REPORTS

The Genome Sequence of Tauroine Cattle: A Window to Ruminant Biology and Evolution

The Bovine Genome Sequencing and Analysis Consortium,1 Christine G. Elsik,1 Ross L. Tellam,2 Kim C. Worley3

To understand the biology and evolution of ruminants, the cattle genome was sequenced to about sevenfold coverage. The cattle genome contains a minimum of 22,000 genes, with a core set of 14,345 orthologs shared among seven mammalian species of which 1217 are absent or undetected in noneutherian (marsupial or monotreme) genomes. Cattle-specific evolutionary breakpoint regions in chromosomes have a higher density of segmental duplications, enrichment of repetitive elements, and species-specific variations in genes associated with lactation and immune responsiveness. Genes involved in metabolism are generally highly conserved, although five metabolic genes are deleted or extensively diverged from their human orthologs. The cattle genome sequence thus provides a resource for understanding mammalian evolution and accelerating livestock genetic improvement for milk and meat production.

References and Notes

2. Previous IMR studies of game-playing include Gallagher et al. (43) and Bhatt and Camerer (44), but they address different issues. In particular, Bhatt and Camerer found higher insula and ACC activity when comparing choices to first-order beliefs in dominance-solvable games.
3. We are considering here coordination without visual or other contact. Nonhuman primates seem able to coordinate their actions (simultaneously pulling on bars to obtain food) when they are in visual contact (45).
9. See (46). In our experiment, the average number of steps required to find the game-theoretic solution for all 40 dominance-solvable games is 3.675.
17. In coordination games, the participant has to encode and hold this information as well. However, because the targets of both players are the same, the demand on this capacity should be smaller.
39. See (47). The NCI can be interpreted as the probability that two randomly chosen individuals make the same choice relative to the probability of successful coordination if all choose randomly (SOM, S2.3).
44. M. Bhatt, C. Camerer, Games Econ. Behav. 52, 424 (2005).
48. We thank M. Hsu for helpful comments on the manuscript and J.-Y. Leu, J.-Y. Wang, D. Niddam, and participants at many seminars for discussions. Technical assistance from C.-R. Chou, C.-T. Chen, C.-H. Lan, S.-C. Lin, K.-L. Chen, Y.-Y. Chung, W.-Y. Lin, S. Hsu, R. Chen, and the National Taiwan University Hospital MRI Laboratory is greatly appreciated. This work was supported by the National Science Council of Taiwan (grant NSC 94-2415-H-002-004).

Supporting Online Material

www.sciencemag.org/cgi/content/full/324/5926/519/DC1 Materials and Methods

Figs. S1 to S9 Tables S1 to S18 References

www.sciencemag.org on July 24, 2017 Downloaded from
first appeared ~60 million years ago (1). Cattle represent the Ruminantia, which occupy diverse terrestrial environments with their ability to efficiently convert low-quality forage into energy-dense fat, muscle, and milk. These biological processes have been exploited by humans since domestication, which began in the Near East some 8000 to 10,000 years ago (2). Since then, over 800 cattle breeds have been established, representing an important world heritage and a scientific resource for understanding the genetics of complex traits.

The cattle genome was assembled with methods similar to those used for the rat and sea urchin genomes (3, 4). The most recent assemblies, Btau3.1 and Btau4.0, combined bacterial artificial chromosome (BAC) and whole-genome shotgun (WGS) sequences. Btau3.1 was used for gene-specific analyses. Btau4.0, which includes finished sequence data and used different mapping methods to place the sequence on chromosomes, was used for all global analyses other than gene prediction. The contig N50 (50% of the genome is in contigs of this size or greater) is 48.7 kb for both assemblies; the scaffold N50 for Btau4.0 is 1.9 Mb. In the Btau4.0 assembly, 90% of the total genome sequence was placed on the 29 autosomes and X chromosome and validated (3). Of 1.04 million expressed sequence tag (EST) sequences, 95.0% were contained in the assembled contigs. With an equivalent gene distribution in the remaining 5% of the genome, the estimated genome size is 2.87 Gbp. Comparison with 73 finished BACs and single-nucleotide polymorphism (SNP) linkage data (5, 6) confirmed this assembly quality with greater than 92% genomic coverage, and fewer than 0.8% of SNPs were incorrectly positioned at the resolution of these maps (3, 4).

