

second model, the two main conditions were parametrically modulated by the two categories, respectively (SOM, S5.1). The activation of the precuneus was higher for hard dominance-solvable games than for easy ones (Fig. 4A and table S10). The activation of the insula was higher for the highly focal coordination games than for less focal ones (Fig. 4B and table S11). Previous studies also found that precuneus activity increased when the number of planned moves increased (40, 41). The higher demand for memory-related imagery and memory retrieval may explain the greater precuneus activation in hard dominance-solvable games. In highly focal coordination games, the participants may have felt quite strongly that the pool students must notice the same salient feature. This may explain why insula activation correlates with NCI.

Participants might have disagreed about which games were difficult. We built a third model to investigate whether the frontoparietal activation correlates with how hard a dominance-solvable game is and whether the activation in insula and ACC correlates with how easy a coordination game is. Here, the two main conditions were parametrically modulated by each participant's probability of obtaining a reward in each game (SOM, S2.2 and S5.2). We found a negative correlation between the activation of the precuneus and the participant's probability of obtaining a reward in dominance-solvable games (Fig. 4C and table S12), which suggests that dominance-solvable games that yielded lower payoffs presented harder mental challenges. In a previous study on working memory, precuneus activity positively correlated with response times, a measure of mental effort (24). Both findings are consistent with the interpretation that subjective measures reflecting harder tasks (higher efforts) correlate with activation in precuneus. A positive correlation between insula activation and the participant's probability of obtaining a reward again suggests that coordination games with a highly salient feature strongly activated the "gut feeling" reported by many participants (Fig. 4D and table S13). A previous study found that the subjective rating of "chills intensity" in music correlates with activation of insula (42). Both findings are consistent with the interpretation that the subjective intensity of how salient a stimulus is correlates with activation in insula.

As mentioned, choices were made significantly faster in coordination games than in dominance-solvable games. The results of the second and third models provide additional support for the idea that intuitive and deliberative mental processes have quite different properties. The "slow and effortful" process was more heavily taxed when the dominance-solvable games were harder. The "fast and effortless" process was more strongly activated when coordination was easy.

References and Notes

1. J. Schaeffer *et al.*, *Science* **317**, 1518 (2007).
2. Previous fMRI 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).
4. J. Mehta, C. Starmer, R. Sugden, *Am. Econ. Rev.* **84**, 658 (1994).
5. T. Schelling, *J. Conflict Resolution* **2**, 203 (1958), p. 211.
6. D. Kahneman, *Am. Psychol.* **58**, 697 (2003).
7. K. Stanovich, R. West, *Behav. Brain Sci.* **23**, 645 (2000).
8. A. Rubinstein, *Econ. J.* **117**, 1243 (2007).
9. See (46). In our experiment, the average number of steps required to find out the game-theoretic solution for all 40 dominance-solvable games is 3.675.
10. R. Jung, R. Haier, *Behav. Brain Sci.* **30**, 135 (2007).
11. V. Goel, R. Dolan, *Neuropsychologia* **39**, 901 (2001).
12. I. Noveck, V. Goel, K. Smith, *Cortex* **40**, 613 (2004).
13. M. Atherton *et al.*, *Brain Res. Cogn. Brain Res.* **16**, 26 (2003).
14. P. Kyllonen, R. Christal, *Intelligence* **14**, 389 (1990).
15. M. D'Esposito, *Philos. Trans. R. Soc. London Ser. B* **362**, 761 (2007).
16. A. Baddeley, *Nat. Rev. Neurosci.* **4**, 829 (2003).
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.
18. E. Smith, J. Jonides, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12061 (1998).
19. N. Müller, R. Knight, *Neuroscience* **139**, 51 (2006).
20. E. Smith, J. Jonides, *Science* **283**, 1657 (1999).
21. T. Wager, E. Smith, *Cogn. Affect. Behav. Neurosci.* **3**, 255 (2003).
22. M. Berryhill, I. Olson, *Neuropsychologia* **46**, 1775 (2008).
23. A. Cavanna, M. Trimble, *Brain* **129**, 564 (2006).
24. M. Wallentin, A. Roepstorff, R. Glover, N. Burgess, *Neuroimage* **32**, 1850 (2006).
25. M. Wallentin, E. Weed, L. Østergaard, K. Mouridsen, A. Roepstorff, *Hum. Brain Mapp.* **29**, 524 (2008).
26. A. D. Craig, *Nat. Rev. Neurosci.* **3**, 655 (2002).
27. A. MacDonald III, J. Cohen, A. Stenger, C. Carter, *Science* **288**, 1835 (2000).
28. J. Decety *et al.*, *Neuroimage* **23**, 744 (2004).
29. J. S. Winston *et al.*, *Nat. Neurosci.* **5**, 277 (2002).
30. T. Singer *et al.*, *Science* **303**, 1157 (2004).
31. A. Bartels, S. Zeki, *Neuroreport* **11**, 3829 (2000).
32. J. Woodward, J. Allman, *J. Physiol. (Paris)* **101**, 179 (2007).
33. A. D. Craig, *Nat. Rev. Neurosci.* **10**, 59 (2009).

34. W. Seeley *et al.*, *J. Neurosci.* **27**, 2349 (2007).
35. J. Downar, A. Crawley, D. Mikulis, K. Davis, *Nat. Neurosci.* **3**, 277 (2000).
36. J. Downar, A. Crawley, D. Mikulis, K. Davis, *J. Neurophysiol.* **87**, 615 (2002).
37. K. Davis *et al.*, *J. Neurosci.* **25**, 8402 (2005).
38. K. Taylor, D. Seminowicz, K. Davis, *Hum. Brain Mapp.*, in press; published online 15 December 2008; 10.1002/hbm.20705.
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, S2.5).
40. S. Newman, P. Carpenter, S. Varma, M. Just, *Neuropsychologia* **41**, 1668 (2003).
41. J. Fincham *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3346 (2002).
42. A. Blood, R. Zatorre, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 11818 (2001).
43. H. Gallagher, A. Jack, A. Roepstorff, C. Frith, *Neuroimage* **16**, 814 (2002).
44. M. Bhatt, C. Camerer, *Games Econ. Behav.* **52**, 424 (2005).
45. K. Mendres, F. de Waal, *Anim. Behav.* **60**, 523 (2000).
46. C. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton Univ. Press, Princeton, NJ, 2003), Chap. 5.
47. N. Bardsley, J. Mehta, C. Starmer, R. Sugden, CeDEX Discussion Paper No. 2008-17 (Centre for Decision Research and Experimental Economics, Nottingham, UK, 2008); available at www.nottinghamnetlearning.com/economics/cedex/papers/2008-17.pdf.
48. We thank M. Hsu for helpful comments on the manuscript and J.-Y. Leu, J.T.-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

8 September 2008; accepted 24 February 2009
10.1126/science.1165598

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

The Bovine Genome Sequencing and Analysis Consortium,* 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.

