cluster that encode antimicrobial peptides in BTA27. An expanded $\beta$-defensin cluster is found with an artiodactyl-specific EBR and large segmental duplication. Other genes that are overrepresented in the cattle genome compared to human and mouse include mature cathelicidin peptides, interferon genes, and other genes involved in adaptive immune responses in cattle, suggesting that these adaptive changes could be connected to the increased amount of microorganisms present in the rumen.

In general, segmental duplications in the cattle genome are enriched for the genes involved in reproduction. These families encode the intercellular signaling proteins, pregnancy-associated glycoproteins (on BTA29), trophoblast Kunitz domain proteins (on BTA13), and interferon tau (*IFNT*) (on BTA8). In addition, it was demonstrated that in cattle the gene families encoding milk proteins have been significantly rearranged compared to other mammals. One example is histatherin *(HSTN)*, the gene from casein cluster on BTA6. In cattle, *HSTN* was moved to a regulatory element important for $\beta$-casein expression, and as a probable consequence, *HSTN* is regulated like the casein genes during the lactation cycle (Bovine Genome Sequencing and Analysis Consortium 2009).

Overall, the milk proteome studies (Lemay et al. 2009) have identified 197 unique genes expressed in milk. This gene set was shown to be conserved in all mammals including cattle. Of all milk genes that are shared by cattle and platypus, only ten are not present in other mammals suggesting that the system of milk production is very stable.

Also, in the bovine lineage, milk-related proteins were shown to have significantly lower rate of nonsynonymous ($d_N$) to synonymous ($d_S$) substitutions than other genes in the cattle genome ($P < 0.05$), suggesting a stronger selective constrain for these genes in cattle (Lemay et al. 2009). This might indicate that structural changes in chromosomes, gene duplications or losses, and changes in noncoding regulatory sequences could play the major role in evolution of the milk and mammary gland proteomics in cattle and other mammals. Indeed, in addition to the examples mentioned previously, the cattle serum amyloid A (*SAA*) gene cluster arose from a segmental duplication associated with a cattle-specific EBR, resulting in three mammary gland-expressed *SAA3*-like genes on BTA29 and BTA15 (Bovine Genome Sequencing and Analysis Consortium 2009).

Comparison of the orthologous genes in cattle, dog, mouse, rat, human, opossum, and platypus has identified 14,345 orthologous groups of which 12,592 are single copy orthologs present in human, cattle or dog, mouse or rat, opossum, or platypus. A total of 1,217 groups were identified as placental mammal-specific. Around 1,000 gene groups shared by cattle or dog and mouse or rat were not identified in the human genome, suggesting their elimination in evolution or misassembly in the human genome. Many of dog/cattle-specific groups are G-protein coupled receptors (Bovine Genome Sequencing and Analysis Consortium 2009). In cattle and dog genomes, there are 147 orthologous groups that are not present in other mammals, which is less than 1,112 groups specific to rodent genomes. This difference can be explained by a higher mutation rate in the murid rodent genes than in human, dog, and cattle genomes. However, the cattle genome contains 71 genes that were subject to positive selection (based on $d_N/d_S$ ratio). Of these, ten are related to the immune system (Bovine Genome Sequencing and Analysis Consortium 2009). Comparison of the cattle and human metabolic pathways has revealed a high degree of conservation. However, five human metabolic genes were deleted or highly diverged in the cattle genome. Based on the functions of the genes in the human genome, their deletion in the cattle genome could be adaptive and impact fatty acid metabolism, the mevalonate pathway, detoxification, and pyrimidine metabolism.

In conclusion, the cattle genome demonstrates a large variety of adaptive features connected to cattle's unique adaptation to environment. Some of them were gained in the ancestral ferungulate or cetartiodactyl lineages, and the comparative genomics provides us with a unique opportunity to detect and date them without having access to ancient DNA and ancestral genomes. Other features, such as lineage-specific gene deletions or births were formed more recently in the lineage leading to all ruminants or cattle. With the availability of other ruminant genomes in the nearest future it will become feasible to date these events as well. Many of them are directly connected to the phenotypic features that make cattle an important livestock species.

## References

Alekseyev, M.A. and Pevzner, P.A. (2010) Comparative genomics reveals birth and death of fragile regions in mammalian evolution. *Genome Biology* **11**: R117.

Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D., Eichler, E.E. (2004) Hotspots of mammalian chromosomal evolution. *Genome Biology* **5**: R23.

Band, M.R., et al. (2000) An ordered comparative map of the cattle and human genomes. *Genome Research* **10**: 1359–1368.

Barendse, W., et al. (1997) A medium-density genetic linkage map of the bovine genome. *Mammalian Genome* **8**: 21–28.

Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522–528.

Brown, J.D. and O'Neill, R.J. (2010) Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *Annual Review of Genomics and Human Genetics* **11**: 291–316.

Chowdhary, B.P., Raudsepp, T., Fronicke, L., Scherthan, H. (1998) Emerging patterns of comparative genome organization in some mammalian species as revealed by Zoo-FISH. *Genome Research* **8**: 577–589.

De, S., Teichmann, S.A., Babu, M.M. (2009) The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. *Genome Research* **19**: 785–794.

Everts-van der Wind, A., et al. (2004) A 1463 gene cattle-human comparative map with anchor points defined by human genome sequence coordinates. *Genome Research* **14**: 1424–1437.

Everts-van der Wind, A., et al. (2005) A high-resolution whole-genome cattle-human comparative map reveals details of mammalian chromosome evolution. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 18526–18531.

Hayes, H. (1995) Chromosome painting with human chromosome-specific DNA libraries reveals the extent and distribution of conserved segments in bovine chromosomes. *Cytogenetics and Cell Genetics* **71**: 168–174.

Heuertz, S. and Hors-Cayla, M.C. (1978) Bovine chromosome mapping with the cell hybridization technic. Localization on the x chromosome of glucose-6-phosphate dehydrogenase, phosphoglycerate kinase, alpha-galactosidase A and hypoxanthine phosphoribosyltransferase. *Annals of Human Genetics* **21**: 197–202.

Ihara, N., et al. (2004) A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Research* **14**: 1987–1998.

Itoh, T., Takasuga, A., Watanabe, T., Sugimoto, Y. (2003) Mapping of 1400 expressed sequence tags in the bovine genome using a somatic cell hybrid panel. *Animal Genetics* **34**: 362–370.

Larkin, D.M., et al. (2003) A cattle-human comparative map built with cattle BAC-ends and human genome sequence. *Genome Research* **13**: 1966–1972.

Larkin, D.M., et al. (2009) Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Research* **19**: 770–777.

Lemay, D.G., et al. (2009) The bovine lactation genome: insights into the evolution of mammalian milk. *Genome Biology* **10**: R43.

Lewin, H.A., Larkin, D.M., Pontius, J., O'Brien, S.J. (2009) Every genome sequence needs a good map. *Genome Research* **19**: 1925–1928.

Murphy, W.J., et al. (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* **309**: 613–617.

Piontkivska, H., et al. (2009) Cross-species mapping of bidirectional promoters enables prediction of unannotated 5′ UTRs and identification of species-specific transcripts. *BMC Genomics* **10**: 189.

Rebeiz, M. and Lewin, H.A. (2000) Compass of 47,787 cattle ESTs. *Animal Biotechnology* **11**: 75–241.

Schmitz, A., et al. (1998) Comparative karyotype of pig and cattle using whole chromosome painting probes. *Hereditas* **128**: 257–263.

Snelling, W.M., et al. (2007) A physical map of the bovine genome. *Genome Biology* **8**: R165.

Snelling, W.M., et al. (2010) Genome-wide association study of growth in crossbred beef cattle. *Journal of Animal Science* **88**: 837–848.

Springer, M.S., Murphy, W.J., Eizirik, E., O'Brien, S.J. (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 1056–1061.

Tsipouri, V., et al. (2008) Comparative sequence analyses reveal sites of ancestral chromosomal fusions in the Indian muntjac genome. *Genome Biology* **9**: R155.

Womack, J.E. and Moll, Y.D. (1986) Gene map of the cow: conservation of linkage with mouse and man. *Journal of Heredity* **77**: 2–7.

Womack, J.E., et al. (1997) A whole-genome radiation hybrid panel for bovine gene mapping. *Mammalian Genome* **8**: 854–856.

Zimin, A.V., et al. (2009) A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology* **10**: R42.

# Chapter 9
# Sequencing the Bovine Genome

*Kim C. Worley and Richard A. Gibbs*

## Introduction

The large project of sequencing and assembly of the bovine genome leveraged the many resources available and provided the pivotal substrate for research in the future. The assembly method combines the bacterial artificial chromosome (BAC) plus whole genome shotgun (WGS) local assembly used for the rat and sea urchin with the WGS-only assembly used for many other animal genomes, including the rhesus macaque.

## Background on Genome Assembly

Mammalian genomes are sequences typically about three billion base pairs (3 gigabases or Gb) in total length with the average individual chromosome about one hundred million base pairs (100 Mb) long. Sequencing technologies produce sequences with read lengths of between 25 and 1000 base pairs. Genome assembly is the process of combining many short sequences (reads) into a long consensus sequence. This process is always a compromise, since assembly methods are selected that can be applied in a uniform manner to the entire genome. The sequences can be aggressively merged, which can create false joins in some cases. Or, the sequences can be conservatively merged, leaving some sequences unjoined but creating fewer false joins.

Artificial or random sequences are easier to assemble correctly than real genomic sequence due to the nonrandom nature of real genomic sequence that contains repeats, duplications, and polymorphisms. The major methods for genome assembly include the hierarchical approach, the WGS approach, and the combined approach. The hierarchical approach, where BACs are isolated, mapped to the genome, and individually sequenced, was used for the human genome (Lander et al. 2001). The advantage of the hierarchical approach is that the individual BACs contain DNA from a single haplotype, and the assembly within a BAC avoids conflicts due to polymorphisms and is more contiguous and correct for a given amount of sequence coverage. The whole genome shotgun or WGS method reduces the cost of genome sequencing by avoiding the BAC cloning and library construction costs, and avoids biased representation in sequences that do not clone in BACs. These advantages come at a cost,

*Bovine Genomics*, First Edition. Edited by James E. Womack.

however, since the WGS method has greater difficulties dealing with the genomic features including repeats and polymorphisms that make assembly of real genomic sequence more difficult than random sequence assembly. A number of mammalian genomes have used the WGS method, including the first mouse genome, the macaque (Gibbs et al. 2007), dog (Kirkness et al. 2003), opossum (Mikkelsen et al. 2007), platypus (Warren et al. 2008), chimpanzee (CSAC 2005), and the low coverage genome sequences including cat (Pontius et al. 2007).

## Bovine Sequencing Strategy, DNA Sources

The bovine genome was assembled at the Baylor College of Medicine Human Genome Sequencing Center. The sequence assembly strategy was a hybrid of the WGS and the hierarchical BAC clone approaches, and used methods (Liu et al. 2009) similar to those used to assemble the rat (RGSC 2004) and sea urchin genomes (SUGSC 2006). As with the sea urchin (SUGS Consortium 2006), to reduce the cost, many of the BACs for the bovine project were sequenced in groups or pools, rather than individually.

In addition to using the advantage of the separation of a small data set for local assembly provided by a BAC-based assembly, the bovine assembly took advantage of the local assembly further by tuning the assembly parameters for each BAC to address local differences in sequence characteristics (e.g., repeat content and degree of polymorphism compared to the WGS sequence) to produce the best assembly within each enriched BAC (eBAC).

The DNA for the small insert WGS libraries was isolated from leukocytes from a Hereford cow (L1 Dominette 01449; American Hereford Association registration number 42190680, provided by Dr. Timothy Smith, US Meat Animal Research Center, Clay Center, NE; 30% inbreeding coefficient). The DNA for the BAC library was isolated from her sire (L1 Domino 99375; 31% inbreeding coefficient).

The sequence production took place over an extended period of time. During that time, the individual sequence reads were available in the international sequence databases trace archives. The project was carried out in accordance with policies from the NHGRI (National Human Genome Research Institute) concerning Community Resource Projects (http://www.genome.gov/page.cfm?pageID=10506537).

Table 9.1 lists the variety of sequence data used in the bovine genome project. The majority of the sequencing was performed using the Sanger sequencing method

**Table 9.1** Read data summary.

| | Number of reads | | |
|---|---|---|---|
| Insert size (kb) | Sequenced | Trimmed | Assembled Btau_4.0 |
| 2–6 | 26,978,021 | 23,094,990 | 18,741,158 |
| 200 | 207,901 | 168,766 | 113,857 |
| 2–6 | 6,422,870 | 5,107,693 | 4,279,498 |
| 2–6 | 5,377,386 | 5,116,216 | 2,917,875 |

(Sanger et al. 1977) produced on the ABI 3730 capillary sequencing machines. More than half of the BAC shotgun sequence was produced from individual BAC libraries and the remainder was produced from pools of BACs sequenced together. The initial WGS assembly (Btau_2.0) and the BAC plus WGS assemblies (Btau_3.1, Btau_4.0) used all Sanger sequences with average trimmed read length of 730 bp. The later Btau_4.5 assembly incorporated SOLiD sequence data from 25 bp reads.

## The Different Genome Assemblies

A number of assembly versions were produced with different data and methods and used for different analyses as the project proceeded. Table 9.2 lists the assemblies and a brief description of each. The initial assemblies of the small insert WGS data (Btau_1.0 and Btau_2.0) provided early access to preliminary assembled sequence. Later whole genome assemblies (Btau_3.1, Btau_4.0, and Btau_4.5) incorporated BAC sequences as well as WGS sequences. Btau_3.1 combined individual BAC sequences assembled as individual BACs with overlapping WGS sequence, with sequences from the Btau_2.0 WGS-only assembly. Sequences were placed in Btau_3.1 using preliminary physical mapping data (Liu et al. 2009), and Btau_3.1 was used for most of the genome analyses (BGSC 2009). Subsequently, Btau_4.0 was constructed by placing the sequence using different mapping information (Liu et al. 2009), though most sequence contigs remain unchanged between Btau_3.1 and Btau_4.0. Btau_4.0 was used for many of the global analyses (BGSC 2009), including analyses of the GC content, repeats, homologous synteny blocks, and segmental duplications. The Btau_4.5, created after the publication of the genome sequence, used additional sequence from the SOLiD sequencing platform to scaffold sequences and incorporate more of the small, WGS-only contigs into scaffolds. The Btau_4.2 and Btau_4.6 assembly versions incorporated the available high-quality finished BAC sequences into the genome assembly versions Btau_4.0 and Btau_4.5, respectively, replacing the corresponding whole genome assembly sequences.

In all cases, the assembled sequences are found in contiguous sequence pieces termed contigs. Individual contigs may be linked by mate pair or other information into scaffolds with gaps of estimated sizes separating the contigs within the scaffold.

**Table 9.2** Assembly versions.

| Assembly version | Description |
|---|---|
| Btau_1.0 | Low-coverage WGS-only assembly provided early access |
| Btau_2.0 | WGS-only assembly provided early access |
| Btau_3.1 | Combined BAC and WGS assembly used for bulk of genome analyses |
| Btau_4.0 | Incorporated additional sequence, positioned with different map information |
| Btau_4.2 | Finished BACs replaced corresponding draft sequences from Btau_4.0 |
| Btau_4.5 | SOLiD data incorporated for scaffolding, added contigs |
| Btau_4.6 | Finished BACs replaced corresponding draft sequences from Btau_4.5 |

WGS, whole genome shotgun; BAC, bacterial artificial chromosome.

The order and orientation of the contigs and scaffolds in the genome is described in an AGP file. The consensus sequence for the genome is presented in sequence files where the contigs that are placed in the genome but lack information about correct orientation are arbitrarily oriented. For this reason, the consensus sequence may appear to have errors in orientation that would be ambiguous in the AGP file representation of the genome assembly.

## Bovine Genome Assembly Methods

The genome assembly used the Atlas genome assembly system (Havlak et al. 2004), the details are provided in Liu et al. (2009). The assembly process consisted of multiple phases, outlined in Figure 9.1. The boxes in Figure 9.1 indicate the major parts of the assembly process. Box 1 surrounds the WGS data and the two WGS assemblies. Box 2 surrounds the BAC clone insert sequence generation, the BAC-based assembly steps. Box 3 encloses the assembly steps involved with combining the BAC-based and WGS assemblies. Boxes 4 and 5 indicate the mapping processes. Box 6 in Figure 9.1 indicates the additional sequencing using the SOLiD technology and assembly improvements performed after the genome assembly publication.

As outlined within Box 2 in Figure 9.1, the BAC data was generated from individual BAC clone libraries (on the right) or libraries from pools of arrayed BAC clones (on the left). The majority of the Sanger sequence data was generated from WGS libraries. The sequence data is summarized in Table 9.1.

## Description of the WGS-Only Assembly

Two assembly versions were prepared using only WGS reads from small insert plasmid libraries (Figure 9.1, Box 1). These WGS assemblies did not include sequence from the BAC clones. Btau_1.0 (September 2004) was produced using small insert plasmid clones and BAC end sequences (BES) with about 3×WGS coverage. Btau_2.0 (June 2005) was produced with about 6.2×WGS coverage.

The Btau_2.0 assembly release was produced by assembling WGS reads with the Atlas genome assembly system (Havlak et al. 2004). Several WGS libraries, with inserts of 2–4 kb, and 4–6 kb, were used to produce the data. About 23 million reads were assembled, representing about 17.7 Gb of sequence and about 6.2× coverage of the (clonable) bovine genome. BES were used for scaffolding.

## Description of the BAC-Based Assembly

Individual BAC sequences were assembled with Phrap (www.phrap.org; de la Bastide and McCombie 2007) as outlined in the Box 2 in Figure 9.1, first with just the BAC generated sequences, then in combination with the WGS reads that overlap the BAC as an eBAC. 19,667 BAC projects (12,549 individual sequenced clones and 7118 clones from BAC pools) were sequenced and assembled.

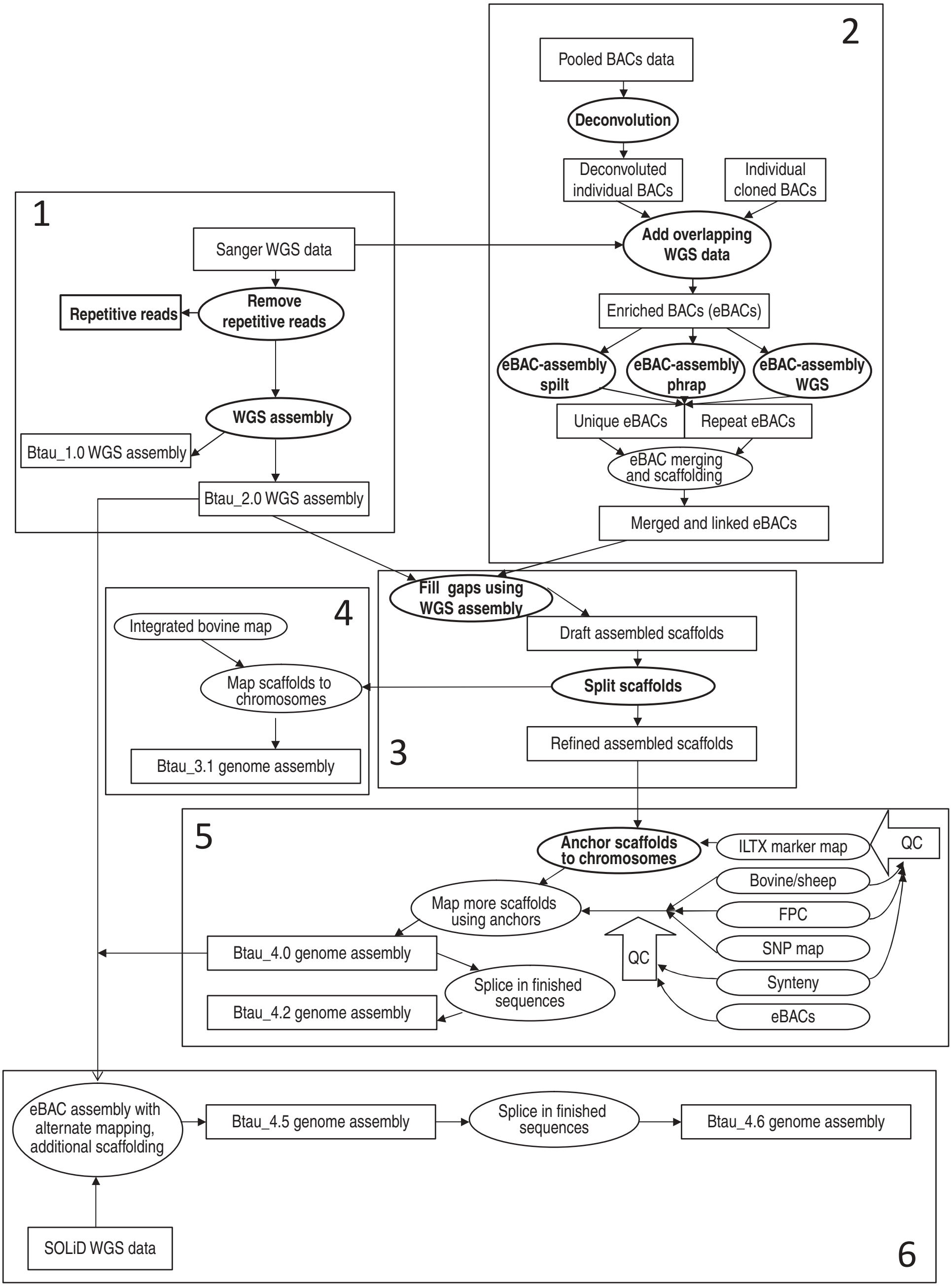


**Figure 9.1** Sequence assembly.

Three assembly methods were applied to each individual eBAC using the BAC reads and the WGS reads that overlapped with the BAC reads:

1. PHRAP: eBAC assemblies were produced by Phrap (www.phrap.org) using either raw or trimmed reads. The better assembly result from the two read sets was determined based on contig and scaffold size statistics.
2. SPLIT: The positions of potential misjoins in the contigs generated from method 1 were detected when a region in a contig had a lack of clone coverage and contained conflicting clone links with the other contigs. The reads in this region were removed and Phrap (www.phrap.org) assembly was performed again to split the original contig.
3. WGS: Each individual eBAC was treated as a mini-genome and the standard ATLAS-WGS assembly procedure was applied, including detecting overlaps among the reads, filtering conflicting overlaps based on overlap patterns, clustering reads into bins based on their overlaps, and PHRAP assembly in each bin.

These three assembly methods were implemented as new components that have been added to the Atlas assembly system.

For any BAC, the assembly using one of the aforementioned three methods was selected (based on the sequence alignment of this BAC against the BACs that overlapped with it) and used in the next step of BAC merging (Liu et al. 2009). Briefly, the combined read set assemblies for each BAC were refined by contig merging and scaffolding based on clone-end mate-pair constraints. Sets of overlapping BAC clones were identified and merged based on shared WGS reads and sequence overlaps of individual BAC assemblies. The merged BAC assemblies were further scaffolded using information from mate pairs, BAC clone vector locations, and BAC assembly sequences.

## Combining BAC and WGS Assemblies and Mapping to Chromosomes

Contigs from the Btau_2.0 WGS assembly were used to fill in the gaps in the BAC-based assembly (e.g., those due to gaps in the BAC tiling path), creating the combined assembly, Btau_3.1 and the initial assembly scaffolds for later assemblies (Box 3, Figure 9.1).

The major difference between assembly versions Btau_4.0 and Btau_3.1 is the position of sequence scaffolds on the chromosomes (detailed later). The Btau_3.1 scaffolds were placed in the genome using an early version of the Integrated Bovine Map (Snelling et al. 2007) that merged data from several independent maps (Box 4, Figure 9.1). Details about the methods used to address inconsistencies between the genome scaffold marker order and the Integrated Bovine Map marker order are described (Liu et al. 2009). In contrast to Btau_3.1, the Btau_4.0 used the available mapping information in a staged approach, relying upon the more consistent data instead of all of the available data (Box 5, Figure 9.1). This assembly used the ILTX (Everts-van der Wind et al. 2005) and BAC fingerprint contig (Snelling et al. 2007) maps to place contigs and split scaffolds based upon consistent bovine and ovine BES data (Dalrymple et al. 2007). These methods resulted in more accurately assembled chromosomes with 90% of the total genome sequence placed on the 29 autosomes and X chromosome and validated Liu et al. (2009).

## Description of Mapping and Placement for Btau_3.1

The assembled contigs and scaffolds of the Btau_3.1 assembly were placed on the chromosomes using an early version of the Integrated Bovine Map (Snelling et al. 2007) that represents merged data from several independent maps. A total of 21,971 bovine markers were compared to the Btau_3.1 scaffolds using MegaBLASTN (Zhang et al. 2000). The vast majority of the markers (21,666, 98.6%) have matches to the assembly. The results were first filtered by requiring matches to at least 40% of the marker length with at least 90% match identity. Repeated filtering removed markers with second match location scores of the top hits that were within 50 points of each other.

After filtering, scaffolds with markers were anchored onto the chromosomes according to the marker orders provided in the Integrated Bovine Map. In the cases where a scaffold had markers from different chromosomes, the scaffold was checked for dog and human synteny. If the synteny information confirmed that the scaffold should be on different chromosomes, the scaffold was split. Otherwise, the minor group(s) of the markers were ignored. In the cases where a scaffold had markers from a single chromosome but the markers were far apart, the scaffold was anchored by the major group of the markers. In the cases where the markers were on a single chromosome but the integrated map marker order was not consistent with the mapping on the genome scaffold assemblies, the marker order was rearranged according to the scaffold sequences. The scaffold orientation on the chromosome was determined by the order of the markers. When it was impossible to determine the orientation (e.g., a scaffold with a single marker), the scaffolds were labeled as unoriented.

## Description of Refined Mapping and Placement for Btau_4.0

The contigs and scaffolds are not significantly changed from the Btau_3.1 assembly to the Btau_4.0 assembly, but different map information was used to place the contigs and scaffolds in the genome, resulting in more accurate chromosome structures in Btau_4.0. The mapping procedure is described here.

BES reads from both Hereford (189,587) and Non-Hereford (131,700) breeds were aligned to the scaffolds using BLASTN, and clone links were used to generate a set of larger scaffolds. Scaffolds that had potential misassemblies were split based on bovine and sheep BES links (Dalrymple et al. 2007) when the bovine and sheep BES consistently indicated that the parts of the scaffold mapped to different regions. After splitting, the scaffolds were mapped to the chromosomes based on the ILTX marker map (Everts-van der Wind et al. 2005) where the positions of the markers on the scaffolds were determined by BLASTN alignment.

The order of the scaffolds on the chromosomes was refined based on the information from three sources: (1) the fingerprint contig map (FPC) (British Columbia Cancer Agency, Canada's Michael Smith Genome Sciences Centre, unpublished), (2) human and dog synteny, and (3) links by sheep BAC clones (Dalrymple et al. 2007). When any three adjacent scaffolds had order information from at least two of the three sources and the order was consistent among these sources but in conflict with the ILTX map (Everts-van der Wind et al. 2005), the order of the scaffolds was modified from the ILTX map order (Everts-van der Wind et al. 2005). The scaffolds that were not oriented by the ILTX map (Everts-van der Wind et al. 2005) were oriented using the FPC information when such information was available.

Additional scaffolds were placed if two adjacent scaffolds from before were present in the FPC map (British Columbia Cancer Agency, Canada's Michael Smith Genome Sciences Centre, unpublished), and there were additional scaffolds in the FPC map between them. These additional scaffolds from FPC were filled in on the chromosomes.

The remaining unoriented scaffolds were further oriented based on human synteny. This step oriented ~9% of the scaffolds. Additional scaffolds were mapped to the chromosomes based on the bovine and sheep BES links with the supporting evidence from the FPC (British Columbia Cancer Agency, Canada's Michael Smith Genome Sciences Centre, unpublished data), and single nucleotide polymorphism (SNP) maps. Finally, when various sources suggested different locations of scaffolds, the ambiguity was resolved where possible by checking the synteny and the individual eBAC assemblies. Overall, 90% of the total genome was placed on chromosomes.

## Assembly Metrics

The final products of the Atlas assembler are a set of contigs (contiguous blocks of sequence) and scaffolds. Scaffolds include sequence contigs that can be ordered and oriented with respect to each other as well as isolated contigs that could not be linked (single contig scaffolds or singletons). Reads that clustered into groups of three or fewer were not assembled. The metrics for the major assemblies are given in Table 9.3. The N50 size of the contigs in the Btau_2.0 assembly is 18.9 kb and the N50 of the scaffolds is 434.7 kb. The N50 size is the length such that 50% of the assembled genome lies in blocks of the N50 size or longer. The total length of all contigs in Btau_2.0 is 2.62 Gb. When the gaps between contigs in scaffolds are included, the total span of the Btau_2.0 assembly is 3.1 Gb (some scaffolds with large gaps may artificially increased the assembly size). The total length of all contigs in the Btau_3.1 and Btau_4.0 assemblies is 2.73 Gb, while the total span of the assembly is 2.87 Gb. The combined assemblies (Btau_3.1, Btau_4.0, Btau_4.5) include a total of 26,052,388 reads, which yields about 7.0×-sequence coverage (using the average trimmed read length of 730 bp and the assembly size as 2.73 Gb).

## Assembly Validation

The genome assemblies were tested against available bovine sequence data sets (expressed sequence tag (EST) sequences and finished BAC sequences) to measure the extent of coverage or completeness. When assembled contigs from the combined assemblies (Btau_3.1, Btau_4.0) or the assembled contigs and unassembled reads from the WGS assembly (Btau_2.0) were tested, over 95% of the sequences in these data sets were found to be represented, indicating that the shotgun libraries used to sequence the genome were comprehensive (Liu et al. 2009). Of the 1.04 million EST sequences, 95.0% were contained in the assembled contigs of the published combined assemblies (Btau_3.1, Btau_4.0) (Liu et al. 2009). Assuming the ESTs are uniformly distributed throughout the genome, the estimated genome size is

**Table 9.3** Statistics for whole genome assemblies.

| | Btau_2.0 | Btau_3.1 | Btau_4.0 | Btau_4.5 |
|---|---|---|---|---|
| Contigs | | | | |
| Number | 321,107 | 131,620 | 131,620 | 81,945 |
| N50 (kb) | 18.9 | 48.7 | 48.7 | 81.9 |
| Bases + gaps (Gb) | 2.62 | 2.73 | 2.73 | 2.77 |
| Bases (Gb) | 2.62 | 2.73 | 2.73 | 2.77 |
| Percentage | 85 | 95 | 95 | 94 |
| Anchored scaffolds | | | | |
| Number | 4,409 | 3,053 | 2,331 | 1,717 |
| N50 (kb) | 1,247 | 1,940 | 2,687 | 2,872 |
| Bases + Gaps (Gb) | 1.7 | 2.40 | 2.58 | 2.63 |
| Bases (Gb) | 1.4 | 2.29 | 2.47 | 2.48 |
| Percentage | 54.8 | 84 | 90 | 89 |
| Unanchored scaffolds | | | | |
| Number | 98,058 | 13,045 | 11,830 | 11,861 |
| N50 (kb) | 189 | 166 | 94 | 80 |
| Bases + gaps (Gb) | 1.4 | 0.47 | 0.28 | 0.31 |
| Bases (Gb) | 1.2 | 0.44 | 0.26 | 0.29 |
| Percentage | 45.2 | 16 | 10 | 11 |
| Total scaffolds | | | | |
| Number | 102,467 | 16,098 | 14,161 | 13,578 |
| N50 (kb) | 434 | 997 | 1,922 | 2,573 |
| Bases + gaps (Gb) | 3.1 | 2.87 | 2.87 | 2.94 |
| Bases (Gb) | 2.62 | 2.73 | 2.73 | 2.77 |
| Percentage | 100 | 100 | 100 | 100 |

2.73 Gb/95% = 2.87 Gb. The quality of the assembly was also tested by aligning it to the 73 finished BACs (Liu et al. 2009). The genomic coverage in the BACs was high, between 92.5% and 100.0% (average of 98.5%) of the BAC sequence in the assembly. The assembled contigs and scaffolds were aligned linearly to the finished BACs, suggesting that misassemblies are rare. When compared to 317 BACs including the 73 finished BACs and 243 enhanced phase 2 ordered and oriented BACs, there were a total of 31 inconsistencies between the draft genome sequence and the BACs. These inconsistencies included sequences that did not match the BAC, misoriented sequences, and sequences that were not adjacent.

## Mapping QC

The accuracy of marker positions in the genome is reflected by the order of scaffolds on the chromosomes, as scaffolds were placed on chromosomes based on their alignments to markers. SNP linkage data that was initially used by two independent groups to order scaffolds on particular chromosomes with high confidence. One group used SNP linkage data to order scaffolds on Chr6 (Nilsen et al. 2008) and another placed scaffolds on Chr19 and Chr29 (Prasad et al. 2007). For these three chromosomes the

**Table 9.4** Mapping comparison.

| | Total shared scaffolds | Misplaced scaffolds | | |
|---|---|---|---|---|
| Chromosome 6 | 61 | 0 | 15 | 7 |
| Chromosome 19 | 45 | 0 | 6 | 9 |
| Chromosome 29 | 28 | 0 | 7 | 7 |

order of scaffolds was compared with the independent mapping evidence for three datasets: (1) Btau_3.1, which used an early version of the Integrated Bovine Map (Snelling et al. 2007), (2) Btau_4.0, and (3) the scaffold order using the published version of the Integrated Bovine Map (Snelling et al. 2007). The comparison showed the consistency between the evidence and Btau_4.0 where all the scaffolds in Btau_4.0 were in increasing order. In contrast, conflicts occurred when comparing the evidence with Btau_3.1. Most of the inconsistencies occurred between neighboring scaffolds, suggesting that errors in the order of Btau_3.1 markers were primarily local errors. Chr6 clearly had many more errors in Btau_3.1 than Chr19 and Chr29. The published version of the Integrated Bovine Map showed fewer conflicts with the evidence overall (e.g., Chr6) than the version of the Integrated Bovine Map used in Btau_3.1, although the differences did not necessarily solve the conflicts and in some cases even generated new inconsistencies (e.g., Chr19).

Table 9.4 summarizes the number of misplaced scaffolds in three data sets (Btau_4.0; Btau_3.1; and the Integrated Bovine Map (Snelling et al. 2007)) for three chromosomes when compared with the independent mapping evidence.

## Quality Assessment of the Assembly by Linkage Analysis

Further assessment of the Btau_4.0 assembly was performed by comparing dense SNP linkage maps constructed from genotyping 17,482 SNPs in 2637 Norwegian Red bulls belonging to 108 half-sib families with the physical positioning of the SNPs on all autosomal chromosomes (Liu et al. 2009). The analysis revealed few SNPs (134, <0.8%) were incorrectly positioned within assembly indicating the high degree of precision in the Btau_4.0 assembly. These misplaced SNPs were relocated in the linkage map to a position corresponding to the most closely linked, correctly assigned SNP. Additionally, 568 SNPs from 321 unplaced scaffolds were mapped to linkage groups.

## Sex Chromosomes and Autosome Assemblies

The majority of the sequence in the project is from the female animal, so the genome sequence is described for the 29 autosomes and the X chromosome. However, as the BAC library was prepared from a male animal, and the BAC fingerprint contigs were built from random clones from that library, both the X and Y chromosomes are represented in the BAC fingerprint contigs. Representative BACs in all of the BAC fingerprint contigs were sequenced to low coverage, including Y chromosome

BACs. Since the clone coverage on the sex chromosomes in the BAC library is half that of the autosomes, there will be less depth of clone coverage on the sex chromosomes and this may result in more gaps in the coverage of the sex chromosomes by BAC clones. The WGS sequence was from the female animal, so there is no additional WGS sequence to assemble with the low-coverage BAC skim sequences for the Y chromosome, unless it is pseudoautosomal sequence from the X chromosome or autosomal sequence that is similar to the Y sequence. Since the BAC fingerprint contigs were used to build the combined BAC + WGS assemblies, there are genome sequence scaffolds from both sex chromosomes as well as the autosomes. These 43 Y chromosome BACs map to 281 WGS contigs from the genome assembly, 276 of which were not placed on the autosomes or the X chromosome. The five contigs that were mapped to chromosomes in the assembly may share duplicated gene or gene motif, or repeat sequences with the Y chromosome. The Y chromosome scaffolds from these BACs are unlabeled in the unplaced chromosome. There is an ongoing project to improve the bovine Y chromosome to the level of high-quality finished sequence.

## Later Improvements

For the Btau_4.2 assembly and the Btau_4.6 assembly, the available finished sequence from GenBank was spliced into the Btau_4.0 and Btau_4.5 assemblies, respectively. There were 74 BACs finished to human grade finishing standards (HTGS_PHASE3) and 243 BACs finished to enhanced phase 2 ordered and oriented condition. Enhanced phase 2 submissions have order and orientation of contigs established using one or more of the following: read-pair data from individual subclones, overlaps with neighboring BAC clones, alignment to the available reference sequence (e.g., human), or confirmation by PCR testing. A total of 58 Mb of these available finished sequence data was substituted in the following manner. The finished sequences were mapped to the draft (Btau_4.0/Btau_4.5) assembly and the overlapping draft sequence contigs were removed if they were completely contained within the finished BAC. Draft sequence contigs that partially overlapped finished sequences were moved to the unassigned chromosome (ChrUn) to maintain the representation of the unique sequence not found in the finished clones.

The majority of the global genome analyses used the Btau_4.0 assembly. The Btau_4.5 assembly incorporated more WGS contigs, removed duplicated sequences, and improved scaffolding using ~100× clone coverage in SOLiD paired-end data from 25 bp reads with 1–2-kb inserts. An additional 9 Mb was mapped to chromosomes and 27 Mb included on ChrUn for the Btau_4.5 assembly using these methods. Table 9.5 gives the comparisons of the Btau_4.0 and Btau_4.5 assemblies to the available mRNA sequences.

## Data Availability

The genome assembly version Btau_4.0 is available in GenBank under accession number AAFC0000000.3. Other genome versions are also available in GenBank under

**Table 9.5** Comparison of version 4.0 and 4.5 using mRNA data.

| | | Chromosomes | | Chromosomes + ChrUn | | Chromosomes + ChrUn + omitted WGS | |
|---|---|---|---|---|---|---|---|
| Assembled genome size (Gb) | Btau_4.0 | 2.47 | | 2.73 | | 2.87 | |
| | Btau_4.5 | 2.48 | | 2.77 | | | |
| | | Number | Percentage | Number | Percentage | Number | Percentage |
| Matches to mRNA RefSeqs | Btau_4.0 | 9,469 | 97.49 | 9,631 | 99.16 | 9,695 | 99.81 |
| | Btau_4.5 | 9,465 | 97.45 | 9,712 | 99.99 | | |
| >95% length matched | Btau_4.0 | 7,833 | 80.64 | 8,053 | 82.91 | 8,121 | 83.61 |
| | Btau_4.5 | 8,015 | 82.52 | 8,291 | 85.36 | | |
| >90% length matched | Btau_4.0 | 8,614 | 88.69 | 8,853 | 91.15 | 8,938 | 92.02 |
| | Btau_4.5 | 8,778 | 90.37 | 9,090 | 93.59 | | |
| >80% length matched | Btau_4.0 | 9,023 | 92.9 | 9,288 | 95.62 | 9,387 | 96.64 |
| | Btau_4.5 | 9,160 | 94.31 | 9,504 | 97.85 | | |
| >50% length matched | Btau_4.0 | 9,247 | 95.2 | 9,525 | 98.06 | 9,641 | 99.26 |
| | Btau_4.5 | 9,328 | 96.04 | 9,694 | 99.8 | | |

WGS, whole genome shotgun.

the same accession number prefix with different version number suffixes. Since the process of genome assembly involves decisions about which sequences to include and which sequences to exclude, there are sequences from this project that were omitted from the genome assembly. Some of the omitted sequences are highly repetitive sequence reads, others may have enough sequencing errors that they did not match the assembled sequences, others are assembled sequence contigs that appear to be duplicates of sequences in the assembly (perhaps from the second haplotype). These excluded sequence are available from the BCM-HGSC ftp site.

## Acknowledgment

Much of the data for this chapter is taken from the report in Liu et al. 2009, which is licensed under the creative commons license (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## References

Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**(5926): 522–528.

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Dalrymple, B.P., et al. (2007) Using comparative genomics to reorder the human genome sequence into a virtual sheep genome. *Genome Biology* **8**: R152.

de la Bastide, M. and McCombie, W.R. (2007) Assembling genomic DNA sequences with PHRAP. *Current Protocols in Bioinformatics*, Chapter 11, Unit 11.14.

Everts-van der Wind, A., et al. (2005) A high-resolution whole-genome cattle-human comparative map reveals details of mammalian chromosome evolution. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 18526–18531.

Gibbs, R.A., et al. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**: 222–234.

Havlak, P., et al. (2004) The Atlas genome assembly system. *Genome Research* **14**: 721–732.

Kirkness, E.F., et al. (2003). The dog genome: survey sequencing and comparative analysis. *Science* **301**(5641): 1898–1903.

Lander, E.S., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Liu, Y., et al. (2009) *Bos taurus* genome assembly. *BMC Genomics* **10**: 180.

Mikkelsen, T.S., et al. (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**: 167–177.

Nilsen, H., et al. (2008) Construction of a dense SNP map for bovine chromosome 6 to assist the assembly of the bovine genome sequence. *Animal Genetics* **39**: 97–104.

Phrap, {HYPERLINK "www.phrap.org"}.

Pontius, J.U., et al. (2007) Initial sequence and comparative analysis of the cat genome. *Genome Research* **17**: 1675–1689.

Prasad, A., et al. (2007) High resolution radiation hybrid maps of bovine chromosomes 19 and 29: comparison with the bovine genome sequence assembly. *BMC Genomics* **8**: 310.

Rat Genome Sequencing Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**: 493–521.

Sanger, F., Nicklen, S., Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**(12): 5463–5467.

Sea Urchin Genome Sequencing Consortium (2006) The genome of the sea urchin Strongylocentrotus purpuratus. *Science* **314**: 941–952.

Snelling, W.M., et al. (2007) A physical map of the bovine genome. *Genome Biology* **8**: R165.

Warren, W.C., et al. (2008) Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**: 175–183.

Zhang, Z., et al. (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**: 203–214.

# Chapter 10
# Bovine Genome Architecture

*David L. Adelson*

## Introduction

Mammals vary widely in their appearance and physiology, yet are very similar based on comparisons of their genes. The core mammalian genome consists of approximately 20,000 protein-coding genes, with the vast majority conserved across species (Lander et al. 2001; Venter et al. 2001; Metzker et al. 2004; Lindblad-Toh et al. 2005). However, these protein-coding genes account for only ~2% of a typical mammalian genome. The rest of the genome is nonprotein coding and, for the most part, not transcribed. While there is still debate on how much of the genome is in fact transcribed, almost half of a typical mammalian genome comes from repetitive DNA dubbed by some as "junk DNA," derived from self-propagating mobile elements and retroviruses (RTE) (Jurka et al. 2007). More recently, it has become clear that evolution has made use of these repetitive sequences to wire new regulatory circuits (Mikkelsen et al. 2007). This has resulted from the incorporation of RTE into promoters, miRNA precursors, and coding exons (Babushok et al. 2007; Gentles et al. 2007). Species-specific RTE can also contain regulatory elements such as the P53 tumor suppressor binding motif, and thus influence transcriptional regulatory networks genome-wide (Wang et al. 2007). Therefore, while the protein-coding genetic complement of mammals is virtually identical, the remainder of these genomes is both highly repetitive yet variable. Genome architecture is thus largely dependent on RTE-derived sequences. We believe that RTE are important sources of genetic variation and over time have been subject to selection such that they have been incorporated into a number of crucial, yet cryptic functions. Genome architecture can be viewed as the relationship between genome structure and function, and one of biology's grand challenges is to determine how RTE contribute to the structure and regulation of the genome both during the life cycle of an organism, and within an evolutionary context.

While most repetitive elements have previously been well characterized in mammals (Jurka et al. 2005), our work on the bovine (Elsik et al. 2009) and equine (Wade et al. 2009) genome projects has shown that a significant amount of the genome is erroneously considered to be nonrepetitive based on generic repeat libraries (Table 10.1).

While a 3%–4% difference in repeat annotation might not seem catastrophic, it is worth pointing out that this exceeds the total protein-coding sequence of the

*Bovine Genomics*, First Edition. Edited by James E. Womack.

**Table 10.1** Effect of de novo repeat libraries on repeat detection.

| Genome | RepBase bp detected | RepBase percentage of genome | De novo lib bp detected | De novo percentage of genome |
|---|---|---|---|---|
| Cow (Btau_3.1) | 1,439,833,224 | 49.5 | 1,542,500,559 | 53 |
| Horse (ECv1) | 1,029,835,986 | 42 | 1,105,289,360 | 45 |

bovine genome by twofold and can have major implications for analyses of genome architecture. If repetitive elements are poorly or incompletely masked, they can create significant problems for the identification of segmental duplications (SDs) because SD analysis depends on the results of whole genome self-alignment. If the self-alignment used for SD analysis contains unmasked repeats it can both increase the time required for computational analysis of the alignments and promote the inclusion of spurious SD.

Figure 10.1 graphically demonstrates how much of an improvement in a self-alignment can result from stringent masking of repeats. The repeat masking carried out with de novo identified repeats in Figure 10.1C demonstrates how stringent repeat masking reduces the number of spurious off-axis self-alignments and helps identify potential SD as hits clustered close to the axis.

These data demonstrate the necessity of de novo repeat identification and annotation. Furthermore, in spite of significant analysis of individual repeats and their association with various genome features, there have been no global, comprehensive analyses of repeat correlations. Finally, while retrotransposition events have been shown to be important sources of mutation in mice and humans (Kazazian 1999;

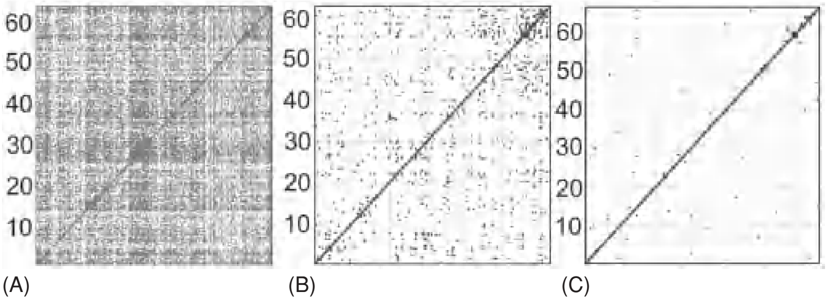


**Figure 10.1** Effect of repeat masking with a de novo identified repeat library. (A) Self-alignment of unmasked bovine chromosome 18 sequence. (B) Self-alignment of bovine chromosome 18 sequence masked with RepeatMasker using the default library. (C) Self-alignment of bovine chromosome 18 sequence masked with RepeatMasker using a library of de novo identified bovine repeat consensus sequences. Off-axis points arise from interspersed repeats or segmental duplications (SDs). Stringent repeat masking as in panel (C) highlights the presence of bona fide SDs. Alignments were carried out using mummer v3.2. Axis scale in millions of base pairs.

Desmarais et al. 2006; Korbel et al. 2007), they are only recently beginning to be analyzed with respect to their effects on genome structural variation (Cordaux and Batzer 2009; Xing et al. 2009).

Biological mechanisms underlying genome evolution are believed to originate with RTE insertions that ultimately lead to segmental (gene) duplications/deletions, incorporation of RTE into protein-coding genes (exaptation), or gene duplication via retrogene formation (Baertsch et al. 2008). The resulting "churning" of both nonprotein-coding regions and protein domains are two of the major forces that drive adaptation and speciation. Evidence of selection should exist in associations of RTE-derived repeats that are both conserved across mammalian genomes or are species specific. Because evolutionary conservation is a hallmark of functional importance, these associations uncover novel, functionally important aspects of genome architecture. While the main focus of this monograph is on evolutionary questions, RTE insertions are believed to be frequent events that give rise to novel mutations (Cordaux and Batzer 2009; Xing et al. 2009). This is an important research problem both in terms of our understanding of evolutionary mechanisms and processes, and also due to the fact that these processes frequently give rise to mutations or structural variation affecting gene regulation and function that can result in disease or influence economically important agricultural traits.

## De Novo Repeat Identification and Annotation

The Bovine Genome Project was the first to employ a two-pronged approach to de novo repeat identification. One prong was based on self-alignment of the genome as the initial step for detection of repeats, followed by clustering into families using a PALS/PILER pipeline (Edgar and Myers 2005). The second prong used RepeatScout (Price et al. 2005), which generates repeat consensus sequences based on a greedy extension of sequence seed matches. By combining the output of these two methods we were able to generate repeat consensus sequences with both high sensitivity and high specificity (Adelson et al. 2009).

### *Coverage*

About 40%–45% of a typical mammalian genome is made up of interspersed repeats. The bovine genome is no exception, with ~46.5% interspersed repeats, the majority of which are of retrotransposon origin (Adelson et al. 2009; Table 10.2).

Retrotransposons are not efficient genomic parasites when one considers that the vast majority of long interspersed nucleotide elements (LINE) insertions in the genome are 5′ truncated sequences resulting from incomplete reverse transcription and only about 1 in 500 LINE elements is full length (see Section "Clade-Specific Repeats"). Obvious differences in interspersed repeat coverage are the very high percentage of short interspersed nucleotide elements (SINE), including a large number of tRNA, and very low percentage of endogenous retrovirus (ERV) in the bovine genome, compared to horse, human, and mouse. The tRNA-derived SINE

**Table 10.2** Genomic repeat content.

| | | | Percent coverage of genome | | | |
|---|---|---|---|---|---|---|
| Group | Number | Total bp | *Bos taurus* | Horse | Human | Mouse |
| Non-LTR retrotransposons (LINE) | | | | | | |
| L1 | 616,259 | 328,664,804 | 11.26352 | 16.25 | 17.07 | 19.14 |
| RTE (BovB) | 376,067 | 313,409,818 | 10.74072 | 0.18 | NA | 0.02 |
| L2 | 132,485 | 34,553,185 | 1.18416 | 2.87 | 3.07 | 0.37 |
| CR1 | 14,524 | 3,083,954 | 0.10569 | 0.25 | 0.27 | 0.06 |
| | 1,139,335 | 679,711,761 | 23.29409 | 19.55 | 20.4 | 19.59 |
| SINEs | | | | | | |
| BOV-A2 | 377,697 | 68,880,046 | 2.360556 | NA | NA | NA |
| Bov-tA | 1,461,800 | 225,579,571 | 7.730733 | NA | NA | NA |
| ART2A | 348,768 | 121,997,595 | 4.18092 | NA | NA | NA |
| tRNA | 388,920 | 57,981,206 | 1.98705 | 0.02 | NA | 0 |
| MIR | 301,335 | 40,569,445 | 1.39034 | 2.66 | 2.43 | 0.55 |
| Other | 4322 | 432,334 | 0.01482 | 4.3 | 10.68 | 6.78 |
| | 2,882,842 | 515,440,197 | 17.66441 | 6.98 | 13.11 | 7.34 |
| ERVs | | | | | | |
| MaLR | 135,536 | 42,285,673 | 1.44915 | 2.68 | 3.72 | 4 |
| ERVL | 69,540 | 25,833,994 | 0.88534 | 1.82 | 1.56 | 1.03 |
| ERV1 | 68,518 | 23,706,917 | 0.81245 | 1.53 | 3 | 0.8 |
| ERVK | 4038 | 1,536,800 | 0.05267 | 0.07 | 0.29 | 4.02 |
| | 277,632 | 93,363,384 | 3.19961 | 6.1 | 8.56 | 9.84 |
| DNA transposons | | | | | | |
| DNA All | 244,174 | 57,157,641 | 1.95882 | 3.15 | 3 | 0.89 |
| LTR other | | | | | | |
| BTLTR1 | 11,338 | 6,494,236 | 0.22256 | NA | NA | NA |
| ARLTR2 | 14,358 | 4,127,734 | 0.14146 | NA | NA | NA |
| Other | 8656 | 1,773,440 | 0.06078 | 0.17 | 0 | 0.01 |
| | 34,352 | 12,395,410 | 0.4248 | 0.17 | 0 | 0.01 |
| Dinucleotide SSR | | | | | | |
| di AC | 539,678 | 6,835,776 | 0.23426 | 0.21 | 0.14 | 0.76 |
| di AT | 440,644 | 4,957,167 | 0.16988 | 0.13 | 0.08 | 0.2 |
| di AG | 375,243 | 3,537,184 | 0.12122 | 0.19 | 0.05 | 0.43 |
| di CG | 9081 | 85,400 | 0.00293 | 0 | 0 | 0.01 |
| | 1,364,646 | 15,415,527 | 0.5283 | 0.53 | 0.28 | 1.4 |
| Trinucleotide SSR | | | | | | |
| tri AGC | 285,325 | 3,910,867 | 0.13403 | 0.05 | 0 | 0.06 |
| tri AAT | 231,133 | 2,361,839 | 0.08094 | 0.1 | 0.04 | 0.09 |
| tri AGG | 199,279 | 1,945,722 | 0.06668 | 0.07 | 0.01 | 0.12 |
| tri AAG | 194,774 | 1,894,234 | 0.06491 | 0.08 | 0.01 | 0.12 |
| tri AAC | 163,282 | 1,793,155 | 0.06145 | 0.05 | 0.02 | 0.09 |
| tri ACC | 100,462 | 1,043,099 | 0.03575 | 0.04 | 0.01 | 0.06 |
| tri ATC | 81,511 | 799,361 | 0.02739 | 0.04 | 0.01 | 0.04 |
| tri ACT | 32,644 | 334,552 | 0.01147 | 0.01 | 0 | 0.01 |

**Table 10.2** (*Continued*)

| | | | Percent coverage of genome | | | |
|---|---|---|---|---|---|---|
| Group | Number | Total bp | *Bos taurus* | Horse | Human | Mouse |
| Trinucleotide SSR (*continued*) | | | | | | |
| tri CCG | 19,735 | 219,746 | 0.00753 | 0.01 | 0 | 0.01 |
| tri ACG | 1769 | 17,524 | 0.0006 | 0 | 0 | 0 |
| | 1,309,914 | 14,320,099 | 0.49075 | 0.45 | 0.1 | 0.6 |
| Tetra/pentanucleotide SSR | | | | | | |
| tetra, penta All | 2,979,022 | 36,540,157 | 1.25225 | 1.22 | 0.39 | 2.16 |
| Unclassified | | | | 11.27 | | |
| Interspersed Repeat Total | 4,578,335 | 1,358,068,393 | 46.54174 | 47.22 | 45.08 | 37.65 |
| SSR Total | 5,653,575 | 66,275,552 | 2.2713 | 1.75 | 0.78 | 4.16 |

found in the bovine genome include hundreds of thousands of very highly conserved tRNA, many of which appear to be perfectly functional. In general, SINE are clade-specific repeats derived from truncated LINE and help define species-specific genome architecture. In addition, the bovine genome has more simple sequence repeats (SSR) than the horse or human, but the significance of this finding is unclear.