We used the cattle genome to catalog protein-coding genes, microRNA (miRNA) genes, and ruminant-specific interspersed repeats, and we manually annotated over 4000 genes. The consensus protein-coding gene set for Btau3.1 (OGSv1), from six predicted gene sets (4), consists of 26,835 genes with a validation rate of 82% (4). On this basis, we estimate that the cattle genome contains at least 22,000 protein-coding genes. We identified 496 miRNA genes of which 135 were unannotated miRNAs (4). About half of the cattle miRNA occur in 60 genomic miRNA clusters, containing two to seven miRNA genes separated by less than 10 kbp (fig. S2). The overall GC content of the cattle genome is 41.7%, with an observed-to-expected CpG ratio of 0.234, similar to that of other mammals.

The cattle genome has transposable element classes like those of other mammals, as well as large numbers of ruminant-specific repeats (table S4) that compose 27% of its genome. The

---

**Fig. 1.** Protein orthology comparison among genomes of cattle, dog, human, mouse, and rat (*Bos taurus*, *Canis familiaris*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, representing placental mammals), opossum (*Monodelphis domestica*, marsupial), and platypus (*Ornithorhynchus anatinus*, monotreme). (A) The majority of mammalian genes are orthologous, with more than half preserved as single copies (dark blue); a few thousand have species-specific duplications (blue); another few thousand have been lost in specific lineages (orange). We also show those lacking confident orthology assignment (green), and those that are apparently lineage specific [unique (white)]. Placental-specific orthologs are shown in pink. Single- or multiple-copy genes were defined on the basis of representatives in human, bovine, or dog; mouse or rat; and opossum or platypus. (B) Venn diagram showing shared orthologous groups (duplicated genes were counted as one) between laurasiatherians (cattle and dog), human, rodents (mouse and rat), and nonplacental mammals (opossum and platypus) on the basis of the presence of a representative gene in at least one of the grouped species [as in (A)]. (C) Distribution of ortholog protein identities between human and the other species for a subset of strictly conserved single-copy orthologs. (D) A maximum likelihood phylogenetic tree using all single-copy orthologs supports the accepted phylogeny and quantifies the relative rates of molecular evolution expressed as the branch lengths.
consensus sequence of Bov-B, a long interspersed nuclear element (LINE) lacked a functional open reading frame (ORF), which suggested that it was inactive (7). However, Bov-B repeats with intact ORF were identified in the genome, and their phylogeny (fig. S4) indicates that some are still actively expanding and evolving. Mapping chromosomal segments of high- and low-density ancient repeat content, L2/MIR [a LINE/SINE (short interspersed nuclear element) pair] and Bov-B, and more recent repeats, Bov-B/ART2A (Bov-B–derived SINE pair), revealed that the genome consists of ancient regions enriched for L2/MIR and recent regions enriched for Bov-B/ART2A (fig. S7). Exclusion of Bov-B/ART2A from contiguous blocks of ancient repeats suggests that evolution of the ruminant or cattle genome experienced invasions of new repeats into regions lacking ancient repeats. Alternatively, older repeats may have been destroyed by insertion of ruminant- or cattle-specific repeats. AGC trinucleotide repeats, the most common simple-sequence repeat (SSR) in artiodactyls (which include cattle, pigs, and sheep), are 90- and 142-fold overrepresented in cattle compared with human and dog, respectively (fig. S10). Of the AGC repeats in the cattle genome, 39% were associated with Bov-A2 SINE elements.