Domesticated cattle (*Bos taurus* and *Bos taurus indicus*) provide a significant source of nutrition and livelihood to nearly 6.6

billion humans. Cattle belong to a clade phylogenetically distant from humans and rodents, the Cetartiodactyl order of eutherian mammals, which

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 unpublished 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

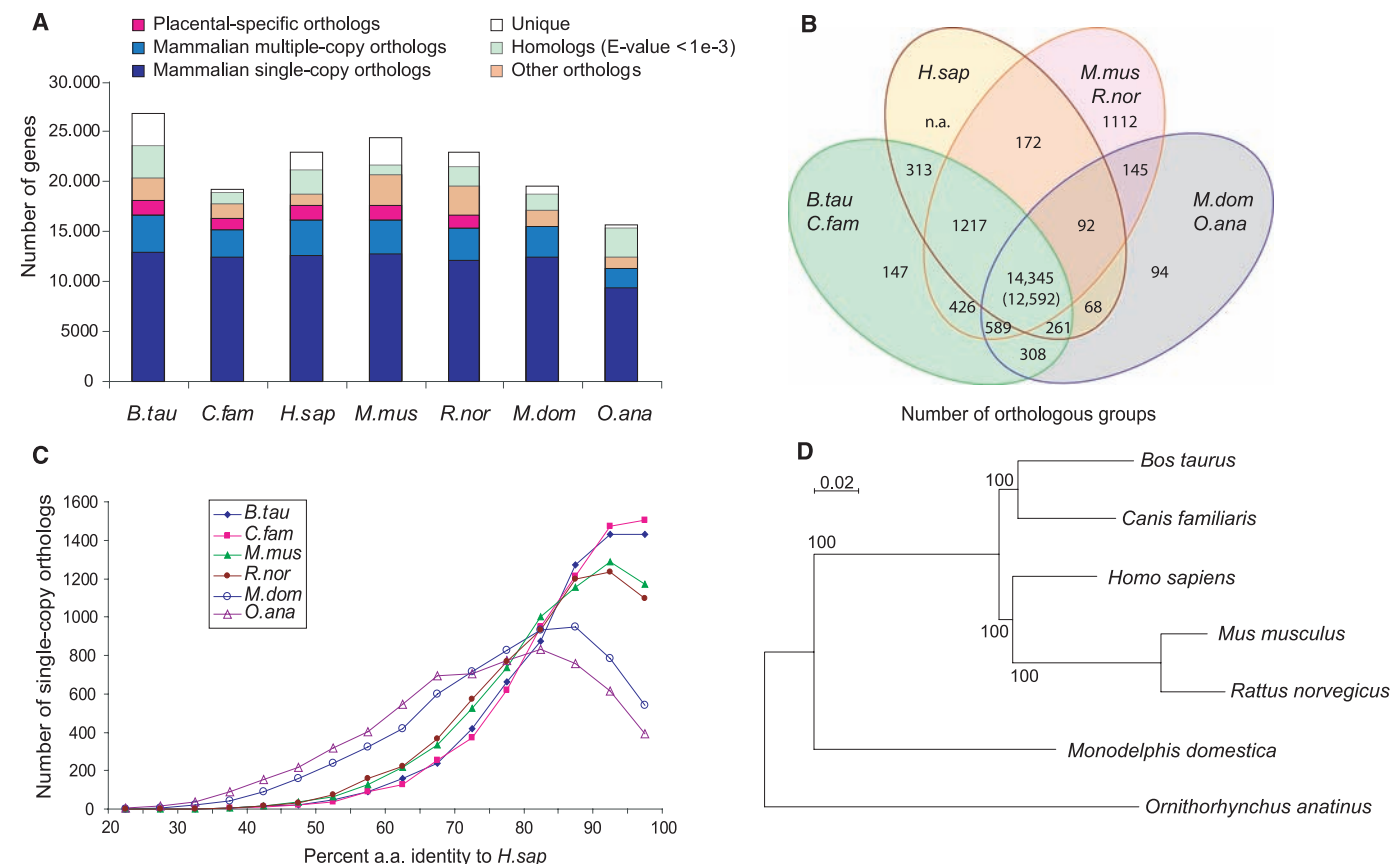


Fig. 1. Protein orthology comparison among genomes of cattle, dog, human, mouse, and rat (*Bos taurus*, *Canis familiaris*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, representing placental mammals), opossum (*Monodelphis domestica*, marsupial), and platypus (*Ornithorhynchus anatinus*, monotreme). (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 placental 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)].

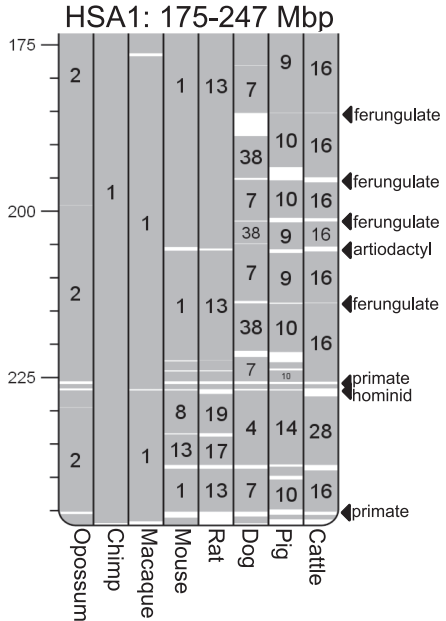


Fig. 2. Examples of EBRs. Ferungulate-, artiodactyl-, and primate-specific EBRs on HSA1 at 175 to 247 Mbp (other lineage-specific EBRs not shown). Homologous synteny blocks constructed for the macaque, chimp, cattle, dog, mouse, rat, and pig genomes were used for pairwise comparisons (4). White areas 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 chimp); and hominoid-specific (human and chimp) rearrangements. Opossum is shown as an outgroup to the eutherian clade, which allows classification of ferungulate-specific EBRs.

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 IFNX genes represent a newly discovered subfamily of IFN and are so named for convenience. BPI, Bactericidal and/or permeability-increasing; RNase, ribonuclease; LBP, lipopolysaccharide-binding protein; ULBP, UL16-binding protein.

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

¹(31).

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- or 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.

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-RTE repeat family (12) in cattle-specific EBRs relative to the remainder of the cattle genome (table S6). In contrast, the SINE-BovA repeat family and the more ancient tRNA^{Glu}-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^{Glu}-derived and LTR-ERV1 repeats, which are at higher densities in artiodactyl EBRs compared with the rest of the genome.