## *Clade-Specific Repeats*

The primary feature that distinguishes the bovine genome from those of other eutherian mammals is the presence of ruminant-specific repeats (Lenstra et al. 1993; Kordis and Gubensek 1999). These consist of the BovB/LINE RTE, and the ART2A, BovA, and Bov-tA SINE derived from LINE RTE. LINE RTE, while specific to ruminants in eutheria, are also found in marsupials (Gentles et al. 2007), monotremes (Jurka 2000), squamates (Kordis and Gubensek 1998), and echinoderms (Jurka 2000).

The current hypothesis used to explain the patchy taxonomic distribution of LINE RTE is that it has been laterally transferred across taxa. By reconstructing the phylogeny of the LINE RTE based on consensus sequences, it is clear that while squamate LINE RTE is most similar to Marsupial LINE RTE, it is also very similar to bovine LINE RTE (Figure 10.2). As previously discussed by Gentles et al. (2007), we are at present unable to resolve whether lateral transmission of LINE RTE has occurred solely based on the phylogenetic data, but that is the most parsimonious explanation.

If LINE RTE did transfer from reptiles to a ruminant ancestor, about a quarter of the bovine genome can be attributed to expansion of this LINE lineage and derivation of associated SINE. In any case, the impact of LINE RTE and derived SINE on bovine genome architecture is significant, and we might expect ruminant/bovine regulatory

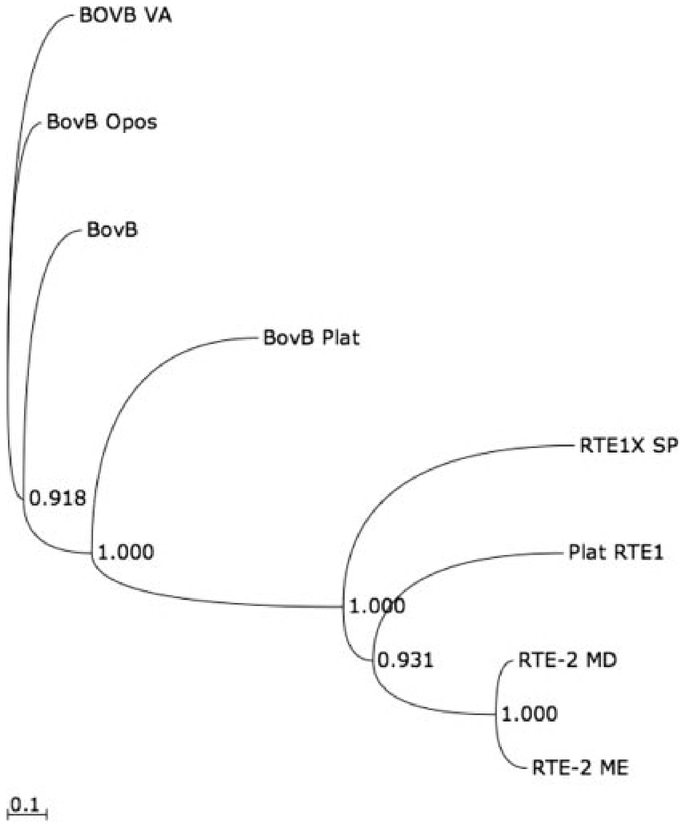


**Figure 10.2** Phylogenetic tree of selected LINE RTE (BovB) repeat consensus sequences. The tree topology supports the possible lateral transmission of BovB containing repeats from squamata to mammals. Consensus sequences were aligned using MUSCLE and the tree was generated using FastTree. Support values are shown at node positions. Nomenclature of consensus sequences is according to RepeatMasker. BOVB VA (*Vipera ammodites*), BovB Opos (*Monodelphis domestica*), BovB (*Bos taurus*), BovB Plat (*Ornithorhynchus anatinus*), RTE1X SP (*Strongylocentrotus purpuratus*), Plat RTE1 (*Ornithorhynchus anatinus*), RTE-2 MD (*Monodelphis domestica*), and RTE-2 ME (*Macropus eugenii*).

networks to be quite different to those from other eutheria, based on changes to promoters caused by RTE insertion.

LINE RTE are still presumed to be active, because there are approximately ten intact LINE RTE with apparently functional open reading frame (ORF) in the bovine genome (Adelson et al. 2009; Figure 10.3). There are 1248 full length LINE RTE in the bovine genome with a per site substitution rate twice that of L1 LINE. This difference in substitution rates may indicate that the reverse transcriptase encoded in the LINE RTE ORF is more error prone than the reverse transcriptase encoded in L1 LINE. An analysis of Opossum LINE RTE revealed only 26 full-length LINE RTE, with a higher per site substitution rate than bovine LINE RTE, none of which had a functional ORF. This supports the notion that LINE RTE in marsupials are older than in ruminants. In spite of this higher mutation rate and relatively small number

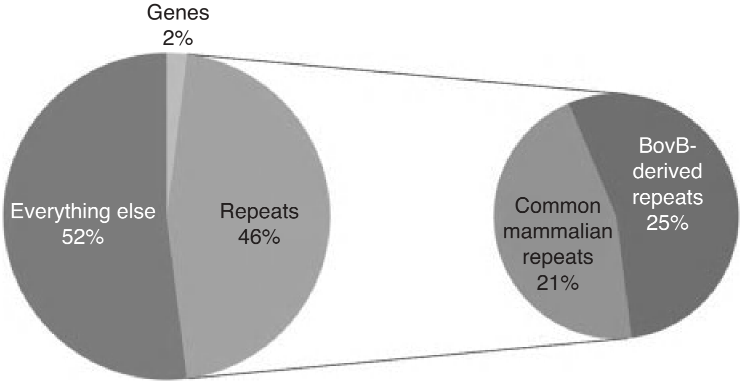


**Figure 10.3** Percentage of repeated sequences in the bovine genome. BovB-derived repeats include LINE RTE, ART2A, BovA, and BovtA repeats.

of potentially active copies, LINE RTE are probably still influencing bovine genome architecture.

## *Common Mammalian Repeats*

In addition to active clade-specific repeats, the bovine genome has a large number of L1 LINE retrotransposons. L1 LINE are ubiquitously distributed in eutherian mammals and are still contributing to human genome structural variation (Beck et al. 2010; Huang et al. 2010). In bovine, we have previously identified 811 full-length L1, with >70 of these presumptive active L1 based on ORF composition (Adelson et al. 2009). This number is comparable to the number of "hot" L1 found in the human genome (Beck et al. 2010), an indication that L1 LINE are probably responsible for ongoing structural variation in the bovine genome. This fact has implications for the causes of genetic diversity in cattle and how we map such diversity. Because both microsatellite and single nucleotide polymorphism (SNP)-based production trait mapping approaches are blind to retrotransposon insertion site polymorphism, current quantitative trait locus (QTL), and genome-wide association studies (GWAS), analyses probably miss potentially interesting loci. The extent of this problem will be clearer once structural variation-based trait mapping becomes available in cattle.

## *Degree of Exaptation*

The evidence for recent exaptation of repeat sequences in the bovine genome is slim, despite evidence of potentially active repeats. Neither L1 nor LINE RTE appear to have significantly contributed to bovine-specific genes via exaptation. The only good examples of recent exaptation of BovB are in the *CFDP2* gene (Takahashi et al. 1998) and *FASTKD3* (Almeida et al. 2007); we detected no additional evidence for recent exaptation of either of these LINE elements in any NCBI bovine refseq.

## Mitochondrial Insertion Sequences

Mitochondrial DNA can be inserted into nuclear genomes to create nuclear mitochondrial insertions (NUMTs) via double strand break (DSB) repair and nonhomologous end joining (NHEJ) in a similar fashion to retrotransposon insertion (Pace et al. 2009). De novo insertion of mitochondrial sequence into the human genome has been shown to cause disease (Turner et al. 2003). A previous report of bovine NUMTs indicated that there were 279 NUMTs (identified by BLASTN) in the bovine genome (Hazkani-Covo et al. 2010). We carried out a similar analysis using LASTZ (Harris and Riemer 2010) that should be more sensitive, and have found evidence for 421 NUMTs genome wide, of which 372 are on chromosome scaffolds (Figure 10.4). The number of NUMTs is only weakly correlated with chromosome length ($R^2 \sim 0.3$), and it is apparent from Figure 10.4 that the distribution of NUMTs is nonrandom with respect to chromosomes. Reports of NUMTs in draft genome sequences have to be viewed with some skepticism; however, as most genome assemblers discard sequences with perfect or near-perfect identity to mitochondrial genomes, because mitochondrial DNA can be a significant source of contamination during library preparation. While the NUMTs identified range in size from 37 bp to 5219 bp and from 58% identity to 100% identity compared to the bovine mitochondrial genome sequence (NC_006853.1), they almost certainly represent an underestimate of the true frequency of NUMTs in the bovine genome. Furthermore, the presence of at least one NUMT with 100% identity to the mitochondrial reference genome indicates that integration of NUMTs is ongoing and may be polymorphic both within and across cattle breeds.

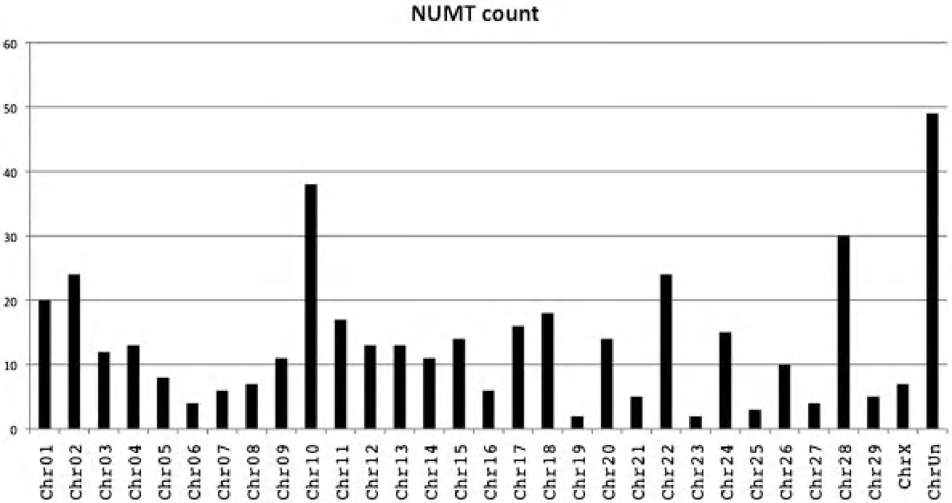


**Figure 10.4** Genomic distribution of nuclear mitochondrial insertion sequences (NUMT). If insertion sequences result from unbiased insertion, they should be distributed randomly across the genome and we would expect the number of NUMT to be strongly positively correlated with chromosome length. The observed chromosomal distribution of NUMT frequency does not support a random distribution of NUMT across the genome.

## Arrays, Duplications, and Correlations

Sequence arrays, duplications, and spatial correlations are additional features of genome architecture. Tandem arrays of sequences include telomeric, centromeric, and pericentromeric repeats. Unfortunately, the draft assembly of the bovine genome relegates most of these regions to the ChrUn or "unknown chromosome" set of contigs. At the present time, all analyses pertaining to arrays, duplications, and correlations have only been performed on the Btau_4 genome assembly.

### *Tandem Arrays*

Tandem arrays of sequences are believed to result from illegitimate recombination events, such as unequal crossing-over or intramolecular crossing-over, followed by gene conversion. We have identified moderately large tandem arrays of satellite sequences within bovine chromosomes 7 and 18. The repeated motif in each tandem array is a 720-bp sequence virtually identical to BTSATII (genbank X03116.1), a satellite sequence described in goats, sheep, and cattle (Buckland 1985). These arrays are characterized by very high degrees of sequence similarity between the BTSATII repeats exceeding 95% identity (Table 10.3). The near-perfect nature of these repeat arrays is consistent with unequal crossing-over and gene conversion.

The locations of these satellite arrays in the middle of the chromosome arms is curious, given that fluorescent hybridization of BTSATII probes to sheep chromosomes only occurs at centromeres or on the short arm of acrocentric chromosomes (D'Aiuto et al. 1997). The additional arrays of BTSATII found on ChrUn contigs probably

**Table 10.3** BTSATII tandem arrays in the bovine genome.

| Chromosome | Start | Stop | Length | Copies | Avg percentage id |
|---|---|---|---|---|---|
| Chr7 | 58773604 | 59009430 | 235826 | 345 | 97.93% |
| Chr18 | 21521005 | 21651903 | 130898 | 192 | 97.78% |
| ChrUn.004.945 | 0 | 57723 | 57723 | 85 | 97.80% |
| ChrUn.004.2568 | 0 | 18611 | 18611 | 28 | 97.78% |
| ChrUn.004.2298 | 7 | 21643 | 21636 | 26 | 94.29% |
| ChrUn.004.3072 | 0 | 15053 | 15053 | 23 | 97.15% |
| ChrUn.004.3649 | 323 | 12696 | 12373 | 19 | 98.18% |
| ChrUn.004.4930 | 0 | 10100 | 10100 | 16 | 97.98% |
| ChrUn.004.5361 | 0 | 8599 | 8599 | 13 | 98.02% |
| ChrUn.004.665 | 68308 | 76529 | 8221 | 11 | 65.99% |
| ChrUn.004.1895 | 16937 | 22037 | 5100 | 8 | 65.83% |
| ChrUn.004.38 | 357202 | 383445 | 26243 | 7 | 66.44% |
| ChrUn.004.665 | 62398 | 66815 | 4417 | 7 | 65.29% |
| ChrUn.004.9407 | 0 | 2393 | 2393 | 4 | 97.08% |
| ChrUn.004.214 | 76334 | 78562 | 2228 | 4 | 76.15% |
| ChrUn.004.9531 | 4 | 2269 | 2265 | 4 | 97.43% |
| ChrUn.004.10463 | 0 | 1602 | 1602 | 3 | 96.70% |

represent centromeric or pericentromeric sequences that are difficult to position within the genome assembly.

## *Segmental Duplication/Copy Number Variation*

While arrays of repeats can arise from nonallelic homologous recombination (NAHR), the same mechanism can also drive duplication of nonrepetitive DNA, giving rise to SDs and copy number variation (CNV) (Inoue and Lupski 2002; Lupski and Stankiewicz 2005). SDs, once established, can then drive CNV, presumably via NAHR (Sharp et al. 2005).

There are two methods that can be used to identify SDs, one is based on detecting duplications within the assembly, whole-genome assembly comparison (WGAC), and the other is based on excess read depth from the whole-genome shotgun sequence data (WSSD). Bovine SDs reported by the Bovine Genome Sequencing Consortium (Elsik et al. 2009) were based on WGAC SDs, supported by WSSD results.

There are 1020 SDs in the bovine genome, accounting for 3.1% of the genome sequence (Elsik et al. 2009; Liu et al. 2009). Almost half (47%) of the SDs map to ChrUn contigs and exhibit similarity to SDs mapped to chromosome assemblies, it is therefore likely that most of the ChrUn SDs represent tandem SDs that cannot be mapped onto the current assembly. Where SDs have been mapped to chromosomes, the duplications are largely intrachromosomal, and large duplications (>300 kbp) tend to occur in regions with clustered tandem duplications (Elsik et al. 2009; Liu et al. 2009). This pattern is similar to what has been observed in rat, mouse, and dog, but different to primates (Bailey et al. 2002; Bailey and Eichler 2006). Subtelomeric and pericentromeric regions are about twofold enriched for SDs compared to the rest of the genome and this result is also consistent with what has been observed in dog, mouse, rat, and human. There are also specific enrichments of satellite repeats within duplicated regions, in particular BTSAT4 and OSSAT2. These differences in patterns of SD may be significantly influenced by the sequencing, assembly strategy, and assembly status, so inferences of biological significance from these differences should be treated with some skepticism (Liu et al. 2009). Most SDs in cattle are associated with gene-containing regions, with 76% of the SDs containing gene duplications. As a result, the genes within SDs are highly similar, and these data are consistent with a role for SDs as drivers of artiodactyl-specific gene formation. Perhaps, most telling in this regard is that the vast majority of pairwise alignments used to identify SDs are less than 1 Mbp apart. Analysis of functional annotation of genes in SDs shows that genes involved in detoxification, innate immunity, and signaling have been recently duplicated in cattle, as they have been in other mammals. This type of evolution is perhaps most evident in bovine-specific clusters of $\beta$-defensin and T-cell receptor variable region genes (Liu et al. 2009). SDs in cattle are also more prevalent in evolutionarily conserved breakpoints, indicating that SDs may promote chromosomal rearrangements via NAHR (Elsik et al. 2009).

SDs are believed to act as seeds for CNV formation (Emanuel and Shaikh 2001; Sharp et al. 2005; Goidts et al. 2006; Marques-Bonet and Eichler 2009) and CNVs in turn are major sources of structural variation in humans, with over 57,000 CNVs at over 14,000 loci identified in the human Database of Genomic Variants (Iafrate et al. 2004) as of July 2010. Because CNVs can affect multiple genes both in terms of regulation and function, they are believed to represent a substantial source of genetic variation. This is supported by evidence that ~18% of the genetic variance in human gene expression is attributable to CNVs (Stranger et al. 2007).

CNVs in cattle have been mapped via comparative genomic hybridization (CGH) using ~400,000 nucleotide probes spaced evenly over the bovine assembly from 90 animals of various breeds (Liu et al. 2010). Over 1000 CNVs were identified in this study, and these were aggregated into more than 200 CNV regions (CNVR) based on overlaps, with 177 CNVR positioned on chromosomes. Total coverage of chromosomal CNVR was 28 Mbp, or ~1% of the bovine genome, with a median size of 89 kbp. Consistent with what has been observed in other mammals, 61% of the CNVs found in this study overlapped with SDs. CNVs are nonrandomly localized across the genome, both in terms of overrepresentation on some chromosomes and in terms of enrichment in pericentromeric and telomeric regions. Half of the chromosomal CNVR were unique to single individuals, but of the remaining CNVR, 49 were present in >5% of the population, making them candidate copy number polymorphism loci. Given the relatively small sample size from this study, it is likely that many more cattle CNVR remain to be discovered. There is, however, already evidence from this relatively small sample that CNVs are arising within breeds and are not of ancestral origin, and are therefore, breed-specific sources of genetic variation.

## Spatial Correlations

Spatial correlations of genomic features can intuitively be ascribed to clustering for positive correlations or exclusion for negative correlations. Perhaps, the earliest mention of genome feature clustering was by R.A. Fisher (Fisher 1930) who showed that interacting genes tend to become more closely linked. More recently, genomic clustering has been shown for clustering of tissue-specific genes to chromosomal expression domains (Yamashita et al. 2004). In cattle, there is evidence for functional clustering of genes within QTL regions (Salih and Adelson 2009), supporting Fisher's prediction. Such clustering could also be the result of SD/CNV leading to gene family expansion. Further analysis of specific, well-mapped QTL may help identify the mechanisms associated with this type of gene clustering.

Nongene-based clustering analysis of the genome has also been carried out. In the original analysis of the human, rat, and mouse genomes (Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002; Gibbs et al. 2004), significant analysis was devoted to repetitive DNA, in particular substitution rates of fossil repeats in order to estimate the neutral substitution rate, and correlations of repeat location with genome features such as gene density and G+C content. However, until the analysis of the bovine genome was done, there was no comprehensive measurement of correlations of repetitive DNA in order to determine if spatial correlations between repeat types and other genome features might exist. We carried out the first such comprehensive correlation analysis (Adelson et al. 2009; Elsik et al. 2009) aimed at identifying spatially correlated features in the bovine genome. In order to carry out this analysis, we partitioned the genome into small segments or bins within which we could identify and count all DNA repeats and genes. Because the segment size for this analysis is critical, we tried a wide range of bin sizes (Adelson et al. 2010) before settling on the optimal size of 1.5 Mbp/bin.

Ruminant-specific interspersed repeats LINE RTE, BovA, BovtA, and ART2A behaved differently to panmammalian interspersed repeats LINE L1, SINE tRNA, and fossil repeats LINE L2 and SINE MIR. Generally speaking, ruminant-specific repeats are negatively correlated with G+C content and gene density, LINE L1 are not correlated with gene density or G+C content, and SINE tRNA and fossil repeats are strongly positively correlated with both. While LINE/SINE pairs in cattle have similar correlations with gene density and G+C content, in human and rodents LINE L1 and paired SINE do not behave the same with respect to G+C content. Ruminant genomes are unique among Eutheria in having two active pairs/sets of LINE/SINE retrotransposons; LINE L1 and SINE tRNA, and LINE RTE and SINE ART2A, BovA, and BovtA. By comparing the correlations based on these two lineages of repeats, one derived from paneutherian LINE L1 and the other from ruminant-specific LINE RTE, we can see two different patterns of spatial correlations with respect to genes, G+C content, and SDs. The paneutherian LINE L1 insertions are not correlated with gene density or G+C content, but are positively correlated with SDs. The ruminant-specific LINE RTE, on the other hand, are not correlated with SDs, but are negatively correlated with gene density and G+C content. While these two LINE types are distinct in terms of their spatial correlations with SDs, genes, and G+C content, they are positively correlated with each other. Because both of these LINE types have presumptive active copies in the genome and have been or are still retrotransposing, it is tempting to speculate that the positive correlations observed between LINE L1 and LINE RTE/SINE ART2A could be the result of active repeats being more likely to be inserted or accumulated in particular genomic regions that are not defined by G+C content or gene density.

In contrast to the two lineages of active LINE elements, the molecular fossils LINE L2 and SINE MIR are strongly positively correlated with each other and positively correlated with gene density and G+C content. These retrotransposon fossils are also negatively correlated with LINE L1 and LINE RTE/SINE ART2A/SINE BovA2, but are positively correlated with SINE tRNA and a number of SSR. These negative correlations of fossil repeats with most active repeats suggest that for these active retrotransposons certain genomic regions depleted of fossil repeats are more likely to be insertion targets. Of particular interest is that the strong positive correlations observed in the top corner of Figure 10.5 are conserved between cattle, horse, and human (Adelson et al. 2009, 2010). Furthermore, the strongest pairwise correlation (LINE L2/SINE MIR; Figure 10.6) is also the strongest pairwise correlation in all eutherian mammals and in marsupials (Adelson, unpublished). Taken together, these clusters of spatial correlations are indicative of genomic regions determined by SSR and interspersed repeat content.

It is possible to determine if regions enriched or depleted in ancient repeats also differ in clade-specific repeat content. This can be clearly demonstrated by identifying the extreme tails of the rank correlation plot shown in Figure 10.6 and analyzing clade-specific repeat content in these bins. The high rank tail corresponds to genome bins where the density of both LINE L2 and SINE MIR is high, and the low rank tail corresponds to bins where these repeats are present at low density. Box plots of clade-specific repeat density (Figure 10.7) illustrate the inverse relationship between ancient repeat density and clade-specific repeat density in these regions.

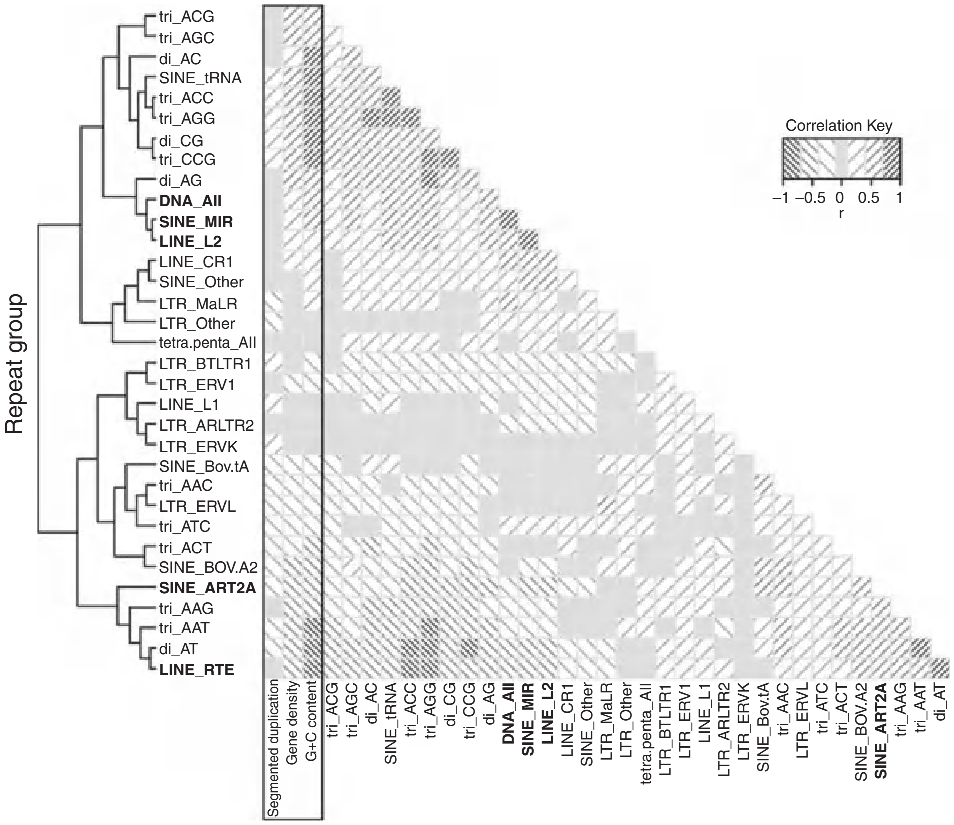


**Figure 10.5** Global pairwise correlations for simple sequence and interspersed repeats. Pairwise correlations among the repeat groups and between the repeat groups and segmental duplication (SD), gene density, and G+C content are shown. Repeat groups are clustered on the basis of all their correlations. Gray cells have nonsignificant correlations (5% 2-tailed test after Bonferroni correction). The right-hand hashed cells indicate significant positive correlations, and the left-hand hashed cells indicate significant negative correlations. Each chromosome was divided into 1.5 Mbp segments (bins) beginning at the 5′ end. For each bin, we calculated the number of repeats from each repeat group based on our repeat analysis that were entirely within the bin, the number of consensus gene models that started in the bin (gene density), the G+C content, and the number of SDs entirely within the bin. All bins with at least 1 Mbp non-N-specified bp were used to calculate Spearman rank correlations between each repeat group and the other repeat groups, as well as gene density, G+C content, and SD. The repeat groups were clustered on the basis of the correlations among the repeat groups, gene density, G+C content, and SD.

Figure 10.7, panels A and B, show that LINE RTE-derived clade-specific repeats vary significantly in their accumulation depending on the density of ancient repeats. This provides further support for the argument that certain regions of the genome differ in their retrotransposon content depending on the age of the retrotransposons. Retrotransposon insertion bias can also be used as a means of determining if SINEs are mobilized/inserted in the male germline or female germline. Overrepresentation on the Y chromosome and under representation on the X chromosome are characteristic

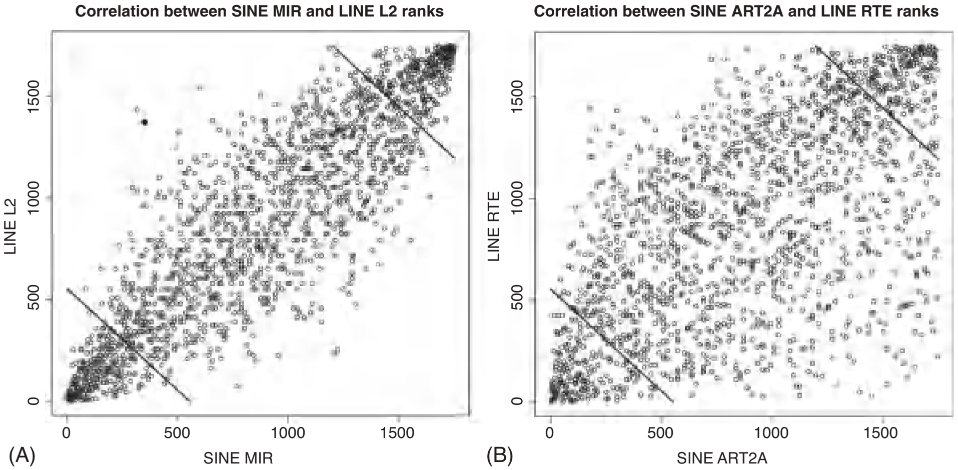


**Figure 10.6** Rank correlations for ancestral and recent LINE/SINE pairs. (A) Ranks of ancestral LINE L2 and SINE MIR counts for each 1.5 Mbp bin. (B) Ranks of recent LINE RTE (BovB) and SINE ART2A counts for each 1.5 Mbp bin. Lines in the upper right and lower left corners indicate the cutoff for the high- and low-density bins, respectively, and are based on the expected 5% tails from the random distribution of the sum of the ranks.

of male germline retrotransposon mobilization and this is the rule for primates (Jurka et al. 2002). This is also the case for cattle, with an X chromosome to autosome SINE density ratio of ~0.89 (Adelson, unpublished data).

## Genome Territories

The concept of genome territories is not new (Cremer et al. 2006), but has largely been explored with respect to cell-type-specific repositioning of nuclear chromosome territories during development, either with respect to gene expression or gene density (Kupper et al. 2007). In addition, regional variation in G+C content gives rise to isochores (Gardiner 1996), and the positive correlation of G+C content with gene density (Federico et al. 2000) has prompted speculation on the significance of such regional variation and its origins. We have also observed positive correlations, not only of gene density with G+C content, but of G+C content with both SSR and interspersed repeats, including LINE L2 and SINE MIR (Figure 10.5). While the significance of these correlations is unclear, it is probable that isochores arise as a result of biased gene conversion associated with recombination (Duret and Arndt 2008) and, therefore, that the correlations we observe with G+C content are unlikely to be causally related.

We identified another type of genome territory by plotting the locations of the bins in the high rank tail from Figure 10.6 (Adelson et al. 2009). We also carried out a similar process with the positively correlated LINE RTE/SINE ART2A pair and plotted the locations of both high and low rank tails on the bovine assembly (Figure 10.8).

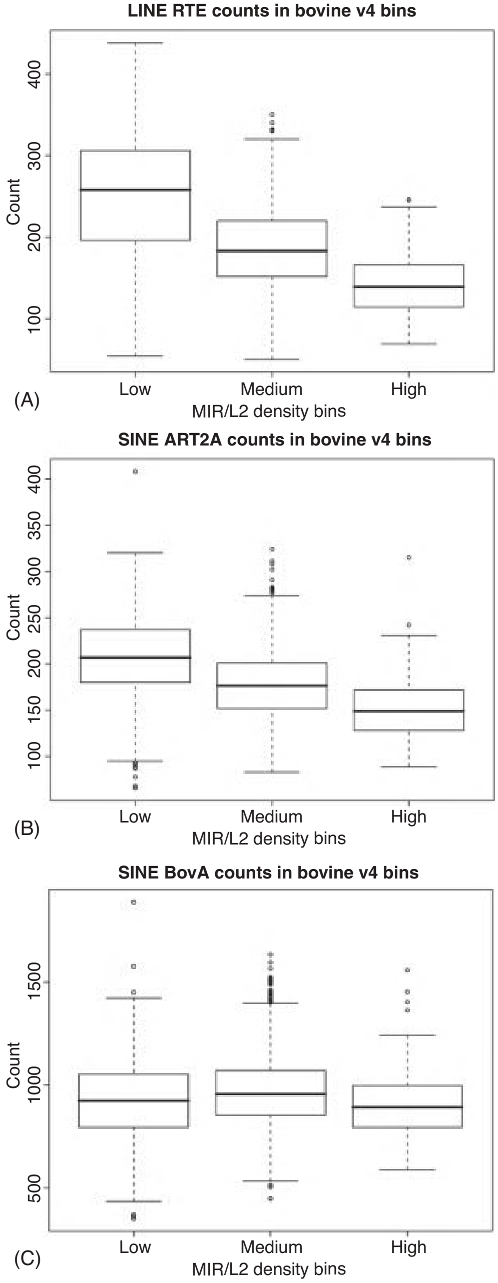


**Figure 10.7** Distribution of clade-specific repeats as a function of ancestral repeat density. (A) LINE RTE: high vs. low, $p$-value $<$ 2.2e-16; high/low vs. medium, $p$-value $=$ 0.5813. (B) ART2A: high vs. low, $p$-value $<$ 2.2e-16; high/low vs. medium, $p$-value $=$ 0.8267. (C) BovA: high vs. low, $p$-value $=$ 0.1787; high/low vs. medium, $p$-value $=$ 1.877e-08. The bins from Figure 10.6A were classified as having low, medium, or high MIR/L2 density. The cutoff between the groups was the 2-tail 10% significance level cutoff for the sum of the MIR and L2 ranks. For the LINE RTE, SINE ART2A, and SINE BovA repeat groups, the statistical package R was used to generate box plots of the number of repeats in the bins in each of the MIR/L2 density categories, and perform Wilcoxon rank sum tests with continuity correction between the high- and low-density groups, and between the medium group and the high and low groups combined, to test for linear and quadratic trends, respectively.

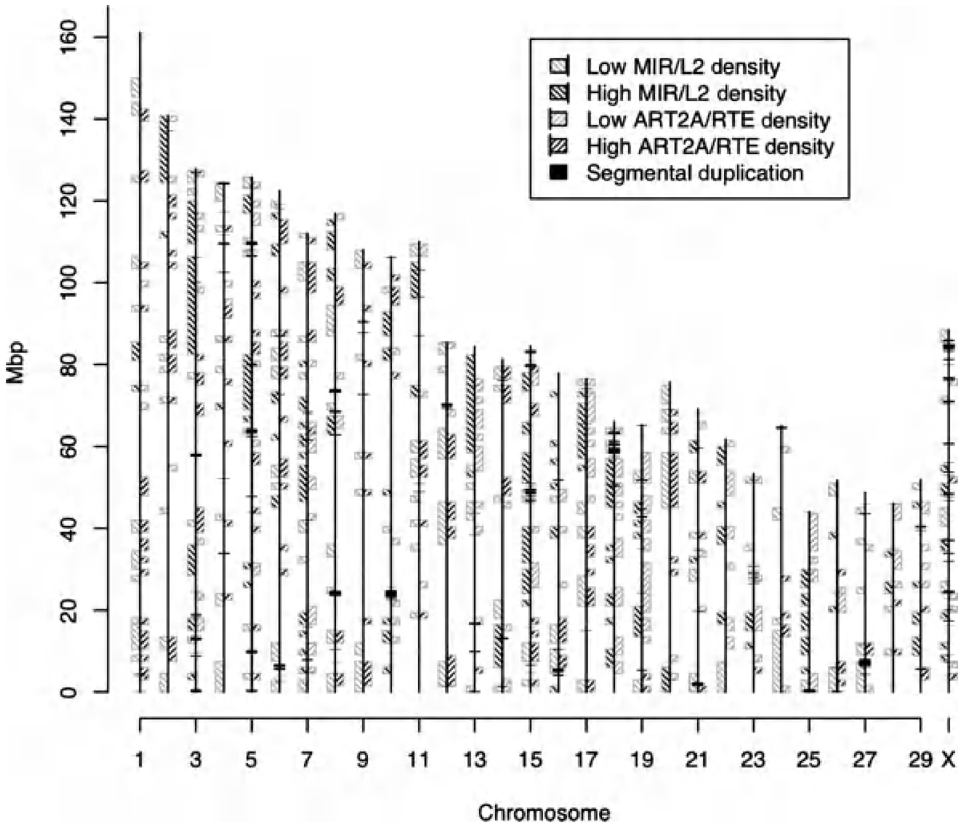


**Figure 10.8** Ancestral and new repeat groups define different genomic territories. Locations of 1.5 Mbp bins with extreme (high and low) ancestral (L2/MIR) and recent RTE/ART2A repeat densities are shown on the Btau_4 assembly, along with segmental duplications (SDs). Ancestral repeats tend to occur in blocks, while recent repeats generally do not. There is no overlap between high-density ancestral repeat blocks and high-density recent repeat blocks. SDs do not appear to colocalize with either high- or low-density regions of either repeat class.

It is apparent from Figure 10.8 that bins from the high rank tail of LINE L2/SINE MIR tend to cluster into larger blocks, while the other extreme bins do not do so. We have dubbed these clusters ancestral genome territories based on their repeat content. Furthermore, close inspection of Figure 10.8 also reveals that the high rank LINE RTE/SINE ART2A bins never overlap with the ancestral genome territories.

The LINE L2/SINE MIR correlation is also the strongest interspersed repeat correlation in the horse genome (Adelson et al. 2010) and the strongest in other eutheria (Adelson, unpublished data). In order to determine if the bovine ancestral genome territories were conserved, we plotted them against human ancestral territories aligned against the bovine genome (Figure 10.9).

We found that the ancestral genome territories are largely conserved, with ~80% of the bovine ancestral genome territories overlapping with the aligned human ancestral genome territories. Furthermore, we also observed the same 80% overlap between

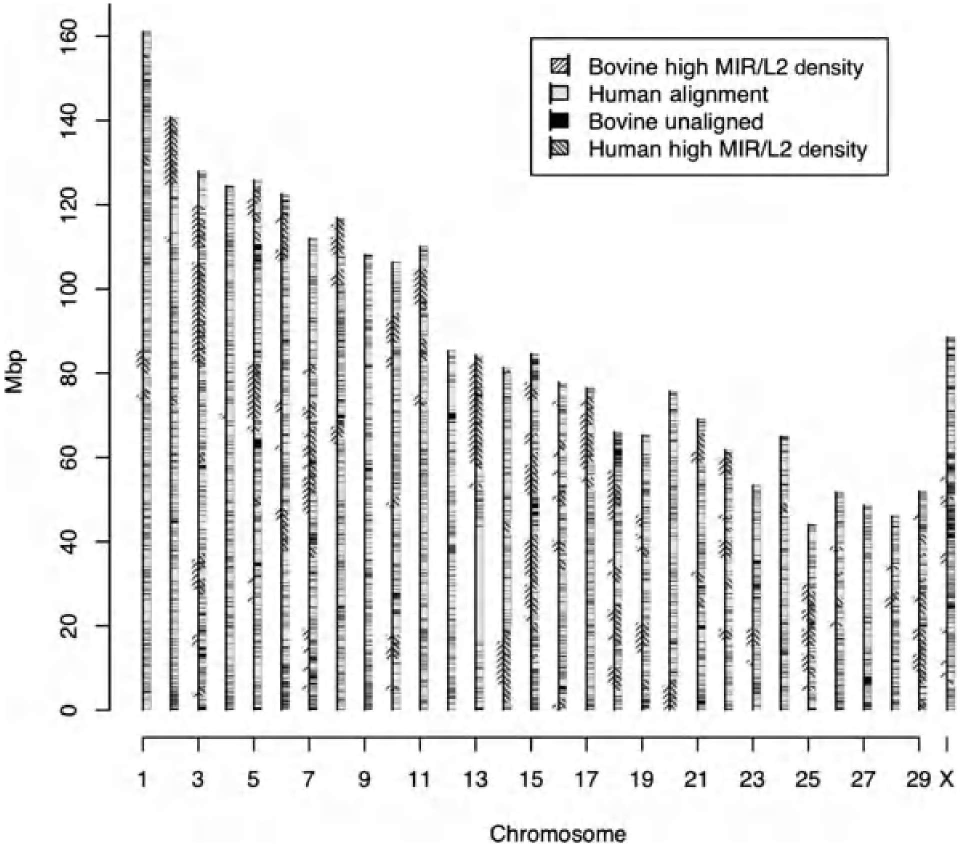


**Figure 10.9** Ancestral repeat domains are evolutionarily conserved. Ancestral (L2/MIR) high-density bins for bovine and human are shown on the bovine assembly, along with the overall bovine/human alignment. Note that the "top" of the chromosomes corresponds to the end near the x-axis on our plot. The y-axis corresponds to nucleotide coordinates in mega base pairs (Mbp) from the bovine assembly Btau_4.

equine ancestral genome territories and human ancestral genome territories (Adelson et al. 2010). The existence of conserved ancestral genome territories based solely on noncoding genome features is a somewhat surprising result.

Fossil retrotransposon densities (L2 and MIR), therefore, appear to define a general feature of mammalian genomes, namely conserved, syntenic ancestral genome domains. Because L2 and MIR have been inactive since the mammalian radiation, the persistence of such domains can only be explained by two alternate scenarios: (1) negative selection that preserved ancestral territories or (2) protection from new retrotransposition events. There is evidence that many LINE L2 and SINE MIR have undergone strong negative selection because they have been co-opted to regulate gene expression (Silva et al. 2003; Lowe et al. 2007). This suggests that the conserved ancestral repeat-enriched genome territories we have discovered are the result of purifying selection or of chromatin structural constraints and are probably of functional significance.

## Conclusions

The architecture of the bovine genome has probably been influenced by lateral transfer of LINE RTE from squamata, which has led to very different noncoding DNA compared to nonruminant eutheria. Furthermore, the presence of two independent, active LINE families makes the bovine genome a useful system for studying genome evolution in eutheria.

Elements that influence bovine genome structural variation, such as SD, CNVR, and retrotransposon insertion site polymorphism may be important determinants of phenotypic variation. At present, these types of polymorphism are not assessed by commercial genotyping platforms for cattle. Future progress in mapping traits of economic importance may depend on new genotyping technologies that can detect these types of polymorphism.

## References

Adelson, D.L., Raison, J.M., Edgar, R.C. (2009) Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 12855–12860.

Adelson, D.L., Raison, J.M., Garber, M., Edgar, R.C. (2010) Interspersed Repeats in the horse (Equus caballus); spatial correlations highlight conserved chromosomal domains. *Animal Genetics* **41**: 91–99.

Almeida, L.M., Silva, I.T., Silva, W.A., Jr., Castro, J.P., Riggs, P.K., Carareto, C.M., Amaral, M.E. (2007) The contribution of transposable elements to Bos taurus gene structure. *Gene* **390**: 180–189.

Babushok, D.V., Ohshima, K., Ostertag, E.M., Chen, X., Wang, Y., Mandal, P.K., Okada, N., Abrams, C.S., Kazazian, H.H., Jr. (2007) A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. *Genome Research* **17**: 1129–1138.

Baertsch, R., Diekhans, M., Kent, W.J., Haussler, D., Brosius, J. (2008) Retrocopy contributions to the evolution of the human genome. *BMC Genomics* **9**: 466.

Bailey, J.A. and Eichler, E.E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews* **7**: 552–564.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., Eichler, E.E. (2002) Recent segmental duplications in the human genome. *Science* **297**: 1003–1007.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., Moran, J.V. (2010) LINE-1 retrotransposition activity in human genomes. *Cell* **141**: 1159–1170.

Buckland, R.A. (1985) Sequence and evolution of related bovine and caprine satellite DNAs. Identification of a short DNA sequence potentially involved in satellite DNA amplification. *Journal of Molecular Biology* **186**: 25–30.

Cordaux, R. and Batzer, M.A. (2009) The impact of retrotransposons on human genome evolution. *Nature Reviews* **10**: 691–703.

Cremer, T., Cremer, M., Dietzel, S., Muller, S., Solovei, I., Fakan, S. (2006) Chromosome territories–a functional nuclear landscape. *Current Opinion in Cell Biology* **18**: 307–316.

D'Aiuto, L., Barsanti, P., Mauro, S., Cserpan, I., Lanave, C., Ciccarese, S. (1997) Physical relationship between satellite I and II DNA in centromeric regions of sheep chromosomes. *Chromosome Research* **5**: 375–381.

Desmarais, E., Belkhir, K., Garza, J.C., Bonhomme, F. (2006) Local mutagenic impact of insertions of LTR retrotransposons on the mouse genome. *Journal of Molecular Evolution* **63**: 662–675.

Duret, L. and Arndt, P.F. (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genetics* **4**: e1000071.

Edgar, R.C. and Myers, E.W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics* **21**: I152–I158.

Elsik, C.G. et al. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522–528.

Emanuel, B.S. and Shaikh, T.H. (2001) Segmental duplications: an 'expanding' role in genomic instability and disease. *Nature Reviews* **2**: 791–800.

Federico, C., Andreozzi, L., Saccone, S., Bernardi, G. (2000) Gene density in the Giemsa bands of human chromosomes. *Chromosome Research* **8**: 737–746.

Fisher, R.A. (1930) *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.

Gardiner, K. (1996) Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends in Genetics* **12**: 519–524.

Gentles, A.J., Wakefield, M.J., Kohany, O., Gu, W., Batzer, M.A., Pollock, D.D., Jurka, J. (2007) Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. *Genome Research* **17**: 992–1004.

Gibbs, R.A. et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**: 493–521.

Goidts, V., Cooper, D.N., Armengol, L., Schempp, W., Conroy, J., Estivill, X., Nowak, N., Hameister, H., Kehrer-Sawatzki, H. (2006) Complex patterns of copy number variation at sites of segmental duplications: an important category of structural variation in the human genome. *Human Genetics* **120**: 270–284.

Harris, R. and Riemer, C. (2010) LASTZ (http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html)

Hazkani-Covo, E., Zeller, R.M., Martin, W. (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genetics* **6**: e1000834.

Huang, C.R., et al. (2010) Mobile interspersed repeats are major structural variants in the human genome. *Cell* **141**: 1171–1182.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., Lee, C. (2004) Detection of large-scale variation in the human genome. *Nature Genetics* **36**: 949–951.

Inoue, K. and Lupski, J.R. (2002) Molecular mechanisms for genomic disorders. *Annual Review of Genomics and Human Genetics* **3**: 199–242.

Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends in Genetics* **16**: 418–420.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**: 462–467.

Jurka, J., Kapitonov, V.V., Kohany, O., Jurka, M.V. (2007) Repetitive sequences in complex genomes: Structure and evolution. *Annual Review of Genomics and Human Genetics* **8**: 241–259.

Jurka, J., Krnjajic, M., Kapitonov, V.V., Stenger, J.E., Kokhanyy, O. (2002) Active Alu elements are passed primarily through paternal germlines. *Theoretical Population Biology* **61**: 519–530.

Kazazian, H.H., Jr. (1999) An estimated frequency of endogenous insertional mutations in humans. *Nature Genetics* **22**: 130.

Korbel, J.O., et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.

Kordis, D. and Gubensek, F. (1998) Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 10704–10709.

Kordis, D. and Gubensek, F. (1999) Horizontal transfer of non-LTR retrotransposons in vertebrates. *Genetica* **107**: 121–128.

Kupper, K., et al. (2007) Radial chromatin positioning is shaped by local gene density, not by gene expression. *Chromosoma* **116**: 285–306.

Lander, E.S. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lenstra, J.A., van Boxtel, J.A., Zwaagstra, K.A., Schwerin, M. (1993) Short interspersed nuclear element (SINE) sequences of the Bovidae. *Animal Genetics* **24**: 33–39.

Lindblad-Toh, K., et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803–819.

Liu, G.E., Ventura, M., Cellamare, A., Chen, L., Cheng, Z., Zhu, B., Li, C., Song, J., Eichler, E.E. (2009) Analysis of recent segmental duplications in the bovine genome. *BMC Genomics* **10**: 571.

Liu, G.E., et al. (2010) Analysis of copy number variations among diverse cattle breeds. *Genome Research* **20**: 693–703.

Lowe, C.B., Bejerano, G., Haussler, D. (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 8005–8010.

Lupski, J.R. and Stankiewicz, P. (2005) Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genetics* **1**: e49.

Marques-Bonet, T. and Eichler, E.E. (2009) The evolution of human segmental duplications and the core duplicon hypothesis. *Cold Spring Harbor Symposia on Quantitative Biology* **74**: 355–362.

Metzker, M.L. et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**: 493–521.

Mikkelsen, T. et al. (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**: 167–177.

Pace, J.K., 2nd, Sen, S.K., Batzer, M.A., Feschotte, C. (2009) Repair-mediated duplication by capture of proximal chromosomal DNA has shaped vertebrate genome evolution. *PLoS Genetics* **5**: e1000469.

Price, A.L., Jones, N.C., Pevzner, P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics* **21**: I351–I358.

Salih, H. and Adelson, D.L. (2009) QTL global meta-analysis: are trait determining genes clustered? *BMC Genomics* **10**: 184.

Sharp, A.J., et al. (2005) Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics* **77**: 78–88.

Silva, J.C., Shabalina, S.A., Harris, D.G., Spouge, J.L., Kondrashovi, A.S. (2003) Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genetics Research* **82**: 1–18.

Stranger, B.E., et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**: 848–853.

Takahashi, I., Nobukuni, T., Ohmori, H., Kobayashi, M., Tanaka, S., Ohshima, K., Okada, N., Masui, T., Hashimoto, K., Iwashita, S. (1998) Existence of a bovine LINE repetitive insert that appears in the cDNA of bovine protein BCNT in ruminant, but not in human, genomes. *Gene* **211**: 387–394.

Turner, C., Killoran, C., Thomas, N.S., Rosenberg, M., Chuzhanova, N.A., Johnston, J., Kemel, Y., Cooper, D.N., Biesecker, L.G. (2003) Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. *Human Genetics* **112**: 303–309.

Venter, J.C. et al. (2001) The sequence of the human genome. *Science* **291**: 1304–1351.

Wade, C.M., et al. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**: 865–867.

Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K., Haussler, D. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 18613–18618.

Waterston, R.H. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 520–562.

Xing, J., et al. (2009) Mobile elements create structural variation: analysis of a complete human genome. *Genome Research* **19**: 1516–1526.

Yamashita, T., Honda, M., Takatori, H., Nishino, R., Hoshino, N., Kaneko, S. (2004) Genome-wide transcriptome mapping analysis identifies organ-specific gene expression patterns along human chromosomes. *Genomics* **84**: 867–875.

# Chapter 11
# Bovine Epigenetics and Epigenomics

*Xiuchun (Cindy) Tian*

## Definitions of Epigenetics and Epigenomics

The term "epi-" ($\varepsilon\pi\acute{\iota}$- in Greek), means "over or above" and suggests that epigenetics is different from inheritable genetic phenomena. It is defined as the study of transmittable changes in phenotype or gene expression caused by mechanisms other than changes in the underlying DNA sequences. These changes continue to be manifested in subsequent cell divisions for the remainder of the cell's life. In 1940, Waddington used the famous epigenetic landscape model (Figure 11.1) to describe how gene regulation modulates development (Waddington 1940). It is generally believed that epigenetic changes do not pass from one generation to the next but are transmittable from the mother cell to the daughter cells. However, some epigenetic aberrations caused by deleterious environmental effects or malnutrition have been reported to last for several generations. Epigenomics is the study of all epigenetic elements in a particular cell or tissue. It is a relatively new field made possible by the development of next-generation sequencing technologies.

## Mechanisms of Epigenetics

The exact molecular basis for epigenetics has been intensively researched. Several mechanisms have been discovered including modifications of DNA and histones, chromatin remodeling, noncoding RNA as well as others. Among these, DNA methylation and histone methylation/acetylation are the most understood.

### *DNA Methylation and CpG Islands*

In mammalian chromosomes, the cytosine residue (C) 5′ to a guanine residue (G) can be methylated by DNA methyl transferases and becomes 5-methyl-2′-deoxycytidine ($5^{m}$C), and the CG dinucleotides are often referred to as CpG (p = phosphate). Using high performance liquid chromatography (HPLC), the overall $5^{m}$C/C in the bovine genome was found to be 3%–5% (Hiendleder et al. 2004; Sandhu et al. 2009). The

*Bovine Genomics*, First Edition. Edited by James E. Womack.

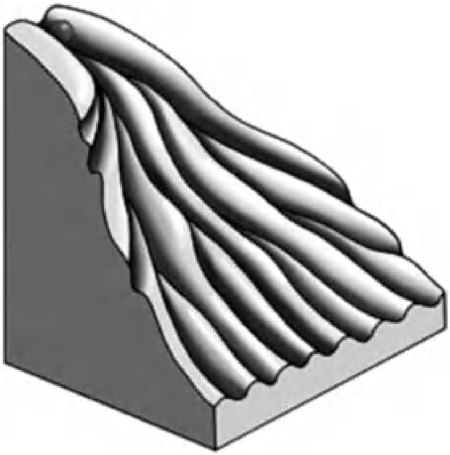


**Figure 11.1** The landscape model of mammalian cell differentiation (modified from Keeton and Gould 1984). The process of mammalian cell differentiation is described as a ball rolling down a hill with many valleys. When the ball is on the top of the hill, it can roll down through any valleys below; this represents the process of a totipotent cell that can differentiate into any tissue of the body. However, as the ball rolls past an intersection, the available valleys for the ball to roll down become limited. When the ball reaches to the bottom of the hill, it can no longer move to another valley or back to the top of the hill. This model was used to illustrate a totipotent cell choosing among different developmental paths; when the cell's fate is partially determined, its differentiation potential becomes limited. Once the cell is terminally differentiated, it cannot trans-differentiate into another cell type or become totipotent again.

bovine placenta is more hypomethylated with approximately 2% of $5^{m}C/C$. These numbers are comparable to those found in other mammalian species such as the human and mouse.