A comparative analysis examined the rate of protein evolution and the conservation of gene repertoires among orthologs in the genomes of dog, human, mouse, and rat (representing placentatal mammals); opossum (marsupial); and platypus (monotreme). Orthology was resolved for >75% of cattle and >80% of human genes (Fig. 1A). There were 14,345 orthologous groups with representatives in human, cattle, or dog; mouse or rat; and opossum or platypus, which represent 16,749 cattle and 16,177 human genes, respectively, of which 12,592 are single-copy orthologs. We also identified 1217 placental mammal–specific orthologous groups with genes present in human, cattle, or dog; mouse or rat; but not opossum or platypus. About 1000 orthologs shared between rodents and laurasiatherians (cattle and dog), many of which encode G protein–coupled receptors, appear to have been lost or may be misannotated in the human genome (Fig. 1B). Gene repertoire conservation among these mammals correlates with conservation at the amino acid–sequence level (Fig. 1C). The elevated rate of evolution in rodents relative to other mammals (8) was supported by the higher amino acid sequence identity between human and dog or cattle proteins relative to that between human and rodent proteins. However, maximum-likelihood analysis of amino acid substitutions in single-copy orthologs supports the accepted sister lineage relation of primates and rodents (1) (Fig. 1D).

Alternative splicing is a major mechanism for transcript diversification (9), yet the extent of its evolutionary conservation and functional impact remain unclear. We used the cattle genome to analyze the conservation of the most common form of alternative splicing, exon skipping, defined as a triplet of exons in which the middle exon is absent in some transcripts, in a set of 1930 exon-skipping events across human, mouse, dog, and cattle (4). We examined 277 cases, with different conservation patterns between human and mouse, in 16 different cattle tissues with reverse transcription polymerase chain reaction (4). These splicing events were divided into a shared set (163 in both human and mouse) and a nonshared set (114 in human but not in mouse). Of the 277, we detected exon-skipping for 188 cases in cattle (table S5), which suggested that the majority of genes with exon-skipping in human were present and regulated in cattle and that, if an event is shared between human and mouse, it was more likely to be found in cattle. It was estimated that at most 40% of exon-skipping is conserved among mammals; thus, our data agree with the upper bound from previous analyses with human and rodents [e.g., (10)].

We constructed a cattle-human Oxford grid (fig. S12) (4) to conduct synteny-based chromosomal comparisons, which reinforced that human genome organization is more similar to cattle’s than rodents’ because most cattle chromosomes primarily correspond to part of one human chromosome, albeit with multiple rearrangements [e.g., (11)]. In contrast, the cattle-mouse Oxford grid shows poorer chromosomal correspondence. Lineage-specific evolutionary breakpoints were identified for cattle, artiodactyls, and ferungulates (a group encompassing artiodactyls and carnivores, represented by cattle, pig, and dog) and are shown with cattle (fig. S11) and human sequence coordinates (Fig. 2) (4). Primate, dog, rodent, mouse, and rat lineage-specific breakpoint positions were similarly identified. A total of 124 evolutionary breakpoint regions (EBRs) were identified in the cattle lineage, of which 100 were cattle–ruminant-specific and 24 were artiodactyl-specific (e.g., Fig. 2). Nine additional EBRs represent presumptive ferungulate-specific rearrangements. Bos taurus chromosome 16 (BTA16) is populated with four ferungulate-specific EBRs, which suggests that this region was rearranged before the Artiodactyla and Carnivora divergence (Fig. 2). Such conserved regions demonstrate that many inversions that occurred before the divergence of the carnivores and artiodactyls have probably been retained in the ancestral form within the human genome. In contrast to the cattle genome, a pig physical map identified only 77 lineage-specific EBRs. Interchromosomal rearrangements and inversions characterize most of the lineage-specific rearrangements observed in the cattle, dog, and pig genomes.

**Table 1.** Changes in the number of genes in innate immune gene families. Many of the β-defensin genes are present in unassigned scaffolds, i.e., they are not yet part of the current assembly. The exact number of β-defensin genes is uncertain. Interferon subfamily pseudogenes predicted on the basis of frame-shift mutations or stop codons within the first 100 amino acids of the coding sequence have been excluded from the table. The IFNα genes represent a newly discovered subfamily of IFNα and are named for convenience. BPI, Bactericidal and/or permeability-increasing; RNase, ribonuclease; LBP, lipopolysaccharide-binding protein; ULBP, UL16-binding protein.