The tRNA^{Glu}-derived SINEs originated in the common ancestor of Suina (pigs and peccaries), Ruminantia, and Cetacea (whales) (13), which suggests that tRNA^{Glu}-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 segmental 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% ($P < 1 \times 10^{-12}$). 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 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 gene family 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), *IFNT* 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., *IFNAR2*, *IFNG*, *CD34*, *TREM1*, *TREML1*, *FCER1A*, *IL23R*, *IL24*, *IL15*, and *LEAP2*). As previously mentioned, immune genes are overrepresented in SDs (see Table 1 and fig. S20). Examples of genes varying in cattle relative to mouse include a 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 immunity may 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 histatherin (*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 α -S2-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 (*SAA*) gene cluster arose from both a laurasiatherian SD and a cattle-specific EBR, which resulted in two mammary gland-expressed *SAA3*-like genes, *SAA3.1* and *SAA3.2* 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 (*CATHL1*) and β_2 -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 hydrolase 2), *ID12* (isopentenyl-diphosphate delta isomerase 2), *GSTT2* (glutathione *S*-transferase

theta 2), and *TYMP* (thymidine phosphorylase), which may be adaptations that impact on fatty acid metabolism (*PLA2G4C* and *FAAH2*); the mevalonate pathway (synthesis of dolichols, vitamins, steroid hormones, and cholesterol) (*IDI2*); detoxification (*GSTT2*); and pyrimidine metabolism (*TYMP*). Phylogenetic analysis shows that *PLA2G4C* was deleted ~87 to 97 million years ago in the laurasiatherian lineages (fig. S22). Strikingly, ~20% of the sequences from two abomasum (last chamber of the cattle stomach) EST libraries (a total of 2392 sequences) correspond to three C-type lysozyme genes. Lysozyme primarily functions in animals as an antibacterial protein, which suggests that they probably function in the abomasum (similar to the monogastric stomach) to degrade the cell walls of bacteria entering from the foregut (30). The cattle genome contains 10 C-type lysozyme genes (table S14 and fig. S23), and EST evidence (fig. S23) shows 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 ruminant digestion.

In summary, the biological systems most affected by changes in the number and organization of genes in the cattle lineage include reproduction, immunity, lactation, and digestion. We highlighted the evolutionary activity associated with chromosomal breakpoint regions and their propensity for promoting gene birth and rearrangement. These changes in the cattle lineage probably reflect metabolic, physiologic, and immune adaptations due to microbial fermentation in the rumen, the herd environment and its influence on disease transmission, and the reproductive strategy of cattle. The cattle genome and associated resources will facilitate the identification of novel functions and regulatory systems of general importance in mammals and may provide an enabling tool for genetic improvement within the beef and dairy industries.

References and Notes

- W. J. Murphy, P. A. Pevzner, S. J. O'Brien, *Trends Genet.* **20**, 631 (2004).
- R. L. Willham, *J. Anim. Sci.* **62**, 1742 (1986).
- Y. Liu et al., *BMC Genomics* **10**, 180 (2009).
- Materials, methods, and additional discussion are available on Science online.
- H. Nilsen et al., *Anim. Genet.* **39**, 97 (2008).
- A. Prasad et al., *BMC Genomics* **8**, 310 (2007).
- H. S. Malik, T. H. Eickbush, *Mol. Biol. Evol.* **15**, 1123 (1998).
- C. I. Wu, W. H. Li, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1741 (1985).
- B. Modrek, C. J. Lee, *Nat. Genet.* **34**, 177 (2003).
- R. Sorek, R. Shamir, G. Ast, *Trends Genet.* **20**, 68 (2004).
- A. Everts-van der Wind et al., *Proc. Natl. Acad. Sci. U.S.A.* **102**, 18526 (2005).
- D. Kordis, F. Gubensek, *Gene* **238**, 171 (1999).
- M. Shimamura, H. Abe, M. Nikaido, K. Ohshima, N. Okada, *Mol. Biol. Evol.* **16**, 1046 (1999).
- J. A. Bailey, E. E. Eichler, *Nat. Rev. Genet.* **7**, 552 (2006).
- J. A. Bailey et al., *Science* **297**, 1003 (2002).
- W. J. Murphy et al., *Science* **309**, 613 (2005).
- K. Hashizume et al., *Reprod. Fertil. Dev.* **19**, 79 (2007).
- J. H. Larson et al., *Physiol. Genomics* **25**, 405 (2006).
- S. Y. Zhang et al., *Immunol. Rev.* **226**, 29 (2008).
- R. M. Roberts, Y. Chen, T. Ezashi, A. M. Walker, *Semin. Cell Dev. Biol.* **19**, 170 (2008).
- M. Scocchi, S. Wang, M. Zanetti, *FEBS Lett.* **417**, 311 (1997).
- M. G. Katze, Y. He, M. Gale Jr., *Nat. Rev. Immunol.* **2**, 675 (2002).
- C. Schmidhauser et al., *Mol. Biol. Cell* **3**, 699 (1992).
- M. Rijinkels, L. Elnitski, W. Miller, J. M. Rosen, *Genomics* **82**, 417 (2003).
- A. J. Molenaar et al., *Biomarkers* **14**, 26 (2009).
- B. Mayer et al., *J. Dairy Res.* **72** (suppl. S1), 107 (2005).
- D. C. Roopenian, S. Akilesh, *Nat. Rev. Immunol.* **7**, 715 (2007).
- T. J. Newby, C. R. Stokes, F. J. Bourne, *Vet. Immunol. Immunopathol.* **3**, 67 (1982).
- M. L. Clawson et al., *Mamm. Genome* **15**, 227 (2004).
- D. M. Irwin, *J. Mol. Evol.* **41**, 299 (1995).
- J. H. Larson et al., *BMC Genomics* **7**, 227 (2006).
- Funded by the National Human Genome Research Institute (NHGRI U54 HG003273); the U.S. Department of Agriculture's Agricultural Research Service (USDA-ARS agreement no. 59-0790-3-196) 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 AgResearch Ltd., all of New Zealand; the Research Council of Norway; the Kleberg Foundation; and the National, Texas, and South Dakota Beef Check-off Funds. The master accession for this WGS sequencing project is AAF03000000. The individual WGS sequences are AAF030000001 to AAF030000002, and the scaffold records are CM000177 to CM000206 (chromosomes) and DS490632 to DS495890 (unplaced scaffolds).