A deficit for CpG dinucleotides exists in mammalian genomes. This low occurrence of CpG is believed to have been caused by the spontaneous mutation of methylated CpGs to TpGs. In cattle, the G+C content is approximately 42% and the ratio of observed over the expected CpGs ($Obs_{CpG}/Exp_{CpG}$) is approximately 0.24 (Han et al. 2008; Table 11.1). This means that CpG dinucleotides constitute roughly 1% of the

**Table 11.1** Comparisons of CpG islands of cattle to three other mammalian species.

| Species | Genome size (Gb) | GC content (%) | $Obs_{CpG}/Exp_{CpG}$ | CpG island (no. of CpG island) | CpG island density (/Mb) | GC content (%) | $Obs_{CpG}/Exp_{CpG}$ |
|---|---|---|---|---|---|---|---|
| Mouse | 2.48 | 41.7 | 0.192 | 20,458 | 8.2 | 60.6 | 0.756 |
| Dog | 2.31 | 41.0 | 0.244 | 58,327 | 25.3 | 62.2 | 0.753 |
| Human | 2.85 | 40.9 | 0.236 | 37,531 | 13.2 | 62.0 | 0.743 |
| Cattle | 2.29 | 41.9 | 0.236 | 36,729 | 16.0 | 61.2 | 0.740 |

Adapted from Han et al. (2008).

cattle genome instead of the expected 6.25% among all the 16 possible dinucleotides combinations. Of all the CpGs, approximately 70%–80% are methylated.

CpG dinucleotides often exist in clusters in the genome. The term "CpG island" was developed to describe DNA elements of at least 500 bp (base pairs) that contain clusters of CpG dinucleotides. These fragments also have 50% or higher C+G content and a ratio of 0.60 or higher for $Obs_{CpG}/Exp_{CpG}$ (Takai and Jones 2002). By comparing ten sequenced mammalian genomes, Han et al. (2008) reported that CpG islands are poorly conserved among species in both number and density, and even these genomes may encode similar number of genes. Interestingly, cattle and humans are very similar in the number of CpG islands and C+G content in both the entire genome and in CpG islands. They are also similar in the ratio of $Obs_{CpG}/Exp_{CpG}$ (Table 11.1). Rodents and dogs, however, are either much lower or higher in these parameters.

CpG islands are present in the promoter regions of most housekeeping and imprinted genes. These CpGs are mostly hypomethylated. DNA methylation is maintained during cell replication by the primary DNA methyl transferase, *DNMT1*, which is responsible for transferring methyl groups to semimethylated, newly synthesized DNA strands (Figure 11.2A). Unless otherwise stated, DNA methylation in literature and also in this chapter refers to the methylation of the cytosine residue. Methylation can also occur in the adenine residue, but it is a relatively rare occurrence and its role in cellular function, if any, is still poorly defined. In addition to *DNMT1*, which is a maintenance methyl transferase, *DNMT3a* and *DNMT3b* are de novo methyl transferases that set up DNA methylation patterns early in embryonic development.

CpG dinucleotides that are methylated are usually found in the repetitive regions such as centromeres. Methylation of DNA plays a critical role in transcriptional regulations. Unmethylated or hypomethylated CpG islands in promoters are often

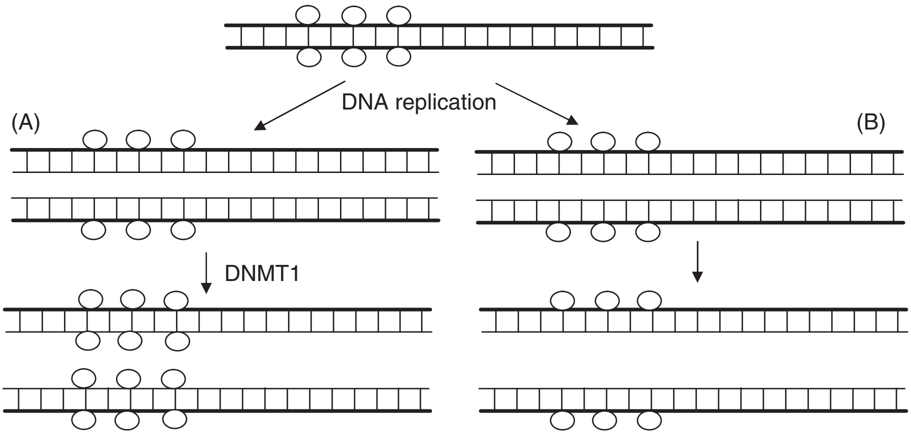


**Figure 11.2** Schematic illustrations of DNA methylation maintenance and passive demethylation. During replication of methylated (circles = methylation or $-CH_3$) DNA, the newly synthesized strand (thin line) of the DNA will be methylated by DNA methyl transferase 1 (DNMT1) in the hemimethylated DNA according to the template of the mother strand (thick line, A). In the absence of the functional DNMT1, the newly synthesized DNA strand will remain unmethylated (B). In the next cycles of DNA replication, the DNA will be unmethylated, hence the term "passive demethylation."

associated with active transcription. Conversely, hypermethylated DNA suppresses transcription. A good example is the silencing of parasitic retrotransposons and viral DNA insertions in the mammalian genome by methylation. DNA methylation is, therefore, a defense mechanism of the genome to repress the expression of foreign DNA insertions. Due to its role in gene expression regulation, it is easy to understand that DNA methylation is involved in many processes during development when it is necessary to turn off a specific subset of genes. Methylated DNA, for example, is involved in maintaining the silencing of a specific parental allele in genomic imprinting (see Section "Genomic Imprinting"). The mechanisms by which DNA methylation impacts gene expression may be twofold. First, the methyl group may itself physically impede the binding of transcriptional proteins to the gene, thus blocking transcription. Second, and perhaps more importantly, methylated DNA may be bound by proteins known as Methyl-CpG-binding domain proteins. These proteins recruit additional factors such as histone deacetylases (HDACs) and other chromatin-remodeling proteins that modify histones, thereby forming compact, inactive chromatin.

DNA methylation is a dynamic event. The most dramatic changes occur during early embryo development and tissue differentiation, when specific regions of DNA are targeted by methylation through mechanisms yet to be understood. It is believed that once established, these specific methylation patterns remain fairly stable. Loss of DNA methylation in a nonspecific fashion, however, has been reported during aging and long-term cell culture. Kang et al. (2001b), for example, reported that after long periods of culture, DNA methylation in bovine fetal fibroblast cells was reduced at euchromatic repeats as well as protein-coding genes such as cytokeratin, L-lactoglobulin, and interleukin 13.

DNA methylation is generally not inherited by the next generation because epigenetic signals are erased and reestablished in gametes. This is why regulations by DNA methylation are considered epigenetic, but not genetic. Environmental genotoxins and maternal dietary constraints, such as methionine deficiency during pregnancy (lack of substrate for DNA methylation), have been shown to affect epigenetics of future generations (reviewed by Skinner and Guerrero-Bosagna 2009; Burdge and Lillycrop 2010;). Pharmaceutical reagents, such as 5-aza-2′-deoxycytidine, an inhibitor of DNA methyl transferase, have been widely used to study the functions of DNA methylation.

Numerous methods have been developed to identify the location of methylated CpG and to quantify the levels of DNA methylation. Different conclusions can be reached when analyzing the same material using different methods. Among the frequently used approaches, immunostaining of DNA methylation, HPLC, and bisulfite sequencing with nonspecific primers can be used to study global DNA methylation including methylation of mainly repetitive sequences. Bisulfite sequencing with specific primers, DNA methylation profiling, and next-generation sequencing can generate methylation information of specific CpG islands.

## *Histone Modifications: Acetylation, Methylation, and Histone Variants*

The core histone molecules, H2A, H2B, H3, and H4, can be modified throughout their entire sequences. The unstructured N-termini of histones, the histone tails, are highly modified in particular. These modifications include acetylation, methylation, ubiquitination, phosphorylation, and poly-ADP-ribosylation. Approximately 60 different residues in core histones have been shown to be modified. The total number of the combinations of these modifications in a nucleosome is so large that it was compared to the number of genes in the mammalian genome (Kerppola 2009).

Acetylation and methylation of histones have been intensely studied. Lysine residues of histones can be modified for acetylation by specific histone acetyl transferases (HATs) and HDACs. On histone H3, four lysine (K) residues, K9, K14, K18, and K56, can be acetylated. On histone H4, these are K5, K8, K13, and K16 (Kouzarides and Berger 2006). The lysines of histones can also be methylated by posttranslational modifications. Five lysine residues on H3, K4, K9, K27, K36, and K79, and one residue on H4, K20, can be methylated by the relevant histone methyl transferases, and demethylated by histone demethylases (Kouzarides and Berger 2006).

Histone modifications are associated with gene expression regulation, DNA replication, and DNA recombination in a systematic and reproducible way. The term "histone code" has been used to describe how histone modifications affect these cellular functions. It is important to note that the same type of modifications at different locations of the histone molecule can induce great variations in a histone's association with the DNA molecule and thus produce dramatically different effects on transcription. Additionally, multiple modifications may occur at the same amino acid residue, and these modifications may work together to change the behavior of the nucleosome. Among all forms of histone modifications, the role of acetylation is the best understood. For instance, acetylation at K14 and K9 of the tail of histone H3 is generally correlated with active RNA transcriptions.

Direct evidence that histone acetylation upregulates gene expression also exists. For example, Sakurai et al. (2009) reported that in ruminant ungulates the interferon tao gene (*IFNT*) is expressed only by the mononuclear trophectoderm cells. This is associated with higher histone K3K18 acetylation and lower histone H3K9 methylation in these cells. Treatment of cells that do not normally express *IFNT* with an HDAC inhibitor, trichostatin A, partially induced *IFNT* expression.

The roles of histone methylation, however, are more complex. It is generally believed that methylated H3K4, H3K36, and H3K79 are associated with active transcription, while methylated H3K9, H3K27, and H4K20 are associated with a transcriptionally inactive state (Briggs et al. 2001). Moreover, it was demonstrated that H3K9 methylation is mechanistically linked to DNA methylation (Soppe et al. 2002). This is crucial for heterochromatin assembly and specific binding of heterochromatin protein 1 (Bannister et al. 2001; Lachner et al. 2001).

Similar to DNA methylation, the modifications of histones also appear to be a dynamic event. A good example in the bovine is the change of acetylated histone (H4K5) in bovine fetal fibroblast cells during the cell cycle. During interphase, the immunostain of acetylated H4K5 is distributed throughout the entire nucleus. During mitosis, acetylated H4K5 stain appears to concentrate around the chromosomes but is absent from them. The level of H4K5 stain is lower from early prometaphase to late telophase than in the interphase, possibly caused by the reduction of acetylation of H4 (Wee et al. 2006).

In addition to the posttranslational modifications to each of the five major histone types (H1 and core histones), variants of histones exist, which can also be modified posttranslationally. Histone variants can be classified as homomorphous and heteromorphous, depending on how much their sequences vary from the main canonical isoforms (Ausió 2006). The homomorphous variants are those with slight amino acid alterations such as H2A.1 and H2A.2 and H3.1, H3.2, and H3.3, while the heteromorphous variants have larger changes from the histone molecule such as H2A.X, H2A.Z, macroH2A (mH2A), H2A Barr body-deficient, and centromeric protein A. The roles of these variants range from destabilizing the histone octamer, maintaining chromosome integrity, chromatin remodeling (thus transcription regulation), to specific roles in defining the nucleosome structure of the centromeres (Ausió 2006). Perhaps, the most intensely studied isoform is macrohistone H2A (mH2A). It has an N-terminal region with high sequence homology to H2A but it also contains a 25-kDa nonhistone macrodomain of unknown function (Aravind 2000). MacroH2A is a chromatin silencer and was first identified by its occurrence in inactivated X chromosome ($X_i$) of mammalian females (see Section "X-Chromosome Inactivation").

## *Noncoding RNA*

In addition to chromatin modifications, many other forms of epigenetic regulations exist in the cell. One of these is noncoding RNAs (ncRNAs). Most genomes studied to date undergo widespread transcription. The majority of the transcripts, however, are not translated into proteins (see review by Nagano and Fraser 2009). ncRNAs are involved in gene expression regulation at multiple stages. For example, they are directly involved in protein synthesis, RNA maturation and transport, and in gene silencing through regulating chromatin structure and mRNA degradation. These regulatory roles are carried out through either base-pairing or nonbase-pairing mechanisms. Short ncRNAs, such as small interfering RNA (siRNAs), micro RNA (miRNAs), and piwi-interacting RNAs (piRNAs), are highly conserved at the sequence level and inhibit gene expression through specific base pairing with their targets. Long ncRNAs (lncRNAs), ranging in size from 50 kb to several hundred kb, are poorly conserved and regulate transcriptional silencing ranging from a single gene to an entire chromosome. They mediate the silencing of multiple genes in cis, despite lacking homology with their target genes (Mohammad et al. 2009).

It has been recognized that lncRNAs, such as *XIST, AIR* (antisense *IGF2R* RNA), and *KCNQ1OT1* (*KCNQ1* overlapping transcript 1; *KCNQ1* = potassium voltage-gated channel, *KQT*-like subfamily, member 1), play a functional role as organizers of chromatin structure and suppressor of gene activities. For example, *XIST* has been found to coat or paint the entire X chromosome as an early event in its inactivation (Clemson et al. 1996). *AIR* is an imprinted gene expressed from the paternal allele. It silences the paternal allele of *IGF2R*, possibly by binding to the chromatin at the paternal *IGF2R* and adjacent gene loci. The proximity of lncRNAs and the genes they silence on the chromosomes suggest a common mechanism in this action. It is proposed that these lncRNAs interact with chromatin by associating with histones modified for gene silencing (Nagano and Fraser 2009). For example, *KCNQ1OT1* was colocalized with trimethylated H3K9 and H3K27, which are known hallmarks of gene suppression.

Because of their effect in gene expression regulation, it has been suggested that lncRNAs play a central role in tissue differentiation during which coordinated activation and repression of specific subsets of genes occur.

## *Chromatin/Chromosome Remodeling*

Modifications to chromatin architecture are well-known epigenetic mechanisms for transcriptional activation and silencing. Chromatin structural changes can be induced by posttranslational modifications of histone proteins, substitution with histone variants, remodeling of nucleosome positions and structures, alterations of chromatin compaction, and chromatin looping and folding. Many factors, such as CCCTC-binding factor (CTCF), Polycomb group proteins, and SWItch/Sucrose NonFermentable (SWI/SNF) complex, have been shown to affect chromatin structures. However, direct evidence of how specific chromatin structures are transmitted through cell division is lacking.

Polycomb group proteins control the expression of a variety of genes from early embryogenesis through birth and to adulthood. It is believed that they maintain transcription repression by catalyzing methylation of H3K27 (Schuettengruber et al. 2007), but such changes in histones are not always required. The mechanisms of gene-specific recruitment, transcription repression, and selective derepression of genes by the Polycomb group proteins are still largely not understood (Kerppola 2009). In cattle, members of Polycomb repressive complex 2 were expressed throughout preimplantation development (Ross et al. 2008), and their presence in the nucleus is associated with changes in trimethylated H3K27 in early bovine embryos and is implicated in maternal zygotic transition (Ruddock-D’Cruz et al. 2008).

CTCF is a highly conserved zinc finger protein that participates in diverse regulatory functions. In addition to transcriptional activation/repression, CTCF also serves as an enhancer-insulator. CTCF mediates the formation of intrachromosomal loops and interchromosomal contacts. As thousands of CTCF-binding sites have been identified in the mammalian genome, CTCF has been suggested to act as a master organizer of the genome. It has been proposed that CTCF plays a primary role in the formation of the complex chromatin web of interactions, allowing it to be transmitted through cell division (reviewed by Phillips and Corces 2009).

Chromatin remodeling by ATP-dependent mechanisms is another form of epigenetic regulation. In the bovine, ATP-dependent chromatin-remodeling factors are found to be involved in oocyte maturation (Wee et al. 2010). Inhibition of these activities with apyrase led to retarded chromatin remodeling in bovine oocytes and resulted in poor development of fertilized embryos.

It has been observed that location of a chromosome in the nucleus can also be an epigenetic property that is associated with gene expression. For example, gene-dense chromosomes are typically located more interior while gene-poor chromosomes are more peripheral. Using the most gene-rich and gene-poor chromosomes in cattle, chromosomes 19 and 20, respectively, Koehler et al. (2009) observed that the radial arrangements of these chromosomes were the same in embryos up to the 8-cell stage. At the 10- to 16-cell stage, chromosome 19 translocated significantly more internally while chromosome 20 more peripherally. These changes correspond to genomic activation in bovine embryos and the distribution patterns persisted to adulthood in all cell types (fibroblasts and lymphocytes) examined.

## Examples of Epigenetic Regulations: Genetic Imprinting and X-Chromosome Inactivation

Epigenetics play an important role in three major processes during fetal development: (1) genomic imprinting, (2) X-chromosome inactivation (XCI) in females, and (3) tissue differentiation. It has also been shown to be highly involved in the development of cancer where cells acquire abnormal methylation patterns on tumor suppressor genes. Two widely studied epigenetic phenomena, (1) genomic imprinting and (2) XCI, are discussed in this section with regard to cattle embryonic development.

### *Genomic Imprinting*

Both the maternal and paternal genomes are required for normal development. For the majority of genes in mammalian species including cattle, both the maternal and paternal alleles are expressed. However, for a small group of genes, one parental allele is preferentially or exclusively expressed (Figure 11.3). This is termed "genomic or genetic imprinting." The majority of imprinted genes have roles in fetal growth and development.

Genomic imprinting renders diploid mammals functional hemizygous at the imprinted loci. This is deleterious because recessive mutations, which decrease fitness and survivability, are easily revealed. However, imprinting has been maintained through millions of years of evolution, suggesting that it has some evolutionary advantages. To date, over a dozen theories have been postulated to account for the evolutionary advantage of genomic imprinting. One of these, "the conflict theory," is the most plausible (Moore and Haig 1991). It is based on the premise that female mammals mate with several partners in a lifetime, and this creates a genetic

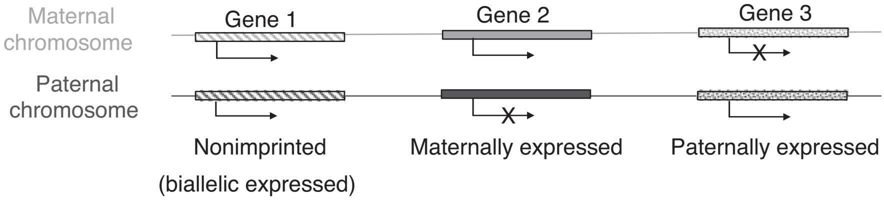


**Figure 11.3** Schematic illustration of expression patterns of nonimprinted vs. imprinted genes. **Gene 1** (hatched boxes) is nonimprinted and is expressed from both the maternal (grey line) and paternal chromosomes (black line), that is, biallelic expressed. **Gene 2** (solid boxes) is maternally expressed; it is only transcribed into mRNA (bent arrow) from the maternal copy of the gene (grey solid box), located on the maternal chromosome; while the same gene (black solid box) on the paternal chromosome is silenced (crossed bent arrow; monoallelic expression). **Gene 3** (dotted boxes) is paternally expressed.

contention between the male and female genomes. The conflict arises during pregnancy when imprinted genes affect fetal growth and nutrient acquisition. Paternally inherited genes are selected to extract the maximum resources possible from the mother to benefit his offspring's growth and fitness. Whereas maternally inherited genes are selected to conserve their resources in order to divide them between her current and future offspring and therefore, maximize her own reproductive potential. Consistent with this theory, paternally expressed genes tend to promote growth, while maternally expressed genes inhibit growth. Additionally, many genes are only imprinted in the placenta (Tycko and Morison 2002).

Parental-specific allelic expression of imprinted gene is regulated by distinct DNA elements that exhibit allele-specific epigenetic modifications, such as differential DNA methylation between the two parental alleles. There are normally CpG islands in or near imprinted genes. Furthermore, imprinting control regions and secondary differentially methylated regions are characterized by an overlapping pattern of H3K4 trimethylation (active chromatin) and H3K9 trimethylation (repressive chromatin) in bovine somatic tissue (Dindot et al. 2009), suggesting that histone modification is also involved in genomic imprinting.

### *Genomic Imprinting in Cattle*

To date, 143 and at least 63 imprinted genes, imprinted small nucleolar RNAs (snoRNAs), microRNAs (miRNAs) have been identified in the mouse (http://www.har.mrc.ac.uk/research/genomic_imprinting/) and human (http://igc.otago.ac.nz/home.html), respectively. Although there is a general belief that genomic imprinting is conserved among mammalian species, emerging evidence suggests that such an assumption is false. For example, only 39 of the previously mentioned genes are found to be imprinted in both mice and humans. A search of the literature yielded only 18 genes that have been confirmed to be imprinted in the bovine (Table 11.2). Many genes imprinted either in the mouse or human are not imprinted in cattle.

The time course of imprinting establishment has been best studied in the mouse: distinct allelic expression of imprinted genes is seen as early as the 2-cell stage, and by the blastocyst stage, monoallelic expression of most imprinted genes is observed. This time course closely resembles the reestablishment of genomewide DNA methylation in early development (Latham 1999; Monk and Salpekar 2001). The onset of monoallelic expression of imprinted genes in the bovine has not been systematically examined. Studies utilizing materials from scattered developmental stages showed that monoallelic expression pattern of confirmed imprinted genes in cattle was not exhibited by the blastocyst stage with the exception of the *XIST* gene (Cruz et al. 2008). When day-14 parthenogenetic embryos (containing only the maternal genome) and naturally fertilized embryos (containing both parental genomes) were compared, bovine genes *MAGEL2* and *MEST*, which have been confirmed to be imprinted in cattle, did not exhibit monoallelic expression (Tveden-Nyborg et al. 2008). At day 21, monoallelic expression was observed for bovine *MEST* (Tveden-Nyborg et al. 2008). Another early time point studied was day 17. At this stage, the *SNRPN* gene was exclusively paternally expressed (Suzuki et al. 2009) and this continued at day 40 of gestation in liver, muscle, and brain. Slight leaky expression of the silent maternal allele was seen in heart and placenta.

**Table 11.2** Genes confirmed to be imprinted in the bovine.

| Gene | Chromosome | Expressed allele | References |
|---|---|---|---|
| *PEG10* (Paternal expressed gene 10) | 04 | P | Khatib et al. 2007 |
| *MEST* (mesoderm-specific transcript, also *PEG1*) | 04 | P | Ruddock et al. 2004 |
| *NAP1L5* (nucleosome assembly protein 1-like 5) | 06 | P | Zaitoun and Khatib 2006 |
| *IGF2R* (Insulin-like growth factor receptor 2; or *M6PR*, mannose-6-phosphate receptor) | 09 | M | Killian et al. 2001; Long and Cai 2003; Suteevun-Phermthai et al. 2009 |
| *NESP55* (neuroendocrine secretory protein) | 13 | M | Khatib 2004 |
| *NNAT* (neuronatin) | 13 | P | Ruddock et al. 2004; Zaitoun and Khatib 2006 |
| *DGAT1* (acyl CoA-diacylglycerol-acyltransferase) | 14 | * | Kuehn et al. 2007 |
| *MIMT1* (mitochondrial import protein 1) | 18 | P | Kim et al. 2007 |
| *USP29* (ubiquitin-specific peptidase 29) | 18 | P | Kim et al. 2007 |
| *PEG3* (paternally expressed gene 3) | 18 | P | Kim et al. 2004 |
| *GTL2* (gene trap locus 2; or *MEG3*, maternal expressed gene 3) | 21 | M | Dindot et al. 2004 |
| *RTL1* (retrotransposon-like 1 or *PEG11*) | 21 | P | Khatib et al. 2007 |
| *MAGEL2* (melanoma antigen, family L, 2) | 21 | P | Khatib et al. 2007 |
| *SNRPN* (small nuclear ribonucleoprotein polypeptide N) | 21 | M | Lucifero et al. 2006 |
| *H19* (H19 fetal liver mRNA) | 29 | M | Zhang et al. 2004; Curchoe et al. 2009 |
| *IGF2* (insulin-like growth factor 2) | 29 | P | Dindot et al. 2004; Curchoe et al. 2005 |
| *TSSC4* (tumor suppressing subtransferable candidate 4) | 29 | M | Khatib et al. 2007 |
| *XIST* (X inactivation-specific transcript) | X | P | Dindot et al. 2004 |
| *MAOA* (monoamine oxidase type A) | X | M in placenta | Xue et al. 2002 |

P, paternal; M, maternal; *, parent-of-origin effect on milk production.

With the increasing number of genes confirmed to be imprinted in the bovine, genomic characteristics common to bovine imprinted genes were searched. Using 11 imprinted and control (nonimprinted) genes, Khatib et al. (2007) analyzed the occurrence of CpG islands, G+C content, tandem repeats, and retrotransposable elements. They found that bovine imprinted genes have a higher G+C content, more CpG islands, and tandem repeats than nonimprinted genes. Fewer short interspersed nuclear elements (SINEs) were located in imprinted cattle genes than control genes, consistent with findings in humans and mice. Long interspersed nuclear elements (LINEs) and long terminal repeats, however, were found to be significantly underrepresented in imprinted genes compared to controls, contrary to findings in humans and mice. Interestingly, highly conserved tandem repeats in nine of the genes imprinted in all three species were identified in this study, suggesting conservation of epigenetic regulatory mechanisms for the allelic-specific expression of imprinted genes in these species.

### *Epigenetic Status of Bovine Imprinted Genes*

Few imprinted genes have been characterized for the methylation status on their CpG islands. Direct evidence that DNA methylation and histone acetylation regulate the monoallelic expression of these genes has not been reported. This section contains all available data published to date on the characterization of DNA methylation in bovine imprinted genes.

Epigenetic mechanisms have been shown to be involved in the expression level of *IGF2R* in cattle (Long and Cai 2007). Treatment of cattle cell lines with inhibitors of DNA methylation (5-aza-2′-deoxycytidine) and histone deacetylation (trichostatin A) caused an increase and decrease, respectively, in the level of *IGF2R* expression. A CpG island was found in intron 2 of the bovine *IGF2R* gene, conservative with the mouse where it was found to be the imprinting control element. This putative imprinting control region is nearly completely unmethylated in the sperm and has significant variation in DNA methylation in blood, liver, brain, and heart, suggesting that *IGF2R* imprinting in cattle may be tissue-specific.

The DNA methylation status of four bovine imprinted genes, (1) *PEG3,* (2) *XIST,* (3) *PEG10*, and (4) *MAOA* was studied in the skin and sperm of adult cattle (Liu et al. 2008b). The CpG islands in *PEG3*, *PEG10*, and *XIST* were monoallelically methylated (a 50:50 ratio of methylated vs. unmethylated DNA strands; Figure 11.4). The fact that the sperm DNA was completely unmethyated in these regions suggests that differential methylation may be involved in the allelic expression of these genes in bovine somatic cells.

The paternally expressed *IGF2* is the first imprinted genes identified and its imprinting status has been found to be conserved in all mammalian species studied to date. In cattle, Gebert et al. (2006) identified a CpG-rich region in exon 10 of bovine *IGF2* that was differentially methylated in mature oocytes and sperm. Furthermore, they (Gebert et al. 2009) reported that methylation signals from the silenced maternal allele was removed from this intragenic CpG-rich region after fertilization, but partially replaced by the time the embryo reached blastocyst stage. This pattern of DNA methylation changed by midgestation. At day 130 of gestation, the bovine *IGF2* exon 10 was found to be mainly hypermethylated in adrenal, kidney, and liver (Couldrey and Lee 2010). It is possible that at this stage DNA methylation at exon 10 is no longer

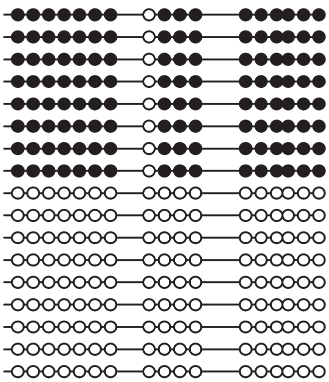


**Figure 11.4** Schematic illustration of DNA methylation in the intergenic region of the bovine H19 and IGF2 in the liver of a newborn calf (Curchoe et al. 2009). Each line represents a DNA fragment analyzed. Open and closed circles indicate either unmethylated or methylated CpG sites. When approximately 50% of the analyzed strands are hypomethylated and the other 50% hypermethylated, the region is called monoallelically methylated. If the strands can be distinguished as maternal or paternal by the use of DNA polymorphisms, the strands are then called differentially methylated between parental alleles.

associated with monoallelic expression of *IGF2*. The imprinting status of the bovine *IGF2* gene may also be coregulated with *H19* under another differentially methylated region. Curchoe et al. (2009) reported that a CpG island intergenic of *H19* and *IGF2* was also monoallelically methylated in multiple tissues of newborn calves that showed paternal expression of *IGF2*.

Similar to *IGF2*, the bovine *SNRPN* gene is also differentially methylated in bovine sperm and oocytes in a CpG island at the promoter region. Additionally, DNA from somatic cells is monoallelically methylated in day-17 embryos, and in liver samples from day-60 fetuses and adult cattle (Lucifero et al. 2006). This methylation pattern continued to midgestation (day 130, Couldrey and Lee 2010) in the three tissues examined, (1) adrenal, (2) kidney, and (3) liver. These data suggest that although the sperm and oocyte carry differentially methylated alleles to the fertilized embryos, monoallelic expression does not start until much later in embryo development. Different epigenetic regulation of allelic expression of *SNRPN* must exist before and after day 17 of embryo development. Another interesting feature of *SNRPN* is that monoallelic methylation of its promoter was correlated with the exclusive/nearly exclusive paternal expression in all somatic tissues studied in day-40 fetuses with the exception of the heart where DNA was hypomethylated but monoallelic expression was maintained (Suzuki et al. 2009). This suggests that in the heart, the imprinting status of the *SNRPN* gene may be maintained by mechanisms other than DNA methylation. At day 130 of gestation, another imprinted gene in cattle, *KCNQ1OT1*, was also found to be monoallelic in adrenal, kidney, and liver (Couldrey and Lee 2010).

Many imprinted genes are clustered on the chromosome because they are coregulated by the same imprinting control region. This has been well characterized in the mouse but, to date, only one such study has been conducted in the bovine. The bovine *PEG3-MIMT1-USP29* gene domain was characterized by Kim et al. (2007). The imprinting status of all three genes is believed to be controlled by the 4-kb CpG island surrounding the first exons of *PEG3* and *MIMT1*. It is differentially methylated between parental alleles, which may be the mechanism that renders all three genes to be only expressed from the paternal allele.

## X-Chromosome Inactivation

In all eutherian mammalian species with the exception of X-monosomic mutants (Scott et al. 2006), XCI is used to achieve an equality of expression of X-linked genes between males and females (Lyon 1961) (Figure 11.5). Inactivation of the X chromosome and maintenance of the inactive state are achieved through epigenetic mechanisms. The *XIST* gene encodes an ncRNA that is expressed in *cis* from the chromosome that is to be inactivated ($X_i$; Borsani et al. 1991). The *XIST* transcripts coat the X chromosome and recruit chromatin-modifying proteins that convert the X chromosome into a heterochromatic, silenced state (Okamoto et al. 2004). In addition to the coating of $X_i$ by *XIST* transcripts, other epigenetic mechanisms are also involved. In the bovine, trimethylated H3K9 and H3K27, as well as macroH2A1, are preferentially concentrated on the $X_i$, whereas the histone variant macroH2A2 is not a marker for this chromosome. Interestingly, different heterochromatin regions on the bovine $X_i$ can be identified by their unique histone isoform composition (Coppola et al. 2008), suggesting specific epigenetic modifications of the same chromosome for inactivation.

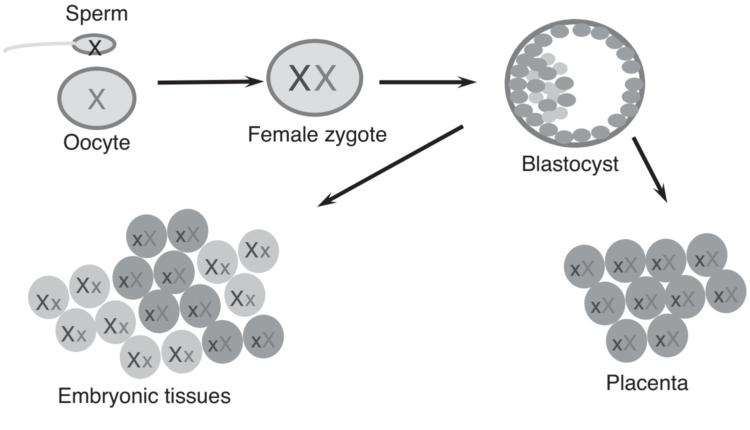


**Figure 11.5** X-chromosome inactivation (XCI) in early fertilized embryos. During fertilization, the sperm carries an inactive x (black, lowercase; blue in zygotes) while the egg carries an active X (grey uppercase). Both X are active after the formation of the female zygote (XX both uppercase). At the time of blastocyst formation, cells in the inner cell mass randomly inactivate one X, either of the paternal (black) or maternal (grey) origin, resulting in random XCI (the dark and light grey circles represent cells maintaining active paternal or maternal X chromosome respectively). This XCI pattern is transmitted to all tissues in the fetus. In cells of the trophectoderm, which will become the placenta, the paternal X chromosome (black) is preferentially inactivated (lower case), resulting in imprinted XCI (light grey circles).

XCI is also subjected to genomic imprinting. In mice, the best-studied species, imprinted XCI occurs during preimplantation development, with the paternal X chromosome ($X_p$) being preferentially silenced (Takagi and Sasaki 1975). This pattern persists in the trophectoderm lineage, such that only the maternal X ($X_m$) is expressed in the placenta. This is accomplished by preferential paternal *XIST* expression at the time of zygotic genome activation (Okamoto et al. 2005). In the inner cell mass (ICM), however, $X_p$ is reactivated, after which $X_p$ and $X_m$ are subject to random inactivation in the developing embryo (Okamoto et al. 2005).

A systematic analysis of XCI in cattle has yet to be completed. Available data indicate that XCI in cattle is very similar to that in the mouse. It has been shown that the trimethylated H3K27 is asymmetrically distributed between the male and female pronuclei so that only one of the pronuclei is stained (Breton et al. 2010), demonstrating that the male X contains mainly inactive chromatin as is also reported in the mouse. The *XIST* transcripts have been detected in bovine embryos as early as the 2-cell stage (De La Fuente et al. 1999). Both the maternal and paternal alleles of the X-linked gene *MAOA*, which has been shown to be subjected to XCI in cattle (Xue et al. 2002), were present in the 4-, 8- to 16-cell, blastocyst, and expanded blastocyst embryos, but only the maternal allele was present in the morula stage. It was, therefore, confirmed that XCI is established at the morula stage and $X_p$ is reactivated at the blastocyst stage (Ferreira et al. 2010). Interestingly, the late-replicating (and presumptive) $X_i$ was not observed until the early blastocyst stage, suggesting that late replication of the $X_i$ may not be an early event in bovine XCI. Additionally, as in the mouse, evidence of imprinted XCI in cattle placentas has also been observed (Xue et al. 2002).

Not all genes on $X_i$ are inactive. Approximately 15% of X-linked genes escape XCI in humans, while far fewer genes do so in the mouse. Methylation of the CpG islands at the 5′ ends of X-linked genes is a general feature of the inactive X throughout eutherian species (Kaslow and Migeon 1987). Using DNA methylation as an indicator for activity status of X-linked genes, Yen et al. (2007) analyzed seven X-linked genes, (1) *ZFX,* (2) *CRSP2,* (3) *UTX,* (4) *UBe1,* (5) *JARID1C,* (6) *AR*, and (7) *FMR1* in cattle. It was found that *FMR1* and *AR* are subject to inactivation, while the *UTX* gene escapes XCI. The genes *ZFX, CRSP2, UBE1,* and *JARID1C* showed a pattern of lack of methylation, suggesting that they also escape XCI. Interestingly, for *JARID1C*, analysis of two cows showed one had methylation and one did not. This suggests that *JARID1C* may be variable in its inactivation status among female cattle as has been reported in humans. Taken together, these data clearly demonstrate that as in humans, cattle also have a large number of genes that escape XCI.

## Epigenomics of the Early Bovine Embryos

### *DNA Methylation in Early Bovine Embryos*

The most dramatic changes in DNA methylation occur during gametogenesis and early embryo development. Gametes—ooyctes and sperms—have relatively low levels of DNA methylation compared to those in the differentiated somatic cells (Dean et al. 2001; Phutikanit et al. 2010). Shortly after fertilization, these relative low levels of methylation undergo further loss in both the male and female pronuclei in most mammalian species. The mechanisms and speed of demethylation, however, are dramatically different. The male pronucleus loses methylation very rapidly and this occurs in the absence of transcription or DNA replication and is thus termed active demethylation. The female pronuclues, on the other hand, undergoes step-wise decreases in DNA methylation with each round of DNA replication as a result of the absence of functional *DNMT1*. This renders the newly replicated DNA strand devoid of methylation and a reduction in the overall level of DNA methylation. This replication-dependent demethylation is referred to as passive demethylation (Figure 11.2B).

Park et al. (2007) compared dynamics of global DNA methylation in zygotes from mice, rats, rabbits, goats, pigs, sheep, and cattle. They classified these species into three distinct categories according to DNA methylation states of the male pronucleus. In type-I species, the male pronucleus is actively demethylated to near completion (mouse and rat). In type-II species, the paternal DNA methylation is largely maintained (sheep and pig). Finally, in type-III species, the male pronucleus undergoes partial demethylation (cattle and goat). Similar findings were also reported by others (Dean et al. 2001; Beaujean et al. 2004; Lepikhov et al. 2008).

Detailed changes in DNA methylation have been well described up to the elongated stage of bovine embryos. Between 10 and 14 hours postinsemination, the male pronucleus decondenses and is ~40% more demethylated than the female pronuclei (Bourc'his et al. 2001; Abdalla et al. 2009). Global DNA methylation is further reduced between the 2- and 4-cell stages with de novo methylation occurring after the 8-cell stage (Santos et al. 2003), concurrent with zygotic genome activation (Dean et al. 2001). Around the time of blastocyst formation, there is a marked increase in the methylation of both DNA and histones. At the blastocyst stage, global DNA was more hypermethylated in ICM than the trophectoderm. At the elongated stage of embryo development, methylation at the satellite I sequence continued to increase in both embryonic disc and trophectoderm (Sawai et al. 2010).

Despite the existence of an overall dynamics of global methylation during early embryo development, the regions of the genome that contribute to the global DNA methylation stains have different transformation pattern. For example, Kang et al. (2005) reported that the overall DNA methylation was maintained in Satellite 1 and Bov-B LINE (Kang et al. 2001b), decreased in alpha satellites, and increased in Satellite II sequences from bovine zygote to blastocyst. These observations suggest that even modifications of DNA at repetitive, noncoding regions are differentially regulated.

Protein-coding regions of the genome also undergo methylation reprogramming during embryonic development. Niemann et al. (2010) analyzed 41 DNA regions from 25 developmentally important genes on 15 different chromosomes. It was revealed that the bovine blastocysts have dramatically lower levels of DNA methylation in these regions than somatic cells such as fibroblasts or peripheral blood mononuclear cells. The dynamics of methylation on protein-coding genes from gametes to blastocysts were delineated using two single-copy genes, bovine epidermal cytokeratin and mammary gland-specific $\beta$-lactoglobulin genes as examples (Kang et al. 2002). They were methylated in sperm, mature oocytes, and zygotes. This methylation status was maintained until 4- to 8-cell stage while some demethylation occurred. Additional and extensive demethylation occurred at the morula and blastocyst stages.

The first lineage-specific asymmetry of DNA methylation in postfertilization development (Santos et al. 2003) occurs at the ICM and trophectoderm differentiation. The lower level of methylation in trophectoderm continues to the bovine placenta, which is approximately 40% lower in global DNA methylated than fetal tissues at the same stage of development (Hiendleder et al. 2004). Specific genes in the bovine placenta, such as adenomatous polyposis coli (*APC*), secreted frizzled-related protein 2 (*SFRP2*) (both negative regulator of WNT signaling), vitamin D catabolic 24-hydroxylase (*CYP24A1*), and *DNMT1* also have extremely low levels of DNA methylation (Ng et al. 2010).

Dynamics of DNA methylation can be brought about by changes in DNA methyl transferases. *DNMT1* was found to be present in the cytoplasm in metaphase II stage oocytes and in zygotes; it entered the nuclei at the 8–16 cell stage bovine embryos, coincident with the increase in DNA methylation (Lodde et al. 2009). These enzymes themselves are subjected to epigenetic modifications. For example, the 5′ regions of *DNMT1* and *DNMT2* were found to be nearly completely unmethylated in all normal adults, IVF fetuses, and sperm. *DNMT3a* and *DNMT3b*, however, were nearly completely methylated in adult skin, hypermethylated in skin of IVF fetuses, and completely methylated for *3a* and nearly monoallelically methylated for *3b* in sperms (Liu et al. 2008a). These data, while fragmented at present, demonstrated the complexity of the regulations of DNA methylation during development.

## *Histone Acetylation and Methylation During Bovine Embryo Development*

Similar to DNA methylation, dramatic changes also occur in the modifications of histones during gamete and early embryo development. While present in germinal vesicle (GV) stage oocytes, acetylation signals are absent in matured bovine oocytes or sperm on histone H4, including K5, K8, K9, K12, and K16 (Wee et al. 2006; Maalouf et al. 2008; Racedo et al. 2009). In early bovine embryos, a reverse correlation between global DNA methylation and histone acetylation was observed for H4K5, K8, K12, and K16 (Maalouf et al. 2008). The male pronucleus start to gain acetylated H4K5 signals at the time of pronuclei formation (7–8 hours after insemination). This is coincident with the beginning of decondensation and demethylation of the male pronucleus. The female pronucleus also becomes transiently hyperacetylated. The acetylated H4K5 signals were detected both in male and female pronuclei 10 hours after insemination and thereafter further increased as the zygote developed (Wee et al. 2006). Interestingly, the intensity of histone acetylation peaks at the 8-cell stage, when DNA methylation is the lowest. Acetylated H3K9 has a similar time course as H4K5. At the blastocyst stage, trophectoderm cells are more intensely stained for acetylated lysine while ICM cells were stained weakly.

The pattern of distribution and intensity of H3K9 methylation very closely parallel that of DNA methylation (Wee et al. 2006). Both dimethylated and trimethylated histone H3K9 were present during the entire process of oocyte maturation (Park et al. 2007; Racedo et al. 2009). During the pronucleus stage, methylated H3K9 disappears from the male pronuclei when it is rapidly demethylated but is present in female pronuclei (Park et al. 2007; Lepikhov et al. 2008). Around the time of blastocyst formation, there is a marked increase in the methylation of H3, similar to changes in DNA methylation. Trimethylated H3K4 stain, however, does not differ between male and female pronuclei (Lepikhov et al. 2008).

It remains unclear why the early embryos undergo such dramatic epigenomic changes. It possible that these events are essential to remove differences in gamete-specific DNA methylation and histone modification patterns and to reformat the genome prior to initiation of normal development (Han et al. 2003).

## Epigenetics of Bovine Nonimprinted Protein-coding Genes

Sparse and fragmented data have been reported in the status of DNA methylation and histone modification in protein-coding genes in cattle. In a small number of cases, correlations were made between the epigenetic status of the gene and their expression levels. Direct proof that DNA methylation regulates the expression of genes is even rarer. The following examples represent the majority of nonimprinted genes, if not all, that have been characterized in the bovine.

Three published studies contain convincing data for a regulatory role of DNA methylation in gene expression in cattle. Nakaya et al. (2009) reported that the bovine placental lactogen gene (*bPL*) was hypomethylated in the cotyledonary tissue and treatment of bovine trophoblast cell lines with 5-aza-2′-deoxycytidine, increased the expression of *bPL*. When studying the effect of DNA methylation on milk production, Vanselow et al. (2006) observed that the promoter of the bovine alphaS1-casein gene was hypomethylated in the lactating udder only. Infection of the fully lactating cows with a pathogenic *E. coli* strain remethylated the promoter and experimentally elicited acute shutdown of casein synthesis, suggesting that DNA methylation is involved in turning off the casein gene. Lastly, when the CpG island in the promoter of a bovine nonclassical *MHC-I* gene (*NC1*) was artificially methylated, *NC1* expression was completely abrogated of its constitutive expression (O’Gorman et al. 2010).

A number of studies described the methylation status of protein-coding genes although no cause–effect data were obtained. The methylation status of five pluripotent genes, (1) *OCT4,* (2) *SOX2,* (3) *NANOG,* (4) *REX1*, and (5) *FGF4*, was determined in fetal fibroblast cells, and fertilized embryos at the 8-cell and morula stages (Lan et al. 2010). It was found that *OCT4* and *REX1* were nearly completely methylated in fibroblast cells, while *NANOG* had low levels of methylation, and *SOX2* and *FGF4* were nearly completely unmethylated. In early embryos at either the 8-cell or morula stage, *SOX2* and *REX1* had similar methylation pattern as in fibroblasts. *OCT4* lost while *Nanog* gained methylated in early embryos compared to fibroblast cells, both became mosaic in methylation in early embryos. The CpG island in the promoter of *OCT4* continued its mosaic methylation pattern in the fetal brain and intercotyledonary membranes at days 48 and 59 of gestation (Kremenskoy et al. 2006). It appears that the *OCT4* gene became completely methylated in a later stage in fibroblast cells. Kremenskoy et al. (2006) also reported that the CpG island in the promoter region of the bovine leptin gene, which is involved in the regulation of fetal and placental growth, was nearly completely unmethylated in the tissues.

At midgestation (day 130), most bovine nonimprinted genes studied by Couldrey and Lee (2010) including *HAND1, ASCL2, KCNQ1, CDKN1C, GR, CSF-1*, and *STAT5a* were hypomethylated in adrenal, liver, and kidney. The gene *DIO3*, however, and all repetitive (Satellite I/II/alpha) were hypermethylated in all three tissues. In bovine adult blood and skin, the bovine repetitive sequence, Satellite 1 and promoter regions of two single-coy gene, interleukin 3, and epidermal cytokeratin were found to be hypermethylated (Chen et al. 2005).

It has been known that genes that contain more than one transcription start sites may be subjected to regulation by different CpG islands in their multiple promoters. One such study has been conducted in the bovine. The key enzyme of estrogen biosynthesis, aromatase cytochrome P450, is encoded by the *CYP19* gene that has two promoters (promoter-1.1 and -2) that contain CpGs, albeit at low densities. Bovine granulose cells that produce high amount of estrogen were largely unmethylated at both promoters but only expressed from promoter-2. Both promoters were methylated in corpora lutea of pregnancy that produced low levels of transcripts from promoter 1.1. It, therefore, appears that DNA methylation is inversely related to transcription activity in promoter-2 but not promoter-1.1 (Vanselow et al. 2005).

Aberrant DNA methylation in tumor suppressor genes has been associated with the development of cancer in humans. One such study is available in cattle. The fragile histidine triad (FHIT) gene is a tumor suppressor known to be inactivated in many human tumors. When cattle with chronic enzootic hematuria were studied, it was found that, unlike in human tumors, FHIT in vesical tumors was largely unmethylated. Furthermore, the same levels and mRNA isoforms of FHIT were detected in tumors and in healthy tissues. Although Guidi et al. (2008) suggested that further studies and larger sets of cases would be useful to confirm their finding, the data seem to suggest that altered epigenetic modifications of FHIT is not a hallmark of bovine vesical tumors.

## Effect of Biotechnology on Epigenomics

In cattle, a wide range of congenital abnormalities collected termed “the large offspring syndrome” including symptoms such as large birth weight, enlarged placenta (mainly due to placental hydrops), and reduced number of cotyledons, reluctant to suckle, difficulty breathing and standing, and hypothermia, are frequently observed as a result of embryo culture and several forms of biotechnological manipulations (reviewed by Young and Fairburn 2000). The early embryos are very vulnerable to environment that alters their epigenetic elements because this is the stage when the most dramatic changes of occur on these elements. Artificial manipulations, therefore, render the embryos unsuitable for further development due to the fact that aberrant epigenetic changes can be stably transmitted through cell division to fetal stages and even postnatal development. In somatic cell nuclear transfer (cloning), high rate (>90%) of developmental failure (abortion) is common place. Even cloned animals that do develop to term, large calf syndrome is frequently observed in cattle and other species. These developmental failures can result from two sources of faulty epigenetic modifications. First, an incomplete erasure of the somatic cell epigenetic marks by the early cloned embryos will lead to abnormal expression of many developmentally important genes. Second, the culturing of the cloned embryos in vitro results in additional aberrations. As a result, it has been demonstrated that reprogramming of DNA methylation, expression of imprinted and developmentally important genes, X-chromosome inactivation, and telomerase activity are incomplete in cloned embryos as compared with naturally fertilized embryos.

Numerous studies have been conducted on the epigenetic status of cloned animals. A few examples are given in the following text to illustrate the severity of the problems at the present time. Although significant reprogramming do occur in both the DNA methylation of protein-coding genes and gene expression profiles by the blastocyst stage (Han et al. 2003; Kang et al. 2003; Smith et al. 2005; Niemann et al. 2010), globally, DNA methylation is adversely affected by incomplete reprogramming (Kang et al. 2001a), and embryo culture. The presence of 10% serum in embryo culture medium causes significant increase in body and organ weights and the percentage of DNA methylation (Hiendleder et al. 2006). Abnormal expression levels as well as leaky expression of the silenced allele of imprinted genes have been widely reported (Long and Cai 2007; Yang et al. 2005; Curchoe et al. 2009; Suzuki et al. 2009; Suteevun-Phermthai et al. 2009).

Interestingly, the few cloned fetuses that do survive to midgestation or to term do have quite normal gene expression and epigenetics for the few genes analyzed. SCNT clones in midgestation (day 130) were found to have similar methylation patterns in nearly all genes except *SNRPN* and *KCNQ1OT1*, which are confirmed imprinted genes in cattle (Couldrey and Lee 2010). Similarly, in live born clones, relatively few DNA aberrations could be found in genes studied (*β-ACTIN*, *VEGF*, *OCT4*, *TERT*, *H19*, and *IGF2*) and a repetitive sequence (*ART2*) in five organs (heart, liver, spleen, lung, and kidney) (Lin et al. 2008).

In addition to embryo culture and cloning, gene-targeting also alters epigenetic signals because insertion of foreign DNA molecules and prolonged cell culture for positive cell selection can all affect DNA methylation. Lin et al. (2009) reported that targeting of bovine fetal cells by the *β*-glucosyl transferase caused widespread changes in DNA methylation in the six genes studied: (1) *β-actin*, (2) *VEGF4*, (3) *TERT*, (4) *H19*, (5) *IGF2*, and (6) a repetitive sequence *art2*. These results call for caution in the interpretation of data obtained from gene knockout studies.

## Conclusion

Each mammalian tissue has its own unique epigenetic modifications, which undergo dramatic changes during development, differentiation, and aging. The current understanding of epigenomics in the bovine is equivalent to a giant puzzle that is missing the majority of the pieces. The study of epigenomics of bovine development is, therefore, an enormous project ahead of us. The human epigenome project in which genomewide DNA methylation patterns of all human genes in all major organs will be identified, cataloged, and interpreted is a multination effort. With the completion of the bovine genome sequencing and application of next-generation sequencing technology, bovine epigenomics will be completely uncovered with combined efforts in the not too distant future.

## References

Abdalla, H., Hirabayashi, M., Hochi, S. (2009) Demethylation dynamics of the paternal genome in pronuclear-stage bovine zygotes produced by in vitro fertilization and ooplasmic injection of freeze-thawed or freeze-dried spermatozoa. *Journal of Reproduction and Development* **55**(4): 433–439.

Aravind, L. (2000) Exploring histones and their relatives with the Histone Sequence Database. *Trends in Genetics* **16**(11): 517–518.

Ausió, J. (2006) Histone variants–the structure behind the function. *Functional Genomics and Proteomics* **5**(3): 228–243.

Bannister, A.J., Zegerman P., Partridge J.F., Miska, E.A., Thomas, J.O., Allshire, R.C., Kouzarides, T. (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**: 120–124.

Beaujean, N., Taylor, J., Gardner, J., Wilmut, I., Meehan, R., Young, L. (2004) Effect of limited DNA methylation reprogramming in the normal sheep embryo on somatic cell nuclear transfer. *Biology of Reproduction* **71**(1): 185–193.

Borsani, G., et al. (1991) Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351**: 325–329.

Bourc'his, D., Le Bourhis, D., Patin, D., Niveleau, A., Comizzoli, P., Renard, J.P., Viegas-Péquignot, E. (2001) Delayed and incomplete reprogramming of chromosome methylation patterns in bovine cloned embryos. *Current Biology* **11**(19): 1542–1546.