<table>
<thead>
<tr>
<th>Gene family</th>
<th>Bovine</th>
<th>Human</th>
<th>Murine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cathelicidin</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>RNase</td>
<td>21</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>BPI-like</td>
<td>13</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>BPI/LBP</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>β-Defensin</td>
<td>~106</td>
<td>39</td>
<td>52</td>
</tr>
<tr>
<td>Interferon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subfamilies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFNK</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IFNE</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IFNB</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IFNA</td>
<td>13</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>IFNW</td>
<td>24</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IFNT</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IFNX</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IFNL</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>IFNZ</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>C-type lysozyme</td>
<td>10</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>ULBP</td>
<td>30</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

![Fig. 2. Examples of EBRs. Ferungulate-, artiodactyl-, and primate-specific EBRs on HSA1 at 175 to 247 Mbp (other lineage-specific EBRs not shown). Homologous synteny blocks constructed for the macaque, chimpanzee, cattle, dog, mouse, rat, and pig genomes were used for pairwise comparisons (4). White areas correspond to EBRs. Arrows to the right of the chromosome ideogram indicate positions of representative cattle-specific; artiodactyl-specific (specific to the chromosomes of pigs and cattle); ferungulate-specific (cattle, dog, and pig); primate-specific (human, macaque, and chimpanzee); and hominoid-specific (human and chimpanzee) rearrangements. Opossum is shown as an outgroup to the eutherian clade, which allows classification of ferungulate-specific EBRs.](http://science.sciencemag.org/content/324/5946/524/F1)
An examination of repeat families and individual transposable elements within cattle-, artiodactyl- and ferungulate-specific EBRs showed a significantly higher density of LINE-L1 elements and the ruminant-specific LINE-RT1 repeat family (12) in cattle-specific EBRs relative to the remainder of the genome (table S6). In contrast, the SINE-BovA repeat family and the more ancient tRNA<sup>Glu</sup>-derived SINE repeats (13) were present in lower density in cattle-specific EBRs, similar to other LINEs and SINEs (table S7). The differences in repeat densities were generally consistent in cattle-, artiodactyl- and ferungulate-specific EBRs, with the exception of the tRNA<sup>Glu</sup>-derived and LTR-ERV1 repeats, which are at higher densities in artiodactyl EBRs compared with the rest of the genome.

The tRNA<sup>Glu</sup>-derived SINEs originated in the common ancestor of Suina (pigs and peccaries), Ruminantia, and Cetacea (whales) (13), which suggests that tRNA<sup>Glu</sup>-derived SINEs were involved in ancestral artiodactyl chromosome rearrangements. Furthermore, the lower density of the more ancient repeat families in cattle-specific EBRs suggests that either more recently arising repeat elements were inserted into regions lacking ancient repeats or that older repeats were destroyed by this insertion (table S7). The repeat elements differing in density in EBRs were also found in regions of homologous synteny, which suggests that repeats may promote evolutionary rearrangements (see below). Differences in repeat density in cattle-specific EBRs are thus unlikely to be caused by the accumulation of repeats in EBRs after such rearrangements occur. We identified a cattle-specific EBR associated with a bidirectional promoter (figs. S14 and S15) that may affect control of the expression of the CYB5R4 gene, which has been implicated in human diabetes and, therefore, may be important in the regulation of energy flow in cattle (4).

We identified 1020 segregational duplications (SDs) corresponding to 3.1% (94.4 Mbp) of the cattle genome (4). Duplications assigned to a chromosome showed a bipartite distribution with respect to length and percent identity (fig. S16), and interchromosomal duplications were shorter (median length 2.5 kbp) and more divergent (<94% identity) relative to intrachromosomal duplications (median length 20 kbp, ~97% identity) and tended to be locally clustered (fig. S17). Twenty-one of these duplications were >300 kbp and located in regions enriched for tandem duplications (e.g., BTA18) (fig. S18). This pattern is reminiscent of the duplication pattern of the dog, rat, and mouse but different from that of primate and great-ape genomes (14, 15). On average, cattle SDs >10 kbp represent 11.7% of base pairs in 10-kbp intervals located within cattle-specific EBRs and 23.0% of base pairs located within the artiodactyl-specific EBRs. By contrast, in the remainder of the genome sequence assigned to chromosomes the fraction of SDs was 1.7% (P < 1 × 10<sup>-12</sup>). These data indicate that SDs play a role in promoting chromosome rearrangements by nonallelic homologous recombination [e.g., (16)] and suggest that either a significant fraction of the SDs observed in cattle occurred before the Ruminant-Suina split, and/or that the sites for accumulation of SDs are nonrandomly distributed in artiodactyl genomes.