The Bovine Genome Sequencing and Analysis Consortium

Principal Investigator: Richard A. Gibbs¹

Analysis project leadership: Christine G. Elsik,^{2,3} Ross L. Tellam⁴

Sequencing project leadership: Richard A. Gibbs,¹ Donna M. Muzny,¹ George M. Weinstock^{5,1}

Analysis group organization: David L. Adelson,⁶ Evan E. Eichler,^{7,8} Laura Elnitski,⁹ Christine G. Elsik,^{2,3} Roderic Guigó,¹⁰ Debora L. Hamernik,¹¹ Steve M. Kappes,¹² Harris A. Lewin,^{13,14} David J. Lynn,¹⁵ Frank W. Nicholas,¹⁶ Alexandre Reymond,¹⁷ Monique Rijinkels,¹⁸ Loren C. Skow,¹⁹ Ross L. Tellam,⁴ Kim C. Worley,¹ Evgeny M. Zdobnov^{20,21,22}

Sequencing project white paper: Richard A. Gibbs,¹ Steve M. Kappes,¹² Lawrence Schook,¹³ Loren C. Skow,¹⁹ George M. Weinstock,^{5,1} James Womack²³

Gene prediction and consensus gene set: Tyler Alioto,¹⁰ Stylianos E. Antonarakis,²⁰ Alex Astashyn,²⁴ Charles E. Chapple,¹⁰ Hsiu-Chuan Chen,²⁴ Jacqueline Chrast,¹⁷ Francisco Câmara,¹⁰ Christine G. Elsik^{2,3} (leader), Olga Ermolaeva,²⁴ Roderic Guigó,¹⁰ Charlotte N. Henriksen,¹⁷ Wratko Hlavina,²⁴ Yuri Kapustin,²⁴ Boris Kiryutin,²⁴ Paul Kitts,²⁴ Felix Kokocinski,²⁵ Melissa Landrum,²⁴ Donna Maglott,²⁴ Kim Pruitt,²⁴ Alexandre Reymond,¹⁷ Victor Sapojnikov,²⁴ Stephen M. Searle,²⁵ Victor Solovyev,²⁶ Alexandre Souvorov,²⁴ Catherine Ucla,²⁰ George M. Weinstock,^{5,1} Carine Wyss²⁰

Experimental validation of gene set: Tyler Alioto,¹⁰ Stylianos E. Antonarakis,²⁰ Charles E. Chapple,¹⁰ Jacqueline Chrast,¹⁷ Francisco Câmara,¹⁰ Roderic Guigó¹⁰ (leader), Charlotte N. Henriksen,¹⁷ Alexandre Reymond,¹⁷ Catherine Ucla,²⁰ Carine Wyss²⁰

MicroRNA analysis: Juan M. Anzola,³ Daniel Gerlach,^{20,21} Evgeny M. Zdobnov^{20,21,22} (leader)

GC composition analysis: Eran Elhaik,^{27,28} Christine G. Elsik^{2,3} (leader), Dan Graur,²⁷ Justin T. Reese²

Repeat analysis: David L. Adelson⁶ (leader), Robert C. Edgar,²⁹ John C. McEwan,³⁰ Gemma M. Payne,³⁰ Joy M. Raison³¹

Protein ortholog analysis: Thomas Junier,^{19,20} Evgenia V. Kriventseva,³² Evgeny M. Zdobnov^{20,21,22} (leader)

Exon-skipping analysis: Jacqueline Chrast,¹⁷ Eduardo Eyraes,^{33,34} Charlotte N. Henriksen,¹⁷ Mireya Plass,³⁴ Alexandre Reymond¹⁷ (leader)

Evolutionary breakpoint analysis and Oxford grid: Ravikiran Donthu,¹³ Denis M. Larkin,^{13,14} Harris A. Lewin^{13,14} (leader), Frank W. Nicholas¹⁶

Bidirectional promoter analysis: Laura Elnitski⁹ (leader), Denis M. Larkin,^{13,14} Harris A. Lewin,^{13,14} James Reecy,³⁵ Mary Q. Yang⁹

Segmental duplication analysis: David L. Adelson,⁶ Lin Chen,⁷ Ze Cheng,⁷ Carol G. Chitko-McKown,³⁶ Evan E. Eichler^{7,8} (leader), Laura Elnitski,⁹ Christine G. Elsik,^{2,3} George E. Liu,³⁷ Lakshmi K. Matukumalli,^{38,37} Juzhou Song,³⁹ Bin Zhu³⁹

Analysis of gene ontology in segmental duplications: Christine G.

Elsik,^{2,3} David J. Lynn¹⁵ (leader), Justin T. Reese²

Adaptive evolution: Daniel G. Bradley,⁴⁰ Fiona S.L. Brinkman,¹⁵ Lilian P.L. Lau,⁴⁰ David J. Lynn¹⁵ (leader), Matthew D. Whiteside¹⁵

Innate immunity: Ross L. Tellam⁴ (leader), Angela Walker,⁴¹ Thomas T. Wheeler⁴²

Lactation: Theresa Casey,⁴³ J. Bruce German,^{44,45} Danielle G. Lemay,⁴⁵ David J. Lynn,¹⁵ Nauman J. Maqbool,⁴⁶ Adrian J. Molenaar,⁴² Monique Rijinkels¹⁸ (leader)

Metabolism: Harris A. Lewin^{13,14} (leader), Seongwon Seo,⁴⁷ Paul Stothard⁴⁸

Adaptive immunity: Cynthia L. Baldwin,⁴⁹ Rebecca Baxter,⁵⁰ Candice L. Brinkmeyer-Langford,¹⁹ Wendy C. Brown,⁵¹ Christopher P. Childers,² Timothy Connelley,⁵² Shirley A. Ellis,⁵³ Krista Fritz,¹⁹ Elizabeth J. Glass,⁵⁰ Carolyn T.A. Herzog,⁴⁹ Antti Iivanainen,⁵⁴ Kevin K. Lahmers,⁵¹ Loren C. Skow¹⁹ (leader)

Annotation data management: Anna K. Bennett,² Christopher P. Childers,² C. Michael Dickens,³ Christine G. Elsik^{2,3} (leader), James G.R. Gilbert,²⁵ Darren E. Hagen,² Justin T. Reese,² Hanni Salih³

Manual annotation organization: Jan Aerts,⁵⁵ Alexandre R. Caetano,⁵⁶ Brian Dalrymple,⁴ Christine G. Elsik,^{2,3} Jose Fernando Garcia,⁵⁷ Richard A. Gibbs,¹ Clare A. Gill,^{3,58} Debora L. Hamernik,¹¹ Stefan G. Hiendler,⁵⁹ Erdogan Memili,⁶⁰ Frank W. Nicholas,¹⁶ James Reecy,³⁵ Monique Rijinkels,¹⁸ Loren C. Skow,¹⁹ Diane Spurlock,³⁵ Paul Stothard,⁴⁸ Ross L. Tellam,⁴ George M. Weinstock,^{5,1} John L. Williams,⁶¹ Kim C. Worley¹

cDNA tissues, libraries, and sequencing: Lee Alexander,⁶² Michael J. Brownstein,⁶³ Leluo Guan,⁶⁴ Robert A. Holt⁶⁴ (leader), Steven J.M. Jones⁶⁴ (leader), Marco A. Marra⁶⁴ (leader), Richard Moore,⁶⁴ Stephen S. Moore⁴⁸ (leader), Andy Roberts,⁶² Masaaki Taniguchi,^{65,48} Richard C. Waterman⁶²