Breton, A., LE Bourhis, D., Audouard, C., Vignon, X., Lelievre, J.M. (2010) Nuclear profiles of H3 histones trimethylated on Lys27 in bovine (Bos taurus) embryos obtained after in vitro fertilization or somatic cell nuclear transfer. *Journal of Reproduction and Development* **56**(4): 379–388.

Briggs, S.D., Bryk, M., Strahl, B.D., Cheung, W.L., Davie, J.K., Dent, S.Y., Winston, F., Allis, C.D. (2001) Histone H3 lysine 4 methylation is mediated by *Set1* and required for cell growth and rDNA silencing in *Saccharomyces cerevisiae*. *Genes & Development* **15**: 3286–3295.

Burdge, G.C. and Lillycrop, K.A. (2010) Nutrition, epigenetics, and developmental plasticity: implications for understanding human disease. *Annual Review of Nutrition* **30**: 7.1–7.25.

Chen, T., Jiang, Y., Zhang, Y.L., Liu, J.H., Hou, Y., Schatten, H., Chen, D.Y., Sun, Q.Y. (2005) DNA hypomethylation of individual sequences in aborted cloned bovine fetuses. *Frontiers in Bioscience* **10**: 3002–3008.

Clemson, C.M., McNeil, J.A., Willard, H.F., Lawrence, J.B. (1996) XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *Journal of Cell Biology* **132**: 259–275.

Coppola, G., Pinton, A., Joudrey, E.M., Basrur, P.K., King, W.A. (2008) Spatial distribution of histone isoforms on the bovine active and inactive X chromosomes. *Sexual Development* **2**(1): 12–23.

Couldrey, C. and Lee, R.S. (2010) DNA methylation patterns in tissues from mid-gestation bovine foetuses produced by somatic cell nuclear transfer show subtle abnormalities in nuclear reprogramming. *BMC Developmental Biology* **10**: 27–43.

Cruz, N.T., Wilson, K.J., Cooney, M.A., Tecirlioglu, R.T., Lagutina, I., Galli, C., Holland, M.K., French, A.J. (2008) Putative imprinted gene expression in uniparental bovine embryo models. *Reproduction, Fertility and Development* **20**(5): 589–597.

Curchoe, C., Zhang, S., Bin, Y., Zhang, X., Yang, L., Feng, D., O'Neill, M., Tian, X.C. (2005) Promoter-specific expression of the imprinted IGF2 gene in cattle (Bos taurus). *Biology of Reproduction* **73**(6): 1275–1281.

Curchoe, C.L., Zhang, S., Yang, L., Page, R., Tian, X.C. (2009) Hypomethylation trends in the intergenic region of the imprinted IGF2 and H19 genes in cloned cattle. *Animal Reproduction Science* **116**(3–4): 213–225.

Dean, W., Santos, F., Stojkovic, M., Zakhartchenko, V., Walter, J., Wolf, E., Reik, W. (2001) Conservation of methylation reprogramming in mammalian development: aberrant reprogramming in cloned embryos. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 13734–13738.

De La Fuente, R., Hahnel, A., Basrur, P.K., King, W.A. (1999) X inactive-specific transcript (Xist) expression and X chromosome inactivation in the preattachment bovine embryo. *Biology of Reproduction* **60**: 769–775.

Dindot, S.V., Kent, K.C., Evers, B., Loskutoff, N., Womack, J., Piedrahita, J.A. (2004) Conservation of genomic imprinting at the XIST, IGF2, and GTL2 loci in the bovine. *Mammalian Genome* **15**(12): 966–974.

Dindot, S.V., Person, R., Strivens, M., Garcia, R., Beaudet, A.L. (2009) Epigenetic profiling at mouse imprinted gene clusters reveals novel epigenetic and genetic features at differentially methylated regions. *Genome Research* **19**(8): 1374–1383.

Ferreira, A.R., Machado, G.M., Diesel, T.O., Carvalho, J.O., Rumpf, R., Melo, E.O., Dode, M.A., Franco, M.M. (2010) Allele-specific expression of the MAOA gene and X chromosome inactivation in in vitro produced bovine embryos. *Molecular Reproduction and Development* **77**(7): 615–621.

Gebert, C., Wrenzycki, C., Herrmann, D., Gröger, D., Reinhardt, R., Hajkova, P., Lucas-Hahn, A., Carnwath, J., Lehrach, H., Niemann, H. (2006) The bovine IGF2 gene is differentially methylated in oocyte and sperm DNA. *Genomics* **88**(2): 222–229.

Gebert, C., et al. (2009) DNA methylation in the IGF2 intragenic DMR is re-established in a sex-specific manner in bovine blastocysts after somatic cloning. *Genomics* **94**(1): 63–69.

Guidi, E., Uboldi, C., Ferretti, L. (2008) Molecular analysis of the fragile histidine triad (FHIT) tumor suppressor gene in vesical tumors of cattle with chronic enzootic hematuria (CEH). *Cytogenetic and Genome Research* **120**(1–2): 173–177.

Han, Y.M., Kang, Y.K., Koo, D.B., Lee, K.K. (2003) Nuclear reprogramming of cloned embryos produced in vitro. *Theriogenology* **59**(1): 33–44.

Han, L., Su, B., Li, W.H., Zhao, Z. (2008) CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biology* **9**(5): R79.1–R79.12.

Hiendleder, S., Mund, C., Reichenbach, H.D., Wenigerkind, H., Brem, G., Zakhartchenko, V., Lyko, F., Wolf, E. (2004) Tissue-specific elevated genomic cytosine methylation levels are associated with an overgrowth phenotype of bovine fetuses derived by in vitro techniques. *Biology of Reproduction* **71**(1): 217–223.

Hiendleder, S., et al. (2006) Tissue-specific effects of in vitro fertilization procedures on genomic cytosine methylation levels in overgrown and normal sized bovine fetuses. *Biology of Reproduction* **75**(1): 17–23.

Kang, Y.K., Koo, D.B., Park, J.S., Choi, Y.H., Chung, A.S., Lee, K.K., Han, Y.M. (2001a) Aberrant methylation of donor genome in cloned bovine embryos. *Nature Genetics* **28**(2): 173–177.

Kang, Y.K., Koo, D.B., Park, J.S., Choi, Y.H., Lee, K.K., Han, Y.M. (2001b) Differential inheritance modes of DNA methylation between euchromatic and heterochromatic DNA sequences in ageing fetal bovine fibroblasts. *FEBS Letters* **498**(1): 1–5.

Kang, Y.K., Park, J.S., Koo, D.B., Choi, Y.H., Kim, S.U., Lee, K.K., Han, Y.M. (2002) Limited demethylation leaves mosaic-type methylation states in cloned bovine pre-implantation embryos. *EMBO Journal* **21**(5): 1092–1100.

Kang, Y.K., Yeo, S., Kim, S.H., Koo, D.B., Park, J.S., Wee, G., Han, J.S., Oh, K.B., Lee, K.K., Han, Y.M. (2003) Precise recapitulation of methylation change in early cloned embryos. *Molecular Reproduction and Development* **66**(1): 32–37.

Kang, Y.K., Lee, H.J., Shim, J.J., Yeo, S., Kim, S.H., Koo, D.B., Lee, K.K., Beyhan, Z., First, N.L., Han, Y.M. (2005) Varied patterns of DNA methylation change between different satellite regions in bovine preimplantation development. *Molecular Reproduction and Development* **71**(1): 29–35.

Kaslow, D.C. and Migeon, B.R. (1987) DNA methylation stabilizes X chromosome inactivation in eutherians but not in marsupials: evidence for multistep maintenance of mammalian X dosage compensation. *Proceedings of the National Academy of Sciences of the United States of America* **84**: 6210–6214.

Keeton, W. and Gould, J. (1984) *Biological Science*. New York: WW Norton and Company, Inc.

Kerppola, T.K. (2009) Polycomb group complexes–many combinations, many functions. *Trends in Cell Biology* **19**(12): 692–704.

Khatib, H. (2004) Imprinting of Nesp55 gene in cattle. *Mammalian Genome* **15**(8): 663–667.

Khatib, H., Zaitoun, I., Kim, E.S. (2007) Comparative analysis of sequence characteristics of imprinted genes in human, mouse, and cattle. *Mammalian Genome* **18**(6–7): 538–547.

Killian, J.K., Nolan, C.M., Wylie, A.A., Li, T., Vu, T.H., Hoffman, A.R., Jirtle, R.L. (2001) Divergent evolution in M6P/IGF2R imprinting from the Jurassic to the Quaternary. *Human Molecular Genetics* **10**(17): 1721–1728.

Kim, J., Bergmann, A., Lucas, S., Stone, R., Stubbs, L. (2004) Lineage-specific imprinting and evolution of the zinc-finger gene ZIM2. *Genomics* **84**(1): 47–58.

Kim, J., Bergmann, A., Choo, J.H., Stubbs, L. (2007) Genomic organization and imprinting of the Peg3 domain in bovine. *Genomics* **90**(1): 85–92.

Koehler, D., Zakhartchenko, V., Froenicke, L., Stone, G., Stanyon, R., Wolf, E., Cremer, T., Brero, A. (2009) Changes of higher order chromatin arrangements during major genome activation in bovine preimplantation embryos. *Experimental Cell Research* **315**(12): 2053–2063.

Kouzarides, T. and Berger, S.L. (2006) Chromatin modifications and their mechanism of action. In: *Epigenetics*, edited by C.D. Allis, T. Jenuwein, D. Reinberg, and M.L. Caparros, pp. 191–209. New York: Cold Spring Harbor Press.

Kremenskoy, M., Kremenska, Y., Suzuki, M., Imai, K., Takahashi, S., Hashizume, K., Yagi, S., Shiota, K. (2006) Epigenetic characterization of the CpG islands of bovine leptin and POU5F1 genes in cloned bovine fetuses. *Journal of Reproduction and Development* **52**(2): 277–285.

Kuehn, C., Edel, C., Weikard, R., Thaller, G. (2007) Dominance and parent-of-origin effects of coding and non-coding alleles at the acylCoA-diacylglycerol-acyltransferase (DGAT1) gene on milk production traits in German Holstein cows. *BMC Genetics* **8**: 62–70.

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., Jenuwein, T. (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**: 116–120.

Lan, J., Hua, S., Zhang, H., Song, Y., Liu, J., Zhang, Y. (2010) Methylation patterns in 5′ terminal regions of pluripotency-related genes in bovine in vitro fertilized and cloned embryos. *Journal of Genetics and Genomics* **37**(5): 297–304.

Latham, K.E. (1999) Epigenetic modification and imprinting of the mammalian genome during development. *Current Topics in Developmental Biology* **43**: 1–49.

Lepikhov, K., Zakhartchenko, V., Hao, R., Yang, F., Wrenzycki, C., Niemann, H., Wolf, E., Walter, J. (2008) Evidence for conserved DNA and histone H3 methylation reprogramming in mouse, bovine and rabbit zygotes. *Epigenetics & Chromatin* **1**(1): 8–18.

Lin, L., Li, Q., Zhang, L., Zhao, D., Dai, Y., Li, N. (2008) Aberrant epigenetic changes and gene expression in cloned cattle dying around birth. *BMC Developmental Biology* **8**: 14–23.

Lin, L., Xu, W., Dai, Y., Li, N. (2009) DNA methylation changes in cell line from $\beta$-lactoglobulin gene targeted fetus. *Animal Reproduction Science* **112**: 402–408.

Liu, J., Liang, X., Zhu, J., Wei, L., Hou, Y., Chen, D.Y., Sun, Q.Y. (2008a) Aberrant DNA methylation in 5′ regions of DNA methyltransferase genes in aborted bovine clones. *Journal of Genetics and Genomics* **35**(9): 559–568.

Liu J.-H., Yin, S., Xiong, B., Hou, Y., Chen, D.-Y., Sun, Q.-Y. (2008b) Aberrant DNA methylation imprints in aborted bovine clones. *Molecular Reproduction and Development* **75**: 598–607.

Lodde, V., Modina, S.C., Franciosi, F., Zuccari, E., Tessaro, I., Luciano, A.M. (2009) Localization of DNA methyltransferase-1 during oocyte differentiation, in vitro maturation and early embryonic development in cow. *European Journal of Histochemistry* **53**(4): 199–207.

Long, J.E. and Cai, X. (2007) Igf-2r expression regulated by epigenetic modification and the locus of gene imprinting disrupted in cloned cattle. *Gene* **388**(1–2): 125–134.

Lucifero, D., Suzuki, J., Bordignon, V., Martel, J., Vigneault, C., Therrien, J., Filion, F., Smith, L.C., Trasler, J.M. (2006) Bovine SNRPN methylation imprint in oocytes and day 17 in vitro-produced and somatic cell nuclear transfer embryos. *Biology of Reproduction* **75**(4): 531–538.

Lyon, M.F. (1961) Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**: 372–373.

Maalouf, W.E., Alberio, R., Campbell, K.H. (2008) Differential acetylation of histone H4 lysine during development of in vitro fertilized, cloned and parthenogenetically activated bovine embryos. *Epigenetics* **3**(4): 199–209.

Mohammad, F., Mondal, T., Kanduri, C. (2009) Epigenetics of imprinted long noncoding RNAs. *Epigenetics* **4**(5): 277–286.

Monk, M. and Salpekar, A. (2001) Expression of imprinted genes in human preimplantation development. *Molecular and Cellular Endocrinology* **183**(Suppl.): S35–S40.

Moore, T. and Haig, D. (1991) Genomic imprinting in mammalian development: a parental tug-of-war. *Trends in Genetics* **7**(2): 45–49.

Nagano, T. and Fraser, A.E. (2009) Emerging similarities in epigenetic gene silencing by long noncoding RNAs. *Mammalian Genome* **20**: 557–562.

Nakaya, Y., Kizaki, K., Takahashi, T., Patel, O.V., Hashizume, K. (2009) The characterization of DNA methylation-mediated regulation of bovine placental lactogen and bovine prolactin-related protein-1 genes. *BMC Molecular Biology* **10**: 19–33.

Ng, H.K., Novakovic, B., Hiendleder, S., Craig, J.M., Roberts, C.T., Saffery, R. (2010) Distinct patterns of gene-specific methylation in mammalian placentas: implications for placental evolution and function. *Placenta* **31**(4): 259–268.

Niemann, H., Carnwath, J.W., Herrmann, D., Wieczorek, G., Lemme, E., Lucas-Hahn, A., Olek, S. (2010) DNA methylation patterns reflect epigenetic reprogramming in bovine embryos. *Cellular Reprogramming* **12**(1): 33–42.

O'Gorman, G.M., Al Naib, A., Ellis, S.A., Mamo, S., O'Doherty, A.M., Lonergan, P., Fair, T. (2010) Regulation of a bovine nonclassical major histocompatibility complex class I gene promoter. *Biology of Reproduction* **83**(2): 296–306.

Okamoto, I., Otte, A.P., Allis, C.D., Reinberg, D., Heard, E. (2004) Epigenetic dynamics of imprinted X inactivation during early mouse development. *Science* **303**: 644–649.

Okamoto, I., Arnaud, D., Le Baccon, P., Otte, A.P., Disteche, C.M., Avner, P., Heard, E. (2005) Evidence for *de novo* imprinted X-chromosome inactivation independent of meiotic inactivation in mice. *Nature* **438**: 369–373.

Park, J.S., Jeong, Y.S., Shin, S.T., Lee, K.K., Kang, Y.K. (2007) Dynamic DNA methylation reprogramming: active demethylation and immediate remethylation in the male pronucleus of bovine zygotes. *Developmental Dynamics* **236**(9): 2523–2533.

Phillips, J.E. and Corces, V.G. (2009) CTCF: master weaver of the genome. *Cell* **137**(7): 1194–1211.

Phutikanit, N., Suwimonteerabutr, J., Harrison, D., D'Occhio, M., Carroll, B., Techakumphu, M. (2010) Different DNA methylation patterns detected by the Amplified Methylation Polymorphism Polymerase Chain Reaction (AMP PCR) technique among various cell types of bulls. *Acta Veterinaria Scandinavica* **52**: 18–26.

Racedo, S.E., Wrenzycki, C., Lepikhov, K., Salamone, D., Walter, J., Niemann, H. (2009) Epigenetic modifications and related mRNA expression during bovine oocyte in vitro maturation. *Reproduction, Fertility, and Development* **21**(6): 738–748.

Ross, P.J., Ragina, N.P., Rodriguez, R.M., Iager, A.E., Siripattarapravat, K., Lopez-Corrales, N., Cibelli, J.B. (2008) Polycomb gene expression and histone H3 lysine 27 trimethylation changes during bovine preimplantation development. *Reproduction* **136**(6): 777–785.

Ruddock, N.T., Wilson, K.J., Cooney, M.A., Korfiatis, N.A., Tecirlioglu, R.T., French, A.J. (2004) Analysis of imprinted messenger RNA expression during bovine preimplantation development. *Biology of Reproduction* **70**(4): 1131–1135.

Ruddock-D'Cruz, N.T., Prashadkumar, S., Wilson, K.J., Heffernan, C., Cooney, M.A., French, A.J., Jans, D.A., Verma, P.J., Holland, M.K. (2008) Dynamic changes in localization of Chromobox (Cbx) family members during the maternal to embryonic transition. *Molecular Reproduction and Development* **75**(3): 477–488.

Sakurai, T., et al. (2009) Induction of endogenous interferon tau gene transcription by CDX2 and high acetylation in bovine nontrophoblast cells. *Biology of Reproduction* **80**(6): 1223–1231.

Sandhu, J., Kaur, B., Armstrong, C., Talbot, C.J., Steward, W.P., Farmer, P.B., Singh, R. (2009) Determination of 5-methyl-2′-deoxycytidine in genomic DNA using high performance liquid chromatography-ultraviolet detection. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **877**(20–21): 1957–1961.

Santos, F., Zakhartchenko, V., Stojkovic, M., Peters, A., Jenuwein, T., Wolf, E., Reik, W., Dean, W. (2003) Epigenetic marking correlates with developmental potential in cloned bovine preimplantation embryos. *Current Biology* **13**: 1116–1121.

Sawai, K., Takahashi, M., Moriyasu, S., Hirayama, H., Minamihashi, A., Hashizume, T., Onoe, S. (2010) Changes in the DNA methylation status of bovine embryos from the blastocyst to elongated stage derived from somatic cell nuclear transfer. *Cellular Reprogramming* **12**(1): 15–22.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., Cavalli, G. (2007) Genome regulation by polycomb and trithorax proteins. *Cell* **128**: 735–745.

Scott, L.A., Kuroiwa, A., Matsuda, Y., Wichman, H.A. (2006) X accumulation of LINE-1 retrotransposons in Tokudaia osimensis, a spiny rat with the karyotype XO. *Cytogenetic and Genome Research* **112**: 261–269.

Skinner, M.K. and Guerrero-Bosagna, C. (2009) Environmental signals and transgenerational epigenetics. *Epigenomics* **1**(1): 111–117.

Smith, S.L., Everts, R.E., Tian, X.C., Du, F., Sung, L.Y., Rodriguez-Zas, S.L., Jeong, B.S., Renard, J.P., Lewin, H.A., Yang, X. (2005) Global gene expression profiles reveal significant nuclear reprogramming by the blastocyst stage after cloning. *Proceedings of the National Academy of Sciences of the United States of America* **102**(49): 17582–17587.

Soppe, W.J., Jasencakova, Z., Houben, A., Kakutani, T., Meister, A., Huang, M.S., Jacobsen, S.E., Schubert, I., Fransz, P.F. (2002) DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in Arabidopsis. *EMBO Journal* **21**: 6549–6559.

Suteevun-Phermthai, T., et al. (2009) Allelic switching of the imprinted IGF2R gene in cloned bovine fetuses and calves. *Animal Reproduction Science* **116**(1–2): 19–27.

Suzuki, J. Jr, Therrien, J., Filion, F., Lefebvre, R., Goff, A.K., Smith, L.C. (2009) In vitro culture and somatic cell nuclear transfer affect imprinting of SNRPN gene in pre- and post-implantation stages of development in cattle. *BMC Developmental Biology* **9**: 9–22.

Takagi, N. and Sasaki, M. (1975) Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature* **256**: 640–642.

Takai, D. and Jones, P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America* **99**(6): 3740–3745.

Tveden-Nyborg, P.Y., Alexopoulos, N.I., Cooney, M.A., French, A.J., Tecirlioglu, R.T., Holland, M.K., Thomsen, P.D., D'Cruz, N.T. (2008) Analysis of the expression of putatively imprinted genes in bovine peri-implantation embryos. *Theriogenology* **70**(7): 1119–1128.

Tycko, B. and Morison, I.M. (2002) Physiological functions of imprinted genes. *Journal of Cellular Physiology* **192**: 245–258.

Vanselow, J., Pöhland, R., Fürbass, R. (2005) Promoter-2-derived Cyp19 expression in bovine granulosa cells coincides with gene-specific DNA hypo-methylation. *Molecular and Cellular Endocrinology* **233**(1–2): 57–64.

Vanselow, J., Yang, W., Herrmann, J., Zerbe, H., Schuberth, H.J., Petzl, W., Tomek, W., Seyfert, H.M. (2006) DNA-remethylation around a STAT5-binding enhancer in the alphaS1-casein promoter is associated with abrupt shutdown of alphaS1-casein synthesis during acute mastitis. *Journal of Molecular Endocrinology* **37**(3): 463–477.

Waddington, C.H. (1940) *The Temporal Course of Gene Reactions, Organisers and Genes*. p. 69. Cambridge, UK: Cambridge University Press.

Wee, G., Koo, D.B., Song, B.S., Kim, J.S., Kang, M.J., Moon, S.J., Kang, Y.K., Lee, K.K., Han, Y.M. (2006) Inheritable histone H4 acetylation of somatic chromatins in cloned embryos. *Journal of Biological Chemistry* **281**(9): 6048–6057.

Wee, G., Shin, S.T., Koo, D.B., Han, Y.M. (2010) Behaviors of ATP-dependent chromatin remodeling factors during maturation of bovine oocytes in vitro. *Molecular Reproduction and Development* **77**: 126–135.

Xue, F., Tian, X.C., Du, F., Kubota, C., Taneja, M., Dinnyes, A., Dai, Y., Levine, H., Pereira, L.V., Yang, X. (2002) Aberrant patterns of X chromosome inactivation in bovine clones. *Nature Genetics* **31**(2): 216–220.

Yang, L., Chavatte-Palmer, P., Kubota, C., O'neill, M., Hoagland, T., Renard, J.P., Taneja, M., Yang, X., Tian, X.C. (2005) Expression of imprinted genes is aberrant in deceased newborn cloned calves and relatively normal in surviving adult clones. *Molecular Reproduction and Development* **71**(4): 431–438.

Yen, Z.C., Meyer, I.M., Karalic, S., Brown, C.J. (2007) A cross-species comparison of X-chromosome inactivation in Eutheria. *Genomics* **90**(4): 453–463.

Young, L.E. and Fairburn, H.R. (2000) Improving the safety of embryo technologies: possible role of genomic imprinting. *Theriogenology* **53**(2): 627–648.

Zaitoun, I. and Khatib, H. (2006) Assessment of genomic imprinting of SLC38A4, NNAT, NAP1L5, and H19 in cattle. *BMC Genetics* **7**: 49–58.

Zhang, S., Kubota, C., Yang, L., Zhang, Y., Page, R., O'Neill, M., Yang, X., Tian, X.C. (2004) Genomic imprinting of H19 in naturally reproduced and cloned cattle. *Biology of Reproduction* **71**(5): 1540–1544.

# Chapter 12
# Mapping Quantitative Trait Loci

*Joel I. Weller*

## Introduction

As compared to other agricultural species, dairy cattle are unique in the value of each animal, the long generation interval, and the very limited fertility of females. Thus, unlike plant and poultry breeding, most dairy cattle breeding programs are based on selection within the commercial population. Similarly, detection of quantitative trait loci (QTL) and marker-assisted selection (MAS) programs are generally based on analysis of existing populations. The specific requirements of dairy cattle breeding has led to the generation of very large data banks in most developed countries, which are available for analysis. Numerous studies have proposed that accuracy of the evaluations of young sires can be increased by identification of the individual QTL via linkage to genetic markers, and various strategies were proposed for application of MAS (reviewed by Weller 2009).

The important issues in mapping QTL will be outlined, including description of the types of markers currently in use for QTL detection, methods and experimental designs to detect QTL and to estimate QTL effects and location suitable for dairy cattle, the statistical power of the various experimental designs, description of methodologies used to derive genetic evaluations based on genetic markers and pedigree, the current state of QTL detection and MAS in dairy cattle, and methods to identify the actual polymorphisms responsible for observed QTL and description of the reported results.

## DNA-Level Genetic Markers, SSRs vs. SNPs

Short sequence repeats (SSRs), or microsatellites, consist of tandem repeats of a sequence of 1–4 DNA base pairs, the most common being “TG.” SSRs are polymorphic in the number of repeats of the core sequence. In the early 1990s, SSRs became the marker of choice, chiefly because of their high polymorphic information content and relative ease of genotyping (Glowatzki-Mullis et al. 1995). Thousands of SSR sequences are scattered throughout the genomes of all advance organisms. Genetic maps of at least several hundred SSR were developed for nearly all the important agricultural species (indexed at: http://www.animalgenome.org/community/other.html).

Unlike SSRs, single nucleotide polymorphisms (SNPs) are nearly always biallelic. SNPs occur at a frequency of approximately 0.3–1 SNP/kbp throughout the human

*Bovine Genomics*, First Edition. Edited by James E. Womack.

genome (Marth et al. 2001), and apparently at equal frequencies in other mammalian species. Advantages of SNPs are summarized by Werner et al. (2004). Genotyping error rates tend to be lower for SNPs (Kennedy et al. 2003; Bonin et al. 2004), larger numbers of markers can be run jointly and genotype determination is completely automatic, eliminating what is generally the largest cost element of genotyping (Kennedy et al. 2003; Anderson and Graza 2006). In January 2008, Illumina announced release of the Infinium(R) BovineSNP50 BeadChip, which includes 54,001 SNPs approximately evenly spaced across the entire bovine genome (http://www.illumina.com/products/bovine_snp50_whole-genome_genotyping_kits.ilmn).

## Detecting and Mapping of QTL via Within-Family Genetic Linkage

Detection of QTL requires generation of linkage disequilibrium (LD) between the genetic markers and QTL. In plants this is generally accomplished by crosses between inbred lines. For the reasons noted in the introduction, this is not a viable option for dairy cattle, in which all analyses must be based on analysis of the existing population. For advanced commercial populations, the "daughter" and "granddaughter" designs, which make use of the existence of large half-sib families, were the most appropriate for QTL analysis until 2006 (Weller et al. 1990). These designs are diagramed in Figures 12.1 and 12.2.

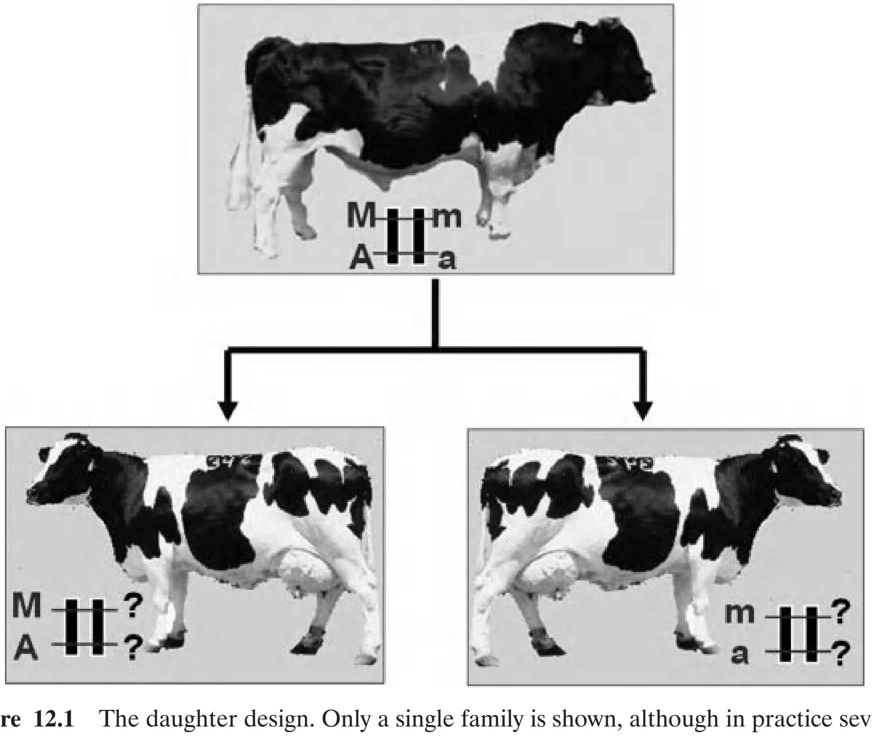


**Figure 12.1** The daughter design. Only a single family is shown, although in practice several families will be analyzed jointly. The sire is assumed to be heterozygous for a quantitative trait loci (QTL) and a linked genetic marker. The two alleles of the marker locus are denoted "M" and "m," and the two alleles of the QTL are denoted "A" and "a." Alleles of maternal origin are denoted by question marks.

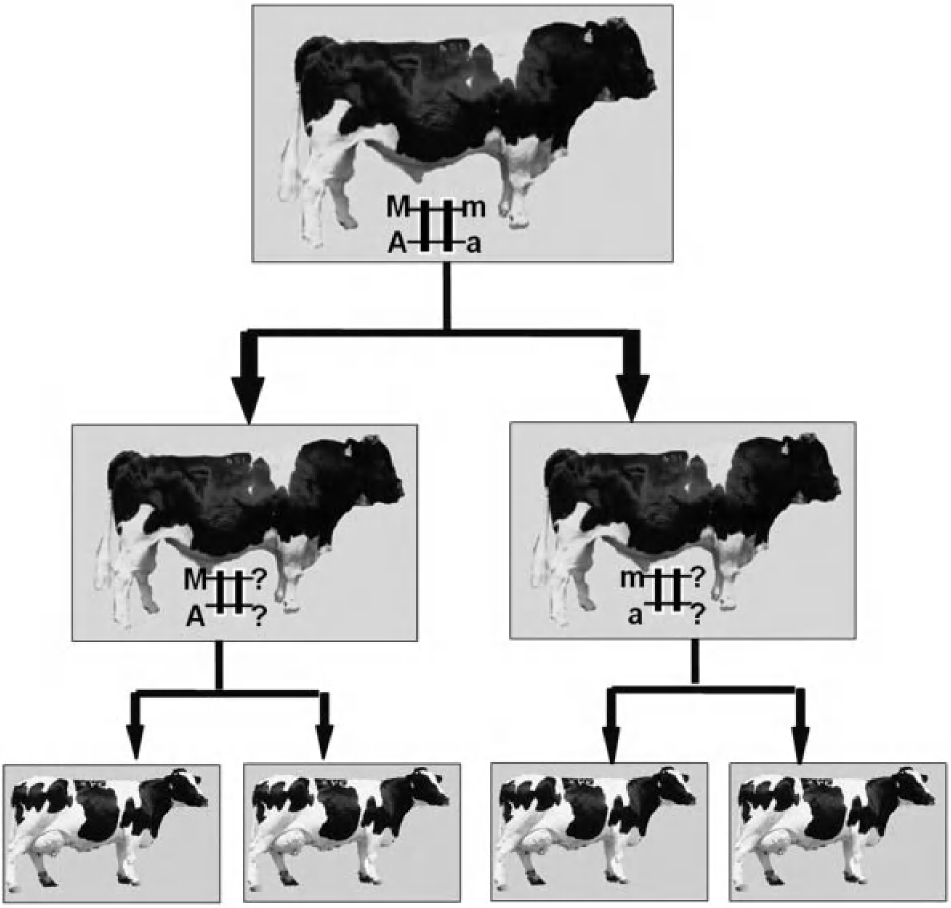


**Figure 12.2** The granddaughter design. The grandsire is assumed to be heterozygous for a quantitative trait loci (QTL) and a linked genetic marker. As in Figure 12.1, only a single family is shown. The two alleles of the marker locus are denoted "M" and "m," and the two alleles of the QTL are denoted "A" and "a." Alleles of maternal origin are denoted by question marks. Genotypes are not listed for the granddaughters because they are not genotyped.

For both designs, only the alleles of the sire are followed in the progeny. Linkage phase between QTL and genetic markers will differ among the families are analyzed. Therefore, any specific QTL will be heterozygous in only a fraction of the families included in the analysis. Thus, QTL effects must be estimated within families, and these designs are, therefore, less powerful per individual genotyped than designs based on crosses between inbred lines (Weller 2009). Furthermore, these designs have the disadvantage that progeny with the same genotype as the sire are uninformative, because the progeny could have received either paternal allele. This problem is alleviated if multiple linked markers are genotype. In this case it should possible to determine for nearly all progeny which paternal haplotype was received.

For the daughter design, power of 0.7, with a type I error of 0.01 is obtained for a QTL with a substitution effect of 0.2 phenotypic standard deviations if 400 daughters of each of 10 sires are analyzed for a trait with heritability of 0.2. This entails genotyping 4000 individuals. Power is maximized when the frequency of the two QTL alleles is equal. The granddaughter design has the advantage of greater statistical power per individual genotyped. Because each genotype is associated with multiple phenotypic records, power per individual genotyped in the granddaughter design can be fourfold the power of the daughter design (Weller et al. 1990). With heritability of 0.2 and a type I error of 0.01, power is 0.74 to detect a segregating QTL with a substitution effect of 0.2 phenotypic standard deviations if genetic markers are analyzed on 100 sons of each of 10 grandsires, with 50 quantitative trait-recorded granddaughters per son. Comparing this example to the previous example for the daughter design, greater power is obtained to detect an effect of the same magnitude with the granddaughter design, even though only one-quarter the number of individuals are genotyped (4000 vs. 1000). The disadvantage of the granddaughter design is that the appropriate data structure; hundreds of progeny tested bulls, sons of a limited number of sires; is found only in the largest dairy cattle populations.

Additional experimental designs have also been proposed. Coppieters et al. (1999) proposed the "great-granddaughter design." One of the disadvantages of the granddaughter design is that the number of progeny-tested sons of most sires is too low to obtain reasonable power to detect QTL of moderate effects. Coppieters et al. (1999) proposed that power can be increased by also genotyping progeny-tested grandsons of the grandsire. Inclusion of the grandsons is complicated by the fact that there is another generation of meiosis between the grandsire and his grandson.

A significant drawback of all the designs considered so far is that they give no indication as to the number of QTL alleles segregating in the population or their relative frequencies. To answer this question, Weller et al. (2002) proposed the "modified granddaughter design" presented in Figure 12.3. Assume that a segregating QTL for a trait of interest has been detected and mapped to a short chromosomal segment using either a daughter or a granddaughter design. Consider the maternal granddaughters of a grandsire with a significant contrast between his two paternal alleles. This grandsire will be denoted the "heterozygous grandsire." Each maternal granddaughter will receive one allele from her sire, who is assumed to be unrelated to the heterozygous grandsire; and one allele from her dam, who is a daughter of the heterozygous grandsire. Of these granddaughters, one-quarter should receive the grandpaternal QTL allele with the positive effect, one-quarter should receive the negative grandpaternal QTL allele, and half should receive neither grandpaternal allele. In the third case, the granddaughter received one of the QTL alleles of her granddam, the mate of the heterozygous grandsire. These granddams can be considered a random sample of the general population with respect to the allelic distribution of the QTL. All genetic and environmental effects not linked to the chromosomal segment in question are assumed to be randomly distributed among the granddaughters, or are included in the analysis model. Thus, unlike the daughter or granddaughter designs, it is possible to compare the effects of the two grandpaternal alleles to the mean QTL population effect.

Assuming that the QTL is "functionally biallelic" (that is, there are only two alleles with differential expression relative to the quantitative trait), and that allele origin can be determined in the granddaughters, the relative frequencies of the two QTL alleles in the population can be determined by comparing the mean values of the three groups

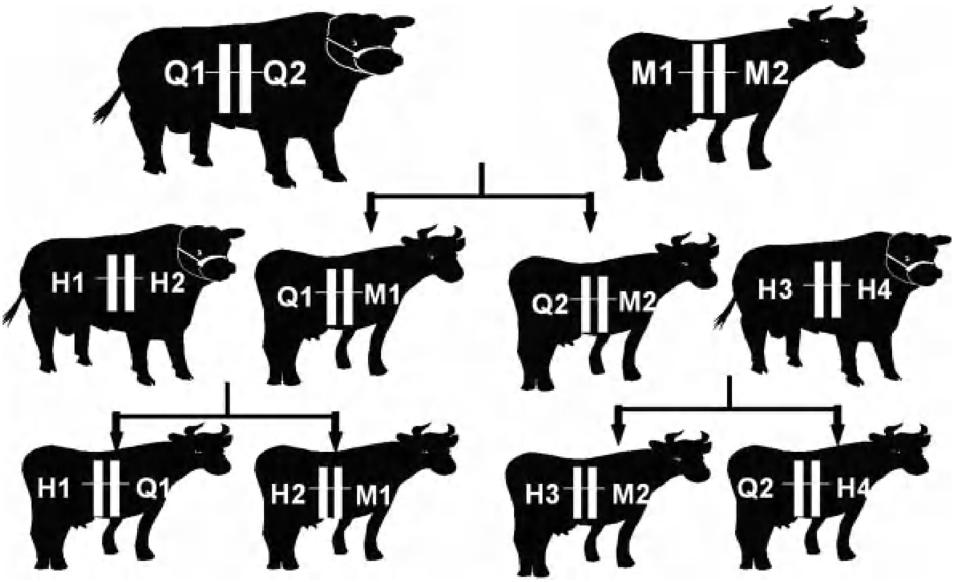


**Figure 12.3** The modified granddaughter design. Only alleles for the quantitative trait loci (QTL) are shown. Alleles originating in the heterozygous grandsire are termed "Q1" and Q2." Alleles originating in the granddams are termed "M1" and "M2." Alleles originating in the sires are termed "H1," "H2," "H3," and "H4."

of granddaughters for the quantitative trait. Using the modified granddaughter design, it is also possible to estimate the number of alleles segregating in the population, and to determine if the same alleles are segregating in different cattle populations. Weller et al. (2002) estimated the frequency of the QTL allele that increases fat and protein concentration on *Bos taurus* chromosome BTA6 in the Israeli Holstein population as 0.69 and 0.63, relative to fat and protein percent, by the modified granddaughter design. This corresponds closely to the frequency of 0.69 estimated for the *Y581* allele of the *ABCG2* gene for cows born during the same time period (Cohen-Zinder et al. 2005).

## Methods to Estimate QTL Effects and Location in Dairy Cattle

If a significant effect on a quantitative trait is associated with a genetic marker, the difference between the means of marker genotype classes will be a biased estimate of the QTL effect, due to recombination between the QTL and the genetic marker. Furthermore, it is not possible to estimate QTL location from the means of the marker classes. Weller (1986) first demonstrated that maximum likelihood (ML) methodology could be used to obtain estimates of QTL location and effect, unbiased by recombination. Lander and Botstein (1989) proposed interval mapping, based on ML for a QTL bracketed between two markers. Haley and Knott (1992), and Martinez and Curnow (1992) proposed an interval mapping method based on nonlinear regression, which was easier to apply than ML, and can readily handle missing genotypes on some of the markers.

Their methods are not directly applicable to half-sib designs, because, as noted previously, linkage relationships between the QTL and the genetic markers will be different across families, and in some families the common ancestor will be homozygous for the QTL. Furthermore, if multiple QTL alleles are segregating in the population, or if the observed effect is due to several tightly linked QTL, the magnitude of the effect will also differ across families. Methods suitable for interval mapping that account for these problems were developed by Knott et al. (1996), and have been applied to nearly all daughter and granddaughter design analyses. Their method assumes a single QTL location for all families, but estimates a separate effect for each family. The analysis model is as follows:

$$Y_{\mathrm{ijk}} = \mu_{1\mathrm{i}}(1 - p_{\mathrm{ij}}) + \mu_{2\mathrm{i}} p_{\mathrm{ij}} + e_{\mathrm{ijk}},$$

where $Y_{\mathrm{ijk}}$ is the trait record for individual k of family i with marker genotype j, $\mu_{1\mathrm{i}}$ and $\mu_{2\mathrm{i}}$ are the means for progeny that received paternal QTL alleles 1 and 2 in family i, $p_{\mathrm{ij}}$ is the probability that a progeny of sire i with marker genotype j received paternal QTL allele 1, and $e_{\mathrm{ijk}}$ is the random residual. Although QTL location is assumed to be the same in all families, $p_{\mathrm{ij}}$ must be computed separately for each individual, because it will depend on which markers are informative in each progeny of each family.

Although estimation of confidence intervals (CI) are important both for parameter estimates of QTL effects and location, the literature has dealt chiefly with estimation of CI for QTL location. Lander and Botstein (1989) proposed the LOD-score (logarithm of the odds to the base 10) drop-off method to estimate CIs for QTL location, but several studies have shown that this method can seriously underestimate the actual value (e.g., Darvasi et al. 1993). Visscher et al. (1996) proposed that CIs could be estimated by the "nonparametric bootstrap" method. In this method, a large number of samples of the same size as the actual data set are drawn from the data *with repeats*. Thus, in a particular bootstrap sample, the first observation may appear twice, and the second observation not at all. The parameters of interest are then estimated from these samples, and the distribution of the parameter estimates can be used to estimate the CI for all of the QTL parameters. This method tends to overestimate the CI for QTL location. Bennewitz et al. (2003) proposed improvements to the bootstrap method that results in shorter CIs that are still unbiased.

## Difficulties and Biases in QTL Analysis

Most studies to detect QTL have considered many markers and multiple traits. In some studies nearly the entire genome was analyzed. This generates a serious problem with respect to the appropriate threshold to declare significance. If normal pointwise significance levels of 5 or 1% are used, many marker–trait combinations will show "significance" by chance. This problem is even more severe if significance is determined for each half-sib family, in addition to multiple markers and traits. Several solutions to this problem have been proposed, none of which are completely satisfactory.

Churchill and Doerge (1994) proposed to empirically estimate rejection thresholds for the null hypothesis of no segregating QTL by a "permutation test." In this method many different samples are generated from the actual data by "shuffling" the trait values with respect to the marker genotypes. That is, each individual genotyped is assigned the trait values of a randomly selected individual from the sample. Since the trait values for all individuals are now random with respect to marker genotypes, the null hypothesis of no linkage between the genetic markers and QTL is correct by definition. The test statistics computed from these "permutation samples" are then used to construct the empirical distribution of the test statistic under the null hypothesis. The appropriate rejection threshold for any desired type I error rate can then be derived from the empirical distribution of the test statistic. This method has the advantage that no assumptions are required with respect to distributional properties of either the quantitative traits or the genetic markers. Rejection thresholds are computed based on the actual number and genomic distribution of markers genotyped. The disadvantage of this method is that thresholds must be computed anew by permutation for each chromosome of each data set analyzed. Most studies that have analyzed multiple chromosomes applied a permutation analysis to a single chromosome of intermediate length to compute a "chromosome-wise" error rate, and then used the "Bonferroni adjustment" (Simes 1986) to compute the "genome-wise" error rate.

The only solution that is able to adequately deal with both multiple traits and families in addition to multiple markers is the "false discovery rate" (FDR) (Weller et al. 1998). The FDR is defined for a specific nominal probability value as the ratio of the expected number of tests for which the null hypothesis is rejected to the observed number of rejected tests. For example, if 1000 tests are performed, it is expected that by chance that 10 will have nominal probability values $<0.01$. If in fact 40 tests had probability values $<0.01$, then the FDR $= 10/40 = 0.25$. That is, in this case 75% of the "significant" test should represent true effects.

The QTL effects derived from either daughter or granddaughter will still be biased for several reasons. First, the usual assumptions of interval mapping, a single QTL segregating within the marker interval and no QTL in adjacent intervals, often do not reflect reality. Second, the dependent variable is generally an "adjusted" record, either daughter yield deviations (DYD) (VanRaden and Wiggans 1991) or estimated breeding value (EBV). EBV computed by best linear unbiased prediction methodology are regressed in proportion to the amount of information available for each animal. Thus if QTL effects are estimated by analysis of EBV, the effects will be underestimated (Thomsen et al. 2001). The problem is somewhat alleviated if DYD, which are unregressed means of daughter records corrected for fixed effects, are analyzed instead of EBV. Unlike EBV, the variances of DYD decrease with increase in the number of daughters. Thus, weighting DYD by the bulls' reliabilities in an analysis of marker effects is in accordance to the generalized least squares principle that records with greater variance should be given smaller weights. Israel and Weller (1998) demonstrated that QTL effects derived from analysis of either EBV or DYD will be underestimated.

In addition to this downward bias, there are two sources of upward biases for QTL effects. First, the direction of the effects is generally arbitrary. Thus, absolute values are retained, and all effects are $>0$. Since in nearly all QTL analyses only the absolute value of the effect is considered, least squares estimates will be inflated due to nonzero residuals. Assuming that the residual and the actual QTL effects are uncorrelated, the variance of the least squares estimates will be equal to the sum of the residual and true QTL effect variance. Thus, positive values for QTL effects will be obtained even in the absence of a segregating QTL.

Finally, only the effects deemed "significant" are retained, and this is a selected sample. Beavis (1994) first noted that the estimates for effects deemed "significant" will be biased, due to the fact that only the largest effects will be reported and retained for further analysis. Bias will be greater for small effects for the following reason. Although large QTL effects will be deemed significant in any case, small or marginal effects will be denoted "significant" only if the estimate is larger than the actual effect (Georges et al. 1995).

## The Current State of QTL Detection in Dairy Cattle by Within-Family Linkage Studies

Genome scans by the granddaughter design have been completed for Holsteins from Canada (Nadesalingam et al. 2001), the Netherlands (Spelman et al. 1996; Schrooten et al. 2000), France (Bennewitz et al. 2003a; Boichard et al. 2003), Germany (Kuhn et al. 2002; Bennewitz et al. 2003a), New Zealand (Spelman et al. 1999a), and the United States (Georges et al. 1995; Ashwell et al. 1996, 1997, 1998, 1998a, 2001, 2004, 2005; Zhang et al. 1998; Ashwell and Van Tassell 1999; Heyen et al. 1999); Finnish Ayrshires (Vilkki et al. 1997; Viitala et al. 2003; Schulman et al. 2004); French Normande and Montbeliarde cattle (Boichard et al. 2003); Norwegian cattle in Norway (Klungland et al. 2001; Olsen et al. 2002); and SRB in Sweden (Holmberg and Andersson-Eklund 2004). Daughter design analyses have been performed for Israeli Holsteins (Mosig et al. 2001; Ron et al. 2004; Weller et al. 2008). Most studies have considered the five economic milk production traits; milk, fat, and protein production; and fat and protein concentration, although a number of studies have also considered somatic cell score (SCS), female fertility, herdlife, calving traits, twinning rate, health traits, temperament, and conformation traits. The SCS is a log function of the concentration of somatic cells, and has been shown to be a useful indicator of udder health.

Khatkar et al. (2004) performed a meta-analysis, combining data from most of these studies, and found significant across-study effects on chromosomes 1, 3, 6, 9, 10, 14, and 20.

Results for milk, fat, and protein production, fat and protein concentration, SCS, and many other traits, including meat production traits are summarized at: http://www.animalgenome.org/cgi-bin/QTLdb/BT/index. Significant effects were found on all 29 autosomes, but most effects were found only in single studies and have not been repeated.

## Genome Scans, Within-Family Linkage vs. Populationwide Linkage Disequilibrium

Although, as noted previously, the current genetic maps for the important livestock species consists of thousands of genetic markers, increasing marker density beyond a marker each 10 cM will generally have very little effect on the length of the CI for within-family mapping. With a marker-saturated genetic map, the length of the 95% CI, $\text{CI}_{(0.95)}$, by linkage analysis for backcross designs between breeds, daughter, or granddaughter designs can be estimated as follows:

$$\text{CI}_{(0.95)} = 3073/(d^2N)$$

Where $d =$ the QTL substitution effect in units of the "trait" standard deviation (EBV or DYD), and $N =$ the number of individuals genotyped (Weller and Soller 2004). Thus, over 1000 individuals must be genotyped to obtain a $\text{CI}_{(0.95)}$ of 10 cM, if the substitution effect is 0.5 standard deviations.

Meuwissen and Goddard (2000) proposed that CIs for QTL location could be dramatically reduced by application of populationwide LD mapping. Unlike analysis of genetic linkage within families that extends over tens of cM, populationwide LD extends in dairy cattle at most only over individual cM (Sargolzaei et al. 2008). Application of LD mapping, therefore, requires much denser genetic maps than required for detection of QTL via linkage, but does not require a specific family structure (Grapes et al. 2004). Since the first reports of genome-wide associations studies (GWAS) based on dense arrays of SNPs in 2006 (Goddard et al. 2006), very few specific results of effects detected and their locations have been published. All studies agree that based on the criterion of the FDR, many more statistically significant SNP effects have been found, as compared to linkage-based genome scans (Loberg and Dürr 2009; Cole et al. 2009; VanRaden et al. 2009). The apparent reasons are as follows:

1. LD effects can be tested by a simple linear model of number of "+" alleles on the bulls' EBV or DYD (VanRaden et al. 2009). This test is inherently more powerful than test of significance for within-family linkage.
2. Genome coverage with tens of thousands of SNPs is more complete.
3. Unlike granddaughter designs, which can only utilize bulls from large half-sib families, all sires with evaluation can be included in LD analyses. Thus, effective sample sizes are greater.

The results of Cole et al. (2009) and VanRaden et al. (2009) for the US Holstein population confirm that at least with respect to the quantitative trait nucleotides (QTN) that have been detected, results of GWAS do correspond to the results obtained previously by granddaughter designs. The largest effects found were for protein concentration on BTA6 and fat concentration on BTA14. The effects on BTA6 flanked the *ABCG2* gene, which has been shown to have a major effect for this trait (Cohen-Zinder et al. 2005), and the effects on BTA14 flanked the *DGAT1* gene (Grisart et al. 2002), which has a major effect on fat concentration with lesser effects on milk and fat yield. Both effects were first discovered by daughter and granddaughter designs.

## Estimation of QTL Effects from Genome Scans

Application of MAS based on GWAS requires solutions to new statistical problems. Specifically, how should information from pedigree, phenotypic records, and genotypes be combined to optimally rank candidates for selection? Goddard and Hayes (2007) proposed that genomic selection (GS) could be divided into the following three steps:

1. Use the markers to deduce the genotype of each animal at each QTL.
2. Estimate the effects of each QTL genotype on the trait.
3. Sum all the QTL effects for selection candidates to obtain their genomic estimated breeding values (GEBV).

Estimation of QTL effects by LD from genome scans is problematic; first, because only a very small fraction of the population will be genotyped. Furthermore, these will generally be males without records on the traits of interest. In addition, there are at least five potential sources of bias. First, estimated effects will be underestimated because LD between any specific marker and the QTL will be incomplete. Thus, as noted previously for linkage mapping, only part of the QTL effect will be detected. Second, if the effects of the SNP are analyzed on sires, then the dependent variable analyzed will generally be the sires' EBV or DYD, as also noted for linkage mapping.

In addition, there are three sources of upward bias. In a simple regression model, animals that are related will tend to have a higher probability to inherit the same marker alleles identical by descent. Since these animals also have a common polygenic variance, the estimated effect will also include a polygenic component due to relationships (Calus and Veerkamp 2007). Various studies have proposed that this problem could be solved by inclusion of the inverse of relationship matrix in the analysis (e.g., VanRaden 2008). However, if EBV or DYD are analyzed, and the QTL effect is assumed to be a random variable, the distributional properties of the model are problematic. If the residual variance in this model represents deviation of the "record" from the animal's true additive variance, a DYD or EBV based on thousands of daughters should then have a residual variance approaching zero.

The second and third sources of upward bias are also the same as for linkage mapping. Only a small fraction of the markers will have "significant" effects on the quantitative traits, and this will be a selected sample. The third source of upward bias results from the fact that when an LOD score or regression effect is maximized over many pointwise tests in interval mapping, the locus-specific effect-size estimate is also maximized, and this will tend to be greater than the actual QTL effect (Goring et al. 2001).

## Studies on the Distribution of QTL Effects

Several studies have proposed that the second source of upward bias could be alleviated by application of Bayesian estimation methods, for example, Hayes and Goddard (2001). Bayesian estimation differs from ML estimation in that in Bayesian estimation the likelihood function is multiplied by the "prior distribution" of the parameters. This generally results in "shrinkage" of the parameter estimates, but requires assumptions about the nature of distribution of QTL effects. If many QTL are analyzed jointly, it should be possible to estimate both the QTL effects and the parameters of the distribution of QTL effects (Weller et al. 2005). The extent of shrinkage will vary inversely with the sample size.

Hayes et al. (2006) estimated that the number of detectable QTL affecting milk production is on the order of 150, based on a whole genome scan with 10,000 SNPs, while Chamberlain et al. (2007) estimated the total number of QTL at 30 by analysis of a daughter design. These studies demonstrate that most of the additive genetic variance can be explained by QTL that can be detected by SNP genome scans, provided that the number of animals analyzed and the SNP densities are sufficient.