SDs involving genic regions may give rise to new functional paralogs. Seventy-six percent (778 out of 1020) of the cattle SDs that correspond to complete or partial gene duplications with high sequence identity (median 98.7%). This suggests that many of these gene duplications are specific to either the artiodactyla or the Bos lineage and tend to encode proteins that often interface with the external environment, particularly immune proteins and sensory and/or olfactory receptors. Several of these gene duplications are also duplicated in other mammalian lineages (e.g., cytochrome P-450, sulfotransferase, ribonuclease A, defensins, and pregnancy-associated glycoproteins). Paralogs located in segmental duplications that are present exclusively in cattle may have functional implications for the unique physiology, environment, and diet of cattle.

An overrepresentation of genes involved in reproduction in cattle SDs (tables S8 and S9) is associated with several gene families expressed in the ruminant placenta. These families encode the intercellular signaling proteins pregnancy-associated glycoproteins (on BTA29), trophoblast Kunitz domain proteins (on BTA13), and interferon tau (IFNT) (on BTA8). A family gene encoding prolactin-related proteins (on BTA23) was only identified in the assembly-dependent analysis of SDs. These genes regulate ruminant-specific aspects of fetal growth, maternal adaptations to pregnancy, and the coordination of parturition (17, 18). Although type I interferon (IFN) genes are primarily involved in host defense (19), IFN<sub>α</sub> prevents regression of the corpus luteum during early pregnancy, which results in a uterine environment receptive to early conceptus development (20).

Signatures of positive selection (obtained by measurement of their rates of synonymous and nonsynonymous substitutions) identified 71 genes (4), including 10 immune-related genes (i.e., IFN<sub>α</sub>R2, IFNG, CD34, TREM1, TREML1, FCER1A, IL23R, IL24, IL15, and LEAP2). As previously mentioned, immune genes are overrepresented in SDs (table S21). In cattle-specific EBRs relative to mouse include the cluster of β-defensin genes, which encode antimicrobial peptides; the antimicrobial cathelicidin genes [which show increased sequence diversity of the mature cathelicidin peptides (21)]; and changes in the numbers of interferon genes (22) and the number and organization of genes involved in adaptive immune responses in cattle compared with human and mouse (4). This extensive duplication and divergence of genes involved in innate immunity may be because of the substantial load of microorganisms present in the rumen of cattle, which increases the risk of opportunistic infections at mucosal surfaces and positive selection for the traits that enabled stronger and more diversified innate immune responses at these locations. Another possibility is that many of these immune genes have been under selection due to the herd structure, which can promote rapid disease transmission. Also, immune function–related duplicated genes have gained nonimmune functions, e.g., IFNT (see above), and the C-class lysozyme genes, which are involved in microbial degradation in the abomasum (see below).

There has been substantial reorganization of gene families encoding proteins present in milk. One such rearrangement affecting milk composition involves the histatatin (HSTN) gene within the casein gene cluster on BTA6 (fig. S21). In the cattle genome, HSTN is juxtaposed to a regulatory element (BCE) important (23) for β-casein (CSN2) expression, and as a probable consequence, HSTN is regulated like the casein genes during the lactation cycle. This rearrangement that led to the juxtaposition of HSTN next to the BCE is also the probable cause of deletion of one of the two copies of αs2–like casein genes (CSN1S2A) present in other mammalian genomes (24). The biological implications of this change in casein gene copy number are not yet clear.