Genome sequence production: Joseph Chacko,¹ Mimi M. Chandrabose,¹ Andy Cree¹ (leader), Marvin Diep Dao,¹ Huyen H. Dinh¹ (leader), Ramatu Ayesha Gabisi,¹ Sandra Hines,¹ Jennifer Hume¹ (leader), Shalini N. Jhangiani,¹ Vandita Joshi,¹ Christie L. Kovar¹ (leader), Lora R. Lewis,¹ Yih-shin Liu,¹ John Lopez,¹ Margaret B. Morgan,¹ Donna M. Muzny¹ (leader), Ngoc Bich Nguyen,¹ Geoffrey O. Okwuonu,¹ San Juana Ruiz,¹ Jireh Santibanez,¹ Rita A. Wright¹

Sequence finishing: Christian Buhay¹ (leader), Yan Ding,¹ Shannon Dugan-Rocha¹ (leader), Judith Herdandez,¹ Michael Holder,¹ Aniko Sabo¹

Automated BAC assembly: Amy Egan,¹ Jason Goodell,¹ Katarzyna Wilczek-Boney¹

Sequence production informatics: Gerald R. Fowler¹ (leader), Matthew Edward Hitchens,¹ Ryan J. Lozado,¹ Charles Moen,¹ David Steffen,^{66,1} James T. Warren,¹ Jingkun Zhang¹

BAC mapping: Readman Chiu,⁶⁴ Steven J.M. Jones,⁶⁴ Marco A. Marra⁶⁴ (leader), Jacqueline E. Schein⁶⁴

Genome assembly: K. James Durbin,^{67,1} Paul Havlak,^{68,1} Huaiyang Jiang,¹ Yue Liu,¹ Xiang Qin,¹ Yanru Ren,¹ Yufeng Shen,^{1,69} Henry Song,¹ George M. Weinstock,^{5,1} Kim C. Worley¹ (leader)

Sequence library production: Stephanie Nicole Bell,¹ Clay Davis,¹ Angela Jolivet Johnson,¹ Sandra Lee,¹ Lynne V. Nazareth¹ (leader), Bella Mayurkumar Patel,¹ Ling-Ling Pu,¹ Selina Vattathil,¹ Rex Lee Williams Jr.¹

BAC production: Stacey Curry,¹ Cerissa Hamilton,¹ Erica Sodergren^{5,1} (leader)

Sequence variation detection: Lynne V. Nazareth,¹ David A. Wheeler¹

Markers and mapping: David L. Adelson,⁶ Jan Aerts,⁵⁵ Wes Barris,⁴ Gary L. Bennett,³⁶ Brian Dalrymple,⁴ André Eggen,⁷⁰ Clare A. Gill,^{3,58} Ronnie D. Green,⁷¹ Gregory P. Harhay,³⁶ Matthew Hobbs,⁷² Oliver Jann,⁵⁰ Steve M. Kappes¹² (leader), John W. Keele,³⁶ Matthew P. Kent,⁷³ Denis M. Larkin,^{13,14} Harris A. Lewin,^{13,14} Sigbjørn Lien,⁷³ John C. McEwan,³⁰ Stephanie D. McKay,⁷⁴ Sean McWilliam,⁴ Stephen S. Moore,⁴⁸ Frank W. Nicholas,¹⁶ Gemma M. Payne,³⁰ Abhirami Ratnakumar,^{75,4} Hanni Salih,³ Robert D. Schnabel,⁷⁴ Timothy Smith,³⁶ Warren M. Snelling,³⁶ Tad S. Sonstegard,³⁷ Roger T. Stone,³⁶ Yoshikazu Sugimoto,⁷⁶ Akiko Takasuga,⁷⁶ Jeremy F. Taylor,⁷⁴ Ross L. Tellam,⁴ Curtis P. Van Tassel,³⁷ John L. Williams⁶¹

Genomic DNA: Michael D. MacNeil⁶²

Manual annotation: Antonio R.R. Abatepaulo,⁷⁷ Colette A. Abbey,³ Jan Aerts,⁵⁵ Virpi Ahola,⁷⁸ Jassudara G. Almeida,⁵⁷ Ariel F. Amadio,⁷⁹ Elen Anatriello,⁷⁷ Suria M. Bahadue,² Cynthia L. Baldwin,⁴⁹ Rebecca Baxter,⁵⁰ Anna K. Bennett,² Fernando H. Biase,¹³ Clayton R. Boldt,³ Candice L. Brinkmeyer-Langford,¹⁹ Wendy C. Brown,⁵¹ Alexandre R. Caetano,⁵⁶ Jeffery A. Carroll,⁸⁰ Wanessa A. Carvalho,⁷⁷ Theresa Casey,⁴³ Eliane P. Cervellati,⁵⁷ Elsa Chacko,⁸¹ Jennifer E. Chapin³, Ye Cheng,³⁵ Christopher P. Childers,² Jungwoo Choi,³ Adam J. Colley,⁸² Timothy Connelley,⁵² Tatiana A. de Campos,⁵⁶ Marcos De Donato,⁸³