A number of studies have considered the question of the appropriate distribution for QTL effects, and in nearly all cases a single-sided distribution was assumed; that is, the QTL effect was assumed to vary from zero to infinity. Hayes and Goddard (2001) estimated the distribution of QTL effects for cattle and swine by combining results from several studies. They assumed a gamma distribution for the QTL effects. The dairy cattle analysis was based on the QTL estimates from three granddaughter design analyses, considering only "significant" effects. Thus, a truncated gamma distribution was assumed. Weller et al. (2005) estimated the parameters of the distributions of QTL effects for nine economic traits in dairy cattle from a daughter design analysis of the Israeli Holstein population including 490 marker-by-sire contrasts (Ron et al. 2004). A separate gamma distribution was derived for each trait. The estimates derived for the individual QTL effects using the gamma distributions for each trait were regressed relative to the least squares estimates, but the regression factor decreased as a function of the least squares estimate. On simulated data, the mean of least squares estimates for effects with nominal 1% significance was more than twice the simulated values, while the mean of the ML estimates was slightly lower than the mean of the simulated values. The coefficient of determination for the Bayesian estimates was fivefold the corresponding value for the least squares estimates.

## Appropriate Criteria for Evaluation of GEBV

Evaluation of methodology to compute GEBV are generally based on analysis of a population of bulls with EBV derived from progeny tests. The population is then divided into two sets. In the "training set" consisting of older bulls, all information including genotypes, genetic relationships, and daughter records are used to estimates the marker effects. In the "validation set" consisting of younger bulls, GEBV are computed based only on the marker genotype effects derived from the training set and pedigree. The GEBV of the validation set of bulls are then compared to the EBV of these bulls based on their daughter records and relationships (Hayes et al. 2009; VanRaden et al. 2009).

Most studies that have compared GEBV to EBV based on daughter records have done so on the basis of coefficients of determination between the two evaluations (Hayes et al. 2009; VanRaden et al. 2009; Su et al. 2010). Although this criterion is important, a second criterion is the bias of GEBV. That is, GEBV are unbiased if the regression of true breeding values on GEBV is not significantly different from unity and the y-intercept is not significantly different from zero (Aguilar et al. 2010). If the regression is less than unity, then the bulls with the highest GEBV will be inflated relative to the true genetic values of these bulls.

## Implementation of Methodology to Compute GEBV for Dairy Cattle

Most studies that have proposed implementation of MAS for dairy cattle have assumed a breeding program based on the progeny test scheme (e.g., Spelman et al. 1999). An example of a progeny test scheme for a moderately sized population is given in Figure 12.4. In the progeny test scheme, a cohort of young bulls is evaluated on the basis of a first crop of 50–100 daughters per bull. About 10% of the bulls with the highest genetic evaluations are then returned to general service. However, at this point the cohort of bulls will be at least 5 years old. If high accuracy genetic evaluations could be obtained for these bulls based on genetic markers at the age of 1 year, then the sum of the generation intervals along the four paths of inheritance could be reduced by about 4 years. This would increase the annual rate of genetic gain by about 20% (Weller 2009).

"Interbull," a subcommittee of the International Committee for Animal Recording, is responsible for promoting the development and execution of international genetic evaluations for cattle. Nine countries, (1) Australia, (2) Canada, (3) France, (4) Germany, (5) Ireland, (6) Israel, (7) New Zealand, (8) the Netherlands, and (9) the United States, responded to the Interbull survey question: "Which methodology is being used to estimate SNP effects?" (Loberg and Dürr 2009). Several different methods are being implemented. Four countries have adopted Bayesian methods, described briefly previously. "Bayes-A" and "Bayes-B" methodology differ in that in Bayes-A all marker included in the analysis are assumed to have a nonzero effect on the trait analyzed, while in Bayes-B methodology it is assumed that most markers have no effect (Meuwissen et al. 2001). Bayes-B methods are clearly closer to reality, but require significantly greater computing time. However, Su et al. (2010) found in the analysis

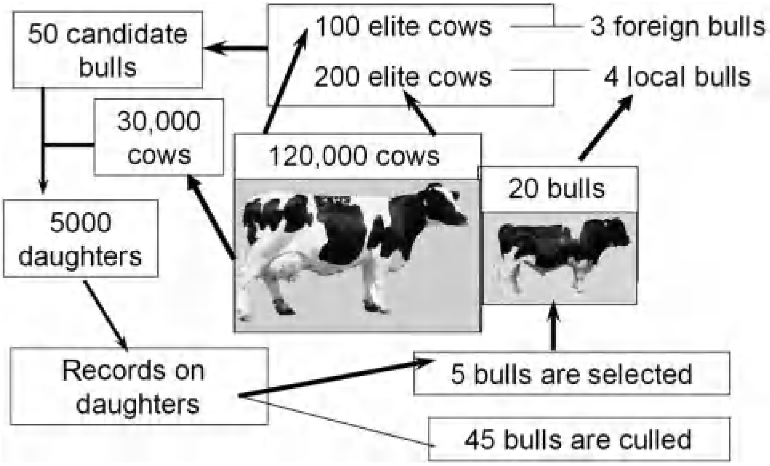


**Figure 12.4** The Israeli Holstein breeding program.

of the Danish Holstein population that the highest coefficient of determination of actual EBV was obtained with a common prior for all markers, that is Bayes-A.

VanRaden (2008) proposed analysis of DYD as the dependent variable with all SNPs included as random effects. Genotypes for 38,416 informative markers of the 54,001 included on the BeadChip, and the August 2003 genetic evaluations for 3576 Holstein bulls born before 1999 were used to predict the January 2008 daughter deviations for 1759 bulls born from 1999 through 2002. GEBV were computed using linear and nonlinear genomic models. For linear predictions, the traditional additive genetic relationship matrix was replaced by a genomic relationship matrix, which is equivalent to assigning equal genetic variance to all markers. Final GEBV combined three terms by selection index:

1. Direct genomic prediction.
2. Parent averages computed from the set of genotyped ancestors using traditional relationships.
3. Published parent averages or pedigree indexes, constructed as 0.5(sire EBV) + 0.25(maternal grandsire EBV) + 0.25(birth year mean EBV).

Combined predictions were more accurate than official parent averages for all 27 traits analyzed (VanRaden et al. 2009). Reliabilities were 0.02–0.38 higher with nonlinear genomic predictions included as compared to parent averages alone. Misztal et al. (2009) found that regressions of EBV based on progeny tests on GEBV derived by this method were less than unity for the trait “final score” (a conformation trait). Regressions for other traits have not been published. Variations of the method of VanRaden (2008) have also been applied to other national dairy cattle populations, for example, Liu et al. (2009). GEBV have been published in the United States since April 2008. Nearly all of the top US bulls are currently young bulls with GEBV, but without progeny tests (http://aipl.arsusda.gov/dynamic/sortnew/current/OHOnm.html).

The same data were also analyzed by Bayes-A and Bayes-B procedures (VanRaden 2008). In the Bayes-A analysis, the prior distribution was a simple, heavy-tailed distribution generated from a normal variable divided by $1.25^{abs(s-2)}$, where $s$ is the number of standard deviations from the mean and 1.25 determines departure from normality. In the Bayes-B analysis, only 38,416 markers of 50,000 included in the analysis were assumed to have nonzero effects. Gains in realized reliabilities were minimal, as compared to the linear model (VanRaden et al. 2009). Similar results were found for Australian dairy cattle (Hayes et al. 2009). This is not surprising considering that very large samples of bulls were analyzed. Therefore, “shrinkage” of estimates by Bayesian methodology should be minimal.

Advantages of the multistage system for genomic evaluation include no change to the regular evaluations and simple steps for predicting genomic values for young genotyped animals. Disadvantages include weighting parameters, such as variance components (Guillaume et al. 2008) or selection index coefficients (VanRaden 2008), loss of information, and biased evaluations (Misztal et al. 2009). Furthermore, the extension to alternative analysis models, such as multitrait evaluations or test-day models, is not obvious. Finally, tracing back anomalies in a two- or three-step procedure might become very complicated. As for the loss of information, several problems exist in the use of DYD for bulls or “yield deviations” for cows (VanRaden and Wiggans 1991). These problems are weights (caused by different amount of information in the original data set), bias (e.g., caused by selection), accuracy (for animals in small herds), and collinearity (e.g., the yield deviations of two cows in the same herd). Finally, the expectation of Mendelian sampling in selected animals is not zero (Party and Ducrocq 2009).

Other studies proposed methods for deriving GEBV from direct analysis of the complete population, even though only a small fraction is actually genotyped. Legarra et al. (2009) assumed that SNP effects are random, with conditioning of the genetic value of ungenotyped animals on the genetic value of genotyped animals via the selection index (e.g., pedigree information), and then used the genomic relationship matrix for the latter. This results in a joint distribution of genotyped and ungenotyped genetic values, with a pedigree-genomic relationship matrix H. In this matrix, genomic information is transmitted to the covariances among all ungenotyped individuals. Matrix H is suitable for iteration on data algorithms that multiply a vector times a matrix, such as preconditioned conjugated gradients (Misztal et al. 2009).

This method was applied to 10,466,066 US Holsteins records for final score (Aguilar et al. 2010). GEBV were computed based on 6508 bulls genotyped for the Illumina BovineSNP50 BeadChip and records up to 2004. GEBV were compared to those obtained by the multistep method (VanRaden 2008) on the same data. Comparisons were based on regressions of 2009 EBV of bulls without daughter records prior to 2005 on GEBV, and coefficients of determination. This approach includes a parameter, $\lambda$, which represents the fraction of the additive variance explained by the genomic information. By estimating $\lambda$ the goodness of “genomic” fit can be determined without creating the training and validation populations. Subsequently, comparisons of different models are simplified. With “optimal” scaling, this method was more accurate and less biased than the multistep method (Aguilar et al. 2010).

Hayes et al. (2009a) proposed to use genotypes to construct a “realized” relationship matrix between individuals, as opposed to the average relationship matrix generally included in animal model evaluations. In the realized relationship, matrix elements are the realized proportion of the genome that is identical by descent between pairs of individuals, based on shared marker haplotypes. They demonstrated that by replacing the average relationship matrix derived from pedigree with the realized relationship matrix, the accuracy of the breeding values can be substantially increased, especially for individuals with no phenotype of their own. Hayes et al. (2009a) also demonstrated that this method of predicting breeding values is exactly equivalent to the GS methodology where the effects of QTL contributing to variation in the trait are assumed to be normally distributed. The accuracy of breeding values predicted using the realized relationship matrix can be deterministically predicted for known family relationships, for example, half-sibs.

## Identification of Quantitative Trait Nucleotides

Neither the earlier MAS programs based on SSR, nor current programs based on high-density arrays of SNP, assumed that the actual QTN responsible for the observed genetic variance were identified. Rather, frequency of the desirable alleles could be increased via selection for linked markers.

Methods developed to find QTN suitable for plants and model animals cannot be applied to most livestock species because of the lack of inbred lines, the long generation interval, the cost of each animal, and difficulty to produce transgenics or "knockouts." Glazier et al. (2002) noted that the most conclusive evidence that the QTN has been found is a demonstration that replacement of the variant nucleotide results in swapping one phenotypic variant for another. Currently, this is not possible for livestock species. Considering these limitations, how does one prove that a candidate polymorphism is in fact the QTN? As noted by Mackay (2001), "The only option . . . is to collect multiple pieces of evidence, no single one of which is convincing, but which together consistently point to a candidate gene."

Despite these limitations, at least four QTN have been identified and verified by multiple studies in farm animals (Ron and Weller 2007). Of these, two are in dairy cattle, *DGAT1* and *ABCG2*. As noted previously, the most significant effects for both genes are on fat and protein concentration, which have the highest heritabilities of all the traits routinely analyzed in dairy cattle.

Ron and Weller (2007) presented a general scheme to progress from QTL identification to QTN determination. Their proposal is based on first reduction of the CI for QTL location to individual map units, and then LD analysis of the polymorphisms within the reduced CI. The final step in determination of the QTN is to find a polymorphism within the CI with "concordance" with the QTL genotypes. Although QTL genotypes of individuals cannot be determined by the observed phenotype for the quantitative traits, in application of daughter or granddaughter designs, it is possible to determine the QTL genotypes of the family patriarchs with a high degree of certainty (Israel and Weller 2004). Complete concordance is obtained only if:

1. all individuals known to be homozygous for the QTL are also homozygous for the polymorphism;
2. all individuals heterozygous for the QTL are also heterozygous for the polymorphism; or
3. the same QTL allele is associated with the same allele of the putative QTN for all the heterozygous animals.

For livestock, concordance is the most impressive proof that the QTN has been detected, and all four QTN detected so far have used this criterion as the primary proof. Of course, it is possible that a QTN may not display complete concordance. First, sire genotypes may be misclassified, especially if either the QTL effect or the number of progeny used to determine the QTL genotype is relatively small. Second, complete concordance is expected only if the QTL effect is due to a single dimorphic site. Kuhn et al. (2004) found that four German Holstein sires segregating for the QTL on BTA14 were homozygous for the K232N polymorphism in *DGAT1*. They conclude that the effect in these sires is due to a SSR polymorphism of the *DGAT1* promoter. Of course, the possibility still exists that neither mutation is the QTN, and that a third, undiscovered site explains all of the variation for this QTL.

Concordance can only be considered a proof of QTN detection if the probability of concordance by chance within the CI is sufficiently low so that this hypothesis can be statistically rejected. Assuming that only two alleles are segregating in the population, the probability that a specific polymorphism will show concordance ($p_c$) is computed as follows (Ron and Weller 2007):

$$p_c = \int_0^1 (2[p(1-p)]^n[1-2p(1-p)]^m)\mathrm{d}p,$$

where $p$ is the probability of one of the two alleles, 1–$p$ is the probability of the other allele, and $m$ and $n$ are the numbers of patriarchs homozygous and heterozygous for the QTL, respectively. The expectation of the number of SNP with complete concordance within the CI can be estimated as: $Sp_c$, where $S$ is the expected number of SNP within the CI. SNPs occur at a frequency of approximately 0.3–1 SNP/kbp throughout the human genome (Marth et al. 2001). Assuming a similar frequency for the cattle genome, a CI of 1 Mbp (~1 cM) will include 1000–3000 SNP. The hypothesis of concordance by chance can then be rejected if $P_{s>0} < p_1$, where $P_{s>0}$ is the probability that any SNP within the CI will display concordance by chance, and $p_1$ is the type I error required for rejection of the null hypothesis. Assuming the standard value of 0.05 for the type I error, the critical $S$ value, $S_c$, for which $P_{s>0} \leq 0.05$ for any given values of $n$ and $m$ can then be estimated as: $0.05/p_c$. Although increasing the number of QTL homozygotes does increase the value of $S_c$, increasing the number of heterozygotes has a much greater effect. Five homozygotes and five heterozygotes are required to obtain an $S_c$ value >1000. That is, assuming a SNP density of one SNP per kbp, for an interval of 1 Mbp or 1 cM, the probability of concordance by chance is <0.05. With ten homozygotes and eight heterozygotes, $S_c$ approaches three million, which covers the entire length of the genome, again assuming one SNP per kbp.

## Validation of Quantitative Trait Nucleotides

Both statistical and physiological methods can be applied to validate a putative QTN in dairy cattle. Statistical methods, summarized by Cohen-Zinder et al. (2005), include demonstrating the following:

1. The effect of the putative QTN accounts for the entire effect observed by interval mapping.
2. No other polymorphisms in LD with the QTL have significant effects in models that also include the effect of the putative QTN.
3. The same QTN is segregating in diverse populations.
4. Changes in the allelic frequencies of the QTN correspond to the changes expected due to selection in the population.

Relative to the most likely QTL location, the effect of the single generation of recombination between the patriarchs and their progeny on the observed QTL effect in a backcross, daughter, or granddaughter design will be minimal. This will not be the case for the effect of a marker on the quantitative trait as estimated from a random sample of individuals from the population. In this case, many generations of recombination will reduce the observed effect, even if the marker is tightly linked to the QTN. Thus, demonstrating that the effect on the quantitative trait associated with the putative QTN in a random sample of individuals is equal to the effect estimated by interval mapping in a daughter or granddaughter design is a strong indication that the QTN has been correctly determined. Similarly, if the QTN has been misidentified, then other linked markers should still have effects on the quantitative trait if a random sample of individuals from the population is analyzed, even though the putative QTN is included in the model. However, even if the QTN has been correctly identified, other linked markers could still have significant effects on the quantitative trait, if other QTL are segregating in the same chromosomal segment.

Functional assays of QTN validation include demonstrating either unequal production of the alternative alleles' products, or differences in protein function. "Knockout" mutations were demonstrated for both *DGAT1* and *ABCG2* in mouse (Smith et al. 2000; Jonker et al. 2005).

## Application of Identified QTL in Marker-Assisted Selection

For a specific identified QTL to be useful in a MAS program, it must fulfill the following criteria:

1. Due to the huge multiple comparison problem, confidence that a segregating QTL has in fact been detected is only obtained if the observation is repeated in several independent studies (Lander and Kruglyak 1995).
2. The effect should be sufficiently large, such that the CI for QTL location is small enough so that a haplotype including the QTL can be determined and followed through pedigrees.
3. The net effect of the QTL on the selection index must be significant.
4. Scope for selection is possible only if the frequency of the positive allele is relatively low.

It is quite difficult to find a single reported QTL that meets all these criteria. For example, of the two QTN that have been determined in cattle, *DGAT1* fails the third criterion, and *ABCG2* fails the fourth criterion. The allele of *DGAT1* that increases fat production reduces protein, thus the net effect on most selection indices will be close to zero (Weller et al. 2003). For *ABCG2*, one allele is clearly economically favorable for most selection indices currently in use, but this allele is already at a very high frequency in all dairy cattle populations analyzed (Ron et al. 2006).

Thus, the arguments against extending significant effort toward QTN detection are daunting, and can be summarized as follows:

1. The infinitesimal model appears to be approximately accurate for the traits of interest. That is, genetic variance is apparently due to the joint effect of many genes, all with very small effects.
2. Even if a QTN is detected, it may not be useful in selection.
3. The methods derived so far for GS appear to work well.
4. Detection of QTN is expensive and time consuming, especially if the effect is due to a more complicated genetic mechanism, such as copy number variation, or DNA methylation.

Despite the arguments against QTN determination given previously, several justifications can be given, such as the following:

1. Once the QTN is determined, this will yield useful information on gene function and QTL architecture. Although the two QTN determined in dairy cattle are both missense mutations, this is not the case for the two other QTN determined in other farm animal species (Ron and Weller 2007). Thus, it is still not clear if missense mutations are the exception or the rule for QTN in commercial animals.
2. Understanding the ties between genetic variation and functional characteristics of specific genes may contribute to drug discovery for the benefit of both farm animals and human.
3. As demonstrated for the case of *ABCG2*, although SNPs in close linkage to a major QTN will generally display highly significant effects, the effect will still be significantly less than the effect obtained with the QTN (Cohen-Zinder et al. 2005; Olsen et al. 2007). Thus, selection on the QTN may be more efficient than selection on a marker in LD.
4. LD relationships change over time, which will reduce the efficiency of selection.
5. LD relationships may be different between populations and thus, MAS may not be applicable in populations without GWAS.
6. Allelic frequencies for a marker in LD will not accurately reflect the allelic frequencies of the QTN. As noted previously, this information is very useful, as it gives a horizon for the gain that can be achieved by selection.
7. If the QTN is determined, then selection can also be applied to other populations and breeds, including those populations that have not been analyzed by a GWAS (e.g., Kaupe et al. 2004; Goddard et al. 2006; Ron et al. 2006).

Should QTN be treated differently than LD markers in genetic evaluation programs? The problem that only a small fraction of the population will be genotyped will still apply, but it would seem that once a QTN is detected, bias is no longer a factor, and the QTN can be treated as a fixed rather than random effect.

Finally, we should note that investment in breeding programs is unlike other investments in that the gains are eternal and cumulative. Thus, a relatively small change in the rate of genetic gain can have a huge economic value. As shown in an example given by Weller (1994), for a program with a constant annual investment, net profit will be positive within a 20-year profit horizon if nominal annual costs are less than threefold the nominal annual gain. Consider the US dairy cattle population with 10,000,000 cows. Current rates of genetic gain are approximately equal to 100 kg per year in terms of economically correct milk production. Thus, an additional 1% increase in the rate of genetic gain is approximately equal to 1 kg per cow per year. The nominal value of this gain is approximately \$0.1 per cow, or \$1 m for the entire industry. Thus, an annual investment of \$3 m in QTN detection can be justified even if it results in only a 1% increase in the rate of genetic gain.

## Conclusions

In the last 20 years, there have been huge advances in both DNA technology and statistical methodology. It can now be stated with near certainty that the technology is available to detect and accurately map segregating QTL in dairy cattle. Furthermore, although many effects reported in the literature are "false positives," there is a wealth of evidence that several QTL are in fact real. A number of effects have been repeated across numerous experiments, and the actual QTN have been identified for at least two QTL in cattle. Although GS can increase rates of genetic gain without determination of the causative QTN, determination of these QTN should increase rates of genetic gain, and aid in the understanding of the mechanisms through which the trait is affected. Investment in breeding programs is unlike other investments in that the gains are eternal and cumulative. Thus, a relatively small change in the rate of genetic gain can have a huge economic value.

## Acknowledgment

This research was supported by grants from the Israel Milk Marketing Board and the European Sixth Research and Technological Development Framework Programme, Proposal No. 016250–2 SABRE.

## References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J. (2010) *Hot topic:* A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* **93**: 743–752.

Anderson, E.C. and Graza, J.C. (2006) The power of single nucleotide polymorphisms for large scale parentage inference. *Genetics* **172**: 2567–2582.

Ashwell, M.S., Da, Y., VanRaden, P.M., Rexroad, C.E., Miller, R.H. (1998) Detection of putative loci affecting conformational type traits in an elite population of United States Holsteins using microsatellite markers. *Journal of Dairy Science* **81**: 1120–1125.

Ashwell, M.S., Da, Y., Van Tassell, C.P., VanRaden, P.M., Miller, R.H., Rexroad, C.E. (1998a) Detection of putative loci affecting milk production and composition, health, and type traits in a United States Holstein population. *Journal of Dairy Science* **81**: 3309–3314.

Ashwell, M.S., Heyen, D.W., Weller, J.I., Ron, M., Sonstegard, T.S., Van Tassell, C.P., Lewin H.A. (2005) Detection of quantitative trait loci influencing conformation traits and calving ease in Holstein-Friesian cattle. *Journal of Dairy Science* **88**: 4111–4119.

Ashwell, M.S., Rexroad, C.E., Miller, R.H., VanRaden, P.M. (1996) Mapping economic trait loci for somatic cell score in Holstein cattle using microsatellite markers and selective genotyping. *Animal Genetics* **27**: 235–242.

Ashwell, M.S., Rexroad, C.E., Miller, R.H., VanRaden, P.M., Da, Y. (1997) Detection of loci affecting milk production and health traits in an elite US Holstein population using microsatellite markers. *Animal Genetics* **28**: 216–222.

Ashwell, M.S. Van Tassell, C.P. (1999) Detection of putative loci affecting milk, health, and type traits in a US Holstein population using 70 microsatellite markers in a genome scan. *Journal of Dairy Science* **82**: 2497–2502.

Ashwell, M.S., Van Tassell, C.P., Sonstegard, T.S. (2001) A genome scan to identify quantitative trait loci affecting economically important traits in a US Holstein population. *Journal of Dairy Science* **84**: 2535–2542.

Ashwell, M.S., et al. (2004) Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. *Journal of Dairy Science* **87**: 468–475.

Beavis, W.D. (1994) The power and deceit of QTL experiments: lessons for comparative QTL studies. *Annual Corn Sorghum Research Conference.* Washington, DC, Vol. 49, pp. 252–268.

Bennewitz, J., Reinsch, N., Kalm, E. (2003) Comparison of several bootstrap methods for bias reduction of QTL effect estimates. *Journal of Animal Breeding and Genetics* **120**: 403–416.

Bennewitz, J., et al. (2003a) Combined analysis of data from two granddaughter designs: a simple strategy for QTL confirmation and increasing experimental power in dairy cattle. *Genetics Selection Evolution* **35**: 319–338.

Boichard, D., et al. (2003) Detection of genes influencing economic traits in three French dairy cattle breeds. *Genetics, Selection, Evolution* **35**: 77–101.

Bonin, A., Bellemain, E., Eidesen, P.B., Pompanon, F., Brochmann, C., Taberlet, P. (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology* **13**: 3271–3273.

Calus, M.P.L. and Veerkamp, R.F. (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics* **124**: 362–368.

Chamberlain, A.J., McPartlan, H.C., Goddard, M.E. (2007) The number of loci that affect milk production traits in dairy cattle. *Genetics* **177**: 1117–1123.

Churchill, G.A. and Doerge, R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.

Cohen-Zinder, M., et al. (2005) Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research* **15**: 936–944.

Cole, J.B., et al. (2009) Distribution and location of genetic effects for dairy traits. *Journal of Dairy Science* **92**: 2931–2946.

Coppieters, W., et al. (1999) The great-grand-daughter design: a simple strategy to increase the power of a grand-daughter design for QTL mapping. *Genetical Research* **74**: 189–199.

Darvasi, A., Vinreb, A., Minke, V., Weller, J.I., Soller, M. (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* **134**: 943–951.

Georges, M., et al. (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**: 907–920.

Glazier, A.M., Nadeau, J.H., Aitman, T.J. (2002) Finding genes that underlie complex traits. *Science* **298**: 2345–2349.

Glowatzki-Mullis, M.L., Gaillard, C., Wigger, G., Fries, R. (1995) Microsatellite based parentage control in cattle. *Animal Genetics* **27**: 7–12.

Goddard, M.E. and Hayes, B.J. (2007) Genomic selection. *Journal of Animal Breeding and Genetics* **124**: 323–330.

Goddard, M.E., Hayes, B., McPartlan, H., Chamberlain, A.J. (2006) Can the same genetic markers be used in multiple breeds? *Proc. 8th World Congress of Genetics Applied to Livestock Production.* Belo Horizonte (MG), Brazil, 22-14.

Goring, H.H., Terwilliger, J.D., Blangero, J. (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics* **69**: 1357–1369.

Grapes, L., Dekkers, J.C., Rothschild, M.F., Fernando, R.L. (2004) Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics* **166**: 1561–1570.

Grisart, B., et al. (2002) Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research* **12**: 222–231.

Guillaume, F., Fritz, S., Boichard, D., Druet, T. (2008) Short communication: correlations of marker-assisted breeding values with progeny-test breeding values for eight hundred ninety-nine French Holstein bulls. *Journal of Dairy Science* **91**: 2520–2522.

Haley, C.S. and Knott, S.A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315–324.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J, Goddard, M.E. (2009) Invited review: Genomic selection in dairy cattle progress and challenges. *Journal of Dairy Science* **92**: 433–443.

Hayes, B.J., Chamberlain, A., Goddard, M.E. (2006) Use of linkage markers in linkage disequilibrium with QTL in breeding programs. *Proc. 8th World Congress of Genetics Applied to Livestock Production*. Belo Horizonte (MG), Brazil, 30-06.

Hayes, B.J. and Goddard, M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics, Selection, Evolution* **33**: 209–229.

Hayes, B.J., Visscher, P.M., Goddard, M.E. (2009a) Increased accuracy of artificial selection by using the realized relationship matrix. *Genetic Research, Cambridge* **91**: 47–60.

Heyen D.W., et al. (1999) A genome scan for QTL influencing milk production and health traits in dairy cattle. *Physiological Genomics* **1**: 165–175.

Holmberg, M. and Andersson-Eklund L. (2004) Quantitative trait loci affecting health traits in Swedish dairy cattle. *Journal of Dairy Science* **87**: 2653–2659.

Israel C. and Weller, J.I. (1998) Estimation of candidate gene effects in dairy cattle populations. *Journal of Dairy Science* **81**: 1653–1662.

Israel, C. and Weller, J.I. (2004) Effect of type I error threshold on marker-assisted selection in dairy cattle. *Livestock Production Science* **85**: 189–199.

Jonker, J.W., Merino, G., Musters, S., van Herwaarden, A.E., Bolscher, E., Wagenaar, E. Mesman, E., Dale, T.C., Schinkel, A.H. (2005) The breast cancer resistance protein BCRP (ABCG2) concentrates drugs and carcinogenic xenotoxins into milk. *Nature Medicine* **11**: 127–129.

Kaupe, B., Winter, A., Fries, R., Erhardt, G. (2004) DGAT1 polymorphism in *Bos indicus* and Bos taurus cattle breeds. *Journal of Dairy Research* **7**: 182–187.

Kennedy, G.C., et al. (2003) Large-scale genotyping of complex DNA. *Nature Biotechnology* **21**: 1233–1237.

Khatkar, M.S., Thomson, P.C., Tammen, I., Raadsma, H.W. (2004) Quantitative trait loci mapping in dairy cattle: review and meta-analysis. *Genetics, Selection, Evolution* **36**: 163–190.

Klungland, H., et al. (2001) Quantitative trait loci affecting clinical mastitis and somatic cell count in dairy cattle. *Mammalian Genome* **12**: 837–842.

Knott, S.A., Elsen, J.M., Haley, C.S. (1996) Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theoretical and Applied Genetics* **93**: 71–80.

Kuhn, C., Thaller, G., Winter, A., Bininda-Emonds, O.R.P., Kaupe, B., Erhardt, G., Bennewitz, J., Schwerin, M., Fries, R. (2004) Evidence for multiple alleles at the DGAT1 locus better explains a quantitative tip trait locus with major effect on milk fat content in cattle. *Genetics* **167**: 1873–1881.

Kuhn, C., et al. (2002) Quantitative trait loci mapping of functional traits in the German Holstein cattle population. *Journal of Dairy Science* **86**: 360–368.

Lander, E.S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.

Lander, E.S. and Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting reporting linkage results. *Nature Genetics* **11**: 241–247.

Legarra, A., Aguilar, I., Misztal, I. (2009) A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science* **92**: 4656–4666.

Liu, Z., Seefried, F., Reinhardt, F., Reents, R. (2009) Dairy cattle genetic evaluation using genomic information. *Interbull Bulletin 39.* Uppsala, Sweden. pp. 23–28.

Loberg, A. and Dürr, J. (2009) Interbull survey on the use of genomic information. *Interbull Bulletin 39*. Uppsala, Sweden. pp. 3–14.

Mackay, T.F. (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics* **35**: 303–339.

Marth, G., et al. (2001) Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genetics* **27**: 371–372.

Martinez, O. and Curnow, R.N. (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**: 480–488.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Meuwissen, T.H.E. and Goddard, M.E. (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**: 421–430.

Misztal, I., Legarra, A., Aguilar, I. (2009) Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science* **92**: 4648–4655.

Mosig, M.O., Lipkin, E., Khutoreskaya, G., Tchourzyna, E., Soller, M., Friedmann, A. (2001) A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics* **157**: 1683–1698.

Nadesalingam, J., Plante, Y., Gibson, J.P. (2001) Detection of QTL for milk production on Chromosomes 1 and 6 of Holstein cattle. *Mammalian Genome* **12**: 27–31.

Olsen, H.G., Nilsen, H., Hayes, B., Berg, P.R., Svendsen, M., Lien, S., Meuwissen, T.H.E. (2007) Genetic support for a quantitative trait nucleotide in the *ABCG2* gene affecting milk composition of dairy cattle. *BMC Genetics* **8**: 32.

Olsen, H.G., et al. (2002) A genome scan for quantitative trait loci affecting milk production in Norwegian dairy cattle. *Journal of Dairy Science* **85**: 3124–3130.

Party, C. and Ducrocq, V. (2009) Bias due to genomic selection. *Interbull Bulletin 39*. pp. 77–82.

Ron, M., Cohen-Zinder, M., Peter, C., Weller, J.I., Erhardt, G. (2006) *ABCG2* polymorphism in *Bos indicus* and *Bos taurus* cattle breeds. *Journal of Dairy Science* **89**: 4921–4923.

Ron, M. and Weller, J.I. (2007) From QTL to QTN identification in livestock – “Winning by points rather than knock-out”: a review. *Animal Genetics* **38**: 429–439.

Ron, M., et al. (2004) A complete genome scan of the Israeli Holstein population for quantitative trait loci by a daughter design. *Journal of Dairy Science* **87**: 476–490.

Sargolzaei, M., Schenkel, F.S., Jansen, G.B., Schaeffer, L.R. (2008) Extent of linkage disequilibrium in Holstein cattle in North America. *Journal of Dairy Science* **91**: 2106–2117.

Schrooten, C., Bovenhuis, H., Coppieters, W., Van Arendonk, J.A.M. (2000) Whole genome scan to detect quantitative trait loci for conformation and functional traits in dairy cattle. *Journal of Dairy Science* **83**: 795–806.

Schulman, N.F., Viitala, S.M., de Koning, D.J., Virta, J., Maki-Tanila, A., Vilkki, J.H. (2004) Quantitative trait loci for health traits in Finnish Ayrshire cattle. *Journal of Dairy Science* **87**: 443–449.

Simes, R.J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**: 751–754.

Smith S.J., Cases, S., Jensen, D.R., Chen, H.C., Sande, E., Tow, B., Sanan, D.A., Raber, J., Eckel, R.H., Farese, R.V., Jr. (2000) Obesity resistance and multiple mechanisms of triglyceride synthesis in mice lacking Dgat. *Nature Genetics* **25**: 87–90.

Spelman, R.J., Coppieters, W., Karim, L., Van Arendonk, J.A.M., Bovenhuis, H. (1996) Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. *Genetics* **144**: 1799–1808.

Spelman, R.J., Garrick, D.J., Van Arendonk, J.A.M. (1999) Utilization of genetic variation by marker assisted selection in commercial dairy cattle populations. *Livestock Production Science* **59**: 51–60.

Spelman, R.J., et al. (1999a) Quantitative trait loci analysis on 17 nonproduction traits in the New Zealand dairy population. *Journal of Dairy Science* **82**: 2514–2516.

Su, G., Guldbrandtsen, B., Gregersen, V.R., and Lund, M.S. (2010) Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. *Journal of Dairy Science* **93**: 1175–1183.

Thomsen, H., et al. (2001) Comparison of estimated breeding values, daughter yield deviations and deregressed proofs within a whole genome scan for QTL. *Journal of Animal Breeding and Genetics* **118**: 357–370.

VanRaden, P.M. (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**: 4414–4423.

VanRaden, P.M. and Wiggans, G.R. (1991) Derivation, calculation and use of national animal model information. *Journal of Dairy Science* **74**: 2737–2746.

VanRaden, P.M., et al. (2009) Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**: 16–24.

Viitala, S.M., et al. (2003) Quantitative trait loci affecting milk production traits in Finnish Ayrshire dairy cattle. *Journal of Dairy Science* **86**: 1828–1836.

Vilkki, H.J., de Koning, D.J., Elo, K.T., Velmala, R., Maki-Tanila, A. (1997) Multiple marker mapping of quantitative trait loci of Finnish dairy cattle by regression. *Journal of Dairy Science* **80**: 198–204.

Visscher, P.M., Thompson, R., Haley, C.S. (1996) Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**: 1013–1020.

Weller, J.I. (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**: 627–640.

Weller, J.I. (1994) *Economic Aspects of Animal Breeding*. London: Chapman & Hall, p. 244.

Weller, J.I. (2009) *Quantitative Trait Loci Analysis in Animals*. 2nd edition. London, UK: CABI Publishing.

Weller, J.I., Golik, M., Reikhav, S., Domochovsky, R., Seroussi, E., Ron, M. (2008) Detection and analysis of QTL affecting production and secondary traits on chromosome 7 in Israeli Holsteins. *Journal of Dairy Science* **91**: 802–813.

Weller, J.I., Golik, M., Seroussi, E., Ezra, E. Ron, M. (2003) Population-wide analysis of a QTL affecting milk-fat production in the Israeli Holstein population. *Journal of Dairy Science* **86**: 2219–2227.

Weller J.I., Kashi Y., Soller M. (1990) Power of “daughter” and “granddaughter” designs for genetic mapping of quantitative traits in dairy cattle using genetic markers. *Journal of Dairy Science* **73**: 2525–2537.

Weller, J.I., Shlezinger, M., Ron, M. 2005. Correcting for bias in estimation of quantitative trait loci effects. *Genetics, Selection, Evolution* **37**: 501–522.

Weller J.I. and Soller M. (2004) An analytical formula to estimate CI of QTL location with a saturated genetic map as a function of experimental design. *Theoretical and Applied Genetics* **109**: 1224–1229.

Weller, J.I., Song, J.Z., Heyen, D.W., Lewin, H.A., Ron, M. (1998) A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150**: 1699–1706.

Weller, J.I., Weller H., Kliger D., Ron M. (2002) Estimation of quantitative trait locus allele frequency via a modified granddaughter design. *Genetics* **162**: 841–849.

Werner, F.A.O, et al. (2004) Detection and characterization of SNPs useful for identity control and parentage testing in major European dairy breeds. *Animal Genetics* **35**: 44–49.

Zhang, Q., et al. (1998) Mapping quantitative trait loci for milk production and health of dairy cattle in a large outbred pedigree. *Genetics* **149**: 1959–1973.

# Chapter 13
# Genome-Wide Association Studies and Linkage Disequilibrium in Cattle

*M. E. Goddard and B. J. Hayes*

## Introduction

In classical genetics, the existence of a gene was recognized by finding a polymorphic phenotype that segregated in a Mendelian manner. With improvements in molecular biology it has become possible to map and identify the variation in DNA sequence that causes many of these simply inherited traits. The key feature of simple Mendelian traits is that one gene causes a large effect on the phenotype so that individuals can be assigned to the correct genotype class based only on their phenotype. In complex or quantitative traits, this is not the case. Here, variation in many genes and environmental factors cause variation in phenotype so the genotype at any one gene cannot be deduced from the phenotype alone. Consequently, success in finding the DNA sequence variation that causes variation in quantitative traits has been limited compared to simple Mendelian traits. However, quantitative traits are of great importance in agriculture, medicine, and evolution, so a great deal of effort is currently being devoted to localizing and identifying polymorphisms underlying such traits. In cattle, there have been more than 400 reported studies that have attempted to localize mutations causing variation in traits as diverse as milk production, growth, susceptibility, to diseases like Johne's disease and external parasites such as ticks, fertility, and resistance to heat stress.

The methods used to find the causal variants underlying quantitative traits are similar to those used for traits controlled by a single gene. Linkage mapping based on genetic markers and study of candidate genes selected based on their role in the physiology of the trait have both been used. For instance, Georges et al. mapped quantitative trait loci (QTL) for milk production traits in Holstein cattle by linkage analysis using microsatellite markers within half-sib families. A gene for a simple Mendelian trait can usually be mapped to within 1 cM by linkage mapping but this is not the case for QTL. The inability to recognize QTL genotype from phenotype means that recombinations between a marker and a QTL cannot be unambiguously identified. In QTL mapping one relies on a decline in the association between marker genotype and phenotype as the distance between the marker and QTL increases.

*Bovine Genomics*, First Edition. Edited by James E. Womack.

Consequently, with linkage mapping the confidence interval for a QTL was often as large as 50 cM. The precision of mapping is increased if the QTL has a large effect and if a large number of individuals are used in the mapping experiment. Not surprisingly, the few cases where the causative mutation has been identified come from large experiments and QTL of large effect (e.g., Grisart et al. 2002; Blott et al. 2003).

A limitation of linkage mapping is that the correlation between a marker and the trait (the linkage phase) varies from one family to another. This reduces the power to find the association. An alternative is to look for an association that is consistent across the whole population. This will occur if the marker studied is the causative polymorphism or if it is in linkage disequilibrium (LD) with the causative polymorphism. LD in most populations only exists over small distances, so we only expect to see an association between a marker and a trait if the marker is close to a QTL. This is an advantage in that it makes QTL mapping more precise than using linkage mapping but it is a disadvantage in that many more markers are necessary to cover the whole genome. Until recently, there was no practical technology for genotyping the thousands of markers needed to cover the whole genome densely enough to detect QTL. Consequently, linkage mapping was usually used to narrow the position of the QTL and then denser markers in that region were used in an association study to map the QTL more precisely and, hopefully, to identify the causative mutation.

The invention of methods to genotype thousands of single nucleotide polymorphisms (SNPs) at low cost combined with the discovery of thousands of SNPs through genome sequencing has made possible an experimental approach to QTL mapping that employs dense, genome-wide markers, and looks for associations between some of these markers and the trait. These studies have been called genome-wide association studies (GWAS). This chapter will review the methodology and findings of GWAS in cattle. However, as GWAS rely on LD we will first describe the nature of LD and the forces that control it.

## The Nature of LD

LD or gametic phase disequilibrium is a correlation between the alleles at two different loci within the same gamete (see Figure 13.1). The strength of LD can be measured by the correlation squared ($r^2$) (Hill and Robertson 1968). LD arises in the following way: when a new mutation creates a new allele at a previously monomorphic site, the new allele occurs on a single chromosome and so it is associated by chance with all other alleles on this chromosome. If this new allele drifts to higher frequency, the other alleles will remain associated with it except where recombination occurs between the mutant allele and other sites on the chromosome. Therefore, the site of the new mutation is most likely to be in LD with other loci if they are close to it on the chromosome as measured by the recombination distance in Morgans ($c$). If the effective population size ($N_e$) is large, drift occurs slowly and there are many generations of recombination so that LD will only occur over short distances. The expected value of $r^2$ as a result of this process is approximately $1/(2 + 4\,N_e c)$ (Tenesa et al. 2007).

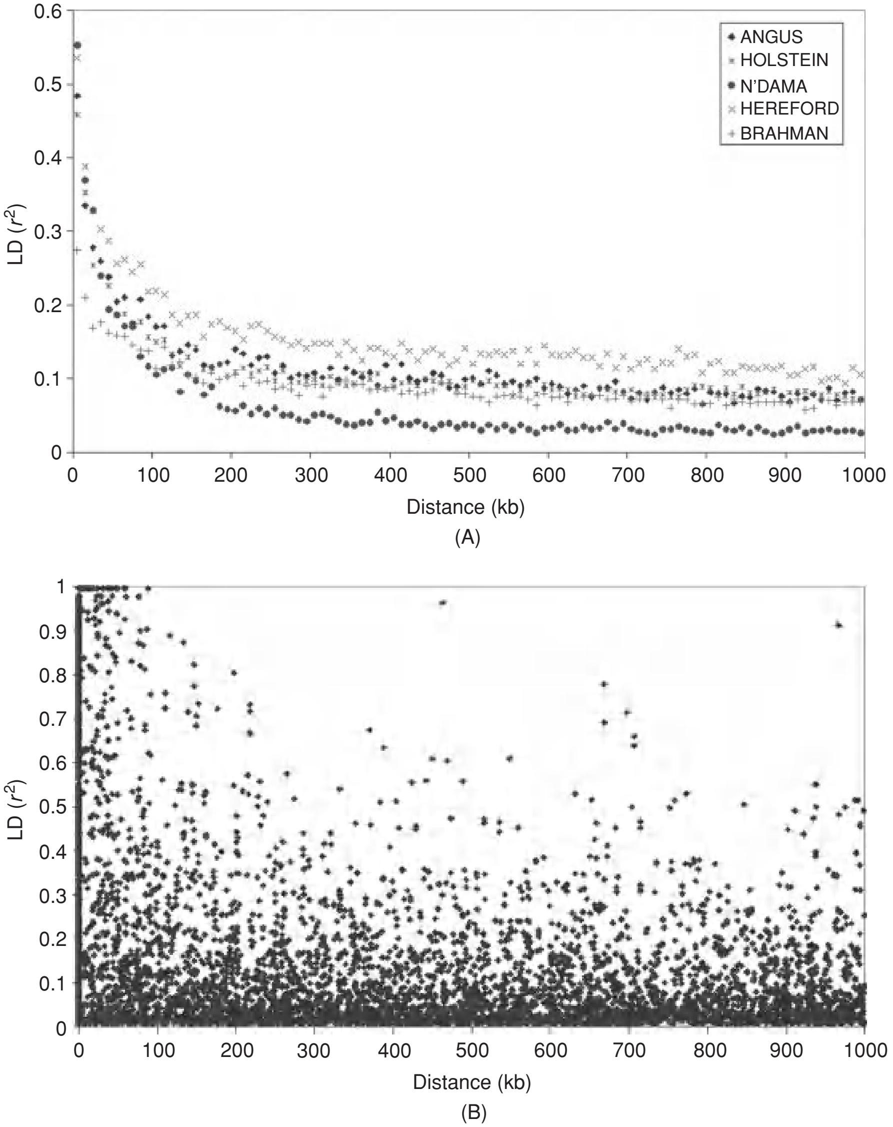


**Figure 13.1** (A) Extent of linkage disequilibrium (LD) in different breeds of cattle (with permission from Bovine HapMap Consortium). (B) Variation in LD at different distances, as demonstrated by pairwise $r^2$ (from 2506 SNPs on chromosome 1 in Australian Holstein). SNP pairs with $r^2 < 0.01$ not shown.

Even if two loci are initially in linkage equilibrium, genetic drift in finite populations will lead to LD. With two alleles at each of two loci, there are four gametic genotypes. Drift will occur in the frequencies of these four types so that they are no longer in linkage equilibrium but recombination will tend to restore the equilibrium. Consequently, LD is again a balance between drift and recombination. In this case, the expected value of $r^2$ is approximately $1/(1 + 4N_ec)$ (Sved 1971). The difference between these two formulae for $E(r^2)$ is that the first is a balance between mutation, drift, and recombination and the second involves only drift and recombination. Thus, the second is more appropriate if the LD has been created mainly by inbreeding due to a reduction in $N_e$, which is exactly what has occurred in cattle.

Both these formulae for $E(r^2)$ assume that $N_e$ is constant over time and this has not been the case for cattle. LD over large distances is mainly controlled by recent $N_e$ because recombination would have broken up associations that arose long ago. Conversely, LD over small distances is still affected by $N_e$ long ago. Consequently, the study of LD over different distances allows us to estimate $N_e$ at various times in the past. This shows that $N_e$ for *Bos taurus* was >50,000 prior to domestication, 1000–2000 after domestication, falling to about 100 in recent times in many breeds. *Bos indicus* cattle have a greater heterozygosity and hence, must have had greater $N_e$ than *B. taurus* at some time in the past (The Bovine HapMap Consortium 2009).

Thus, in *B. taurus* cattle, today, we see a pattern where some LD exists at long distances (<10 cM) but only increases slowly as the distance decreases and eventually reaches high $r^2$ at very small distances (Figure 13.1) (Farnir et al. 2000; Hayes et al. 2003; Gautier et al. 2007; Hayes et al. 2007; McKay et al. 2007; De Roos et al. 2008; Khatkar et al. 2008; Marques et al. 2008; Sargolzaei et al. 2008; Kim and Kirkpatrick 2009; Prasad et al. 2008; The Bovine HapMap Consortium 2009; Bohmanova et al. 2010; Qanbari et al. 2010). This pattern of LD has implications for mapping QTL with GWAS as discussed later. The genetic distances ($c$) used previously are recombination distances but today we often use physical distances in base pairs. On average, 1 cM is equivalent to about 1 Mb but this varies at both large and small scales. At a large scale, some chromosome regions have a higher recombination rate per Mb than others. At a small scale in humans, recombinations tend to occur at specific sites or recombination hot spots (Jeffreys et al. 2001; Myers et al. 2005). This causes loci separated by a hot spot to have lower LD than loci that are the same physical distance apart but not separated by a recombination hot spot. It is assumed that hot spots also exist in cattle but the actual sites appear to evolve very rapidly, so the hot spots in cattle are unlikely to be at sites homologous to those in humans (e.g., Ptak et al. 2005).

The previous description of the average $r^2$ between loci disguises the huge variability in $r^2$ between different pairs of loci even if they are the same recombination distance apart. Thus, an individual pair of loci can be in high LD even though they are separated by 10 cM (Figure 13.1).

Another way to describe the formation of LD is that it is due to the animals in the current population inheriting a segment of chromosome from a common ancestor without any recombination. If $N_e$ is great, the common ancestor is likely to be far in the past and so, only small chromosome segments will have survived intact with no recombination, and therefore, only loci close together will show high LD. If we compare animals in different breeds, they should not share any common ancestors since the breeds diverged. Each breed may have evolved LD due to small $N_e$ since they diverged but the LD will not be the same in both breeds. That is, the phase of LD is likely to be different in different breeds except at distances that are short enough that chromosome segments have not experienced recombination since the breeds diverged. De Roos et al. (2008) found that this occurred among Holstein and Jerseys at distances <10 kb. Consequently, the association between a marker and a

QTL is not likely to be consistent across breeds unless the marker and QTL are very close together.

The $r^2$ statistic is a measure of LD that applies to two loci. It is worth pointing out that there are also statistics measuring LD among multiple loci. For example, "chromosome segment homozygosity" (CSH) is defined as the probability that two chromosome segments in the current population have descended from a common ancestor without a recombination (Hayes et al. 2003). If two chromosome segments are identical by descent in this way, they will have identical alleles at all loci except for new mutations. This will create a haplotype that may be common in the population. The expected value of CSH is $1/(1 + 4N_ec)$. Thus, where $4N_ec$ is small, we expect to see a relatively small number of haplotypes segregating in the population even if we examine many loci within this segment. To date, most GWAS rely on pairwise LD measures, however, multilocus measures like CSH have been used to infer the past effective population size of cattle (e.g., MacLeod et al. 2010).

## Design of GWAS

The basic design is simple: a sample of animals is measured for the trait of interest and genotyped for a genome-wide panel of markers. Two considerations in this design are to avoid false positives and to maximize the power to discover true positives (i.e., minimize false negatives). False positives arise through confounding a variable that affects the trait with the genotypes, and by multiple testing.

The most obvious source of confounding is admixture of populations that differ in allele frequency at the markers and in mean for the trait, but without accounting for this structure in the analysis. For instance, if the population contains both Black Holsteins and Red Holsteins and if the Black Holsteins produce more milk, then there will be an association between the Black allele and milk yield. More subtle forms of admixture or population structure might involve mixing different strains within a breed or even different families. The effect of this population structure can be accounted for by fitting the effect of breed or strain or family in the analysis. For instance, one should fit a polygenic effect to account for the rest of the genome by using a standard animal model. Failure to do this can double the number of false positive SNPs in an analysis (MacLeod et al. 2009). The relationship matrix needed for this animal model can come from the known pedigree, but if this is incomplete, it can be estimated from the markers (a genomic relationship matrix). Another approach that has been used to control for population structure is to create this genomic relationship matrix, take the principal components of the matrix, then include the loadings on the principal components as covariates in the model (Price et al. 2006; Patterson et al. 2006). The assumption here is that the principal components will capture population stratification as these are likely to be the main axes of variation in the genomic relationship matrix. Pausch et al. (2011) applied this approach to a GWAS in Fleckvieh cattle for calving ease and growth-related traits.

The problem of false positives due to the large number of significance tests carried out is addressed under the heading "analysis."

The power of a GWAS depends on the number of animals and the number of markers used. The number of markers should be high enough so that a QTL anywhere in the genome will be in high LD with at least one marker. For instance, in Holsteins, if SNPs are spaced 50 kb apart, the average $r^2$ between adjacent SNPs is 0.2 (Figure 13.1). However, to find SNPs that are in consistent LD with a QTL across several *B. taurus* breeds would require SNPs <10 kb apart (De Roos et al. 2008).

An approximate guide to the number of animals needed can be calculated as follows. The square of the correlation between a marker ($m$) and the phenotype ($p$) ($r^2(m,p)$) is:

$$r^2(m,p) = r^2(p,g) * r^2(g,q) * r^2(q,m),$$

where $g$ is the breeding value, $r^2(p,g)$ is the heritability of the phenotypic measurement, $r^2(g,q)$ is the proportion of genetic variance due to the QTL, and $r^2(q,m)$ is the LD $r^2$ between the QTL and the marker. Therefore if, for instance, $r^2(p,g) = 0.3$, $r^2(g,q) = 0.01$, $r^2(q,m) = 0.7$, then the expected value of $r^2(m,p) = 0.002$. The $F$ statistic for testing the significance of the association between the trait and this marker is approximately $Nr^2(m,p)$ where $N$ is the number of animals. If we require an $F > 10$ (approximately $P < 0.001$), to declare significance, we need $N = 5000$. This emphasizes the need for large experiments if the QTL, which typically have small effects, are to be detected, particularly if heritability is low (Figure 13.2).

It is important to point out that in *B. indicus* cattle, a greater number of genome-wide markers will be required to achieve powerful GWAS than in *B. taurus* breeds, given the lower levels of LD at short distances (Figure 13.1) (The Bovine HapMap Consortium 2009).

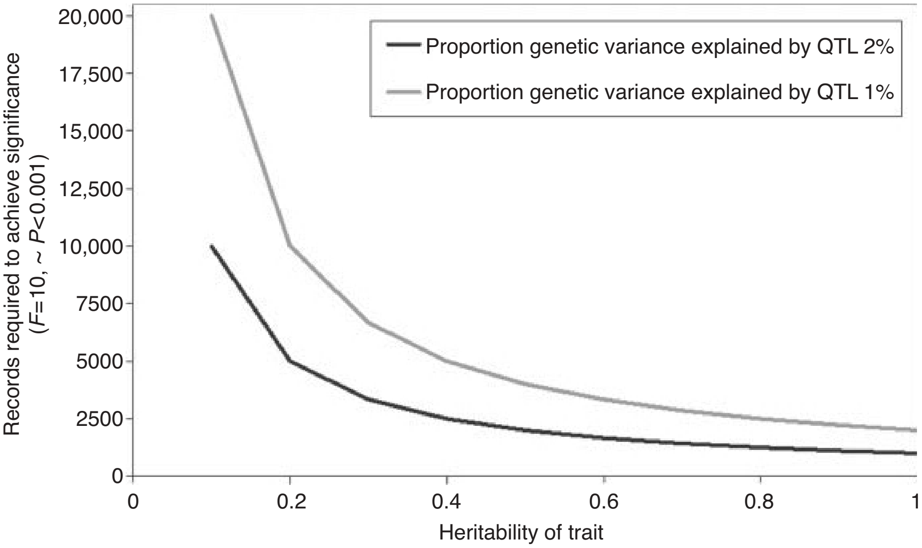


**Figure 13.2** Number of records required in a genome-wide association study (GWAS) to detect QTL explaining 2% or 1% of the genetic variance, at different heritabilties for the trait and significance level set at $F > 10$. Linkage disequilinrium ($r^2$) between the marker and the QTL was 0.5.