Additionally, the cattle serum amyloid A (SAA4) gene cluster arose from both a luarasianther SD and a cattle-specific EBR, which resulted in two mammatory gland–expressed SAA4-like genes, SAA4.1 and SAA4.3 on BTA29, and an SAA3-like gene on BTA15 (fig. S21). SAA3.2 has been shown to inhibit microbial growth (25). Two additional milk protein genes were associated with SDs: cathelicidin (CATHLL1) and β2-microglobulin (B2M)—part of the neonatal Fc receptor (FcRn) that transfers immunoglobulin IgG across epithelial cells of many tissues including the gut and mammary gland (26, 27). IgG is the predominant immunoglobulin in cow’s milk compared with IgA in human milk (28). Unlike humans, who acquire passive immunity from the mother via placental transfer of immunoglobulins during pregnancy, calves acquire passive immunity by ingestion of IgG in milk (28). B2M is also redistributed in epithelial cells upon calving, and it protects IgG from degradation (26). A genetic variant of B2M has negative effects on passive immune transfer (29). The additional copy of the gene encoding B2M might be associated with the abundance of IgG in cows’ milk and an increased capacity for uptake in the neonatal gut. Considering that the passive transfer of immunity to the calf is one of the important functions of milk, it is striking that lactation–related genes affected by genomic rearrangements often encode immune-related proteins in milk.

Cattle metabolic pathways demonstrated a strong degree of conservation among the comprehensive set of genes involved in core mammalian metabolism (4) and permitted an examination of unique genetic events that may be related to ruminant–specific metabolic adaptations. However, among 1032 genes examined from the human metabolic pathways, five were deleted or extensively diverged in cattle: PLA2G4C (phospholipase A2, group IVC), FAAH2 (fatty acid amide hydrolyase 2), ID2 (isopenentyl-diphosphate delta isomerase 2), GSTT2 (glutathione S-transferase 525

www.sciencemag.org  SCIENCE VOL 324  24 APRIL 2009

Downloaded from http://science.sciencemag.org/ on July 24, 2017
theta 2), and TPMP (thymidine phosphorylase), which may be adaptations that impact on fatty acid metabolism (PLA2G4C and FAAH2); the mevalonate pathway (synthesis of dolichols, vitamers, steroid hormones, and cholesterol) (ID2); detoxification (GSTT2); and pyrimidine metabolism (TPMP). Phylogenetic analysis shows that PLAC2G4C was deleted ~87 to 97 million years ago in the laurasiatherian lineages (fig. S22). It was deleted ~87 to 97 million years ago in the laurasiatherian lineages (fig. S22).

In summary, the biological systems most affected by changes in the number and organization of C-type lysozyme genes (table S14 and fig. S23), and EST evidence (fig. S23) show that six of the seven remaining C-type lysozyme genes are expressed primarily in the intestinal tract, which suggests additional roles for the encoded proteins in nutrient digestion.

References and Notes
5. A. Prasad et al., BMC Genomics 8, 310 (2007).
30. Funded by the National Human Genome Research Institute (NHGRI US4 HG003273); the U.S. Department of Agriculture’s Agricultural Research Service (USDA-ARS) agreement no. 59-7090-3-196 and Cooperative State Research, Education, and Extension Service National Research Initiative (grant no. 2004-35216-14163); the state of Texas; Genome Canada through Genome British Columbia; the Alberta Science and Research Authority; the Commonwealth Scientific and Industrial Research Organization of Australia (CSIRO); Agritech Investments Ltd., Dairy Insight, Inc., and Agfiliagre Ltd., all of New Zealand; the Research Council of Norway; the Kiberg Foundation; and the National, and, South Dakota, Beef Check-off Funds. The master accession for this WGS sequencing project is AACF0000000. The individual WGS sequences are AACF03000000 to AACF3311728, and the scaffold records are CM000177 to CM000206 (chromosomal) and DS456032 to DS459580 (unplaced scaffolds).