Isabel K.F. de Miranda Santos,^{56,77} Carlo J.F. de Oliveira,⁷⁷ Heather Deobald,⁸⁴ Eve Devinoy,⁸⁵ C. Michael Dickens,³ Kaitlin E. Donohue,² Peter Dovc,⁶⁶ Annett Eberlein,⁸⁷ Shirley A. Ellis,⁵³ Carolyn J. Fitzsimmons,⁵⁹ Alessandra M. Franzin,⁷⁷ Krista Fritz,¹⁹ Gustavo R. Garcia,⁷⁷ Jose Fernando Garcia,⁵⁷ Sem Genini,⁶¹ J. Bruce German,^{44,45} James G.R. Gilbert,²⁵ Clare A. Gill,³⁵ Cody J. Gladney,³ Elizabeth J. Glass,⁵⁰ Jason R. Grant,⁴⁸ Marion L. Greaser,⁸⁸ Jonathan A. Green,⁷⁴ Darryl L. Hadsell,¹⁸ Darren E. Hagen,² Hatam A. Hakimov,⁸⁹ Rob Halgren,⁴³ Jennifer L. Harrow,²⁵ Elizabeth A. Hart,²⁵ Nicola Hastings,^{90,50} Marta Hernandez,⁹¹ Carolyn T.A. Herzig,⁴⁹ Stefan G. Hiendler,⁵⁹ Matthew Hobbs,⁷² Zhi-Liang Hu,⁵³ Antti Iivanainen,⁵⁴ Aaron Ingham,⁴ Terhi Iso-Touru,⁷⁸ Catherine Jamis,⁵¹ Oliver Jann,⁵⁰ Kirsty Jensen,⁵⁰ Dimos Kapetis,⁶¹ Tovah Kerr,⁵¹ Sari S. Khalil,² Hasan Khatib,³⁵ Dawood Kolbehadi,^{48,93} Charu G. Kumar,¹³ Dinesh Kumar,^{94,35} Richard Leach,⁵⁰ Justin C-M Lee,² Danielle G. Lemay,⁴⁵ Changxi Li,^{95,48} George E. Liu,³⁷ Krystin M. Logan,⁹⁶ Roberto Malinverni,⁶¹ Nauman J. Maqbool,⁴⁶ Elisa Marques,⁴⁶ William F. Martin,⁴⁵ Natalia F. Martins,⁵⁶ Sandra R. Maruyama,⁷⁷ Raffaele Mazza,⁷⁷ Kim L. McLean,⁸⁴ Juan F. Medrano,⁹⁸ Erdogan Memili,⁶⁰ Adrian J. Molenaar,⁴² Barbara T. Moreno,⁵⁷ Daniela D. Moré,⁷⁷ Carl T. Muntean,³ Hari P. Nandakumar,¹⁹ Marcelo F.G. Nogueira,⁹⁹ Ingrid Olsaker,¹⁰⁰ Sameer D. Pant,⁶² Francesca Panzitta,⁶¹ Rosemeire C.P. Pastor,⁵⁷ Mario A. Poli,¹⁰¹ Nathan Poslusny,² Satyanarayana Rachagan,³⁵ Shoba Ranganathan,^{81,102} Andrej Razpet,⁸⁶ James Reedy,³⁵ Penny K. Riggs,^{35,58} Monique Rijnkels,¹⁸ Gonzalo Rincon,⁹⁸ Nelida Rodriguez-Osorio,^{60,103} Sandra L. Rodriguez-Zas,¹³ Natasha E. Romero,³ Anne Rosenwald,² Lillian Sando,⁴ Sheila M. Schmutz,⁸⁴ Seongwon Seo,⁴⁷ Libing Shen,² Laura Sherman,⁴⁸ Loren C. Skow,¹⁹ Bruce R. Southey,¹⁰⁴ Diane Spurlock,³⁵ Ylva Strandberg Lutzow,⁴ Jonathan V. Sweedler,¹⁰⁴ Imke Tammen,⁷² Masaaki Taniguchi,^{65,48} Ross L. Tellam,⁴ Bhanu Prakash V.L. Telugu,⁷⁴ Jennifer M. Urbanski,² Yuri T. Utsunomiya,⁵⁷ Chris P. Verschoor,⁸² Ashley J. Waardenberg,^{4,105} Angela Walker,⁴¹ Zhiquan Wang,⁴⁸ Robert Ward,¹⁰⁶ Rosemarie Weikard,⁸⁷ Thomas H. Welsh Jr.,^{35,58} Thomas T. Wheeler,⁴² Stephen N. White,^{51,107} John L. Williams,⁶¹ Laurens G. Wilming,²⁵ Kris R. Wunderlich,³ Jianqi Yang,¹⁰⁸ Feng-Qi Zhao¹⁰⁹

¹Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. ²Department of Biology, 406 Reiss, Georgetown University, 37th & O Streets NW, Washington, DC 20057, USA. ³Department of Animal Science, Texas A&M University, 2471 TAMU, College Station, TX 77843–2471, USA. ⁴Livestock Industries, Commonwealth Scientific and Industrial Research Organization (CSIRO), 306 Carmody Road, St. Lucia, Queensland, 4067, Australia. ⁵The Genome Center at Washington University, Washington University School of Medicine, 4444 Forest Park Avenue, St. Louis, MO 63108, USA. ⁶School of Molecular and Biomedical Science, School of Agriculture, Food and Wine, The University of Adelaide, Adelaide, SA, 5005, Australia. ⁷Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195–5065, USA. ⁸Howard Hughes Medical Institute, Seattle, WA 98195, USA. ⁹National Human Genome Research Institute, National Institutes of Health, 5625 Fishers Lane, Rockville, MD 20878, USA. ¹⁰Center for Genomic Regulation and Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain. ¹¹U.S. Department of Agriculture (USDA), Cooperative State Research, Education, & Extension Service, 1400 Independence Avenue SW, Stop 2220, Washington, DC 20250–2220, USA. ¹²National Program Staff, USDA–Agricultural Research Service, 5601 Sunnyside Avenue, Beltsville, MD 20705, USA. ¹³Department of Animal Sciences, University of Illinois at Urbana–Champaign, 1201 West Gregory Drive, Urbana, IL 61801, USA. ¹⁴Institute for Genomic Biology, University of Illinois at Urbana–Champaign, 1201 West Gregory Drive, Urbana, IL 61801, USA. ¹⁵Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada. ¹⁶Faculty of Veterinary Science, University of Sydney, Sydney, NSW, 2006, Australia. ¹⁷Center for Integrative Genomics, University of Lausanne, Lausanne, 1015, Switzerland. ¹⁸Children's Nutrition Research Center, USDA–Agricultural Research Service, Department of Pediatrics–Nutrition, Baylor College of Medicine, 1100 Bates Street, Houston, TX 77030–2600, USA. ¹⁹Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA. ²⁰Department of Genetic Medicine and Development, University of Geneva Medical School, 1 rue Michel-Servet,