## Statistical Analysis of GWAS

The conventional method of analysis is to test the significance of one SNP at a time. This can be done by fitting a linear model with all the fixed and random effects appropriate to the data plus an effect of the SNP. Often, only the additive effect of the SNP is tested by coding the SNP genotype as 0 for one homozygote, 1 for the heterozygote, and 2 for the other heterozygote, but a model including dominance can be fitted by estimating a separate effect for each of the 3 genotypes. As pointed out previously, it is important to include any structure in the population of animals used. This can be done by including a polygenic term in the model with the appropriate numerator relationship matrix if this describes all the structure. It is also possible to estimate the relationship matrix from the markers instead of from the pedigree, then to fit this matrix directly or its principal components as previously described.

The significance of the effect of the SNP can be tested by a t- or F-test. However, when thousands of such tests are made, we expect that 5% will reach a threshold for $p < 0.05$ just by chance. If we desire a test that will find only one falsely significant SNP in every 20 experiments (i.e., an experiment-wide $p$-value of 0.05), then we must use a very stringent $p$-value for each individual SNP. The Bonferroni correction is to use a $p$-value per SNP equal to the experiment wide $p$-value divided by the number of SNPs. Therefore, if you want $p < 0.05$ experimental wide and you test 50,000 SNPs, each individual SNP should only be declared significant if $p < 0.000001$. The Bonferroni correction is often too conservative, as it does not take into account the fact that some of the SNPs are capturing the same QTL information through LD. An alternative is to use permutation testing but the $p$-value per SNP will still need to be very stringent if an experiment-wide $p < 0.05$ is to be obtained.

The disadvantage of such stringent hypothesis testing is that few significant SNPs are found. One may question the appropriateness of a null hypothesis that there are no SNPs associated with the trait when we know the trait has some genetic variance and the SNPs cover the whole genome. An alternative approach is to compute a false discovery rate (FDR) (Benjamini and Hochberg 1995). This is the proportion of the SNPs that are declared significant that are false discoveries. FDR can be estimated as:

$$FDR = p(1 - s)/(s(1 - p)),$$

where $s$ is the proportion of SNPs found to be significant at an SNP-wise $p$-value of $p$. If the FDR is 0.05, it means that only 5% of the significant SNPs are false discoveries.

Commonly, the $p$-values calculated are not conservative enough because the data do not fully match the assumptions of the model. For instance, the trait may not be normally distributed. The most important problem is that the sample of animals used for the experiment may be different to the population where we wish to use the association. For instance, the experiment might have used one herd of cattle but we wish to use the SNP in other herds of the same breed. LD might exist between an SNP and QTL in one herd but not in other herds. For this reason, and to overcome both the multiple testing carried out in the original experiment and false positives due to any population stratification not accounted for, it is necessary to confirm the associations declared significant in an independent sample of animals.

An alternative to the conventional analysis of a GWAS described previously is to fit all SNPs simultaneously. Because there are so many SNPs, this is only possible by treating the SNP effects as random effects. The model is the same as that used for genomic selection in which the aim is to estimate the breeding value of individual animals based on all SNP genotypes (Meuwissen et al. 2001). It is necessary to specify a prior distribution of the SNP effects, so they can be treated as random samples from this distribution. Meuwissen et al. (2001) suggested three prior distributions. One considered all SNP effects as drawn from the same normal distribution (the "best linear unbiased prediction (BLUP)" model). This is unsuitable for QTL mapping because it results in all SNPs having small estimated effects. Another prior distribution assumed that only a proportion of SNPs have nonzero effects. A Bayesian analysis estimates a posterior probability that each SNP has an association with the trait and this probability could be used to find SNPs that are in LD with QTL. The advantage of this method is that only SNPs most closely associated with QTL should be included in the final model.

The BLUP model, with all SNPs fitted simultaneously and their effects assumed to come from the same normal distribution, can be used for a different purpose. That is, it can be used to estimate the total genetic variance explained by the SNPs. This will be less than the full genetic variance if the QTL are not in complete LD with a linear combination of the SNPs. This model can be conveniently implemented by an equivalent model in which the variance of the breeding values (i.e., the relationship matrix) is calculated from the SNP genotypes (Yang et al. 2010). For instance, Yang et al. (2010) found that only half the genetic variance for human height was accounted for by the SNPs.

## Results of Cattle GWAS

Since the recent release of high-density SNP "chips," for example the Parallele 10K and Illumina Bovine SNP50, there have been over 30 GWAS reported in cattle, Table 13.1. GWAS in cattle have successfully identified mutations causing single gene abnormalities such as congenital muscular dystrophy (Charlier et al. 2008). As shown in Table 13.1, GWAS have also been carried out for complex traits such as milk production, feed efficiency, fatty acid composition of meat, tolerance to a host of diseases, tolerance to heat stress, and tolerance to external parasites. However, as the table demonstrated, with a few notable exceptions (such as, Hayes et al. 2009; Feugang et al. 2009; Bierman et al. 2010; Pryce et al. 2010a; and Minozzi et al. 2010), the findings have not been confirmed in an independent sample of animals.

One way of "validating" the results would be to compare the location of significant SNP for different studies using the same phenotype. For example, in Table 13.1, there are four studies conducting GWAS for *Mycobacterium avium* subsp. *paratuberculosis* (MAP) infection status or tolerance to MAP infection. The studies were conducted in 966 Italian (Minozzi et al. 2010), 245 US (Settles et al. 2009; Zanella et al. 2010), or 232 Canadian Holstein cattle (Pant et al. 2010). Comparison of the results is complicated by the fact that each study used a different phenotype. The phenotype used in Settles et al. (2009) was animal infected/not infected, where an animal was considered tissue infected if any tissue sample from an animal contained at least one colony forming unit

Table 13.1 Genome-wide associations studies (GWAS) in cattle. Studies are grouped loosely by trait groups. All studies used the Illumina BovSNP50 array unless otherwise stated.

| Trait(s) | Reference | Breed(s) | Number of individuals | Number of SNPs significant | Validation step in an independent population | Comments |
|---|---|---|---|---|---|---|
| Congenital muscular dystonia 1, congenital muscular dystonia 2, ichthyosis fetalis | Charlier et al. (2008) | Belgian Blue, Italian Chianina | CMD 1(12 cases, 14 controls); CMD2 (7 cases, 24 controls); ICF (3 cases, 9 controls) | | Causal mutation identified | Three genes harboring causal mutations identified |
| Dairy production, fertility, reproduction, conformation traits and other key dairy traits | Cole et al. (2009) | Holstein | 5285 | Significance testing not used, see column "Comments" | No, but effects used to predict accurate estimated breeding values for young dairy bulls (Wiggans et al. 2011) | All SNPs fitted simultaneously |
| Dairy production, somatic cell score, herd life, interval of calving to first service, age at first service | Daetwyler et al. (2008) | Holstein | 484p | 144 | No | 9919 SNPs used |
| Dairy production | Jiang et al. 2010 | Holstein | 2093 | 105 | No | |
| Dairy production, somatic cell score, survival, milking speed temperament, likeability milk × feeding level, milk × temperature, persistency | Bolormaa et al. 2010 | Holstein, Jersey | 1533 | 4514 (across 19 traits) | Yes | Multitrait methodology |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Dairy production and fertility traits | Pryce et al. 2010a | Holstein, Jersey | 1533 | 1573 across 6 traits | Yes, 544 SNPs in Holsteins and 159 in Jerseys | Tested single SNPs as well as variable length haplotypes |
| Persistency of lactation | Pryce et al. 2010b | Holstein, Jersey | 1533 | 619 at $P < 0.005$ | Yes |  |
| Resistance to heat stress, ability to milk at low levels of feeding | Hayes et al. 2010 | Holstein, Jersey | 1533 | 362 | Yes |  |
| Dairy production, somatic cell score | Kolbehdari et al. 2009 | Holstein |  | 28 | No | 1536 SNPs only used |
| Dairy production × level of feeding interaction | Lillehammer et al. 2009 | Holstein | 384 | 157 | No | 9918 SNPs used |
| Dairy production | Mai et al. 2010 | Jersey | 1039 | 98 | No |  |
| Conformation and functional traits | Kolbehdari et al. 2008 | Holstein | 462 | 196 | No | 1036 SNPs in introns used |
| Fertility, age at onset of puberty | Fortes et al. 2010 | Tropical composites | 866 | 2799 | No | Used a novel multivariate approach and gene locations to reconstruct gene pathways involved in puberty |
| Bull fertility | Feugang et al. 2009 | Holstein | 20 bulls extreme for fertility | 97 | Yes, four most significant SNPs tested in 210 bulls extreme for fertility, two significant, implicating integrin beta 5 protein | Used 8207 SNPs |

(*continued*)

**Table 13.1** (*Continued*)

| Trait(s) | Reference | Breed(s) | Number of individuals | Number of SNPs significant | Validation step in an independent population | Comments |
|---|---|---|---|---|---|---|
| Female fertility | Sahana et al. 2010 | Holstein | 2531 | 74 | No | |
| Fertility, fertilization rate, and blastocyst rate | Huang et al. 2010 | Holstein | | 27 | No, but pooled results confirmed by individual genotyping | First application of DNA pooling in a GWAS in cattle |
| Stillbirth and dystocia | Olsen et al. (2010) | Norwegian Red | 2552 | 13 | Yes | 17,343 SNPs used |
| Calving ease | Pausch et al. (2010) | Fleckvieh | 1800 | 2 ($P = 5.72 \times 10^{-15}$, $P = 2.27 \times 10^{-8}$) | No | |
| Twinning | Kim et al. (2009) | Holstein | 200 | 174 | Yes, Bierman et al. (2010), 55 SNPs validated in an independent data set of 921 Holstein bulls. Final set of 18 SNPs explain 34% of variation in twinning rate. | 9919 SNPs used |
| Fatty acid composition in beef | Uemoto et al. (2010) | Japanese black cattle | 160 extreme animals | 32 | No | |
| Tolerance to Johne's disease (measured by tolerance to *Mycobacterium avium* subsp. *paratuberculosis* (MAP) infection) | Zanella et al. (2010) | Holstein | 90 plus, 16 cases and 25 controls | 5 SNPs at $P < 1 \times 10^{-5}$ | No | |
| MAP infection status | Settles et al. (2009) | Holstein | 245 | Seven regions at $P < 5 \times 10^{-5}$, two regions $P < 5 \times 10^{-7}$ | No | |

| | | | | | | |
|---|---|---|---|---|---|---|
| MAP status (by ELISA test) | Pant et al. (2010) | Holstein | 232 animals with known MAP status | 12 genomic regions significant | No | |
| Serologically positive (or not) for MAP by ELISA | Minozzi et al. (2010) | Holstein | 483 MAP ELISA positive and 483 ELISA negative | One region significant $P < 1 \times 10^{-6}$ and three regions significant $P < 1 \times 10^{-5}$ | Yes, in a smaller cohort from the same population Five SNPs were validated | |
| Bovine spongiform encephalopathy (BSE) susceptibility | Murdoch et al. (2010) | Holstein | 481 half sibs,<br>149 BSE cases,<br>184 controls | 27 SNPs | No | Used 3072 SNPs |
| Growth | Snelling et al. (2010) | Angus, Charolais, Gelbvieh, Hereford, Limousin, Red Angus, and Simmental crossbreds | 2603 | 231 SNPs | No | |
| Feed efficiency | Barendse et al. (2007) | Angus, Brahman, Belmont Red, Hereford, Murray Grey, Santa Gertrudis, and Shorthorn cattle | 189 extreme animals | 161 SNPs at $P < 0.01$ | Yes, though this used animals from the middle of the distribution from where the extreme animals were derived | |
| Feed efficiency | Sherman et al. (2010) | Angus, Charolais, or Alberta Hybrid | 464 | 23 at $P < 0.01$ | No | 2633 SNP used |
| Tick burden | Turner et al. (2010) | Ancestry from Jersey, Holstein, Aussie Reds, Sahiwal, Illawara, Shorthorn | 189 extreme animals | 27 at $P < 0.05$ | No. | 7397 SNPs used. Recently, Porto Neto et al. (2010) reported significant SNPs in close proximity to the Integrin alpha 11 gene in a follow-up study. |

of MAP per gram of tissue (CFU/g), and they employed the same definition for faecal samples. Both Minozzi et al. (2010) and Pant et al. (2010) used case control designs where cases were defined as animals serologically positive for MAP by ELISA. Zanella et al. (2010) considered a quite different phenotype, tolerance to MAP infection, rather than MAP infection status. These differences in phenotype, and the small numbers of animals in each study resulting in limited power, likely contributed to the surprising result that none of the significant SNPs reported in any of the studies was confirmed in any of the other studies. Even more surprisingly, none of the significant SNPs in one study were within 1 Mb of significant SNPs reported in another paper. This finding highlights the need to design powerful GWAS. It should be noted that Minozzi et al. (2010) did test their significant SNPs in another group of 277 cattle from the same population, which validated five SNP associations ($P < 0.05$). It should also be noted that Zanella et al. (2010) considered a different phenotype (tolerance to MAP infection) to the other studies (presence/absence of MAP infection).

The ultimate validation study would demonstrate that the association is significant in two or more breeds of cattle. False positives due to population stratification are very unlikely to occur in two independent samples from two different breeds. A further attraction if validating across breeds is that if the association persists, the SNP must be very close to the QTL, given the limited across-breed extent of LD (e.g., De Roos et al. 2008). However, the drawback of attempting to validate an association, which has been discovered in one breed, in another breed is that the QTL may not be segregating across breeds. In studies where validation has been attempted across cattle breeds, the results have been mixed. Pryce et al. (2010a) carried out a GWAS in Holsteins and Jerseys for milk production traits. For instance, in Holsteins, they found 461 SNPs of the 39,000 tested significantly associated ($p < 0.001$) with protein concentration in milk, implying an FDR of 8%. When these SNPs were tested in a separate sample of Holsteins, 210 were significant ($p < 0.01$) and 209 of these had an effect in the same direction as in the discovery experiment. This implies a low FRD (2%) among the confirmed SNPs. However, in Jerseys, only 63 SNPs of the 461 were significant ($p < 0.01$) and only 27 of these had an effect in the same direction. Although less SNPs were confirmed in Jerseys, 63 is still more than the $0.01*461 = 4.6$ SNPs expected by chance. The fact that only 27/63 had effects in the same direction as in Holsteins is expected because at this density of SNPs (1 SNP per 60 kb) the LD phase in Jerseys and Holsteins is unlikely to be the same (De Roos et al. 2008). Denser markers are required for across-breed GWAS in cattle. The release of the Bovine HD chip, with over 700,000 SNPs, is an important step in this direction.

In the future, as the cost of whole genome resequencing continue to decline, GWAS using whole genome sequence data are foreseeable, perhaps using imputation of sequence information in a very large number of individuals (e.g., Meuwissen and Goddard 2010). Such GWAS would have the advantage that the actual mutation causing phenotypic variation would be contained in the data set.

The large number of SNPs associated with protein percentage and other traits implies that most QTL explain a small percentage (<1%) of the genetic variance for quantitative traits. Futher, the effects of the SNPs that are significant is likely to be overestimated due to the Beavis effect. However, GWAS of complex traits do sometimes detect genes of moderate effect. For example, Bierman et al. (2010), describe a validated set of 18 SNPs that explained 34% of variation in twinning rate.

Another example is from Hayes et al. (2010), reporting the results of a GWAS of the percentage of white on the coat of black-and-white Holsteins. They found three genes (*KIT*, *MITF*, and *Pax5*) that together explained 24% of the variance for this trait. However, there were also many other parts of the genome containing SNPs significantly associated with percentage white coat color.

Although some traits have some genes with moderate effects, almost all quantitative traits studied appear to have many QTL, each causing a small proportion of the total genetic variance. This conclusion is also true of GWAS in humans.

The failure to confirm all significant SNPs in a separate confirmation study is due to three factors. Firstly, the confirmation study is usually not powerful enough to detect such small effects. Secondly, the LD phase in the confirmation experiment may be different to that in the animals used in the discovery experiment. This is especially likely if the two experiments use different breeds. Thirdly, the FDR in many discovery experiments is high, so one should not expect to confirm all significant SNPs. One way to increase the power of a GWAS is to combine information from multiple traits. If a QTL affects more than one trait, then using this information should increase the power to detect the QTL Bolormaa et al. (2010) found a small increase in the rate of validation of significant SNPs when a multivariate approach was used to detect associations in the discovery population relative to associations detected from a multivariate approach.

Fortes et al. (2010) also used a multivariate approach to identify associations with puberty in female cattle; however, they went a step further by including information from gene ontology in their study. Neibergs et al. (2010) also used gene ontology information in a GWAS for MAP infection status. They argued that by considering associations accumulated across regions containing genes within groups of gene ontology classifications or KEGG pathways rather than individual associations, they increased power to pick up associations of modest effect.

The precision with which GWAS map a QTL is limited by two effects. Firstly, although LD declines rapidly with distance, it is highly variable and so, an SNP some distance from the QTL may be in high LD with it (e.g., Figure 13.1B). Secondly, there is a sampling error associated with the estimate of the effect of an SNP due to the finite number of animals measured for the trait and genotyped for the SNP. Consequently, QTL are difficult to map precisely, especially the majority of QTL that have small effects. MacLeod et al. (2009) found that for a QTL explaining 5% of the variance mapped with 365 animals, the most significant SNP will be <1 Mb away only 13% of the time. Therefore, for SNPs explaining 1% of the variance, an experiment with 1000 animals might still give a 95% confidence interval of several cM. Since there are 100s if not 1000s of QTL for most traits, the confidence interval for one QTL may overlap with that of the nearest QTL on the same chromosome. This phenomenon will make it difficult to map QTL precisely in cattle, unless a multibreed strategy is used (as the across-breed extent of LD is more limited than within-breed LD).

## Conclusion

Genomic selection is already widely applied in dairy cattle (Dalton 2009) and, in the near future, in other livestock as well. This is occurring without the causal genes or mutations being identified and to do so will be difficult because most of their effects are very small. However, in the longer term, identifying the causal genes and mutations will be important for at least three areas: (1) understanding the biology underlying the response from genomic selection, (2) understanding and managing the consequences of selecting for mutations with undesirable pleiotropic effects (e.g., loci that increase protein yield but decrease fertility), and (3) identifying gene pathways that can be targeted to improve a trait, such as resistance to Johne's disease. Identifying the mutations underlying significant results from GWAS will be challenging, particularly given the likely small effect sizes involved, but will be aided by high-throughput sequencing of RNA transcripts from appropriately designed experiments (e.g., Cánovas et al. 2010), and combining GWAS results with gene ontology information (e.g., Neibergs et al. 2010; Fortes et al. 2010).

One major challenge facing researchers working on cattle is the enormous size of experiments that are required (e.g., Figure 13.1). To date, most GWAS in cattle have been underpowered with a few notable exceptions (e.g., Cole et al. 2009; Olsen et al. 2010). In some instances, the number of genotyped individuals can be greatly reduced while retaining close to the same power. For example, in dairy cattle, if the trait of interest is routinely recorded on cows, the number of individuals that need to be genotyped can be reduced by genotyping progeny-tested bulls, as the heritability of the bulls' "phenotype" (daughter averages) can be close to one for some traits. Another attractive approach for increasing power without dramatically increasing the cost of the experiment is selective DNA pooling, which was used by Huang et al. (2010) in a GWAS for cow fertility.

However, these approaches are not relevant in all cases, and keeping in mind the need for an additional experiment to validate the results from GWAS, large numbers of genotyped and phenotyped individuals will still be required. The obvious solution is to combine experiments across research groups and across countries, perhaps using the GIANT consortium as a model. This consortium was established by human geneticists to dissect the genetic variants underlying variation in human height, by combining data from many research groups, following the recognition that the effect of these variants would likely be very small and extremely large GWAS would be required to detect them. For example, in the latest GWAS from this consortium, 249,746 individuals were used (Speliotes et al. 2010).

Finally, it is worth pointing out that, for some traits, cattle are a very useful model species for GWAS, if the aim is to uncover potential genetic pathways affecting such traits. For example, stillbirth and dystocia are disease that can affect many mammals including humans. In most species, the very low heritability of these traits would mean that 100,000s of phenotypes would be required to dissect this trait. In cattle however, Olsen et al. (2010) demonstrated that the power of almost 1 million records for these traits could be captured by genotyping the sires of the recorded cows, and using daughter averages as phenotypes. They identified and validated a small number of SNPs in the region of a cluster of candidate genes expected to affect bone and cartilage formation (i.e., SPP1, IBSP, and MEPE). Olsen et al. (2010) suggested these candidate genes could be investigated in other species suffering these diseases. In general, the routine recording of other health and reproduction traits in very large numbers of cows, and the ability to capture this information by genotyping a much smaller number of sires of these cows, suggest cattle should not be overlooked as model species for complex diseases. Interestingly, Pryce et al. (2011) recently demonstrated that of the genes implicated in harboring polymorphisms affecting height in humans, a proportion of these genes that is much greater than expected by chance were also associated with stature and weight in cattle.

## References

Barendse, W., Reverter, A., Bunch, R.J., Harrison, B.E., Barris, W., Thomas, M.B. (2007) A validated whole-genome association study of efficient food conversion in cattle. *Genetics* **176**(3): 1893–1905.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**(1): 289–300.

Bierman, C.D., Kim, E., Shi, X.W., Weigel, K., Berger, J.P., Kirkpatrick, B.W. (2010) Validation of whole genome linkage-linkage disequilibrium and association results, and identification of markers to predict genetic merit for twinning. *Animal Genetics* **41**(4): 406–416.

Blott, S., et al. (2003) Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics* **163**(1): 253–266.

Bohmanova, J., Sargolzaei, M., Schenkel, F.S. (2010) Characteristics of linkage disequilibrium in North American Holsteins. *BMC Genomics* **11**: 421.

Bolormaa, S., Pryce, J.E., Hayes, B.J., Goddard, M.E. (2010) Multivariate analysis of a genome-wide association study in dairy cattle. *Journal of Dairy Science* **93**: 3818–3833.

Cánovas, A., Rincon, G., Islas-Trejo, A., Wickramasinghe, S., Medrano, J.F. (2010) SNP discovery in the bovine milk transcriptome using RNA-Seq technology. *Mammalian Genome* **21**: 592–598.

Charlier, C., et al. (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nature Genetics* **40**: 449–454.

Cole, J.B., VanRaden, P.M., O'Connell, J.R., Van Tassell, C.P., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Wiggans, G.R. (2009) Distribution and location of genetic effects for dairy traits. *Journal of Dairy Science* **92**: 2931–2946.

Daetwyler, H.D., Schenkel, F.S., Sargolzaei, M., Robinson, J.A. (2008) A genome scan to detect quantitative trait loci for economically important traits in Holstein cattle using two methods and a dense single nucleotide polymorphism map. *Journal of Dairy Science* **91**(8): 3225–3236.

Dalton, R. (2009) No bull: genes for better milk. *Nature* **457**(7228): 369

De Roos, A.P.W., Hayes, B.J., Spelman, R., Goddard, M.E. (2008) Linkage disequilibrium and persistence of phase in Holstein Friesian, Jersey and Angus cattle. *Genetics* **179**: 1503–1512.

Farnir, F., et al. (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* **10**: 220–227.

Feugang, J.M., Kaya, A., Page, G.P., Chen, L., Mehta, T., Hirani, K., Nazareth, L., Topper, E., Gibbs, R., Memili, E. (2009) Two-stage genome-wide association study identifies integrin beta 5 as having potential role in bull fertility. *BMC Genomics* **10**: 176.

Fortes, M.R., Reverter, A., Zhang, Y., Collis, E., Nagaraj, S.H., Jonsson, N.N., Prayaga, K.C., Barris, W., Hawken, R.J. (2010) Association weight matrix for the genetic dissection of puberty in beef cattle. *Proceedings of the National Academy of Sciences of the United States of America* **107**(31): 13642–13647.

Gautier, M., et al. (2007) Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics* **177**(2): 1059–1070.

Grisart, B., et al. (2002) Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research* **12**: 222–231.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., Savin, K., van Tassell, C.O., Sonstegard, T.S., Goddard, M.E. (2009) A validated genome wide association study to breed cattle adapted to an environment altered by climate change. *PLOS One* **4**: e66761–e66768.

Hayes, B.J., Chamberlain, A.J., McPartlan, H., Macleod, I., Sethuraman, L., Goddard, M.E. (2007) Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genetics Research* **89**(4): 215–220.

Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J., Goddard, M.E. (2010) Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics* **6**(9): e1001139.

Hayes, B.J., Visscher, P.M., McPartlan, H., Goddard, M.E. (2003) A novel multi-locus measure of linkage disequilibrium and it use to estimate past effective population size. *Genome Research* **13**: 635–643.

Hill, W.G. and Robertson, A. (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**: 226–231.

Huang, W., Kirkpatrick, B.W., Rosa, G.J., Khatib, H. (2010) A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. *Animal Genetics* **41**: 570–578.

Jeffreys, A.J., Kauppi, L., Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics* **29**: 217–222.

Jiang, L., Liu, J., Sun, D., Ma, P., Ding, X., Yu, Y., Zhang, Q. (2010) Genome wide association studies for milk production traits in Chinese Holstein population. *PLoS One* **5**(10): e13661.

Khatkar, M.S., Nicholas, F.W., Collins, A.R., Zenger, K.R., Cavanagh, J.A., Barris, W., Schnabel, R.D., Taylor, J.F., Raadsma, H.W. (2008) Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics* **9**: 187.

Kim, E.S., Berger, P.J., Kirkpatrick, B.W. (2009) Genome-wide scan for bovine twinning rate QTL using linkage disequilibrium. *Animal Genetics* **40**(3): 300–307.

Kim, E.S. and Kirkpatrick, B.W. (2009) Linkage disequilibrium in the North American Holstein population. *Animal Genetics* **40**(3): 279–288.

Kolbehdari, D., Wang, Z., Grant, J.R., Murdoch, B., Prasad, A., Xiu, Z., Marques, E., Stothard, P., Moore, S.S. (2008) A whole-genome scan to map quantitative trait loci for conformation and functional traits in Canadian Holstein bulls. *Journal of Dairy Science* **91**(7): 2844–2856.

Kolbehdari, D., Wang, Z., Grant, J.R., Murdoch, B., Prasad, A., Xiu, Z., Marques, E., Stothard, P., Moore, S.S. (2009) A whole genome scan to map QTL for milk production traits and somatic cell score in Canadian Holstein bulls. *Journal of Animal Breeding and Genetics* **126**(3): 216–227.

Lillehammar, M., Hayes, B.J., Meuwissen, T.H.E., Goddard, M.E. (2009) Gene by environment interactions for production traits in Australian dairy cattle. *Journal of Dairy Science* **92**: 4008–4017.

MacLeod, I.M., Hayes, B.J., Savin, S., Chamberlain, A.J., McPartlan, H., Goddard, M.E. (2010) Power of dense bovine single nucleotide polymorphisms (SNPs) for genome scans to detect and position quantitative trait loci (QTL). *Journal of Animal Breeding and Genetics* **127**: 133–142.

MacLeod, I.M., Meuwissen, T.H., Hayes, B.J., Goddard, M.E. (2009) A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. *Genetics Research* **91**: 413–426.

Mai, M.D., Sahana, G., Christiansen, F.B., Guldbrandtsen, B. (2010) A genome-wide association study for milk production traits in Danish Jersey cattle using a 50K single nucleotide polymorphism chip. *Journal of Animal Science* **88**(11): 3522–3528.

Marques, E., Schnabel, R.D., Stothard, P., Kolbehdari, D., Wang, Z., Taylor, J.F., Moore, S.S. (2008) High density linkage disequilibrium maps of chromosome 14 in Holstein and Angus cattle. *BMC Genetics* **9**: 45.

McKay, S.D., et al. (2007) Whole genome linkage disequilibrium maps in cattle. *BMC Genetics* **8**: 74.

Meuwissen, T. and Goddard, M. (2010) The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. *Genetics* **185**(4): 1441–1449.

Meuwissen, T.H., Hayes, B.J., Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4): 1819–1829.

Minozzi, G., Buggiotti, L., Stella, A., Strozzi, F., Luini, M., Williams, J.L. (2010) Genetic loci involved in antibody response to Mycobacterium avium ssp. paratuberculosis in cattle. *PLoS One* **5**(6): e11117.

Murdoch, B.M., Clawson, M.L., Laegreid, W.W., Stothard, P., Settles, M., McKay, S., Prasad, A., Wang, Z., Moore, S.S., Williams, J.L. (2010) A 2 cM genome-wide scan of European Holstein cattle affected by classical BSE. *BMC Genetics* **11**: 20.

Myers, S., Bottolo, L., Freeman, C., McVean, G., Donnelly, P. (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**(5746): 321–324.

Neibergs, H.L., Settles, M.L., Whitlock, R.H., Taylor, J.F. (2010) GSEA-SNP identifies genes associated with Johne's disease in cattle. *Mammalian Genome* **21**(7–8): 419–425.

Olsen, H.G., Hayes, B.J., Kent, M.P., Nome, T., Svendsen, M., Lien, S. (2010) A genome wide association study for QTL affecting direct and maternal effects of stillbirth and dystocia in cattle. *Animal Genetics* **41**(3): 273–280.

Pant, S.D., Schenkel, F.S., Verschoor, C.P., You, Q., Kelton, D.F., Moore, S.S., Karrow, N.A. (2010) A principal component regression based genome wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in holstein cattle. *Genomics* **95**(3): 176–182.

Patterson, N., Price, A.L., Reich, D. (2006) Population structure and eigen analysis. *PLoS Genetics* **2**(12): e190.

Pausch, H., Flisikowski, K., Jung, S., Emmerling, R., Edel, C., Götz, K.U., Fries, R. (2011) Genome-wide association study identifies two major loci affecting calving ease and growth related traits in cattle. *Genetics* **187**: 289–297.

Porto Neto, L.R., Bunch, R.J., Harrison, B.E., Prayaga, K.C., Barendse, W. (2010) Haplotypes that include the integrin alpha 11 gene are associated with tick burden in cattle. *BMC Genetics* **11**: 55.

Prasad, A., Schnabel, R.D., McKay, S.D., Murdoch, B., Stothard, P., Kolbehdari, D., Wang, Z., Taylor, J.F., Moore, S.S. (2008) Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle. *Animal Genetics* **39**(6): 597–605.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**(8): 904–909.

Pryce, J.E., Bolormaa, S., Chamberlain, A.J., Bowman, P.J., Savin, K., Goddard, M.E., Hayes, B.J. (2010a) A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. *Journal of Dairy Science* **93**(7): 3331–3345.

Pryce, J.E., Haile-Mariam, M., Verbyla, K., Bowman, P.J., Goddard, M.E., Hayes, B.J. (2010b) Genetic markers for lactation persistency in primiparous Australian dairy cows. *Journal of Dairy Science* **93**(5): 2202–2214.

Pryce, J.E., Hayes, B.J., Bolormaa, S., Goddard, M.E. (2011) Polymorphic regions affecting human height also control stature in cattle. *Genetics* **187**: 981–984.

Ptak, S.E., Hinds, D.A., Koehler, K., Nickel, B., Patil, N., Ballinger, D.G., Przeworski, M., Frazer, K.A., Pääbo, S. (2005) Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genetics* **37**(4): 429–434. Erratum: *Nature Genetics* (2005) **37**(4): 445.

Qanbari, S., Pimentel, E.C., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R., Simianer, H. (2010) The pattern of linkage disequilibrium in German Holstein cattle. *Animal Genetics* **41**(4): 346–356.

Sahana, G., Guldbrandtsen, B., Bendixen, C., Lund, M.S. (2010) Genome-wide association mapping for female fertility traits in Danish and Swedish Holstein cattle. *Animal Genetics* **41**: 579–588.

Sargolzaei, M., Schenkel, F.S., Jansen, G.B., Schaeffer, L.R. (2008) Extent of linkage disequilibrium in Holstein cattle in North America. *Journal of Dairy Science* **91**(5): 2106–2117.

Settles, M., Zanella, R., McKay, S.D., Schnabel, R.D., Taylor, J.F., Whitlock, R., Schukken, Y., Van Kessel, J.S., Smith, J.M., Neibergs, H. (2009) A whole genome association analysis identifies loci associated with Mycobacterium avium subsp. Paratuberculosis infection status in US holstein cattle. *Animal Genetics* **40**(5): 655–662.

Sherman, E.L., Nkrumah, J.D., Moore, S.S. (2010) Whole genome single nucleotide polymorphism associations with feed intake and feed efficiency in beef cattle. *Journal of Animal Science* **88**(1): 16–22.

Snelling, W.M., Allan, M.F., Keele, J.W., Kuehn, L.A., McDaneld, T., Smith, T.P., Sonstegard, T.S., Thallman, R.M., Bennett, G.L. (2010) Genome-wide association study of growth in crossbred beef cattle. *Journal of Animal Science* **88**(3): 837–848.

Speliotes, E.K., et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42**(11): 937–948.

Sved, J.A. (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**: 125–141.

Tenesa, A., Navarro, P., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E., Visscher, P.M. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**: 520–526.

The Bovine HapMap Consortium (2009) The genetic history of cattle. *Science* **324**: 528–532.

Turner, L.B., Harrison, B.E., Bunch, R.J., Neto, L.R.P., Li, Y., Barendse, W. (2010) A genome-wide association study of tick burden and milk composition in cattle. *Animal Production Science* **50**: 235–245.

Uemoto, Y., Abe, T., Tameoka, N., Hasebe, H., Inoue, K., Nakajima, H., Shoji, N., Kobayashi, M., Kobayashi, E. (2010) Whole-genome association study for fatty acid composition of oleic acid in Japanese Black cattle. *Animal Genetics* [Epub ahead of print] PubMed PMID: 20590532.

Wiggans, G.R., Vanraden, P.M., Cooper, T.A. (2011) The genomic evaluation system in the United States: past, present, future. *J Dairy Sci* **94**(6): 3202–3211.

Yang, J., et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**(7): 565–569.

Zanella, R., Settles, M.L., McKay, S.D., Schnabel, R., Taylor, J., Whitlock, R.H., Schukken, Y., Van Kessel, J.S., Smith, J.M., Neibergs, H.L. (2010) Identification of loci associated with tolerance to Johne's disease in Holstein cattle. *Animal Genetics* **42**: 28–38.

# Chapter 14
# Genomic Selection in Beef Cattle

*Jeremy F. Taylor, Stephanie D. McKay, Megan M. Rolf, Holly R. Ramey, Jared E. Decker, and Robert D. Schnabel*

## Introduction

Genomic selection (GS) has been shown to be an extremely effective technology for increasing the rate of genetic improvement in dairy cattle; however, adoption of the technology within beef cattle has been more limited. The US and international dairy industries are largely dominated by the Holstein breed, and the calibration of the marker density on the first available high-density single nucleotide polymorphism (SNP) genotyping assay to the extent of linkage disequilibrium (LD) within a breed has been sufficient to permit the BovineSNP50 assay to be used to generate molecular estimates of breeding value (MEBV) with accuracies of about 70%. In itself, this was not sufficient to ensure rapid uptake of the technology; however, the low cost of the assay relative to the expense of progeny testing young bulls has rapidly driven the adoption of GS in dairy cattle where at least 70% of the females are bred by artificial insemination (AI).

The experience within the US and international beef industries has been somewhat different. Because very few beef cattle are bred by AI (about 7% in the United States), it is much more difficult to assemble the large training panels of DNA samples on animals with accurate estimated breeding values (or expected progeny differences; EPDs), and the cost of testing an animal with a high-density assay to produce a suite of MEBVs exceeds the value generated in selected natural service bulls due to the relatively small numbers of progeny that they produce relative to their dairy AI counterparts. Consequently, the adoption of GS within beef cattle has, to date, been breed-specific using reduced marker panels that produce MEBVs with relatively low accuracies but that can be marketed at low cost. The problem with this model is that very few of the beef breeds have sufficient numbers of animals with high-accuracy EPDs to allow the development of within-breed MEBV prediction equations, and equations developed using 50,000 SNP assays for one breed do not generally work in other breeds. The solution to this dilemma will likely be the application of the second-generation high-density SNP assays with more than 640,000 SNPs, which will provide a sufficient SNP density to capture the much more limited across-breed LD that will

*Bovine Genomics*, First Edition. Edited by James E. Womack.

be represented in training populations formed by pooling breeds. This strategy is expected to allow sufficient numbers within the training populations to achieve high MEBV accuracies when validated across different breeds.

However, the major constraints to the adoption of GS in beef cattle will remain and include the following:

1. The need to deliver accurate MEBVs at low cost.
2. Understanding the need for periodic retraining of the MEBV prediction equations as the number of generations advances between the training and commercialization populations.
3. Determining the optimum breeding program design when the accuracies on selected females can equal those on males.

The need for periodic retraining is due to the serial dilution of the proportion of the genome of an animal that is identical by descent to members of the training population as generations advance, which leads to a significant loss of accuracy of prediction. For traits that are routinely recorded in beef cattle production, periodic retraining may be facilitated by genotyping members of advanced generations. However, this will require additional investment in high-density assays if commercialization has been facilitated with low-density assays. On the other hand, for traits that are not routinely recorded, such as feed intake or disease resistance, entire training populations may need to periodically be recreated to facilitate retraining.

## Industry Structure

Holstein is the dominant dairy breed worldwide and genetic improvement is delivered to the industry by progeny testing young bulls and AI. This industry structure has rapidly enabled the delivery of GS because of the following reasons:

1. DNA on many thousands of Holstein bulls is available from cryopreserved semen for the construction of training populations.
2. The cost of high-density SNP assays is small relative to the cost of progeny testing.
3. GS is delivered to the industry via the same assays that were used to train the MEBV equations.
4. Young bulls selected for progeny testing go on to achieve accurate estimates of genetic merit based upon progeny data and can be used to continually retrain the MEBV equations.

The primary issues that affect the implementation of GS within the dairy industry are to avoid the extinction of the numerically minor breeds that do not have sufficient animals with accurate estimates of genetic merit to develop breed-specific prediction equations, how to balance the increased short-term rate of improvement of genetic progress against longer term response due to the loss of beneficial alleles that occurs under strong selection, and how to optimize the design of the breeding program in view of the fact that the accuracy of MEBVs produced for females is identical to that of males and there are many more available females than males within the industry.

On the other hand, while Angus is the numerically dominant breed within the US beef industry, there are at least 80 beef breeds in the United States and at least a dozen breeds are numerically very important to beef production. The use of AI has been limited within the industry and is used primarily within the registered sector. Archives of semen on historic bulls with accurate EPDs exist but are distributed among AI and semen companies as well as hundreds of registered breeders and are not designed to ensure that DNA samples on historically important animals are preserved. Because the breed associations have yet to establish DNA repositories for their breeds, the process of assembling training populations has been tedious and expensive and it is extremely difficult to assemble DNA samples from semen on more than 2000–3000 animals for any one breed. However, the focus on the use of AI within the registered sector means that the genetic improvement, which is delivered to the much larger commercial sector, occurs via the use of natural service yearling bulls. These bulls have very low accuracies for EPDs on any trait except growth traits and very little selection is currently applied to natural service bulls, because the majority of bulls that are produced by registered breeders are sold to the commercial sector. Consequently, MEBVs produced on these bulls will provide value only to the specific commercial producers who can capture the value created in their cow herds and calves from improvements in, for example, carcass quality and feed efficiency. However, what is obvious is that the limited selection that is practiced among yearling natural service bulls does not justify the cost of MEBVs produced using expensive high-density genotyping assays. One strategy that has been designed to resolve this issue has been to develop low-density (384 or 10,000 SNPs) assays that can be delivered at an appropriate price point. However, the design of these assays is problematic. Rolf et al. (2010) has shown that there is a significant loss in the accuracy of MEBVs as the number of markers decreases below 10,000, suggesting that there will be appreciable losses in the accuracy of MEBVs derived from low-density assays unless higher densities of SNPs can be accurately imputed from these lower density panels. Additionally, because so many more traits are of economic importance in beef production than are important in dairy production, the strategic selection of, for example, the 100 most informative SNPs associated with each trait, will limit the number of traits than can be tested on these platforms. What appears to be necessary to solve this dilemma is the advent of new genotyping technologies that can genotype tens to hundreds of thousands of SNPs for a cost of $10–$20. While this may seem unlikely now, few of us would have foreseen the ability to genotype 50,000 SNPs for under $100, 5 years ago.

A final vital issue that is currently evolving is the determination of the optimal business model for the translation of GS to the beef industry. The current model involves partnerships between universities and technology companies to generate and commercialize intellectual property (IP) and partnerships between the technology companies and industry organizations such as breed associations to commercialize the IP. The IP may exist as patented causal mutations or MEBV prediction equations defined by a set of simultaneously estimated allele-substitution effects for a set of SNPs that are held as a trade secret since it is not yet clear that such multilocus effects can be patented, or even if there is any inherent value in patenting these models and effects. The basis for this business model is that the breed associations themselves do not have the financial resources to pay for the genotyping of a large number of samples, but are the curators of the databases that must frequently be used to provide phenotypes or EPDs for the training populations. As partners in commercialization efforts they receive MEBVs from the technology companies, which are returned to the customers, the registered breeders, and which enter the breed association databases. In the case of the American Angus Association, these MEBVs are incorporated into statistical analyses, which are run every week to update EPDs using all sources of available data, phenotypes, pedigrees, and MEBVs. What is clearly missing from this process are the genotypes themselves, which remain the property of the technology company. Thus, if the breed association could capture a set of genotypes on a sufficiently large number of animals, the entire commercialization process could be captured by the breed association—at least for the traits that are routinely recorded by their members. Thus, the value of the technology companies may be limited to their investment ability to generate MEBVs for traits such as disease resistance, feed intake, and meat tenderness, which are not routinely measured by the industry. If the breed associations could generate sufficient revenues from the commercialization of DNA diagnostics for routinely measured phenotypes, they could invest in the development of populations of extensively phenotyped animals, and the need for the technology companies would dramatically diminish. A great advantage to this model is a reduction in risk due to volatility within the technology sector that leads these companies into and out of market opportunities as a function of expected returns and also the fact that all necessary data for the evolution of a continuous model retraining system would be owned by a single organization, ensuring the long-term sustainability of GS.

## Genomic Selection Theory

While selection has been practiced in the beef industry for decades and producers have utilized EPDs for decades, GS is a relatively new development (Meuwissen et al. 2001). Traditional EPD analyses include only the probabilities that individuals are identical-by-descent via the incorporation of the numerator relationship matrix (NRM), which accounts for the selection of parents (Henderson 1975), but assumes that the expected value of Mendelian sampling effects is zero. However, this is not the case for traits for which selection operates on the Mendelian sampling effects. Within registered beef cattle populations there has been strong historic selection for growth to weaning and yearling ages while avoiding correlated responses in birth weight. This has been accomplished by parental selection on weaning and yearling weight and then registering only those progeny that have lower birth weights than expected based upon the selected parents. This form of two-stage selection (parents to produce progeny, and progeny within families that enter the population as the parents of the next generation) violates the assumption for the generation of the NRM based upon pedigree information. On the other hand, the genomic relationship matrix (GRM) captures the entire selection history of the population. Not only are pedigree relationships captured by the genomic identity-by-descent, but deviations from expectation due to within-family selection are also intrinsically captured within the observed genomic data.

GS operates under the principle that genetic variation in quantitative traits can be statistically modeled using a large number of well-spaced SNP markers spread throughout the genome, which will be in LD with quantitative trait loci (QTL). SNPs are present about every 300 bases within the *Bos taurus* genome and about every 100 bases in the *Bos indicus* genome (The Bovine HapMap Consortium 2009). Any SNP that is sufficiently close to a QTL and has similar allele frequencies may result in a strong correlation between allele frequencies between the loci, and the resultant LD can be exploited in selection for the QTL without knowing the identity of the causal mutation. This means that models can be constructed that will effectively simultaneously select for all of the QTLs in an animal's genome (Hayes et al. 2009) and young animals can be selected (even at the embryonic stage (Seidel 2009)) before progeny have been born (de Roos et al. 2007), which dramatically shortens generation interval and increases genetic progress (Seidel 2009). It has been estimated that GS could increase response to selection by a factor of two within the dairy industry, saving about 92% of the cost of proving young bulls (Schaeffer 2006). The general principles of GS dictate that a large "training" population of animals is used to estimate the marker effects, which are then used to predict the breeding values of individuals within a "validation" population based only upon genotype. The accuracy of the resulting predictions is theoretically dictated by the amount of LD found between the markers assayed and the true QTL, as well as the number of animals and phenotypic records available in the training population (Toosi et al. 2010). In practice, the accuracy is also influenced by the extent of pedigree relationship between the individuals within the training and validation populations.

## *Simulation Studies*

Initial work to evaluate the performance of GS was performed using simulated datasets. The first of which was performed using relatively dense genotypes but with no clear specification of the distribution of true QTL effects, by Meuwissen et al. (2001). A 1000-cM genome was simulated with multiallelic markers occurring every 1 cM and a single QTL located between each marker. The flanking markers for each 1-cM region were combined to form haplotypes spanning the region. In this study, best linear unbiased prediction (BLUP) of haplotype effects yielded MEBV accuracies of 0.732, while Bayesian methods (BayesA, 0.798 and BayesB, 0.848) increased the accuracy. Calus et al. (2008) simulated a 3-M genome that contained between 119 and 2343 SNP markers, which were analyzed either as single SNPs or as haplotype blocks containing from two to ten markers, with or without covariance information between the haplotypes by combining linkage and LD information. They reported a considerable increase in the accuracy of MEBVs when linkage information was included for highly heritable traits. Additionally, assumptions made about QTL location and distribution affected model performance. When SNPs were at a high density allowing some SNPs to be in strong LD with each QTL, the SNP analysis yielded the highest MEBV accuracies, whereas if no SNPs were in very strong LD with the QTL, the haplotype analysis that included identity-by-descent information yielded the highest accuracies.

Many studies have simulated smaller genome sizes and fewer segregating QTL than are theorized to exist in mammalian species. As a consequence, the estimated QTL effects may be substantially larger than for real QTL, which has allowed the simulated effects to be more accurately predicted than may be possible in cattle data (Goddard and Hayes 2007). Additionally, some early studies included the validation population in the original training data, resulting in inflated MEBV accuracies. Subsequent studies corrected this error and produced less biased evaluations of the effectiveness of GS. Even allowing for the overestimation of the effectiveness in simulated data, GS must produce superior estimates of breeding value to predictions derived from pedigree and phenotype data because it better accounts for the Mendelian sampling that occurs in gametogenesis. However, it is imperative that methods for GS be tested in real populations using large numbers of animals with high-quality phenotypes that have been genotyped with dense marker panels to evaluate prediction accuracy gains.

### *Testing Genomic Selection in Dairy Cattle*

One of the first studies using real genotypic and phenotypic data was performed by de Roos et al. (2007) who used estimated breeding values on 1300 Holstein-Friesian bulls and genotypes at 32 loci across chromosome 14 (including genotypes for a causal *DGAT1* mutation) to compare BLUP and Bayesian approaches under a multiple QTL model with haplotype and polygenic effects for fat percentage. The accuracy of predicted MEBVs was 0.75 compared to the traditional pedigree-based BLUP accuracy of 0.51. In a study of 3330 Danish Holstein bulls using 38,134 SNPs, Su et al. (2010) found that Bayesian estimated MEBVs had accuracies ranging from 0.49 to 0.73 in cross-validation. Accuracies for 18 traits were, on average, 0.26 higher than for the parent averages. Luan et al. (2009) tested GS in a population of 500 Norwegian Red bulls with 18,991 genotyped SNPs. Genomic BLUP (G-BLUP), BayesB, and a mixture model were tested with G-BLUP performing best and with accuracies ranging between 0.12 and 0.62.

The gains from GS in beef cattle are expected to be far less than in dairy cattle (Seidel 2009), but the value of the approach is expected to be greatest for traits that are difficult or expensive to record (feed efficiency, health), that require termination for data collection (carcass traits), that are measured late in life (longevity), or that are measured only in one sex (milk production, heifer pregnancy rate), because MEBV accuracies are equivalent in both males and females (VanRaden et al. 2009). Additionally, GS may have future benefits for animal welfare, for the improvement of traits in which phenotyping requires exposure to disease pathogens or invasive techniques (Solberg et al. 2009).

## Methods for Genomic Selection

### *Genomic Relationship Matrices*

The NRM contains the expected relationship coefficients between animals conditional on the pedigree, and the lack of selection on Mendelian sampling terms. Genotypes can be used to reconstruct the coefficients of relationship among animals in the absence of a pedigree and the resulting matrix is known as the GRM (VanRaden 2008), realized relationship matrix, or genetic relationship matrix (Hayes et al. 2009). When the NRM is replaced by the GRM in the mixed model equations, the resulting MEBVs are known as the G-BLUP estimates (Legarra et al. 2009). Solving for the regression of SNP effects on the produced MEBVs yields the G-BLUP solutions for SNP effects and indicates that equivalent mixed linear models can be written for breeding values and SNP effects. Both assume the infinitesimal model for SNP effects, and similarly regress large- and small-effect SNPs toward zero—the average substitution effect (Villanueva et al. 2005). However, the estimation of the GRM makes no assumptions about selection on the Mendelian sampling effects, and the realized coefficients are conditional on the realized Mendelian sampling and progeny selection effects which should, therefore, result in more accurate MEBVs. One attractive feature of G-BLUP is that all animals, regardless of phenotypic status, are included into the GRM and the predictions are obtained in a single step for all animals. Harris et al. (2008) generated MEBVs for milk production traits on approximately 4500 dairy cattle by direct inversion and obtained accuracies of 0.5–0.67 compared to 0.34 for the parental average breeding values.

VanRaden (2008) discussed three different methods for the formation of a GRM. MEBVs predicted using a GRM, formed using true, base population or sample allele frequencies, were compared under an animal model. The best MEBV accuracies were achieved using the true SNP allele frequencies; however, accuracies were almost identical when base population frequencies were used and were similar when sample frequencies were used. Rolf et al. (2010) used the regression-based method (VanRaden 2008) to estimate a GRM for 698 Angus steers with incomplete pedigrees and 1707 Angus AI bulls to predict MEBVs for feed efficiency traits. Because of the lower than expected heritabilities, MEBV accuracies were limited (0.23–0.44).

Legarra et al. (2009) suggested the definition of the relationship matrix in terms of both marker and pedigree data that blends information from genotyped and ungenotyped but pedigree recorded animals. The NRM is partitioned based on generation (ancestors vs. progeny) and the genotype status of the animals so that additional information provided by genotypes can be plugged in to create a relationship matrix based upon all available information. Aguilar et al. (2010) tested this approach with data on 6,232,548 Holstein cows and BovineSNP50 genotypes on 6508 bulls and found that genetic evaluation using an NRM augmented with genotype information afforded similar accuracies and bias to multistep approaches where pedigree and genomic evaluations were incorporated using selection indexes.

## Estimation of Individual Marker or Haplotype effects

When all animals are genotyped, SNP or haplotype block effects may be estimated by a number of methods (Goddard and Hayes 2007; Goddard and Hayes 2009). Since the available cattle SNP genotyping assays were designed using the most common variants present within the genome (Matukumalli et al. 2009), low-frequency QTLs may not be in significant LD with any assay marker, but may be in LD with a haplotype block of markers (Goddard and Hayes 2007; Goddard and Hayes 2009). Calus et al. (2008) found that the advantages in MEBV accuracy achieved by using haplotype blocks over single markers decreased as the LD between adjacent markers increased and was roughly equivalent when $r^2 \geq 0.21$. However, a drawback of haplotype-based methods is that the exact genomic position of each marker must be known to correctly form haplotype blocks, and at present the cattle assemblies contain numerous errors.

### *Linear Estimation of Marker Effects*

Linear methods of estimating marker (or haplotype) effects result in BLUP, with marker effects assumed to be drawn from a normal distribution with a constant variance. This method accounts for pedigree structure, whether known (NRMs based on pedigree information) or inferred (GRMs based on genomic information). The use of G-BLUP is appealing because the only prior information that is required is the additive genetic variance for the trait of interest (Hayes et al. 2009) and this may be simultaneously estimated in the analysis. The main drawback of the method is that all markers are estimated to have an effect and unexplained genetic variance due to incomplete LD between SNPs, and markers is distributed across all of the small-effect markers creating background noise. However, when the underlying genetic architecture of the analyzed trait is infinitesimal (increasingly supported by numerous analyses), the approach works as well as the nonlinear approaches.

### *Nonlinear Estimation of Marker Effects*

Nonlinear methods for the estimation of marker effects have been gaining traction for traits with several large-effect QTL and for which the majority of tested SNPs have no effect on phenotype. When all markers are included into the predictions, the uninformative markers add noise into the evaluation. Nonlinear approaches allow for the inclusion of a prior distribution that can potentially better characterize the true distribution of QTL effects. While VanRaden et al. (2009) found similar results between Bayesian estimation and linear modeling for dairy traits, Hayes et al. (2009) and Harris et al. (2008) found the increase in MEBV accuracy of Bayesian methods to be approximately 2%–7%.