The Bovine Genome Sequencing and Analysis Consortium
Principal Investigator: Richard A. Gibbs
Analysis project leadership: Christine G. Elo, Ross L Tellam
Sequencing project leadership: Richard A. Gibbs, Donna M. Muzy, George M. Weinstock
Gene prediction and consensus gene set: Christine G. Elo, Ross L Tellam
Protein ortholog analysis: Christine G. Elo, Ross L Tellam, and John J. Glanzer
MicroRNA analysis: Alexandre Reymond, Catherine Ucla, Carine Wyss
Sequence production white paper: Richard A. Gibbs, Steve M. Kappes, Lawrence Schook, Loren C. Skow, George M. Weinstock, and James Womack
MicroRNA analysis: Juan A. Moya, Daniel Gerta, and Evgeny M. Zhlobin
GC composition analysis: Eran Elhaik,7,28 Christine G. Elo,2,7 (leader), Dan Graur,7,29,30,31 Justin T. Reese2
Repeat analysis: David L. Adelson,7,28 Richard B. Edgar,7 John C. McWain,7,28 Gemma M. Payne,32 Joy M. Raison3
Protein orthology analysis: Thomas Junier,20,21,22 Evgenia V. Kriventseva,20,21,22 Evgeny M. Zhlobin,20,21,22 (leader)
Exon-skipping analysis: Jacqueline Chrest,17 Edwardlys Yrases,13,34 Charlotte N. Henrichsen,17 Miryeva Pliss,17,34 Alexander Reynolds2 (leader)
Evolutionary breakpoint analysis and Oxford Grid: Rakvinder Donithu,35,36 Denis M. Larkin,35,36 Harris A. Lewin,35,36 (leader), Frank W. Nicholas,35,36 (leader)
Phylogenetic analysis: Laura Elkins,35,36 Harris A. Lewin,35,36 James Reecy,36 Mary Q. Yang3

Microarray analysis: Juan A. Moya,35,36 Daniel Gerta,35,36 and Evgeny M. Zhlobin

Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds

The Bovine HapMap Consortium*

The emergence of modern civilization was accompanied by adaptation, assimilation, and interbreeding of captive animals. In cattle (Bos taurus), this resulted in the development of individual breeds differing in, for example, milk yield, meat quality, draft ability, and tolerance or resistance to disease and pests. However, despite mapping and diversity studies (1–5) and the identification of mutations affecting some quantitative phenotypes (6–8), the detailed genetic structure and history of cattle are not known.

Cattle occur as two major geographic types, the taurine (humpless—European, African, and Asian) and indicine (humped—South Asian, and East African), which diverged >250 thousand years ago (Kya) (3). We sampled individuals representing 14 taurine (n = 376), three indicine (n = 73) (table S1), and two hybrid breeds (n = 48), as well as two individuals each of Bubalus quarlesi and Bubalus bubalis, which diverged from Bos taurus ~1.25 to 2.0 Mya (9, 10). All breeds except Red Angus (n = 12) were represented by at least 24 individuals. We preferred individuals that were unrelated for ≥2 generations; however, each breed had one or two sire, dam, and progeny trios to allow assessment of genotype quality.

Single-nucleotide polymorphisms (SNPs) that were polymorphic in many populations were primarily derived by comparing whole-genome sequence reads representing five taurine and one indicine breed to the reference genome assembly obtained from a Hereford cow (10) (table S2). This led to the ascertainment of SNPs with high minor allele frequencies (MAFs) within the discovery breeds (table S5). Thus, as expected, with trio progeny removed, SNPs discovered within the taurine breeds had higher average MAFs lines. NDA, N’Dama; SHK, Sheko; NEL, Nelore; BRM, Brahman; GJR, Gir; SGT, Santa Gertrudis; BMA, Beefmaster; ANG, Angus; RUG, Red Angus; HFD, Hereford; NRC, Norwegian Red; HOL, Holstein; LMS, Limousin; CHL, Charolais; BSW, Brown Swiss; JER, Jersey; GNS, Guernsey; PMT, Piedmontese; RMG, Romagnola. (B) Principal components PC1 and PC2 from all SNPs. Taurine breeds remain separated from indicine breeds, and admixed breeds are intermediate.

---

*The full list of authors with their contributions and affiliations is included at the end of the manuscript.
Several thousand years ago, human beings realized the virtues of domesticating wild animals as easy meat. Soon other possibilities became apparent, and as revealed in a series of papers in this issue, early pastoralists became selective about breeding for wool, leather, milk, and muscle power. In two papers, Gibbs et al. report on the bovine genome sequence (p. 522; see the cover, the Perspective by Lewin, and the Policy Forum by Roberts) and trace the diversity and genetic history of cattle (p. 528), while Chessa et al. (p. 532) survey the occurrence of endogenous retroviruses in sheep and map their distribution to historical waves of human selection and dispersal across Europe. Finally, Ludwig et al. (p. 485) note the origins of variation in the coat-color of horses and suggest that it is most likely to have been selected for by humans in need of good-looking transport.