Geneva, 1211, Switzerland. ²¹Swiss Institute of Bioinformatics, 1 rue Michel-Servet, Geneva, 1211, Switzerland. ²²Division of Molecular Biosciences, Imperial College London, South Kensington Campus, London, SW7 2AZ, UK. ²³Department of Veterinary Pathobiology, Texas A&M University, College Station, TX 77843, USA. ²⁴National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892, USA. ²⁵Informatics Department, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK. ²⁶Department of Computer Science, University of London, Royal Holloway, Egham, Surrey, TW20 0EX, UK. ²⁷Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA. ²⁸McKusick—Nathans Institute of Genetic Medicine, BRB 579, Johns Hopkins University School of Medicine, 733 North Broadway, Baltimore, MD 21205, USA. ²⁹45 Monterey Drive, Tiburon, CA 94920, USA. ³⁰Animal Genomics, AgResearch, Invermay, PB 50034, Mosgiel, 9053, New Zealand. ³¹eResearch SA, University of Adelaide, North Terrace, Adelaide, SA, 5005, Australia. ³²Department of Structural Biology and Bioinformatics, University of Geneva Medical School, 1 rue Michel-Servet, Geneva, 1211, Switzerland. ³³Catalan Institution for Research and Advanced Studies, 08010 Barcelona, Catalonia, Spain. ³⁴Computational Genomics, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain. ³⁵Department of Animal Science, Iowa State University, 2255 Kildee Hall, Ames, IA 50011–3150, USA. ³⁶Meat Animal Research Center, USDA–Agricultural Research Service, Clay Center, NE 68933, USA. ³⁷Bovine Functional Genomics Laboratory, USDA–Agricultural Research Service, Beltsville Agricultural Research Center (BARC)—East, Beltsville, MD 20705, USA. ³⁸Department of Bioinformatics and Computational Biology, George Mason University, 10900 University Blvd, Manassas, VA 20110, USA. ³⁹Department of Bioengineering, University of Maryland, College Park, MD 20742, USA. ⁴⁰Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland. ⁴¹Department of Veterinary Pathobiology, 245 Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA. ⁴²Dairy Science and Technology Section, AgResearch, Ruakura Research Centre, East Street, Private Bag 3123, Hamilton, 3240, New Zealand. ⁴³Department of Animal Science, Michigan State University, East Lansing, MI 48824–1225, USA. ⁴⁴Nestlé Research Centre, Vers chez les Blanc CH, Lausanne 26, 1000, Switzerland. ⁴⁵Department of Food Science and Technology, University of California–Davis, Davis, CA 95616, USA. ⁴⁶Bioinformatics, Mathematics and Statistics, AgResearch, Ruakura Research Centre, East Street, Private Bag 3123, Hamilton, 3240, New Zealand. ⁴⁷Division of Animal Science and Resource, Chungnam National University, Daejeon, 305-764, Korea. ⁴⁸Department of Agricultural, Food and Nutritional Science, University of Alberta, 410 AgFor Centre, Edmonton, AL, T6G 2P5, Canada. ⁴⁹Department of Veterinary and Animal Sciences, University of Massachusetts, Amherst, MA 01003, USA. ⁵⁰The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, Midlothian, EH25 9PS, UK. ⁵¹Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, WA 99164, USA. ⁵²Division of Infection and Immunity, The Roslin Institute, Royal (Dick) School of Veterinary Science, University of Edinburgh, Roslin, Midlothian, EH25 9RG, UK. ⁵³Immunology Division, Institute for Animal Health, Compton, RG20 7NN, UK. ⁵⁴Department of Basic Veterinary Sciences, University of Helsinki, Post Office Box 66, Helsinki, FIN-00014, Finland. ⁵⁵Genome Dynamics and Evolution, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK. ⁵⁶Embrapa Recursos Genéticos e Biotecnologia, Final Avenida W5 Norte, Brasília, DF, 70770-900, Brazil. ⁵⁷Animal Production and Health Department, UNESP—São Paulo State University, Aracatuba, SP, 16050-680, Brazil. ⁵⁸Texas AgriLife Research, College Station, TX 77843, USA. ⁵⁹JS Davies Epigenetics and Genetics Group, School of Agriculture, Food & Wine and Research Centre for Reproductive Health, The University of Adelaide, Roseworthy Campus, Roseworthy, SA, 5371, Australia. ⁶⁰Department of Animal and Dairy Sciences, Mississippi Agricultural and Forestry Experiment Station, Mississippi State University, Mississippi State, MS 39762, USA. ⁶¹Parco Tecnologico Padano, Via Einstein, Polo Universitario, Lodi, 26900, Italy. ⁶²Fort Keogh Livestock and Range Research Laboratory, USDA–Agricultural Research Service, Miles City, MT 59301, USA. ⁶³Laboratory of Genetics, National Institute of Mental Health, NIH, Building 49, B1EE16, 49 Convent Drive, Bethesda, MD

20892, USA. ⁶⁴Genome Sciences Centre, British Columbia Cancer Agency, 675 West 10th Avenue, Vancouver, BC, V5Z 1L3, Canada. ⁶⁵Division of Animal Sciences, National Institute of Agrobiological Sciences, Tsukuba, Ibaraki, 305-8602, Japan. ⁶⁶Bioinformatics Research Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. ⁶⁷Department of Biomolecular Engineering, University of California at Santa Cruz, Santa Cruz, CA 95064, USA. ⁶⁸Department of Computer Science, University of Houston, Houston, TX 77204–3010, USA. ⁶⁹Department of Computer Science and Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA. ⁷⁰INRA, Animal Genetics and Integrative Biology, Bovine Genetics and Genomics, 78350 Jouy-en-Josas, France. ⁷¹Pfizer Animal Genetics, Pfizer Animal Health, New York, NY 10017, USA. ⁷²Faculty of Veterinary Science, University of Sydney, Camden, NSW, 2570, Australia. ⁷³Centre for Integrative Genetics and Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Arbotreivien 6, Ås, 1432, Norway. ⁷⁴Division of Animal Sciences, University of Missouri, 920 East Campus Drive, Columbia, MO 65211, USA. ⁷⁵Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala Biomedical Centre Husargatan 3, Uppsala, 75 123, Sweden. ⁷⁶Shirakawa Institute of Animal Genetics, Nishigo, Fukushima 961-8061, Japan. ⁷⁷Department of Biochemistry and Immunology, Ribeirão Preto Medical School, University of São Paulo, Av Bandeirantes 3900, Ribeirão Preto, SP, 14049-900, Brazil. ⁷⁸Biotechnology and Food Research, MTT Agrifood Research Finland, Jokioinen, FI-31600, Finland. ⁷⁹EEA Rafaela, Instituto Nacional de Tecnología Agropecuaria (INTA), Ruta 34 Km 227, Rafaela, Santa Fe, 2300, Argentina. ⁸⁰Livestock Issues Research Unit, USDA–Agricultural Research Service, Lubbock, TX 79403, USA. ⁸¹Department of Chemistry and Biomolecular Sciences & ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, 2109, NSW, Australia. ⁸²Department of Animal and Poultry Science, University of Guelph, Guelph, ON, N1G2W1, Canada. ⁸³Instituto de Investigaciones en Biomedicina y Ciencias Aplicadas, Universidad de Oriente, Avenida Universidad, Cumana, Sucre, 6101, Venezuela. ⁸⁴Department of Animal and Poultry Science, University of Saskatchewan, Saskatoon, SK, S7N 5A8, Canada. ⁸⁵INRA–UR1196, Génétique et Physiologie de la Lactation, F78352 Jouy-en-Josas, France. ⁸⁶Department of Animal Science, University of Ljubljana, Groblje 3, Domžale, SI-1230, Slovenia. ⁸⁷Research Unit Molecular Biology, Research Institute for the Biology of Farm Animals (FBN), Dummerstorf, 18196, Germany. ⁸⁸Department of Animal Sciences, University of Wisconsin–Madison, 1805 Linden Drive, Madison, WI 53706, USA. ⁸⁹Department of Molecular and Cellular Biology, University of Guelph, Guelph, ON, N1G 2W1, Canada. ⁹⁰Cell Biology and Biophysics, European Molecular Biology Laboratory (EMBL)–Heidelberg, Meyerhofstrasse 1, Heidelberg, Germany. ⁹¹Laboratory of Molecular Biology, Instituto Tecnológico Agrario de Castilla y León (ITACyL), Carretera de Burgos km 119, Valladolid, 47071, Spain. ⁹²Department of Dairy Science, University of Wisconsin, Madison, WI 53706, USA. ⁹³Monsanto Company, 3302 SE Convenience Blvd, Ankeny, IA 50021, USA. ⁹⁴Genes & Genetic Resources Molecular Analysis Lab, National Bureau of Animal Genetic Resources, Baldi Bye Pass, Karnal, Haryana, 132001, India. ⁹⁵Lacombe Research Centre, Agriculture and Agri-Food Canada, Lacombe, AL, T4L 1W1, Canada. ⁹⁶Biomedical Sciences, University of Guelph, Guelph, ON, N1G 2W6, Canada. ⁹⁷Zootechnics Institute, Università Cattolica del Sacro Cuore, via Emilia Parmense 84, Piacenza, 29100, Italy. ⁹⁸Department of Animal Science, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA. ⁹⁹Departamento de Ciências Biológicas, Faculdade de Ciências e Letras, UNESP—São Paulo State University, Av Dom Antônio 2100, Vila Tênis Clube, Assis, SP, 19806-900, Brazil. ¹⁰⁰Department of Basic Sciences and Aquatic Medicine, Norwegian School of Veterinary Science, Post Office Box 8146 Dep, Oslo, NO-0033, Norway. ¹⁰¹Instituto de Genética Ewald Favret, Instituto Nacional de Tecnología Agropecuaria (INTA), Las Cabañas y de Los Reseros s/n CC25, Castelar, Buenos Aires, B1712WAA, Argentina. ¹⁰²Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore, 117597, Singapore. ¹⁰³Grupo CENTAURO, Universidad de Antioquia, Medellín, Colombia. ¹⁰⁴Department of Chemistry, University of Illinois, Urbana, IL 61801, USA. ¹⁰⁵Esikit Institute for Cell and Molecular Therapies, Griffith University, Nathan, QLD, 4111, Australia.