BayesA (Meuwissen et al. 2001) simultaneously includes all markers in the model (the proportion of markers that do not have an effect on the trait ($\pi$) is assumed to be 0), but the markers are assumed to have a nonconstant variance and the marker variances are individually estimated using Markov chain Monte Carlo methods. This approach allows large SNP effects to be regressed toward zero, less than SNPs with small effects on the trait. BayesB also does not assume a constant marker variance but differs from BayesA in that not all markers fit in the model ($\pi > 0$). In BayesC, $\pi > 0$ but the markers incorporated into the model are all assumed to have been drawn from a population with a constant variance. Because the SNP effect shrinkage is based upon the frequency that SNPs are incorporated into the model, their variances resemble those estimated in BayesB. Finally, BayesC$\pi$ is a form of BayesC in which the parameter $\pi$ is estimated within the analysis (Table 14.1).

## *Selection Index*

Because genotype data are not routinely returned to breed associations that produce EPDs through the analysis of phenotype and pedigree data, it is possible to produce EPDs and molecular estimates of progeny differences (MEPDs) for the same trait, neither of which utilize all available information. When genotype data are not available to the breed associations, but MEPDs are available, the solution has been to blend

**Table 14.1** A comparison of linear and nonlinear methods for the prediction of SNP effects.

| | | Parameter information | | | |
|---|---|---|---|---|---|
| Analysis | | $\pi$ | Marker variance | Training and validation | Special cases |
| Linear | G-BLUP | 0 | Constant | Yes | Similar to BayesC0 |
| Nonlinear | BayesA | 0 | Variable | Yes | Equivalent to BayesB0 |
| | BayesB | Variable and user assigned | Variable | Yes | |
| | BayesC | Variable and user assigned | Constant, except when shrunk | Yes | |
| | BayesC$\pi$ | Estimated from the data | Constant, except when shrunk | Yes | |

the data using selection index methodology. VanRaden et al. (2009) demonstrated that combined predictions for 27 traits in Holstein cattle had 23% greater realized heritabilities than when compared to the parent average predictions. This gain in heritability corresponded to an increase in information equivalent to approximately 11 daughter records.

## *Benefits and Drawbacks*

GS can enhance the rate of genetic progress by decreasing the generation interval, increasing the accuracy of selection, and increasing the intensity of selection of parental candidates. GS can shorten the generation interval by allowing the selection of animals on genomic information before phenotypic information can be recorded, in some cases as early as the embryonic stage, such as in embryo transfer programs (Seidel 2009). Additionally, GS may increase the accuracy of MEBVs over progeny test results if sufficient training data are available (Schaeffer 2006). Finally, selection intensity in females can be dramatically improved, as bulls and females will have equally accurate MEBVs.

It has also been proposed that the collection of phenotypes will no longer be necessary after the implementation of GS, as genotypes become the sole predictor of genetic merit (Habier et al. 2007). However, the more recent opinion is that it will be necessary to continue collecting high-quality phenotypes to enable retraining of the prediction models as allele frequencies, genetic variation, LD patterns, and relationship patterns between the training and validation populations change. Additionally, spurious SNP associations will decay more rapidly than the decay in association effects due to the breakdown of LD between SNPs and QTLs with recombination (Habier et al. 2007; Solberg et al. 2009). High-quality phenotypes, such as those collected in dairy progeny testing schemes or for the large numbers of progeny sired by a particular beef bull, will continue to be essential for the optimal application of GS (Seidel 2009), where models will likely need to be retrained every so often to account for changes in population structure and LD (Schaeffer 2006). The amount of loss in MEBV accuracy due to these effects can be estimated, but requires the analysis of several generations of phenotyped and genotyped animals (Habier et al. 2007). Even when very large populations are available for training, there is a limit to the amount of variation that can be explained based on the history of the population and the design of the utilized SNP assay. The common variants included on these assays cannot detect rare QTL, as the LD between an SNP and a QTL is limited by their difference in allele frequency (Kizilkaya et al. 2010). The use of assays incorporating several hundred thousand SNPs or the imputation of whole genome polymorphism data in training data sets will alleviate this issue.

## *Within- and Across-Breed Applications*

GS works best and provides the largest gains in MEBV accuracy when animals in both the training and validation sets belong to the same breed. Phase relationships have been shown to extend for not more than 10 kb across breeds of cattle (The Bovine HapMap Consortium 2009) but extend for much greater distances within a breed. Thus, there will always be more SNPs available with which to establish the strongest association with a QTL on a within-breed analysis than in an across-breed analysis. On the other hand, within a training population, the accuracy of MEBVs increases more by increasing the number of animals than by increasing the number of markers genotyped (VanRaden et al. 2009). In beef cattle, there are limitations to obtaining large numbers of animals within a breed with high-accuracy EPDs or phenotypes to build the prediction models. One solution to this limitation is to pool animals from different breeds to obtain large numbers of animals with which to build the MEBV prediction models (de Roos et al. 2009). However, it is not yet clear whether the same QTL segregate in all breeds, and the animals will need to be genotyped with an assay of sufficient density to place at least one SNP within the common core haplotype that harbors each QTL to ensure that the directionality of the detected effect is the same for all breeds. For example, we scored 40,645 SNPs in 651 Angus, 695 Charolais, 1095 Hereford, and 516 Simmental steers and performed G-BLUP to estimate SNP allele substitution effects for Warner–Bratzler shear force (Figure 14.1). Despite the relatively small number of animals genotyped within each breed, we detected relatively few SNPs that generated concordant substitution effects across all of the genotyped breeds. Regions on BTA7 and BTA29 harboring *CAST* and *CAPN1* were among these regions of concordance but had been supplementarily genotyped with 64 additional SNPs to those present on the BovineSNP50 assay. The correlations between SNP effects estimated for each of the breeds was very low ($<0.1$ for all comparisons), and MEBV prediction equations developed in each breed performed equally poorly when validated in the other breeds. We have also performed across-breed training analyses in two-thirds of the randomly sampled data and validation in the remaining one-third of the data and have produced MEBVs with accuracies of about 60% across the breeds. Haplotypes

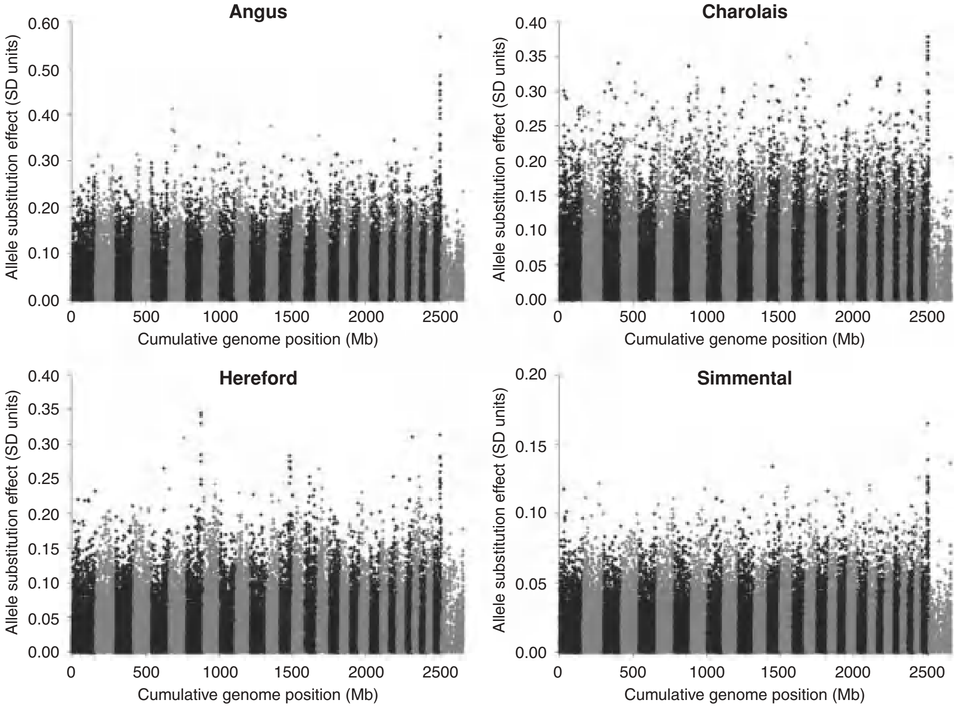


**Figure 14.1** Manhattan plots for Warner–Bratzler shear force produced by G-BLUP using 40,645 SNPs assayed in 651 Angus, 695 Charolais, 1095 Hereford, and 516 Simmental steers.

with strong LD ($r^2 \geq 0.7$) are significantly shorter in these admixed and crossbred populations compared to purebred populations (Toosi et al. 2010), and the extent of LD limits the predictive ability of the BovineSNP50 panel (Kizilkaya et al. 2010). Our high-density genotype data within the *CAST* and *CAPN1* regions detected a QTL in all four breeds (Figure 14.1) and suggested that the new high-density Illumina BovineHD 777K and Affymetrix Axiom BOS 1 640K assays should provide a marker density that would allow the development of effective across-breed MEBV prediction models.

A second issue with the development of across-breed MEBV prediction models is the development of a method to partition animals into the training and validation populations. There is evidence that training in one breed (e.g., Holstein) and validating in another breed (e.g., Jersey) is not effective, and that accuracies in the validation set tend to be very low (Harris et al. 2008). Kizilkaya et al. (2010) has also shown that training in multibreed and validating in purebreds is less effective than training in purebred and validating in multibreed populations due to the greater extent of LD present in purebred populations. If SNP effects cannot be accurately estimated in one breed and applied in another, other strategies for partitioning training and validation populations must be tested to determine those that are the most effective.

Toosi et al. (2010) evaluated different methods of partitioning populations for training and validating MEBV prediction models using simulated data for purebred, admixed, and crossbred populations. They found, without exception, that training and validating within the same breed produced the highest MEBV accuracies. However, training in the admixed populations produced similar accuracies to training and validating in the same purebred population. A 46% decrease in accuracy was observed when validating in a breed differing from that used in the training set, but only a 35% decrease occurred when training was in an F1 population, and validation occurred in a purebred that was not included in the F1. Decreases in MEBV accuracy found from training in three- and four-way crosses were approximately 10%, and training and validating in a crossbred population increased accuracy by 11% relative to training in purebreds and validating in crossbreds. This study also examined the effect of increasing marker density within the known QTL regions on the accuracy of prediction and found that high marker densities were more beneficial when the training population had only a small contribution from the breed used for validation. Finally, larger sample sizes may be needed in multibreed training to achieve MEBV accuracies comparable to those obtained for a purebred population if across-breed and breed-specific effects need to be estimated.

VanRaden et al. (2009) partitioned bulls by birth year with the oldest animals being assigned to the training set and their progeny being assigned to the validation set. This strategy helps to enhance the percentage of variation explained by the MEBVs by creating a large pedigree relationship between the animals in the training and validation sets. Minimizing this relationship leads to a much smaller percentage of variation explained by the MEBV models within the validation population suggesting that the SNPs not only detect LD signals between markers and QTL, but also predict linkage relationships in the population. While the common SNPs on the assays cannot individually detect rare QTL by LD, they may be able to do so by linkage.

Toosi et al. (2010) suggested that it may be beneficial to partition training and validation sets based upon the time since divergence or genetic distance between breeds. Under this approach, animals are selected from different populations to increase the amount of genetic variation present within the training population so that all animals in the validation population have reasonably strong genetic relationships to at least some of the animals in the training set. Using simulated data, Toosi et al. (2010) found that reducing the time since divergence in the training and validation populations greatly increased MEBV accuracies, but that training in admixed rather than crossbred populations resulted in even greater accuracies regardless of time since divergence. De Roos et al. (2009) also simulated two cattle populations with different divergence times. They trained on 1000 individuals from population A and from different subsets of population B. They found that when individuals from population B were omitted from the training set, the accuracy of MEBVs validated in population B was up to 0.77 lower than when validated in population A, and that this effect was most severe when the divergence between populations was greatest. However, training with individuals from both populations resulted in accuracies that were similar to those obtained in population A, as long as the marker density was sufficient for LD relationships to exist across both populations.

## *Reduced SNP Panels*

Small SNP panels are currently commercially available for the prediction of MEBVs in US Angus cattle, and though available, it is not likely that high-density genotyping assays will be widely adopted until their cost decreases. Rolf et al. (2010) used G-BLUP with GRMs formed from subsets of the 50K genotype data generated in commercial Angus steers. Using bootstrap analysis, they found that as few as 1500 SNPs may be sufficient to build a GRM in purebred beef cattle and that very little extra information was obtained when panels larger than 10,000 SNPs were used. Weigel et al. (2009) used a training set of 3305 Holstein bulls and a validation set of 1398 bulls to evaluate the predictive ability of subsets of 300–2000 SNPs, with the largest effects estimated using a Bayesian methodology. $R^2$ values started at 0.064–0.184 (depending on whether the SNPs were evenly spaced or had the largest effects) and increased to 0.291–0.322 with a panel of 2000 SNPs. They concluded that reduced SNP panels could provide a cost-effective way to increase the accuracy of selection of parental candidates and that increasing the number of SNPs from 300 to 1000 or more would provide a significant gain in predictive power.

Other low-density SNP panel applications have been proposed in which parents of large families would be genotyped with high-density panels and their progeny would be genotyped with low-density panels, and haplotype-based imputation (such as fastPHASE; Scheet and Stephens 2006 or findhap.f90; VanRaden 2011) would be used to estimate the missing data with most probable genotypes at each of the missing loci. Hayes et al. (2009) used fastPHASE to impute missing genotypes at every 50th position for 10% of animals along an entire chromosome and found that the program had an accuracy of 98.7%. Mixed-model-based approaches for the estimation of genotypes or haplotypes have also been proposed for animals that have no genotypes at all, based on the use of the NRM between genotyped and ungenotyped animals (Mulder et al. 2010). This approach is not terribly effective if the extent of relationship between genotyped and ungenotyped animals is low.

## SNP Detection and Assay Development

Three major SNP discovery projects have guided the development of the currently marketed bovine assays. The first major SNP discovery effort in cattle can be attributed to the bovine genome sequencing initiative (The Bovine Genome Sequencing and Analysis Consortium 2009), which produced 2.1 M putative SNPs from the Hereford assembly but that had a low (<50%) conversion rate and were not evenly distributed through the genome due to the large inbreeding coefficient (30%) of the sequenced cow (Matukumalli et al. 2009), and 118,249 putative SNPs with an 80% conversion rate from 348,958 shotgun sequence reads from six breeds of cattle that were aligned to the Hereford genome sequence. A relatively uniform number of SNPs was detected among the taurine breeds when SNP frequency was measured per 1000 bases. However, the number of SNPs detected in the indicine breed, Brahman, was approximately double that detected in the taurine breeds. Of the discovered SNPs, 37,470 were assayed by the Bovine HapMap Consortium in 497 cattle representing 14 taurine, three indicine, and two hybrid breeds. These SNPs were primarily detected within taurine cattle and the ascertainment bias due to the breed of discovery resulted in substantive differences in the diversity detected among breeds. The highest average minor allele frequency (MAF) was 0.261 in the sequenced Hereford, while the lowest average MAF was 0.195 in Brahman cattle—a breed shown to harbor at least twice the nucleotide diversity of the taurine breeds. The depth of the coverage and SNP spacing was generally insufficient for the identification of signatures of selection; however, selective sweeps were identified on chromosomes 2, 6, and 14 in regions near genes known to be associated with economically important traits. Mutations in *MSTN* on BTA2 are responsible for double muscling (Grobet et al. 1997), a missense mutation in *ABCG2* on BTA6 has a major effect on milk yield and composition (Cohen-Zinder et al. 2005), and the selective sweep on the proximal end of BTA14 contains *TG* that has been associated with marbling in beef cattle (Barendse et al. 2004).

The Affymetrix GeneChip Bovine Mapping 10K SNP kit was the first commercially available bovine high-density SNP assay. Of the 10,000 SNPs on the assay, 92% were produced as a result of the bovine genome sequencing initiative, while the remaining 8% were obtained from CSIRO, Australia. This assay and an additional 4626 putative SNPs identified from expressed sequence tag data (Hawken et al. 2004) was first used by Khatkar et al. (2007) to characterize haplotype blocks and tag SNPs in 1000 Holstein-Friesian bulls. After excluding monomorphic and unmapped SNPs and also those deviating from Hardy–Weinberg equilibrium, 9195 SNPs with a median spacing of 93.9 kb and average MAF of 0.286 were used for analysis. While this assay was originally designed for high-resolution linkage and association mapping, genomewide LD studies began to indicate that additional SNPs would be required for these purposes. Consequently, Affymetrix released the GeneChip Bovine Mapping 25K SNP kit in May 2007. This GeneChip contained the original 10,000 and an additional 15,000 novel SNPs derived from the sequencing initiative. The assay facilitated the mapping of congenital muscular dystonia type 2 (CMD2), a recessive disorder in Belgian Blue cattle, to a 3.61 Mb region of BTA29 in which a missense mutation in *SLC6A5* was found to be causal for CMD2 (Charlier et al. 2008).

One of the most successful SNP discovery efforts in livestock was catalyzed by the pairing of deep sequencing technology with reduced representation libraries (RRL) (Van Tassell et al. 2008). This SNP detection project was initiated to ensure that sufficient numbers of validated SNPs with known MAF were available for the construction of the Illumina BovineSNP50 assay (Matukumalli et al. 2009). The concept is to produce RRL by pooling the DNA samples of several individuals and performing a size selection of the fragments produced by a complete restriction enzyme digest, avoiding repetitive elements and reducing genomic complexity. More than 71 million sequence reads were generated from three DNA pools on 66 cattle using an Illumina Genome Analyzer. However, the short 25 bp read-length of this nascent sequencing technology did not allow the design of 50-mer probes for the genotyping assay and to overcome this, sequences that flanked each detected SNP were derived from the bovine sequence assembly Btau3.1. After stringent QC thresholds were applied, approximately 50 million sequence reads were utilized to identify 62,042 putative SNPs, which were uniformly distributed across the autosomes and uniquely mapped to Btau3.1. A subset of 24,600 of these SNPs was included on the BovineSNP50 assay and genotypes were produced for the 66 cattle utilized for SNP discovery. Genotypes were produced for 23,357 SNPs yielding a 92% validation rate and an average MAF of 0.27. The BovineSNP50 BeadChip has been the primary driver of GS and genome-wide association studies (GWAS) in the international cattle communities (Settles et al. 2009; VanRaden et al. 2009; Rolf et al. 2010).

Recent advances in sequencing technology have stimulated exhaustive SNP detection efforts in which the whole genomes of individual animals and pools of animals have been sequenced to a total coverage of about 200× resulting in the identification of about 46 M putative SNPs, which are expected to soon be deposited into dbSNP. These SNPs have been used to design both the 777K SNP Illumina BovineHD and the 640K SNP Affymetrix Axiom BOS 1 Array Plate assays. The design of the Affymetrix BOS 1 assay involved the use of three different next-generation sequencing platforms to sequence individuals from 15 breeds of *B. taurus taurus* and *B. taurus indicus* cattle and the development of a prescreening array to validate and estimate MAF for almost 4 M SNPs using the sequenced animals and animals from the HapMap project (The Bovine HapMap Consortium 2009). The result was the validation of approximately 3 million SNPs with 0.7–2.4 M being variable and with an average MAF of from 0.21 to 0.39 across 20 assayed breeds. From these validated SNPs, 640K were sampled to be represented on the commercial BOS 1 assay. The design of the Illumina BovineHD assay also involved the sequencing of similar numbers of animals and breeds resulting in essentially the same set of 46 M putative SNPs as the start point for assay design (although the formal comparison of these data sets has yet to be conducted). The assay contains >749K validated SNPs, of which >99% were mapped to the UMD3.1 bovine assembly (Zimin et al. 2009). The overall average MAF was 0.28 across all breeds and was 0.17, 0.25, and 0.27 for indicine, taurine, and hybrid breeds, respectively. A key difference between the design of these two high-density assays is that the SNPs on the Illumina assay were selected to be evenly physically spaced, while the Affymetrix assay included all SNPs found within transcribed regions and then sampled tagSNPs based upon the LD between loci. This scheme was expected to produce an SNP spacing that was constant on the underlying genetic (recombinational) scale. Array comparison statistics are in Table 14.2. SNP detection and assay development methodologies may change substantially as sequencing technologies increase in throughput and decrease in cost. We expect that future SNP detection chemistries will be based on sequencing rather than the existing multiplex probe hybridization-based chemistries.

**Table 14.2** Comparison of the commercially developed bovine SNP genotyping assays.

| Product | Number of loci | Chemistry | Mean MAF | Mean gap (kb) |
|---|---|---|---|---|
| Illumina | | | | |
| BovineSNP50 | 54,001 | Infinium | 0.26 | 49.4 |
| BovineHD | 777,962 | Infinium | 0.28 | 3.43 |
| Affymetrix | | | | |
| GeneChip Bovine Mapping 10K | 10,000 | MIP[a] | 0.286[b] | 270 |
| GeneChip Bovine Mapping 25K | 25,000 | MIP | 0.24 | 104 |
| Axiom BOS 1 Array Plate | 640,000 | Axiom | 0.21–0.39 | 4.17[c] |

[a] Molecular inversion probe (Hardenbol et al. 2003).

[b] Includes SNPs from the GeneChip Bovine Mapping 10K assay and additional markers from CSIRO, Australia.

[c] These SNPs were not sampled to be physically evenly separated on the assay but to be evenly spaced in terms of expected recombination between the loci. Accordingly, the physical spacing among SNPs is smaller toward the centromeres and telomeres of chromosomes.

There have been three approaches to the commercialization of SNP-based diagnostics in cattle. Pfizer is marketing MEBVs for 14 traits in Angus cattle based upon BovineSNP50 genotypes at a cost of $139. However, this test is not cost effective for commercial producers and many registered breeders, which limits technology application to the elite registered sector. Conversely, Merial is marketing MEBVs for Angus and Limousin based upon filtering the content of the BovineSNP50 assay to the 384 SNPs that best predict EPDs for 12 traits at a cost of $38. Preliminary data suggest that this approach can work well when commercialization occurs within the discovery breed (www.angus.org/AGI/GenomicChoice070811.pdf). The third approach is based upon the development of the Illumina BovineLD BeadChip, which contains 6909 evenly spaced and high-MAF SNPs from the BovineSNP50 assay. While MEBVs can be directly estimated using this assay, the primary use within the US dairy industry has been to impute BovineSNP50 genotypes in Holsteins using the 7K genotypes. This strategy allows the use of MEBV prediction equations developed for the BovineSNP50 assay with accuracies reduced by the extent of genotype imputation error, which is primarily determined by the effective population size and whether the sire and dam or maternal grandsire have previously been genotyped with the BovineSNP50 assay. Strategies for genotype imputation all the way to whole genome polymorphism are currently a hot topic in GS.

## Need to Positionally Clone QTL

Before the development of high-throughput DNA technologies, the use of DNA markers in animal breeding consisted of commercially available tests that examined variants in a single gene, or a very small panel that tested only a limited number of genes (Van Eenennaam et al. 2007). The tested variants were often found using candidate gene approaches or, more favorably, by fine mapping previously identified QTL. However, because there are a large number of genes that underlie variation in quantitative production traits in livestock (Cole et al. 2009; Hayes et al. 2010), this strategy has been shown to be ineffective. The inability of candidate gene and QTL mapping studies to notably impact animal improvement led to the development of GS. However, the statistical approaches that enable the implementation of GS also enable the positional cloning of QTL, and the identification of the causal mutations that underlie a large number of QTL will help remedy some of the difficulties associated with GS.

Kizilkaya et al. (2010) demonstrated that marker panels that exclude causal mutations (quantitative trait nucleotide; QTN) have a limited predictive ability compared to those in which the QTN is included. Thus, the identification of a large number of causal mutations and their inclusion into the marker panels used for GS could have a dramatic impact on the accuracy of MEBV predictions derived from these panels. Additionally, identifying these causal variants may greatly reduce the number of markers that need to be assayed to produce MEBVs. Rather than testing hundreds of thousands of markers per genome, only a few markers per QTL would need to be tested. Even if there are hundreds of moderate- to large-effect QTL that can be detected as influencing selected traits, the number of markers assayed could be a fraction of those required for GS. Furthermore, LD patterns differ and phase relations are not well preserved between breeds (McKay et al. 2007; The Bovine HapMap Consortium 2009), and patterns of LD change with time and after generations of selection (Hartl and Clark 1997). This complicates the estimation of across-breed MEBVs and requires that MEBV prediction equations be periodically updated. However, if causal variants were identified, the same marker panels could be used across breeds and time. When high-density SNP assays are utilized in GS, SNP markers that are in higher LD with QTN will be identified. This will enable the fine-mapping of these QTN, and in the interim, the development and commercialization of revamped reduced representation marker panels. However, changes in commercialized marker panels may require that previously genotyped animals be regenotyped or that sophisticated statistical methods be used to impute genotypes to calibrate the evolving marker panels (Tempelman and Kachman 2008). All of these issues would be resolved with the use of panels comprising causal variants for the prediction of MEBVs.

In addition to resolving issues with GS, the positional cloning of QTL would identify genomic loci warranting additional investigation. After the causal genes have been identified, each should be resequenced in multiple breeds to identify breed-specific variation, which could have functional significance and further enable the improvement of genetic prediction. These causal genes are also important candidates for genetic engineering and small molecule drug targeting. Positional cloning can also assist in the annotation of gene function. As genes with unknown function are identified as being causal for phenotypic variation, biological functions can be applied to these genes and further molecular and cellular functional studies will be suggested. As we better identify the genes and mutations responsible for genetic variance, we will have a more detailed understanding of the relationship between physiology and genetics. Finally, efforts to clone causal mutations in livestock will provide benefits to the greater scientific community including the study of human health, molecular biology, and evolution.

## The Future of Genomic Selection in Beef Cattle

### *Epigenetics*

Epigenetics is the study of heritable changes in gene expression that are not due to changes in the genetic code (Richards 2006). These changes are primarily the result of mechanisms such as DNA methylation and posttranslational histone modification (Jaenisch and Bird 2003; Richards 2006; Morozova and Marra 2008). Methylation events and histone modifications are environmentally influenced and transiently inherited; therefore, epigenetics lies at the boundary of genetic and environmental effects. DNA methylation is most often manifested as the addition of a methyl group to the 5′ position of cytosine residues, which are in CpG (cytosine-phosphate-guanine) rich areas of DNA. This type of methylation can result in altered gene expression including gene silencing due to repressed transcription. Histone modifications are posttranslational alterations such as methylation and acetylation at the N-termini of histone proteins. These modifications cause changes in the chromatin structure that hinder access of the transcriptional machinery to certain DNA sequences or allow access to regions that were once inaccessible resulting in a change of gene expression (Boyes and Bird 1992; Devaskar and Raychaudhuri 2007; Sellner et al. 2007). In mammals, these mechanisms can control gene transcription, X-chromosome inactivation, parental imprinting, and possibly the suppression of transposable element activity (Richards 2006; Lan et al. 2010). Although much has recently been discovered, these mechanisms are not fully understood and much work is needed to uncover their function and potential uses in bovine genomics.

Because epigenetic modifications are stably transmitted over several generations they may play a future role in breeding program design and GS. Sellner et al. (2007) discussed the possibility of integrating imprinted genes and their mutations into MEPDs. Another strategy involves understanding nutritionally and environmentally induced epigenetic effects to alter in utero and early stage development (Tost 2010). A better understanding of the effects of aberrant DNA methylation and histone modification on cloning rates may suggest strategies and technologies that will advance these practices. However, for epigenetic methodologies to benefit beef production, the bovine epigenome or tissue-specific genomewide catalog of DNA methylation patterns must be characterized. This will likely be accomplished using the newly emerged high-throughput sequencing technologies (Suzuki and Bird 2008; Cokus et al. 2008; Morozova and Marra 2008; Schones and Zhao 2008; Wold and Myers 2008; Laird 2010). These approaches are computationally intensive and require the existence of a sequence assembly, but do not require the cloning of samples. They also have single base resolution allowing the precise definition of the boundaries of methylated and unmethylated areas, and methylation patterns in promoter regions.

One of the next-generation technologies incorporates chromatin immunoprecipitation (ChIP) coupled with high-throughput sequencing, termed ChIP-Seq (Mardis 2007; Laird 2010). ChIP-Seq was one of the earliest next-generation technologies to be used for genomewide applications (Park 2009) and can be applied to characterize and map DNA methylation patterns and histone modifications (Laird 2010). Libraries are constructed using immunoprecipitated DNA and are sequenced via a next-generation technology. ChIP-Seq does not require DNA probes, the prior selection of genomic regions of interest, and has low material input requirements. It also has superior sensitivity, lower background noise, and higher resolution than earlier approaches (Barski and Zhao 2009; Park 2009; Schmidt et al. 2009). A second method of analysis is bisulfite sequencing (BS-Seq), which employs the sodium bisulfite conversion of unmethlyated cytosines to uracils followed by PCR amplification, which converts the uracils to thymines, and sequencing to identify the methylation patterns associated with cytosine residues. This approach has been used to identify promoter methylation patterns of candidate genes involved in human cancer (Taylor et al. 2007) and to generate a single nucleotide resolution methylation map of *Arabidopsis thaliana*.

## *Genetic Networks*

Genes interact within networks to regulate phenotypes and consequently mutations in individual genes may produce major effects on phenotype if they are in rate limiting enzymes or regulatory genes (Andersson and Georges 2004), or may behave additively with small cumulative effects within pathways. Since the majority of additive effects in the bovine genome that underlie variation in quantitative traits are small (Goddard and Hayes 2009), interest has recently focused on the use of pathway-based analyses to identify these loci. One such approach in GS is to not fit a GRM based on the entire suite of genotyped SNPs, but to fit a GRM that reflects identity-by-descent among individuals at loci that are involved in specific pathways. Another manifestation of this analysis is to determine if genes within certain pathways are detected as more commonly being associated with trait variation than genes that are selected at random. This form of analysis is known as gene set enrichment analysis, which is a computational method that determines whether an a priori defined set of genes shows statistically significant and concordant associations with a phenotype (Subramanian et al. 2005; Neibergs et al. 2010). The approach has yet to be deeply explored in cattle but can be considered to be a biologically guided Bayesian approach to the identification of the genes that underlie QTL.

On the other hand, nonadditive gene networks are those that require simultaneous interactions between several genes to produce a phenotypic effect when each individual gene may have only a small or even no effect on phenotype (Brazhnik et al. 2002). Discovering the genes within these networks and their epistatic functions can be accomplished in GWAS using nonadditive models but requires large samples to allow sufficient observations within multilocus genotype classes to estimate interaction effects. While these interactions identify the genes, their expression, regulation, and other physical structures of a network in detail (Green et al. 2007), they can also reveal QTN that have nonadditive effects (Flint and Mackay 2009). Since much of US beef production involves the use of crossbreeding to capitalize on breed complementarity and heterosis, efforts should be brought to bear to understand the molecular mechanisms that underlie these phenomena. As these networks become better understood, selection for specific additive × additive genotypes could be facilitated and genotypes within networks that may yield different phenotypes can be examined.

## Conclusions

The genetic determinants of simple Mendelian traits such as coat color and muscle hypertrophy have been localized by GWAS to small genomic regions and the causal mutations have been identified by the sequencing of, usually obvious, candidate genes within these regions (e.g., *MSTN* and *KITL*). However, GWAS of complex traits in human has revealed a significant missing heritability problem, which is likely due to the inability of common variants on genotyping assays to identify the rare variants that create variation in quantitative traits within different families. This phenomenon seems to be less of an issue in cattle suggesting that much of the variation underlying quantitative traits in cattle is common. This is consistent with the large differences in effective population size between cattle ($N_e = 100–500$) and human ($N_e = 7000$) and the recent bottlenecks associated with domestication and breed formation in cattle that should result in the loss of much of the recently evolved variation. While this may provide a powerful argument for the utility and implementation of GS in cattle, the resulting twofold increases in expected response to selection are both exciting and disturbing considering the already low effective population sizes of cattle breeds. Strong selection for production traits will continue to erode the variability within populations and will result in the fixation of haplotypes with large positive effects, but they harbor both favorable and unfavorable alleles, limiting long-term selection response.

Primary issues limiting the efficiency of GS in beef cattle are associated with the assembly of sufficiently large training populations, the need for periodic retraining, and delivering the technology at a price point that justifies adoption. These issues will largely be addressed by advancing technology that will enable very high-density assays to be employed in multibreed training populations and the development of low-cost assays that may be used in the registered and commercial breeding sectors, and within the feedlot sector for marker-assisted management. However, it is not clear if the current business model involving partnerships between academia, genetic technology companies, and breed associations is sustainable for the long-term, and care should be exercised that investments in the development and genotyping of training populations are not lost as the successful business model evolves.

## References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J. (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* **93**: 743–752.

Andersson, L. and Georges, M. (2004) Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics* **5**: 202–212.

Barendse, W., Bunch, R., Thomas, M., Armitage, S., Baud, S., Donaldson, N. (2004) The TG5 thyroglobulin gene test for a marbling quantitative trait evaluated in feedlot cattle. *Australian Journal of Experimental Agriculture* **44**: 669–674.

Barski, A. and Zhao, K. (2009) Genomic location analysis by ChIP-seq. *Journal of Cellular Biochemistry* **107**: 11–18.

Boyes, J. and Bird, A. (1992) Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *The EMBO Journal* **11**: 327–333.

Brazhnik, P., de la Fuente, A., Mendes, P. (2002) Gene networks: how to put the function in genomics. *Trends in Biotechnology* **20**: 467–472.

Calus, M.P.L., Meuwissen, T.H.E., de Roos, A.P.W., Veerkamp, R.F. (2008) Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **178**: 553–561.

Charlier, C., et al. (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nature Genetics* **40**: 449–454.

Cohen-Zinder, M., et al. (2005) Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research* **15**: 936–944.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., Jacobsen, S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**: 215–219.

Cole, J.B., VanRaden, P.M., O'Connell, J.R., Van Tassell, C.P., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Wiggans, G.R. (2009) Distribution and location of genetic effects for dairy traits. *Journal of Dairy Science* **92**: 2931–2946.

de Roos, A.P.W., Hayes, B.J., Goddard, M.E. (2009) Reliability of genomic predictions across multiple populations. *Genetics* **183**: 1545–1553.

de Roos, A.P.W., Schrooten, C., Mullaart, E., Calus, M.P.L., Veerkamp, R.F. (2007) Breeding value estimation for fat percentage using dense markers on *Bos taurus* autosome 14. *Journal of Dairy Science* **90**: 4821–4829.

Devaskar, S.U. and Raychaudhuri, S. (2007) Epigenetics – a science of heritable biological adaptation. *Pediatric Research* **61**: 1R–4R.

Flint, J. and Mackay, T.F.C. (2009) Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research* **19**: 723–733.

Goddard, M.E. and Hayes, B.J. (2007) Genomic selection. *Journal of Animal Breeding and Genetics* **124**: 323–330.

Goddard, M.E. and Hayes, B.J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* **10**: 381–391.

Green, R.D., Qureshi, M.A., Long, J.A., Burfening, P.J., Hamernik, D.L. (2007) Identifying the future needs for long-term USDA efforts in agricultural animal genomics. *International Journal of Biological Sciences* **3**: 185–191.

Grobet, L., et al. (1997) A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nature Genetics* **17**: 71–74.

Habier, D., Fernando, R.L., Dekkers, J.C.M. (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**: 2389–2397.

Hardenbol, P., et al. (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnology* **21**: 673–678.

Harris, B.L., Johnson, D.L., Spelman, R.J. (2008) Genomic selection in New Zealand and the implications for national genetic evaluation. *Proc. ICAR 36th Session*, pp. 325–330.

Hartl, D.L. and Clark, A.G. (1997) *Principles of Population Genetics*. 3rd edition, pp. 99–105. Sunderland: Sinauer Associates, Inc.

Hawken, R.J., Barris, W.C., McWilliam, S.M., Dalrymple, B.P. (2004) An interactive bovine in silico SNP database (IBISS). *Mammalian Genome* **15**: 819–827.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., Goddard, M.E. (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*. **92**: 433–443.

Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J., Goddard, M.E. (2010) Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics* **6**: e1001139.

Henderson, C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**: 423–448.

Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics* **33**: 245–254.

Khatkar, M.S., et al. (2007) A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in Holstein-Friesian cattle. *Genetics* **176**: 763–772.

Kizilkaya, K., Fernando, R.L., Garrick, D.J. (2010) Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *Journal of Animal Science* **88**: 544–551.

Laird, P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics* **11**: 191–203.

Lan, J., Hua, S., He, X., Zhang, Y. (2010) DNA methyltransferases and methyl-binding proteins of mammals. *Acta Biochimica et Biophysica Sinica* **42**: 243–252.

Legarra, A., Aguilar, I., Misztal, I. (2009) A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science* **92**: 4656–4663.

Luan, T., Woolliams, J.A., Lien, S., Kent, M., Svendsen, M., Meuwissen, T.H.E. (2009) The accuracy of genomic selection in Norwegian Red cattle assessed by cross-validation. *Genetics* **183**: 1119–1126.

Mardis, E.R. (2007) ChIP-seq: welcome to the new frontier. *Nature Methods* **4**: 613–614.

Matukumalli, L.K., et al. (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* **4**: e5350.

McKay, S.D., et al. (2007) Whole genome linkage disequilibrium maps in cattle. *BMC Genetics* **8**: 74.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Morozova, O. and Marra, M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**: 255–264.

Mulder, H.A., Calus, M.P.L., Veerkamp, R.F. (2010) Prediction of haplotypes for ungenotyped animals and its effect on marker-assisted breeding value estimation. *Genetics Selection Evolution* **42**: 10.

Neibergs, H.L., Settles, M.L., Whitlock, R.H., Taylor, J.F. (2010) GSEA-SNP identifies genes associated with Johne's disease in cattle. *Mammalian Genome* **21**: 419–425.

Park, P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**: 669–680.

Richards, E.J. (2006) Inherited epigenetic variation – revisiting soft inheritance. *Nature Reviews Genetics* **7**: 395–401.

Rolf, M.M., Taylor, J.F., Schnabel, R.D., McKay, S.D., McClure, M.C., Northcutt, S.L., Kerley, M.S., Weaber, R.L. (2010) Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. *BMC Genetics* **11**: 24.

Schaeffer, L.R. (2006) Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics* **123**: 218–223.

Scheet, P. and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics* **78**: 629–644.

Schmidt, D., Wilson, M.D., Spyrou, C., Brown, G.D., Hadfield, J., Odom, D.T. (2009) ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions. *Methods* **48**: 240–248.

Schones, D.E. and Zhao, K. (2008) Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics* **11**: 179–191.

Seidel, G.E., Jr. (2009) Brief introduction to whole-genome selection in cattle using single nucleotide polymorphisms. *Reproduction, Fertility and Development* **22**: 138–144.

Sellner, E.M., Kim, J.W., McClure, M.C., Taylor, K.H., Schnabel, R.D., Taylor, J.F. (2007) Board-invited review: applications of genomic information in livestock. *Journal of Animal Science* **85**: 3148–3159.

Settles, M., Zanella, R., McKay, S.D., Schnabel, R.D., Taylor, J.F., Whitlock, R., Schukken, Y., Van Kessel, J.S., Smith, J.M., Neibergs, H. (2009) A whole genome association analysis identifies loci associated with Mycobacterium avium subsp. paratuberculosis infection status in US Holstein cattle. *Animal Genetics* **40**: 655–662.

Solberg, T.R., Sonesson, A.K., Woolliams, J.A., Odegard, J., Meuwissen, T.H.E. (2009) Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. *Genetics Selection Evolution* **41**: 53.

Su, G., Guldbrandtsen, B., Gregersen, V.R., Lund, M.S. (2010) Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. *Journal of Dairy Science* **93**: 1175–1183.

Subramanian, A., et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences United States of America* **102**: 15545–15550.

Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics* **9**: 465–476.

Taylor, K.H., Kramer, R.S., Davis, J.W., Guo, J., Duff, D.J., Xu, D., Caldwell, C.W., Shi, H. (2007) Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Research* **67**: 8511–8518.

Tempelman, R.J. and Kachman, S.D. (2008) Integrating genetic evaluations with DNA technologies for the ultimate selection tool. *Journal of Animal Science* **86**(E-Suppl. 2): 103–104.

The Bovine Genome Sequencing and Analysis Consortium (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522–528.

The Bovine HapMap Consortium (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**: 528–532.

Toosi, A., Fernando, R.L., Dekkers, J.C.M. (2010) Genomic selection in admixed and crossbred populations. *Journal of Animal Science* **88**: 32–46.

Tost, J. (2010) DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. *Molecular Biotechnology* **44**: 71–81.

Van Eenennaam, A.L., Li, J., Thallman, R.M., Quaas, R.L., Dikeman, M.E., Gill, C.A., Franke, D.E., Thomas, M.G. (2007) Validation of commercial DNA tests for quantitative beef quality traits. *Journal of Animal Science* **85**: 891–900.

VanRaden, P.M. (2008) Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**: 4414–4423.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Schenkel, F.S. (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**: 16–24.

VanRaden, P. M. (2011) findhap.f90. Accessed January 27, 2012. http://aipl.arsusda.gov/software/findhap/.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., Sonstegard, T.S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* **5**: 247–252.

Villanueva, B., Pong-Wong, R., Fernandez, J., Toro, M.A. (2005) Benefits from marker-assisted selection under an additive polygenic genetic model. *Journal of Animal Science* **83**: 1747–1752.

Weigel, K.A., de los Campos, G., Gonzalez-Recio, O., Naya, H., Wu, X.L., Long, N., Rosa, G.J.M., Gianola, D. (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *Journal of Dairy Science* **92**: 5248–5257.

Wold, B. and Myers, R.M. (2008) Sequence census methods for functional genomics. *Nature Methods* **5**: 19–21.

Zimin, A.V., et al. (2009) A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology* **10**: R42.

# Chapter 15
# Impact of High-Throughput Genotyping and Sequencing on the Identification of Genes and Variants Underlying Phenotypic Variation in Domestic Cattle

*Michel Georges*

## Introduction

The emergence of genomics as a new discipline in the 80s raised the hope that genes and mutations underlying domestic biodiversity could be identified, revealing the molecular architecture of the traits under artificial selection and paving the way to more effective marker-assisted selection (MAS). Twenty-five years later, the arsenal of genomic tools has considerably matured including high-throughput SNP genotyping and next-generation sequencing. The identification of the causative genes and mutations underlying Mendelian traits, including monogenic defects, has become nearly trivial, which allows management of recessive defects with unprecedented efficacy. The availability of medium- and high-density SNP arrays combined with advanced statistical methods allows for genomic selection, which is revolutionizing livestock breeding, starting with dairy cattle. Approaches for the identification of quantitative trait nucleotides (QTN) are being developed and will gain in efficiency as imputation of genome-wide SNP information from resequencing larger cohorts becomes a reality.

## Empirical and Biometrical Selection: Effective Manipulation of a Black Box

As soon as man domesticated plants and animals, he unwittingly modified their genome. By selecting progenitors with desired traits, breeders increased the frequency of favorable alleles at multiple loci. Selected alleles were either sampled from the wild by the domestication process, or appeared postdomestication by neomutation. The remarkable effectiveness of artificial selection is demonstrated by the often larger phenotypic differentiation observed between domestic breeds than between wild species.

*Bovine Genomics*, First Edition. Edited by James E. Womack.

Domestic plant and animal populations seem to conceal sufficient genetic variation to allow sustained response to selection for virtually any trait.

Centuries of empirical, phenotype-based selection were augmented in the twentieth century by biometrical methods rooted in quantitative genetics theory (Lynch and Walsh 1998). Phenotypes were modeled as the outcome of "fixed" environmental and "random" individual animal effects, allowing improved selection of animals with superior breeding values. The spectacular increases in production efficiency achieved during the last 50 years are primarily due to genetic improvement.

One of the remarkable features of both empirical and biometrical selection is that neither requires knowledge of the genes and variants on which they act. Breeders refer to the molecular substrate that they manipulate as the "black box." Quantitative genetics theory assumes that genetic variation for complex traits (including economically important traits in livestock) reflects the addition of mostly tiny allele-substitution effects at a very large number of "polygenes," that is, the infinitesimal model. The distribution of allelic effects is predicted to be exponential: a minority of larger effects dominated by a majority of minute ones (Barton and Keightley 2002; Orr 2005). True "major" gene effects are assumed to be oddities, applying mainly to inherited defects or coat color variation.

## Early Days in Livestock Genomics: Attempting QTL-Based Marker-Assisted Selection

### *Identifying Causative Genes and Variants Influencing Economically Important Traits: Novel Opportunities in Livestock Production*

Approaches based on linkage analysis to singularize (i.e., map) genes underlying phenotypic variation, whether monogenic or complex, were devised very early on (Sax 1923; Thoday 1961). However, in most organisms their implementation was hampered by the lack of genetic markers. This limitation was overcome with the discovery of RFLPs in the 1980s (Bostein et al. 1980) and microsatellites in the 1990s (Weber and May 1989). The same period saw the first positional cloning successes for monogenic diseases in human (Kerem et al. 1989; The Huntington's Disease Collaborative Research Group 1993). These spurred efforts in human and other organisms (including livestock) to identify genes underlying monogenic traits (mostly inherited diseases) as well as quantitative trait loci or QTL (i.e., loci harboring polygenes) influencing complex traits of medical (common diseases) and agronomic relevance. The motivations justifying these efforts were multiple. Identifying the genes and mutations causing genetic defects would allow for the development of new therapies or at least prenatal diagnosis in human, and the elimination of carriers or avoidance of at-risk matings in livestock. Identifying QTL and better QTN (i.e., the causative DNA sequence variants or DSV) influencing disease predisposition would increase our understanding of disease pathogenesis, identify new drug targets, and pave the way toward personalized medicine. Identifying QTL and QTN influencing economically important traits in livestock would allow for more effective MAS, and identify new targets for performance enhancing drugs or for transgenic engineering.

## *Positional Cloning, a Generic Strategy to Identify Causative Genes and Variants, Proceeds in Three Steps*

The most common approach for the identification of genes and variants underlying phenotypes of interest is positional cloning. It was classically viewed as a three-step process. In a first instance, the genes to be identified were mapped by linkage analyses, that is, by extracting within-family information. As linkage extends over tens of centimorgan (corresponding to tens of million of base pairs), panels of 200 to 300 evenly distributed microsatellite markers were sufficient to scan the genome. Likelihood methods were often used to infer which chromosomes in a pedigree were most likely to be identical-by-descent (IBD) at a given map position, as well as to estimate the effect of the segregating alleles on phenotype. Evidence for the presence of a locus affecting phenotype at a given map position was assessed by comparing the likelihood of the data assuming an effect of the locus on phenotype ($H_1$ hypothesis) with that assuming no effect of the locus on phenotype ($H_0$ hypothesis). Significance thresholds were adjusted to account for the realization of ~500 independent tests when scanning a typical mammalian genome (Lander and Kruglyal 1995). Being limited by "current" recombinational events (i.e., occurring in gametes produced by members of the analyzed pedigrees), the mapping resolution was typically limited to several tens of centimorgans, corresponding to several millions of base pairs often encompassing hundreds of genes. Most successful linkage-mapping experiments conducted in cattle took advantage of the large paternal half-sib pedigrees that result from the common use of artificial insemination (AI). In essence, offspring of a given sire are sorted in two groups according to the paternal homolog inherited at the tested map position and the phenotype of interest compared between the two groups. This was, for instance, the main source of information that was exploited to map the *Polled* locus to chromosome 1 (Georges et al. 1993a). It is also the basis of the daughter and granddaughter designs (Weller et al. 1990), and provides the bulk of the linkage signal in most line-crosses (Kim et al. 2003). Occasionally, other pedigree structures have been exploited in cattle, including backcrosses in which the informative F1 parents were cows (Charlier et al. 1995), or complex multigenerational pedigrees (Georges et al. 1993b).

Once one or more loci influencing the phenotype of interest were mapped by linkage analysis, their fine-mapping could be attempted in the second positional cloning step. Fine-mapping requires an increase in (1) local marker density, and (2) local crossover density. Until the recent generation of a reference sequence for the bovine genome and, with it, millions of single nucleotide polymorphisms (SNP) (Bovine Genome Sequencing and Analysis Consortium 2009; Bovine HapMap Consortium 2009), generating locus-specific high-density marker maps was a laborious process. It required either the development and sequence characterization of "comparative anchored tagged sequences" (CATS) to reveal new SNPs, and/or the development of locus-specific YAC/BAC contigs from which new markers (including microsatellites) could be developed (Pirottin et al. 1999; Grisart et al. 2002, 2004). Crossover density was either increased by focusing the analysis on offspring having inherited informative chromosomes recombining in the chromosomal region of interest (Thaller and Hoeschele 2000), or—more commonly and conveniently—by exploiting the nonrandom association expected to exist at the population level between the causative variants and at least some of the nearest markers, that is, by exploiting linkage disequilibrium (LD). The strength of association (measured by $r^2$) between markers and causative mutations decays much faster when exploiting populationwide LD:

$$r^1 = 1/(4Ne\theta + 1),$$

where $\theta$ is the recombination rate;

than when relying on within-family linkage:

$$r^2 = 0.25 + \theta + \theta^2,$$

underpinning the superior resolution of LD-based association mapping.

Assuming that the causative variants could be fine-mapped to chromosome segments encompassing a "tractable" number ($<10$) of "positional candidate genes," the last positional cloning step is typically described as a targeted resequencing effort that is expected to lead to the identification of the causative variants and genes, followed by functional assays to reveal the molecular mechanisms leading to the phenotype of interest. How to achieve this goal exactly, however, is rarely considered in great detail. A particularly poignant issue is that the resequenced mutant and wild-type chromosomes will not only differ for the causative variants, but also at hundreds if not thousands of "passenger" DSV. The question then becomes how to pinpoint the causative needle in a haystack of associated neutral variants and to identify which causative genes they affect. Most causative variants identified so far by positional cloning in livestock are restricted to coding variants that can be easily predicted to have a highly disruptive effect on the structure of a protein with known function in another organisms (typically, human or mouse).

## The Candidate Gene Approach: Pros and Cons

An alternative approach to the three-step positional cloning route for the identification of genes and variants underlying phenotypic variation in livestock has been to skip step one (i.e., linkage mapping) and directly apply principles of association mapping to "physiological candidate genes," that is, genes coding for a protein whose demonstrated functions suggest that it may be involved in the expression of the phenotype or disease of interest. A typical candidate-gene experiment would involve resequencing of the selected gene to identify mutations or polymorphisms, genotyping a phenotyped population, and assessing association between genotype and phenotype. The candidate-gene approach has been successful in identifying the causative gene and mutation for genetic defects for which the molecular cause was known in human or mouse (Shuster et al. 1992; Schwenger et al. 1993; http://omia.angis.org.au/). Moreover, the candidate-gene approach has clearly demonstrated that genetic variation in the major milk proteins (i.e., caseins, $\beta$-lactoglobulin and $\alpha$-lactalbumin) influences physicochemical properties of milk (reviewed in FitzGerald 1997 and Hill et al. 1997).

However, especially when studying complex quantitative traits, the candidate-gene approach as applied in the 1990s suffered several limitations. The first is that it obviously is limited to the study of genes with known function, which even today, only account for a minor fraction of the genome. Hence, the candidate-gene approach is by definition not amenable to uncover novel gene functions. The second is that resequencing was never exhaustive, such that (1) the chance to reveal the actual causative variants was low, and association—if detected–-most likely indirect (i.e., due to LD between the interrogated polymorphisms and the unseen causative variant(s)), and (2) absence of association did not exclude a causative involvement of the selected gene. Moreover, most candidate-gene studies applied liberal significance thresholds providing inappropriate control of the type I error rate (false positives). Indeed, candidate-gene studies very seldom accounted for the actual testing of multiple candidates and markers therein. The nominal $p$-value of 0.05, which was often used as significance threshold in candidate-gene studies, has to be compared with $p$-values $<10^{-6}$–$10^{-8}$, which are required in present-day genomewide association studies (GWAS) (in which admittedly many more test are performed) to declare significance. Finally, most candidate-gene studies did not properly account or correct for stratification, which is very severe in most cattle populations. I suspect that a large fraction of positive associations reported in cattle studies were largely due to population stratification.

Candidate-gene studies applied to quantitative traits also vividly illustrate the difficulty of the third step of positional cloning, that is, the identification of the causative variants. As an example, the effect of the casein gene cluster on milk yield and composition has been extensively studied in cattle, yet a satisfactory understanding of which DSV are causative is still missing.

Of note, prior information about gene function is often used to prioritize positional candidate genes in step three of positional cloning. As an example, when it was realized that the myostatin gene, shown to cause a muscular hypertrophy when knocked-out in the mouse, mapped to the chromosome interval to which the double-muscling gene had been assigned by linkage analysis, it was immediately scanned for mutations in Belgian Blue and other double-muscled breeds, leading to the discovery of an allelic series of loss-of-function mutations (Grobet et al. 1997, 1998; Kambadur et al. 1997; McPherron and Lee 1997). Likewise, the known function of *DGAT1* (i.e., acyl CoA:diacylglycerol acyltransferase) in triglyceride synthesis and the agalactia of *DGAT1* knockout (KO) mice, combined with its colocalization affecting milk fat composition made it a prime resequencing target, leading to the identification of the causative K232A mutation affecting the $V_{\max}$ of the enzyme (Grisart et al. 2002, 2004; Winter et al. 2002). Another interesting example—albeit in dogs—of prioritization of positional candidate genes according to available functional information is the mining of ciliome databases to prioritize 3 of 150 positional candidates as putative causative genes for primary ciliary diskinesia (PCD) leading to the discovery of *CCDC39* loss-of-function mutations in dogs and subsequently in human PCD cases (Merveille et al. 2011).

## *QTL Mapping and MAS: Lukewarm Appraisal*

By the end of the 1990s, a number of mutations underlying Mendelian traits (primarily genetic diseases) had been identified in livestock, providing the means to effectively control the corresponding defects. But even the positional cloning of monogenic traits remained arduous, time consuming, and expensive. A large number of putative QTL, affecting nearly all examined agronomically important traits, were reported (http://www.animalgenome.org/bioinfo/). However, causative DSV had been identified for only three QTL (Grisart et al. 2002, 2004; Winter et al. 2002; Blott et al. 2003; Cohen-Zinder et al. 2005), markers in populationwide LD identified for a few more, while most QTL remained poorly resolved chromosomal regions affecting traits of interest with variable levels of statistical support. MAS schemes aimed at exploiting within-family QTL segregation (Kashi et al. 1990; Mackinnon and Georges 1998) were rightfully considered too tedious, while the trait variance explained by the best characterized QTL was considered insufficient by most breeding organization to justify large-scale implementation of MAS. Funding further QTL mapping and fine-mapping efforts, particularly by the private sector, was under threat.

## Impact of High-Density SNP Genotyping on the Analysis of Monogenic Traits: Highly Effective IBD Mapping

As mentioned previously, positional cloning as practiced in the 1990s was extremely tedious and expensive, especially when dealing with complex traits. Fortunately, spectacular advances in genomics, accrued over the last 10 years, have had a major impact on our ability to identify loci underlying phenotypic variation. Three achievements in particular have been essential. The first is the obtainment of reference sequences for a growing list of species, including cattle (Bovine Genome Sequencing and Analysis Consortium 2009). Draft-quality, annotated sequence information is now available for most of the bovine genome. The second is the emergence of next-generation sequencing (NGS) technology that allows for cost-effective targeted or even whole-genome resequencing of individuals of interest (Mardis 2008). For the bovine, this has resulted in a genomewide collection of SNPs that is now in excess of 40 million (Van Tassel, personal communication). The last is the development of high-throughput SNP genotyping platforms that today permit genotyping of >50,000 SNPs at a cost <$100, and of >700,000 SNP at a cost <$300 (Charlier et al. 2008; Matukumalli et al. 2009). These developments have drastically changed the positional cloning process, essentially merging steps one (linkage mapping) and two (LD-based fine-mapping) in a single step. I will hereafter illustrate these evolutions, first, using specific monogenic and later polygenic examples.

### *Autozygosity Mapping of Recessive Defects in Livestock*

As a result of extensive reliance on AI, popular sires with tens to hundreds of thousands of offspring have become commonplace. While such intense selection may accelerate genetic response for desired traits, it also results in the widespread dissemination of deleterious recessives, of which most individuals carry several (1000 Genomes Project Consortium 2010). This causes frequent bursts of genetic defects in livestock population, including cattle. Well-documented examples of such outbursts in Holstein-Friesian dairy cattle include bovine leukocyte adhesion deficiency (BLAD) (Shuster et al. 1992) and complex vertebral malformation (CVM) (Thomsen et al. 2006). Because of the way these defects emerge in livestock, they are typically allelically homogenous, that is, they involve a unique, IBD mutation that usually traces back to a popular "founder" sire. Affected individuals are thus predicted to be homozygous for an IBD mutation, that is, "autozygous." Cases will not only be autozygous for the mutation, but also for the haplotype upon which the mutation occurred in the founder animal. The expected size of the segment (in centimorgan) shared autozygous by *n* cases can be shown to equal 1/*ng*, where *g* is the number of generations separating the cases from the founder animal (Dunner et al. 1997). For ten cases and ten generations this corresponds to 1 cM or approximately 1 million base pairs. Using the Illumina 50K (respectively 700K) array, such a segment is on average covered by ~20 (respectively 270) SNPs.

Detecting such regions of autozygosity using standard linkage programs that compute exact likelihoods becomes difficult as the large number of available SNPs and ungenotyped pedigree members inflate computing time. Trimming the number of SNPs will accelerate computing but may drastically affect detection power. As an alternative, we have developed software that heuristically scans the genome for regions of autozygosity shared by *n* cases (ASSIST; Charlier et al. 2008). ASSIST computes local *p*-values by phenotype permutation, that is, it generates a distribution of highest "sharing scores" obtained anywhere across the genome when randomly shuffling case-control status. The "sharing scores" account for the SNP's minor allele frequency (MAF) estimated from *m* healthy controls. Local *p*-values of the actual sharing scores are then obtained by comparison with the ranked sharing scores obtained by phenotype permutation. One could argue that cases will on average be more inbred than controls. To account for this, we have developed a sister program that performs "homozygosity mapping" (ASSHOM; Charlier et al. 2008). ASSHOM searches for regions that are homozygous in all cases but not in controls. It generates individual-specific homozygosity scores (*s*) that are combined across cases (*S*). The *s* scores account for allelic frequencies estimated in controls. The *p*-value of the homozygosity scores is determined by locus permutation, that is, random shuffling of *s* scores across the genome, hence accounting for the inbreeding coefficient of the examined cases. Practice shows that, using the 50K array, the probability that *n* randomly selected individuals will be autozygous/homozygous for all SNPs over a segment of size 1/*ng* anywhere in their genome is well below 0.05 (hence providing adequate detection power of such shared segment) with as few as three cases (Charlier et al. 2008). The signal to noise ratio is, of course, considerably enhanced when using the recently developed Illumina 700K rather than 50K array.

Previously, one accumulated as much epidemiological evidence as possible supporting the genetic determinism of a defect before embarking in a tedious positional cloning exercise. With the availability of high-density SNP arrays, association mapping of loci underlying recessive defects has become so cost-effective that it is now the method of choice to determine whether a defect is inherited or not. We have established a heredo-surveillance platform in Belgium that—with the help of a network of veterinarians and farmers—systematically collects samples from abnormal calves. Once a handful of cases with similar phenotypes have been collected, a genome-scan is conducted using SNP arrays to search for regions of autozygosity. Identifying such segments establishes the genetic determinism of the condition and immediately provides the means to select against the defect.

## *Identifying the Causative Mutations Underlying Recessive Defects*

The extreme phenotypes characterizing genetic defects are mostly the outcome of loss-of-function mutations that disrupt protein-encoding genes. A large proportion of these are "structural" variants (i.e., mutations that affect the structure of the encoded protein, as opposed to "regulatory" variants) resulting either from coding mutations (frameshift, nonsense, or nonconservative missense), splice site mutations, or large deletions. Contrary to regulatory mutation, our molecular understanding of transcript processing and translation is sufficient to effectively recognize such loss-of-function mutations. As a consequence, after autozygosity mapping of the disease locus, a logical next step is to sequence the open reading frame and exon boundaries of the protein-encoding genes annotated in the region. These can be prioritized based on known gene function or phenotypic effect of loss-of-function mutations in other organisms (mostly inherited defects in human or KO mice) (Charlier et al. 2008; Fasquelle et al. 2009). However, and as the cost of resequencing by NGS decreases, we are more and more replacing this sequential exon-targeting PCR-based approach by either sequence capture of the entire interval followed by NGS resequencing or even whole genome NGS resequencing (Charlier et al. 2012).

The identification of a predicted loss-of-function mutation in the region of autozygosity is strong evidence that the causative mutation has been identified, but is insufficient proof on its own. Indeed, a nonnegligible proportion of loss-of-function mutations are likely to be asymptomatic as demonstrated by the lack of symptoms often exhibited by KO mice. Especially for missense mutations, the effect on gene function remains difficult to predict. Additional evidence supporting causality is thus required, including: (1) the fact that loss-of-function mutations in the orthologous gene in other species (typically, human or mice) cause related symptoms, (2) the fact that the mutation is not found outside the affected breed, (3) the statistically significant lack of homozygous mutant among healthy animals (requires the genotyping of a large cohort of healthy individuals), (4) the absence of other clear loss-of-function mutations in the autozygous interval (when having complete resequence data), and (5) the demonstration of effects on gene function (nonsense mediated RNA decay in the case of nonsense mutations or altered splicing in the case of splice-site variants). We recently identified a group of three private SNP clustering within a seven bp highly conserved coding region of the *Chloride Channel 7* gene (*CLCN7*) mapping to an interval shared autozygous by 11 hamartoma cases in Belgian Blue (Charlier and Sartelet, unpublished observation). They change an ultraconserved Tyrosine into a Glutamine (Y746Q) in the cystathionine $\beta$-synthase domain 2 (CBS2) of this $Cl^-$ channel. At first glance, and despite the finding of this missense mutation, the link between the *CLCN7* gene and gingival tumors was not obvious. *CLCN7* KO mice primarily exhibit severe osteopetrosis. *Ex post* examination of a bovine calf with hamartoma clearly revealed the previously overlooked osteopetrosis, thereby in essence proving the causality of the Y746Q mutation.

Since 2006 we have mapped 12 recessive defects using autozygosity mapping. Using the approaches described previously, we have now identified the causative mutations from nine of these, providing a rough estimate of the success rate that can be expected using this approach.

### *Dealing with Dominance and Genetic Heterogeneity*

The IBD signal used to map a trait locus is strongest in the case of recessivity and allelic homogeneity. It nevertheless remains detectable in slightly more complex situations. One of these is encountered when tracking dominant traits. Under that scenario, most "cases" are expected to share one IBD haplotype on one of their homologs (rather than two in case of a recessive trait). To map a locus underlying a dominant trait, one therefore searches for regions of the genome for which one cannot exclude that all cases share at least one haplotype IBD. These are segments of the genome for which none of the SNP-genotyped cases have alternate homozygous genotype (say, AA vs. BB). We have developed software (ASSDOM) that generates a score that is inversely proportional to the probability that one could not exclude nonexisting sharing of an IBD haplotype (spanning $k$ markers by $n$ cases). The score corresponds to:

$$\sum\nolimits_{n=1}^{k} -\log(1-p^2)^n,$$

where $p$ is the frequency of the allele for which none of the cases is homozygous estimated in $m$ controls.

The statistical significance of the "nonexclusion" signal is determined by phenotype permutation of the disease status between the $n$ cases and $m$ controls. Care should be taken to avoid that cases and controls have different kinship coefficients. The signal-to-noise ratio is obviously lower than for recessives, but proved sufficient in our hands to map the "color-sided" locus by analyzing 21 cases and 30 controls (Durkin et al. 2012). Using the 700K rather than 50K Illumina array should substantially increase the power to detect dominant trait loci.

Another situation in which ASSIST or ASSHOM would be put in check is in the case of a recessive condition with allelic and/or locus heterogeneity. In such situations, one cannot expect all affected individuals to be autozygous for a single haplotype at a unique locus. Despite the small effective population size, within-breed allelic heterogeneity is sometimes observed. Examples in cattle include double-muscling (Grobet et al. 1998), crooked tail syndrome in Belgian Blue (Sartelet et al. 2012a), and chondrodysplastic dwarfism in Japanese Brown cattle. (Takeda et al. 2002). At least for crooked tail syndrome, the allelic heterogeneity is more than likely related to the fact that animals that are heterozygous for loss-of-function mutations in the *MRC2* gene exhibit enhanced muscularity, which is a desired trait in Belgian Blue (Fasquelle et al. 2009). Whether the occurrence of two distinct chondrodysplasia-causing mutations in the *LIMBIN* gene in Japanese Brown cattle is somehow linked to a selective advantage of heterozygous individuals remains unknown, but is certainly a hypothesis worth testing. We have encountered locus heterogeneity when studying dwarfism in Belgian Blue cattle. In this breed, approximately 40% of dwarfs are autozygous for a splice-site variant in the *RNF11* gene that maps to BTA3, while the remaining cases remain unexplained (Sartelet et al., 2012b).

We have developed software (ASSHAP) that will effectively deal with dominance as well as allelic and locus heterogeneity. Available SNP genotypes are first phased using Phasebook (Druet and Georges 2010), which exploits familial information (when available) as well as populationwide LD information. Phasebook uses a Hidden Markov Model to assign homologs to a predetermined number of ancestral haplotype states. Haplotype state frequencies are then compared between cases and controls using a generalized mixed model that includes a polygenic effect to account for population stratification (Zhang et al. 2012). Hypothesis testing is conducted using a score-test. ASSHAP proved effective at detecting the distinct loci underlying dwarfisms in Belgian Blue cattle. We are now routinely using ASSHAP to map loci influencing binary traits, whether monogenic or complex.

## Impact of High-Density SNP Genotyping on the Analysis of Complex Traits: Genomic Selection & QTN Identification

### *Where Biometry and Genomics Meet: Genomic Selection*

As mentioned before, early QTL mapping efforts mainly provided poorly resolved QTL explaining an insufficient proportion of the genetic variance to justify MAS. Theo Meuwissen and Mike Goddard realized early on that a larger proportion of the genetic variance could be accounted for by: (1) readily exploiting LD rather than linkage when scanning the genome for QTL, and (2) using approaches other than skyscraper significance thresholds to control the type I error rate. The former would become achievable once genomewide medium- and high-density SNP panels would be available, which has since become reality. The latter could be achieved by accounting for the presumed exponential distribution of QTL effects, that is, few large and many small QTL effects. When estimated, QTL effects are assumed to be drawn from such prior distributions using either restricted maximum likelihood (REML) or Bayesian approaches. In the proposed approach, dubbed "genomic selection" (GS) (Meuwissen et al. 2001), QTL effects are first estimated in a "training population" comprising animals with phenotypic and genotypic (SNP) information. Unknown phenotypes are then predicted for genotyped animals (young animals) by addition of the SNP-specific QTL effects estimated in the training population. GS was evaluated as an alternative to progeny testing to select dairy bulls as soon as bovine medium-density SNP arrays became available. The results were close to spectacular (VanRaden et al. 2009). Provided that the training cohort is large enough (i.e., thousands if not tens of thousands of individuals), squared correlations between predicted and realized breeding values are increased by ~35% when compared to traditional parent-average based predictions (largely due to the ability to track Mendelian sampling in the offspring). This added information corresponds to ~30 daughter equivalents for high heritability traits ($h^2$ ~0.3), and to >100 daughter equivalents for low heritability traits ($h^2$ ~0.05). When compared to progeny testing, even moderate drops in accuracy (more pronounced for traits with high $h^2$) are largely compensated for by the fact that genomic EBVs are available at birth (or even for biopsied preimplantation embryos), hence drastically reducing the generation interval. GS was adopted worldwide by the dairy cattle breeding industry in a matter of months and has largely replaced progeny testing in most countries.

The observed distribution of QTL effects confirmed the mainly "quasi-infinitesimal" architecture of most analyzed quantitative traits. With the exception of a handful of larger QTL effects (often coinciding with previously identified QTL, including the *DGAT1* K232A mutation affecting milk fat content), the remainder of the SNP effects are generally minute and evenly distributed across the genome (Hayes et al. 2010). Note that one of the advanced features shared by all GS procedures is that the effects of individual SNPs are estimated conditional on all other SNP effects. For SNP clusters in high LD, this could result in the fractionation of a single true QTL effect among the set of correlated SNPs causing an apparent underestimation of the true size of the QTL effect. This would particularly be the case when using Markov Chain Monte Carlo approaches yielding average SNP effects across chain runs. However, there is no strong evidence thus far that accounting for these dependencies would alter the support for the mostly quasi-infinitesimal architecture of the studied traits. Indeed, for most traits, near-identical genomic EBV accuracies are obtained when fitting a single animal effect with covariances proportionate to genomewide, SNP-based identity-by-state (IBS) metrics. Of note, recent analyses support a similar highly polygenic architecture of human height and predisposition to complex diseases (Yang et al. 2010; Lee et al. 2011).

The successful application of GS in dairy cattle breeding spurred similar efforts in beef cattle and other livestock species. While the contexts are not as obviously in favor of GS in beef cattle as they are in dairy cattle and the common use of breed admixture may impose the user of denser SNP panels, it seems reasonable to predict that GS will soon impact breeding practices across livestock breeds and species.

## *Identifying Causative DSV and Genes: Quixotic Quest?*

The success of GS provides an answer to one of the declared scientific questions justifying early QTL mapping efforts: Fisher's infinitesimal model is not only mathematically convenient but more importantly biologically relevant! Yet, demonstrating the quasi-infinitesimal architecture of complex traits does not provide detailed molecular information of the link between DSV and phenotype. What is the nature of the causative variants and genes? Animal breeders may now use molecular tools, yet still manipulate a black box.

Some of us are still interested in deciphering causality, although realizing that—with the exception of a few larger QTL effects—the demonstrated quasi-infinitesimal architecture of complex traits probably makes this a more arduous task than anticipated. Nevertheless, the availability of high-density SNP genotyping has considerably streamlined QTL fine-mapping and I hereafter describe some of the lessons we learned in efforts to positionally clone QTN.

## *Combined Linkage and LD Fine-Mapping of QTL in Livestock*

As for monogenic traits, QTL mapping is now typically performed by genotyping the target populations with medium- (50K) if not high-density (700K) SNP arrays. This allows for simultaneous extraction of the linkage (within-family) and LD (populationwide) signal. Mixed models provide a convenient framework to that effect when dealing with quantitative traits. An individual's phenotype is modeled as the linear outcome of a series of environmental fixed effects (sex, season . . .), one or more locus-specific QTL effects, a genomewide polygenic effect, and an error term. QTL, polygenic effect, and error terms are modeled as random effects, that is, they are drawn from multivariate normal distributions with constrained covariance structure. Individual error terms are assumed to be uncorrelated. The covariance between pairs of polygenic effects is assumed to be proportionate to twice the kinship coefficient of the respective individuals. The "genomewide" kinship coefficient is either computed from genealogy or from genomewide SNP data (Yang et al. 2010). The covariance between pairs of locus-specific QTL effects is assumed to be proportionate to twice the "locus-specific kinship coefficient." Locus-specific kinship coefficients can be estimated from flanking SNP genotypes using linkage and LD information. IBD-probability of chromosomes segregating in a pedigree (in which most individuals are genotyped) is determined by linkage analysis. The relatedness between founder chromosomes (i.e., the chromosomes of the "top" generation without genotyped ancestors that are considered unrelated by linkage analysis) at a given map position can be quantified using a variety of approaches that exploit LD (Meuwissen and Goddard 2000, 2001). Linkage and LD-derived IBD-probabilities can easily be merged to obtain estimates of the IBD-probability (at a given map position) between all pairs of genotyped chromosomes, yielding "locus-specific kinship coefficients." The respective variance components are estimated from the data by REML analysis, while hypothesis testing is typically performed using a likelihood ratio test (corresponding to two times the logarithm of the likelihood of the data under the full model divided by the likelihood of the data under a reduced model, that is, without QTL effect). Permutation tests can be applied to derive empirical significance accounting for multiple testing, but computation time may be limiting.

This mixed-model approach has many usual properties, including: (1) the seamless, integrated extraction of nearly all positional information embedded in the data (linkage information from both male and female meioses, plus LD information), (2) the fact that it doesn't require assumptions about the number of QTL alleles segregating at a given locus, and (3) effective protection against stratification provided by the polygenic effect and the simultaneous linkage and association testing reminiscent of the transmission disequilibrium test (TDT).

It is worthwhile noting that in livestock, LD can be readily exploited to refine the map position of QTL even in experimental "line-crosses." In model organisms such as mice, F2 pedigrees are generated from inbred parental strains differing for the phenotype of interest. Being inbred, each parental line contributes only one haplotype to the pedigree, all F1 animals having the exact same genotype across the genome. Consequently, there is very little, if any, LD information that can be used to refine the map position of the QTL; all mapping information resides in crossing-over events occurring in gametes produced by F1 animals. In livestock, on the contrary, line-crosses are generated by mating several outbred parents from divergent breeds. As an example, a Holstein-Friesian × Jersey F2 pedigree was generated to identify QTL of interest to the dairy industry involving more than 400 F0 animals of each breed (Karim et al. 2011). As a consequence, more than two haplotypes segregate in the F2 population at most loci, providing potentially useful fine-mapping information.

We have recently developed an efficient variant of this mixed-model approach (Druet and Georges 2010). In this, SNP genotypes are first partially phased using the available familial information. Hidden markov methodology that simultaneously models linkage and LD is then used to assign all homologs in the data set to a predetermined number of "hidden haplotype states." The locus-specific QTL effects in the mixed linear model described before is converted to evaluating the effect on phenotype of the hidden haplotype states. These are still preferably modeled as a random effect (as this tend to yield more conservative "shrunken" haplotype effects when compared to modeling them as fixed effects), although their covariance is set at zero. One of the nice features of the approach is that it conveniently identifies hidden haplotype states with significantly different effects on phenotype. These can become the focus of resequencing efforts to identify the causative DSV.

### *Mendelizing a Polygenic Trait by Marker-Assisted Segregation Analysis (MASA)*

MASA is an alternative approach for the reliable identification of chromosomes that are functionally different at a given QTL. It has been used in several studies that have successfully identified QTN in livestock (Grisart et al. 2002; Van Laere et al. 2003). It takes advantage of the common use of AI in livestock and the ensuing large paternal half-sib pedigrees. Assume a series of large paternal half-sib pedigrees that have both been phenotyped and marker-genotyped at a given QTL location. For each pedigree, one can compute the likelihood of the data assuming that the sire is: (1) heterozygous for the QTL (H1), or (2) homozygous for the QTL (H0). If the likelihood under H1 is significantly larger than under H0, the corresponding sire can confidently be assumed to be heterozygous for the QTL (genotype *Qq*). Thus, its two homologs must differ at the causative QTN. If the likelihood of the data is significantly higher under H0 than under H1, the sires is likely homozygous at the QTL (genotype *QQ* or *qq*) and hence homozygous for the causative QTN.

Having identified a number of sire chromosomes with known QTL genotype (especially *Qq* sires), one can then search for a haplotype that is shared IBS either by all *Q* or all *q* chromosomes. The hypothesis underlying this approach is that either the *Q* or the *q* allele is young, homogenous, and embedded in a unique, long haplotype in the population of interest. This approach allowed fine-mapping of a QTL influencing muscularity in pigs to a 250-Kb SSC2 chromosome segment (Nezer et al. 2003). More recently, it allowed fine-mapping of a QTL influencing bovine stature to a-335 Kb BTA14 segment (Karim et al. 2011). However, the approach is not without pitfalls if some of the hypothesized conditions are not met. We initially erroneously fine-mapped a QTL influencing milk fat composition to a BTA14 chromosome segment, based on a haplotype shared by all studied *Q* chromosomes in both the Dutch and New

Zealand Holstein-Friesian populations (Riquet et al. 1999). Subsequent work showed that the *Q* allele was actually embedded in distinct haplotypes in the Netherlands and New Zealand (Farnir et al. 2002).

An alternative, or potentially complementary approach, is to completely resequence the QTL genotyped chromosomes in the QTL confidence interval (CI) (whether defined by combined linkage + LD fine-mapping, or haplotype sharing). This can be achieved either by direct sequencing of long range PCR products spanning the CI (Karim et al. 2011), by sequence capture of the CI, or—increasingly—by genome-wide resequencing. For CI that typically span hundreds of Kb, this will yield thousands of candidate DSV. However, only a small proportion of those is likely to segregate with QTL genotype among the sequenced chromosomes. As an example, out of >10,000 SNPs identified by resequencing the 780 Kb CI for a bovine stature QTL, only 14 followed the QTL segregating pattern, and all but one of these were clustered in an 80 Kb subregion (Karim et al. 2011). The advantage of resequencing (over the haplotype sharing approach), is that it would be effective even if both QTN alleles were present on multiple marker haplotypes in the studied population(s).

## *Exploiting Between-Breed Haplotype Diversity: Increasing Genetic Resolution*

In the Karim et al. (2011) study, the approaches previously described led to the identification of 13 candidate QTN for a QTL affecting bovine stature. In the two studied breeds (Holstein-Friesian and Jersey), the corresponding DSV were all in perfect LD ($r^2 \sim 1$), precluding further genetic differentiation. Thus, we genotyped an available breed diversity panel in the hope to find recombinant haplotypes segregating at high enough frequency in some populations to be able to study their effect on stature. At least four such haplotypes were observed, segregating in Hereford, Senepol, Simmental, and Wagyu. The phenotypic effect of one of these could be studied in Simmental, allowing us to exclude five from the 13 candidate QTN. The phenotypic effect of the remainder recombinant haplotypes has not been examined so far but could potentially reduce the list of candidate QTN even more. Of interest, the mosaic pattern of several of the observed recombinant haplotypes suggested that they result from gene conversion events rather than from reciprocal homologous recombination.

It has been proposed, particularly in beef cattle, to perform GS across breeds using the high-density (>700K) SNP chips. The purpose of this approach would be to capture LD signals that would be consistent across breeds as they would depend on the causative QTN or very tightly linked DSV. In effect, this corresponds to the genomewide systematization of the approach applied in a targeted way to the BTA14 stature QTL by Karim et al. (2011).

## *Pinpointing the Causative QTN: Need for Multimarker Models*

The approaches previously described may lead to the identification of strong candidate QTN. However, they do not prove the causality of the identified DSV. If causal, the corresponding QTN should be more strongly associated with phenotype than any other DSV in the vicinity. In the Karim et al. (2011) study, this proved to be the case. When assayed for their effect on phenotype, the newly identified candidate QTL yielded a stronger association signal than any other DSV in the vicinity whether considered alone or as haplotypes.

However, there is once again a caveat: the assumption of strongest association is only valid if the QTL reflects the effect of a single QTN or a cluster of perfectly associated QTN. If the observed QTL is allelically heterogeneous, that is, involves multiple QTN that are not in perfect LD, passenger DSV that are in LD with more than one causative QTN may individually yield a stronger signal than any of the truly causative QTN, that is, so-called synthetic or positively misleading association (Dickson et al. 2010; Platt et al. 2010). Overcoming this issue requires the simultaneous inclusion of multiple (if not all) DSV in the statistical model, that is, the effect of a DSV is estimated conditional on that of all other neighbors. This ultimate analysis was not fully conducted in the Karim et al. (2011) paper (although two QTL models were applied). Of note, and as mentioned before, the models underlying GS share this advanced feature to some extent.

Functional tests are sometimes presented as an additional means to distinguish causative from passenger DSV. Common examples are reporter and electrophoretic mobility shift assays to study the effect of DSV predicted to affect promoter strength (Van Laere et al. 2003; Karim et al. 2011). While such assays may be useful to inform us about the molecular modus operandi of DSV that have been proven causative by genetic approaches, I find them generally unconvincing as a support of causality. The exception, of course, is to engineer a candidate DSV in the orthologous position in the mouse genome and demonstrate that it recapitulates the same phenomenology as in the original species. Even then, however, interpretation of the outcome may not be straightforward as experienced for the callipyge phenotype (Davis et al. 2004; Pirottin et al. 2011)

## *Identifying the Causative Genes: Mutational Load and Quantitative Complementation*

In the unlikely scenario that genetics would point to one and only one QTN and that it would be a coding SNP, evidence supporting the causality of the corresponding gene would be very strong. In most cases, however, genetics will provide a limited list of candidate QTN of which some may be coding while most will not. The latter could still alter gene coding capacity by affecting splicing, and this has to be tested by studying the effect of QTN genotype on transcript integrity (Karim et al. 2011). However, there is growing evidence that many QTN will be regulatory, that is, affect gene product in a quantitative rather than qualitative way. Regulatory QTN can affect transcription rate, transcript stability, translation rate, or protein stability. The ovine *c.2360G > A* mutation in the *MSTN* 3′ UTR is an example of a QTN affecting transcript stability and translation rate (Clop et al. 2006). While QTN affecting transcript stability, translation rate, or protein stability reside within the transcript (hence, defining the causative gene), QTN altering transcription rate may act over hundreds of kilobases and affect multiple genes if affecting long-range control elements. Examples of such long-range effects include the *CLPG* mutation (Freking et al. 2002; Smit et al. 2003) and the QTN underlying the BTA14 stature QTL (Karim et al. 2011). Determining which of the affected genes causes the phenotype in such cases is hard. This is well illustrated by the callipyge example. Transgenic mice with ectopic expression of DLK1 protein in skeletal muscle exhibit a muscular hypertrophy, which strongly suggest that *DLK1* is causally involved in the callipyge muscular hypertrophy (Davis et al. 2004). Yet, 10 years after the discovery to the *CLPG* causative mutation, it remains unclear whether ectopic expression of *PEG11* also contributes to the callipyge phenotype (Byrne et al. 2010).

There are two formal tests for gene causality. The first is the demonstration of an effect of phenotype on the mutational burden of the candidate gene. A nice application of this test can be found in the genetics of Crohn's disease (CD). In a classic illustration of positional cloning, Hugot et al. (2001) identified *NOD2/CARD15* as first risk gene for CD. The linkage and LD signal that let to the identification of *NOD2/CARD15* was due to three "common" disruptive mutations (*R702W, G908R, 1007fs*) enriched in cases. Subsequent resequencing of *NOD2* in cases and controls, showed that in addition to the three common mutation, 17% of case chromosomes harbored low frequency or rare missense *NOD2* mutations while the corresponding figure was only 5% in controls. This finding essentially proved the causality of *NOD2* beyond any doubt. Along related lines, a recent scan of positional candidates from GWAS, revealed an enrichment of low-frequency coding variants in the *IL23R* gene in controls. In this case, DSV that are dampening *IL23* signaling thus protect against inflammatory bowel disease (Momozawa et al. 2011). One can imagine that for some genes, some coding variants may be protective, while others will increase risk. The C-alpha score test has recently been adapted to assay the overdispersion that would result from such a situation in the distribution of low-frequency variants among cases and controls (Neale et al. 2011). Related approaches have been applied to continuously distributed traits by resequencing the candidate genes in individuals with extreme phenotype (Romeo et al. 2007).

The identification of an allelic series of disruptive *MSTN* mutations in double-muscled animals of different cattle breeds (Grobet et al. 1998) can be viewed as an application of such burden test (although the causality of the *MSTN* gene was clearly demonstrated before by the hypermuscled phenotype of *MSTN* KO mice). However, in our opinion, the burden test is unlikely to be very powerful in livestock for most complex quantitative traits (unless influenced by a major gene akin to *MSTN*). This is primarily due to the fact that effective population size of most livestock population is very small (in the hundreds). Demonstrating a significant effect of phenotype on mutational burden would therefore entail resequencing of the candidate genes in a prohibitively large sample size.

The second formal test of gene causality is the reciprocal hemizygosity test (Stern 1998; Steinmetz et al. 2002). This very elegant test compares the phenotype of individuals that are heterozygous for the QTL (F1 individuals when the QTL was mapped in an inter- or backcross), yet have been rendered hemizygous for positional candidate genes by knocking-out the madumnal (inherited from their mother), respectively padumnal (inherited from the father) allele. If the KO gene is not involved in the QTL effect, both hemizygous types (padumnal vs. madumnal KO) are functionally equivalent and will have identical phenotype. Note that the phenotype may differ from that of the original F1 as being hemizygous for the tested gene may affect the phenotype, but the reciprocal hemizygotes will be equally affected. If—on the contrary—the KO gene is causally involved in the QTL, the reciprocal hemizygotes will not be functionally equivalent. One strain will only have a functional $Q$ allele, while the other will only have a functional $q$ allele, which will cause their phenotype to differ. All positional candidate genes should be systematically and sequentially tested to assay which (one or several) of the positional candidates contribute(s) to the QTL effect.

Applying the reciprocal hemizygosity test thus requires the generation of two allele-specific KOs per analyzed gene. In practice, this is only achievable in model organisms such as yeast and *Drosophila*. The quantitative complementation assay (QCA) is related to the reciprocal hemizygosity test, yet less demanding: it only requires the generation of one KO per analyzed gene (Mackay 2001). To realize the QCA, one needs: (1) chromosomes carrying, respectively, the $Q$ and $q$ allele for the studied QTL (the two homologs of an F1 parent if the QTL was identified in a back- or intercross), and (2) a pair of chromosomes with, respectively, a wild-type ($+$) and a KO ($\Delta$) copy of the analyzed candidate gene. One then generates individuals of the four possible genotypes, that is, (1) $Q+$, (2) $q+$, (3) $Q\Delta$, and (4) $q\Delta$. The assumption is that if the studied gene is involved in the determinism of the QTL, the $q$ to $Q$ allele substitution effect will be larger when the reference chromosome carries the KO allele ($\Delta$) than when it carries the wild-type allele ($+$), that is:

$$(Q\Delta - q\Delta) > (Q + -q+).$$

The interpretation is thus that in hemizygotes (i.e., in the absence of the buffering effect of a wild-type allele) the $q$ to $Q$ allele substitution effect will be enhanced. The QCA is not as tight as the reciprocal hemizygosity test and its interpretation is somewhat controversial (Service 2004). Nevertheless, it has been used to dissect QTL in *Drosophila* (Mackay 2001), and at least once in the mouse (Yalcin et al. 2004).

Genomewide KO collections are available in *Drosophila* and will soon become available in the mouse (Austin et al. 2004), allowing the use of the QCA in these species. In nonmodel organisms such as human and livestock, generating KO is either unconceivable or unachievable, hence, precluding the systematic use of the QCA. However, naturally occurring null alleles are segregating in outbred populations at sometimes appreciable frequencies (1000 Genomes Project Consortium 2010). These can, in theory, be exploited to perform the QCA. We have used this approach to test the causality of the *CHCHD7* gene in the determinism of the BTA14 QTL on stature (Karim et al. 2011). Following up on eQTL analyses, we identified a splice-site variant predicted to generate a *CHCHD7* null allele. This DSV was segregating at high enough frequency in the Holstein-Friesian dairy cattle population to allow the QCA. There was not the slightest evidence for a difference between the $(Q\Delta - q\Delta)$ and $(Q + -q+)$ contrasts, expected if *CHCHD7* was causally involved in the QTL effect. Both $(Q + -Q\Delta)$ and $(q + -q\Delta)$ were positive, albeit nonsignificantly, and smaller than the $(Q\Delta - q\Delta)$ and $(Q + -q+)$ contrasts despite the fact that the effect of the splice-site variant on *CHCHD7* transcript levels was larger than that of the

QTN. Taken together, these findings did not support a direct role of *CHCHD7* in causing the QTL effect.

### *Genetical Genomics: Distinguishing Correlation from Causation*

The possibility to monitor the expression levels of the entire transcriptome, whether by array technology or RNA sequencing (RNAseq), is offering new opportunities for the identification of causative genes underlying QTL. Assume that a QTL affecting a phenotype of interest (pQTL) has been mapped in an experimental F2 cross for which genomewide transcriptome data is available in a relevant tissue. A gene subject to a *cis*-eQTL effect with CI overlapping that of the pQTL would be considered a prime positional candidate gene (Hubner et al. 2005).

Once again, there is a caveat. *Cis*-eQTL are now known to be very common in all analyzed tissues. Finding colocalized pQTL and eQTL is thus on its own not exceptional at all. While such finding might be due to the fact that the QTN alter(s) the transcript levels of the candidate gene and that this causes variation in the phenotype, an equally likely explanation (but less interesting scenario from the point of view of the positional cloner) is that altered transcript levels are unrelated to phenotypic variation. The observed eQTL could be caused by the same QTN as the pQTL (pQTN) or even (probably, more often) by different eQTN. In all cases, phenotype and transcript levels will be correlated in the F2 population, either because pQTN and eQTN are the same, or because pQTN and eQTN are in strong LD in the F2 population. Correlation between transcript level and phenotype is thus insufficient to conclude for a causal relation. However, one can examine whether the correlation still exists when conditioning on genotype. Thus, one will look at whether phenotype and transcript levels remain correlated within each one of the three possible F2 genotypes. Residual variation in transcript levels is either nongenetic or due to other loci in the genome. If an association with phenotype would still be observed within genotype, this would much more convincingly implicate the positional candidate gene in the determinism of the QTL.

Outbred populations offer additional discriminating power to untangle eQTL and pQTL. Indeed, if pQTL and eQTL are causally related, pQTL and eQTL genotype need to match perfectly for all individuals. Hence, the fact that different F1 sires were segregating (and thus of *Qq* genotype) for the pQTL on stature and the splice-site variant dependent eQTL on *CHCHD7* levels also pleaded against *CHCHD7* being the causative gene (Karim et al. 2011).

## Impact of Next-Generation Sequencing on the Analysis of Monogenic Traits

### *Single-Step Positional Cloning of Genes Underlying Mendelian Defects*

As costs of genomewide resequencing are rapidly diminishing, it is becoming conceivable to search for mutations causing recessive defects in a single step, that is, by directly resequencing the entire genome of a very small number of affected and control individuals. Autozygosity mapping could readily be achieved based on the sequence traces by searching for regions of extended homozygosity shared among affected individuals. Candidate-causative mutations would appear as DSV differentiating case and control genome sequences in autozygous regions, and could be prioritized based on their predicted effect on gene function, known role of affected gene, or location in a highly conserved sequence elements. Candidate DSV can be filtered against the growing database of DSV reported in breeds without the condition. We have successfully applied a closely related approach to identify EMS-induced mutations of interest in zebrafish (Voz et al. 2012).

## *Genotype-Driven Screens for Embryonic Lethals*

We recently used positional cloning to identify the cause of brachyspina syndrome (BSS) in Holstein-Friesian dairy cattle: a 3.3 Kb deletion in the bovine *FANCI* gene. Surprisingly, BS carriers were observed at a much higher frequency (∼7%) than predicted from the incidence of affected calves (≪1/10,000 births). We suspected that this might be due to prenatal death of the majority of homozygous mutant fetuses. To test this hypothesis, we studied the effect of carrier status of sire and dam on the probability that the dam would return into heat after insemination. The probability that the dam would return into heat within 9 months after insemination was increased by ∼12% when sire and dam were BSS carriers when compared to controls. This suggests that at least 12% of concepti, corresponding to half the expected proportions of homozygous mutants, might abort. Preliminary evidence suggests that embryonic survival is even affected in matings where only one of the parents is carrier, which might result from an increase in the proportion of aneuploid gametes produced by carrier animals. The main phenotypic manifestation of the BSS mutations is thus on fertility rather than the genetic defect per se.

The BSS mutation could be identified because at least a fraction of affected calves survive until partum. Recessives that would cause abortion of all affected fetuses would essentially go unnoticed. If several such early recessive lethals were segregating in a breed of interest, they might jointly have a sizeable impact on fertility. Assume that ten such loci would be segregating in the Holstein-Friesian population at a frequency equivalent to BSS, they would cause premature termination of ∼1% of pregnancies.

We have devised a genotype-driven experiment that aims at detecting such embryonic lethals. The aim is to exploit NGS to resequence the exome and exon-intron boundaries for ∼100 distantly related animals from the breed of interest. DSV that are predicted to have a disruptive effect on the protein structure will be identified bioinformatically. An array will be designed to effectively interrogate the ensuing candidate SNPs, and used to genotype a cohort of >5000 healthy individuals from the same breed. True embryonic lethals should (1) never be observed at the homozygous state among healthy animals and the resulting departure from Hardy-Weinberg equilibrium should be statistically significant (including Bonferroni correction for the evaluation of multiple candidates), and (2) carrier status of sire and dam should have a negative effect on fertility traits including nonreturn rates. We are starting to implement this scheme in the Belgian Blue and other cattle breeds.

## Impact of Next-Generation Sequencing on the Analysis of Complex Traits: Imputing Genotype from Sequence Data

Generating genomewide sequence data on cohorts that are large enough for the analysis of complex traits is likely to remain prohibitively costly for some time (although conservative predictions of this kind have increasing probability to be proven wrong). However, what is almost immediately achievable is exploitation of linkage and LD-information to project ("impute") genotype probabilities at a very large number of common- (0.5 > MAF > 0.05) and low-frequency (0.05 > MAF > 0.005) variants from a "reference set" of genomic sequences (similar to the 1000 genomes project in human; 1000 Genomes Project Consortium 2010) upon a "target" population that has been genotyped with medium- (50K) or high-density (700K) SNP arrays. Several efforts toward that goal are ongoing. Highly effective imputation software is available, primarily from developments in human genetics (Marchini and Howie 2010). Association studies and/or GS can be conducted using both real and in-silico predicted genotype probabilities. The expectation is that this will increase detection power and mapping resolution as more causative or markers in tight LD with them will be included in the analyzed set of DSV.

## Conclusions

The development of genomics starting in the 1980s offered perspectives to identify genes and variants underlying phenotypic variation in livestock. While first attempts were arduous, recent technological breakthroughs in genotyping and sequencing technology have greatly accelerated forward genetic dissection of both monogenic and polygenic traits in livestock. Mutations underlying inherited diseases can now be mapped in days and identified in weeks rather than years. This allows for effective management of emerging defects, a recurrent issue in many livestock populations. Cost-effective SNP genotyping has enabled the implementation of GS, which is revolutionizing breeding practices. Combined with genotype imputation from emerging genomewide resequence data, GS-related approaches will increasingly pinpoint causative DSV and genes, improving the accuracy of "genomic breeding values" and revealing the mechanisms linking genotype to phenotype.

## Acknowledgment

Studies described in this review were funded by the Walloon Ministry of Agriculture, the Fonds National de la Recherche Scientifique, the Communauté Française de Belgique, and the European FW7 program. I am grateful to Carole Charlier, Wouter Coppieters, Tom Druet, and Harulo Takeda for critically reviewing this manuscript. The list of references is heavily biased toward work conducted over the years at the UAG. I apologize to those readers who consider that their contribution is insufficiently acknowledged.

## References

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Austin, C.P., et al. (2004) The knockout mouse project. *Nature Genetics* **36**: 921–924.

Barton, N.H. and Keightley, P.D. (2002). Understanding quantitative genetic variation. *Nature Reviews Genetics* **3**: 11–21.

Blott, S., et al. (2003) Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics* **163**: 253–266.

Bostein, D., White, R.L., Skolnick, M., Davis, R.W. (1980) Constrcution of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**: 314–331.

Bovine Genome Sequencing and Analysis Consortium, et al. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522–528.

Bovine HapMap Consortium, et al. (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**: 528–552.

Byrne, K., Colgrave, M.L., Vuocolo, T., Pearson, R., Bidwell, C.A., Cockett, N.E., Lynn, D.J., Fleming-Waddell, J.N., Tellam, R.L. (2010) The imprinted retrotransposon-like gene PEG11 (RTL1) is expressed as a full-length protein in skeletal muscle from Callipyge sheep. *PLoS One* **5**: e8638.

Charlier, C., et al. (1995) The mh gene causing double-muscling in cattle maps to bovine Chromosome 2. *Mammalian Genome* **6**: 788–792.

Charlier, C., et al. (2008) Highly effective SNP-based association mapping and management of recessive defects in livestock. *Nature Genetics* **40**: 449–454.

Charlier, C., et al. (2012) A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina. Submitted for publication.

Clop, A., et al. (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nature Genetics* **38**: 813–818.

Cohen-Zinder, M., et al. (2005) Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research* **15**: 936–944.

Davis, E., Jensen, C.H., Schroder, H.D., Farnir, F., Shay-Hadfield, T., Kliem, A., Cockett, N., Georges, M., Charlier, C. (2004) Ectopic expression of DLK1 protein in skeletal muscle of padumnal heterozygotes causes the callipyge phenotype. *Current Biology* **14**: 1858–1862.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., Goldstein, D.B. (2010) Rare variants create synthetic genome-wide associations. *PLoS Biology* **8**: e1000294.

Druet, T. and Georges, M. (2010) A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* **184**: 789–798

Dunner, S., Charlier, C., Farnir, F., Brouwers, B., Canon, J., Georges, M. (1997) Towards interbreed IBD fine mapping of the mh locus: double-muscling in the Asturiana de los Valles breed involves the same locus as in the Belgian Blue cattle breed. *Mammalian Genome* **8**: 430–435.

Durkin, K., Coppieters, W., Drögemüller, C., Ahariz, N., Cambisano, N., Druet, T., Fasquelle, C., Haile, A., Horin, P., Huang, L., Kamatani, Y., Karim, L., Lathrop, M., Moser, S., Oldenbroek, K., Rieder, S., Sartelet, A., Sölkner, J., Stålhammar, H., Zelenika, D., Zhang, Z., Leeb T., Georges, M., Charlier, C. (2012) Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* **482**: 81–84.

Farnir, F., et al. (2002) Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* **161**: 275–287.

Fasquelle, C., et al. (2009) Balancing selection of a frame-shift mutation in the MRC2 gene accounts for the outbreak of the Crooked Tail Syndrome in Belgian Blue Cattle. *PLoS Genetics* **5**: e1000666.

FitzGerald, R.J. (1997) Exploitation of casein variants. In: *Milk Composition, Production and Biotechnology*, edited by R.A.S. Welch, D.J.W. Burns, S.R. Davis, A.I. Popay, and C.G. Prosser, pp. 153–172. New York: CAB International.

Freking, B.A., Murphy, S.K., Wylie, A.A., Rhodes, S.J., Keele, J.W., Leymaster, K.A., Jirtle, R.L., Smith TP. (2002) Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals. *Genome Research* **12**: 1496–1506.

Georges, M., et al. (1993a) Microsatellite mapping of a gene affecting horn development in Bos taurus. *Nature Genetics* **4**: 206–210.

Georges, M., et al. (1993b) Microsatellite mapping of the gene causing weaver disease in cattle will allow the study of an associated quantitative trait locus. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 1058–1062.

Grisart, B., et al. (2002) Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research* **12**: 222–231.

Grisart, B., et al. (2004) Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 2398–2403.

Grobet, L., Poncelet, D., Royo, L.J., Brouwers, B., Pirottin, D., Michaux, C., Ménissier, F., Zanotti, M., Dunner, S., Georges, M. (1998) Molecular definition of an allelic series of mutations disrupting the myostatin function and causing double-muscling in cattle. *Mammalian Genome* **9**: 210–213.

Grobet, L., et al. (1997) A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nature Genetics* **17**: 71–74.

Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J., Goddard, M.E. (2010) Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics* **6**: e1001139.

Hill, J.P., et al. (1997) In: *Milk Composition, Production and Biotechnology*, edited by R.A.S. Welch, D.J.W. Burns, S.R. Davis, A.I. Popay, and C.G. Prosser, pp. 173–203. New York: CAB International.

Hubner, N., et al. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics* **37**: 243–253.

Hugot, J.P., et al. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**: 599–603

Kambadur, R., Sharma, M., Smith, T.P., Bass, J.J. (1997) Mutations in myostatin (GDF8) in double-muscled Belgian Blue and Piedmontese cattle. *Genome Research* **7**: 910–916.

Karim, L., et al. (2011) Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nature Genetics* **43**: 405–413.

Kashi, Y., Hallerman, E., Soller, M. (1990) Marker-assisted selection of candidate bulls for progeny testing programs. *Animal Production* **51**: 63–74.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., Tsui LC. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**: 1073–1080.

Kim, J.J., Farnir, F., Savell, J., Taylor, J.F. (2003) Detection of quantitative trait loci for growth and beef carcass fatness traits in a cross between Bos taurus (Angus) and Bos indicus (Brahman) cattle. *Journal of Animal Science* **81**: 1933–1942.

Lander, E.S. and Kruglyal, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**: 241–247.

Lee, S.H., Wray, N.R., Goddard, M.E., Visscher, P.M. (2011) Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics* **88**: 294–305.

Lynch, M. and Walsh B. (1998) *Genetic Analysis of Quantitatve Traits*. Sunderland: Sinauer Associates.

Mackay, T.F. (2001) Quantitative trait loci in Drosophila. *Nature Reviews Genetics* **2**: 11–20.

Mackinnon, M. and Georges, M. (1998) A bottom-up approach towards marker assisted selection. *Livestock Production Science* **54**: 229–250.

Marchini, J. and Howie, B. (2010) Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**: 499–511.

Mardis, E.R. (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**: 387–402.

Matukumalli, L.K., et al. (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* **4**: e5350.

McPherron, A.C. and Lee, S.J. (1997) Double muscling in cattle due to mutations in the myostatin gene. *Proceedings of the National Academy of Sciences of the United States of America* **94**: 12457–12461.

Merveille, A.C., et al. (2011) CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs. *Nature Genetics* **43**: 72–78.

Meuwissen, T.H. and Goddard, M.E. (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**: 421–430.

Meuwissen, T.H. and Goddard, M.E. (2001) Prediction of identity by descent probabilities from marker-haplotypes. *Genetics Selection Evolution* **33**: 605–634.

Meuwissen, T.H., Hayes, B.J., Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Momozawa, Y., et al. (2011) Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. *Nature Genetics* **43**: 43–47.

Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., Daly, M.J. (2011) Testing for an unusual distribution of rare variants. *PLoS Genetics* **7**: e1001322.

Nezer, C., Collette, C., Moreau, L., Brouwers, B., Kim, J.J., Giuffra, E., Buys, N., Andersson, L., Georges, M. (2003) Haplotype sharing refines the location of an imprinted quantitative trait locus with major effect on muscle mass to a 250-kb chromosome segment containing the porcine IGF2 gene. *Genetics* **165**: 277–285.

Orr, H.A. (2005). The genetic theory of adpatation: a brief history. *Nature Reviews Genetics* **6**: 119–127.

Pirottin, D., Charlier, C., Georges, M., Takeda, H. (2011). Knocking the CLPG mutation in the mouse genome partially recapitulates the callipyge phenomenology. Submitted for publication.

Pirottin, D., et al. (1999) High-resolution, human-bovine comparative mapping based on a closed YAC contig spanning the bovine mh locus. *Mammalian Genome* **10**: 289–293.

Platt, A., Vilhjálmsson, B.J., Nordborg, M. (2010) Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**: 1045–1052.

Riquet, J., et al. (1999) Fine-mapping of quantitative trait loci by identity by descent in outbred populations: application to milk production in dairy cattle. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 9252–9257.

Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H., Cohen, J.C. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature Genetics* **39**: 513–516.

Sartelet, A., Klingbeil, P., Fasquelle, C., Franklin, C., Géron, S., Isacke, C., Georges, M., Charlier, C. (2012a) Allelic heterogeneity of Crooked Tail Syndrome fits the balancing selection hypothesis. *Animal Genetics* **in press.**

Sartelet, A., Fasquelle, C., Géron, S., Michaux, C., Zhang, Z., Coppieters, W., Georges, M., Druet, T., Charlier, C. (2012b) A splice site mutation in the RNF11 gene causes a deficiency in innate immunity with proportionate dwarfism. *PLOS Genetics* **in press**.

Sax, K. (1923) The association of size differences with seed-coat pattern and igmentation in Phaseolus vulgaris. *Genetics* **8**: 552–560.

Schwenger, B., Schöber, S., Simon, D. (1993) DUMPS cattle carry a point mutation in the uridine monophosphate synthase gene. *Genomics* **16**: 241–244.

Service, P.M. (2004) How Good Are Quantitative Complementation Tests? *Science of Aging Knowledge Environment* **12**: pe13.

Shuster, D.E., Kehrli M.E. Jr, Ackermann, M.R., Gilbert, R.O. (1992) Identification and prevalence of a genetic defect that causes leukocyte adhesion deficiency in Holstein cattle. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 9225–9229.

Smit, M., Segers, K., Carrascosa, L.G., Shay, T., Baraldi, F., Gyapay, G., Snowder, G., Georges, M., Cockett, N., Charlier C. (2003) Mosaicism of Solid Gold supports the causality of a noncoding A-to-G transition in the determinism of the callipyge phenotype. *Genetics* **163**: 453–456.

Steinmetz, L.M., Sinha, H., Richards, D.R., Spiegelman, J.I., Oefner, P.J., McCusker, J.H., Davis, R.W. (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**: 326–330.

Stern, D. (1998) A role of ultrabithorax in morphological differences between Drosophila species. *Nature* **396**: 463–466.

Takeda, H., et al. (2002) Positional cloning of the gene LIMBIN responsible for bovine chondrodysplastic dwarfism. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 10549–10554.

Thaller, G. and Hoeschele, I. (2000) Fine-mapping of quantitative trait loci in half-sib families using current recombinations. *Genetical Research* **76**: 87–104.

The Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**: 971–983.

Thoday, J.M. (1961) Location of polygenes. *Nature* **191**: 368–370.

Thomsen, B., Horn, P., Panitz, F., Bendixen, E., Petersen, A.H., Holm, L.E., Nielsen, V.H., Agerholm, J.S., Arnbjerg, J., Bendixen, C. (2006) A missense mutation in the bovine SLC35A3 gene, encoding a UDP-N-acetylglucosamine transporter, causes complex vertebral malformation. *Genome Research* **16**: 97–105.

Van Laere, A.S., et al. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**: 832–836.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Schenkel, F.S. (2009) Invited review: reliability of genomic predictions for North American Holstein Bulls. *Journal of Dairy Science* **92**: 16–24.

Voz, M., Coppieters, W., Mafroid, I., Baudhuin, A., Cahrlier, C., Meyer, D., Driever, W., Martial, J., Peers, B. (2012) Effective NGS-based identification of ENU-induced mutations in zebrafish. Submitted for publication.

Weber, J.L. and May, P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* **44**: 388–396.

Weller, J.I., Kashi, Y., Soller, M. (1990) Power of daughter and grand-daughter desgns for dtermining linkage between marker loci and quantitatve trait loci in dairy cattle. *Journal of Dairy Science* **73**: 2525–2537.

Winter, A., Krämer, W., Werner, F.A., Kollers, S., Kata, S., Durstewitz, G., Buitkamp, J., Womack, J.E., Thaller, G., Fries R. (2002) ssociation of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 9300–9305.

Yalcin, B., Willis-Owen, S.A., Fullerton, J., Meesaq, A., Deacon, R.M., Rawlins, J.N., Copley, R.R., Morris, A.P., Flint, J., Mott, R. (2004) Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice. *Nature Genetics* **36**: 1197–1202.

Yang, J., et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**: 565–569.

Zhang, Z., Sartelet, A., Charlier, C., Georges, M., Farnir, F., Druet, T. (2012) Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. Submitted for publication.

# Index

*Bovine Genomics*, First Edition. Edited by James E. Womack.