¹⁰⁶Nutrition and Food Sciences, Utah State University, Logan, UT 84322, USA. ¹⁰⁷Animal Disease Research Unit, USDA–Agricultural Research Service, Pullman, WA 99164, USA. ¹⁰⁸Department of Pharmacology, 2-344 BSB, University of Iowa, 51 Newton Road, Iowa City, IA 52242, USA. ¹⁰⁹Department of Animal Science, 211 Terrill, University of Vermont, 570 Main Street, Burlington, VT 05405, USA.

Supporting Online Material

www.sciencemag.org/cgi/content/full/324/5926/522/DC1
Materials and Methods

Figs. S1 to S23
Tables S1 to S14
References

10 December 2008; accepted 16 March 2009
10.1126/science.1169588

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

The Bovine HapMap Consortium*

The imprints of domestication and breed development on the genomes of livestock likely differ from those of companion animals. A deep draft sequence assembly of shotgun reads from a single Hereford female and comparative sequences sampled from six additional breeds were used to develop probes to interrogate 37,470 single-nucleotide polymorphisms (SNPs) in 497 cattle from 19 geographically and biologically diverse breeds. These data show that cattle have undergone a rapid recent decrease in effective population size from a very large ancestral population, possibly due to bottlenecks associated with domestication, selection, and breed formation. Domestication and artificial selection appear to have left detectable signatures of selection within the cattle genome, yet the current levels of diversity within breeds are at least as great as exists within humans.

The emergence of modern civilization was accompanied by adaptation, assimilation, and interbreeding of captive animals. In cattle (*Bos taurus*), this resulted in the develop-

ment 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 ≥ 4 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

*The full list of authors with their contributions and affiliations is included at the end of the manuscript.

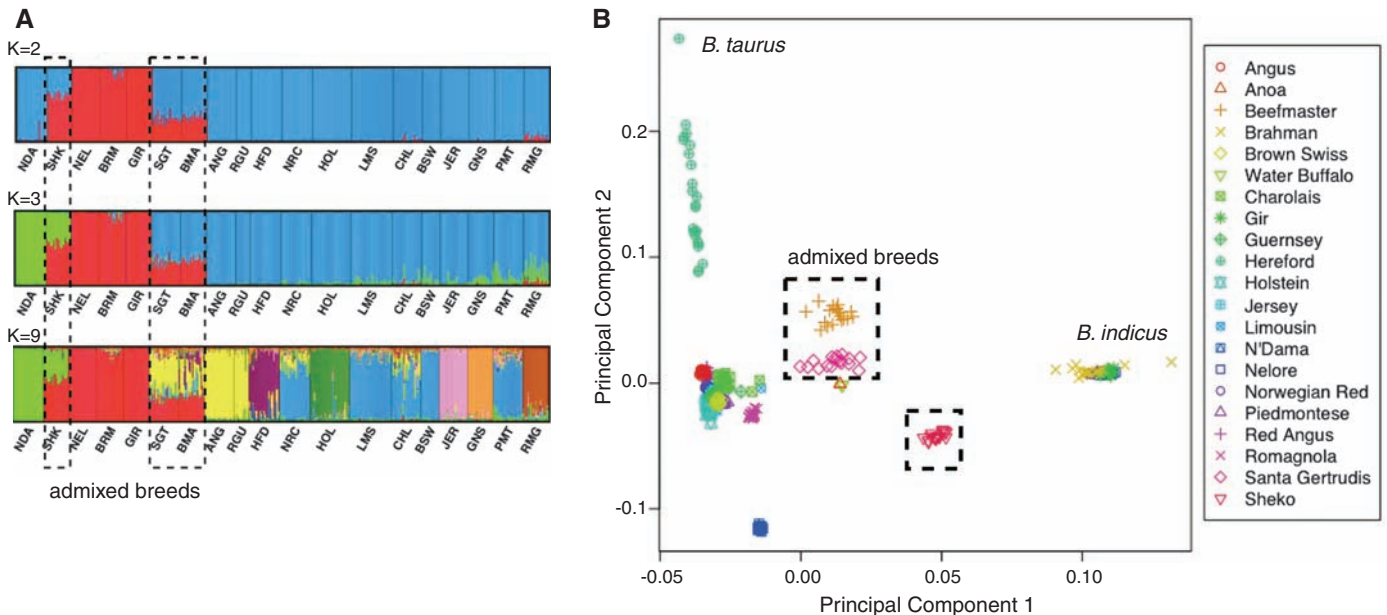


Fig. 1. (A) Population structure assessed by InStruct. Bar plot, generated by DISTRICT, depicts classifications with the highest probability under the model that assumes independent allele frequencies and inbreeding coefficients among assumed clusters. Each individual is represented by a vertical bar, often partitioned into colored segments with the length of each segment representing the proportion of the individual's genome from $K = 2, 3$, or 9 ancestral populations. Breeds are separated by black

lines. NDA, N'Dama; SHK, Sheko; NEL, Nelore; BRM, Brahman; GIR, Gir; SGT, Santa Gertrudis; BMA, Beefmaster; ANG, Angus; RGU, 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.

Originally posted 23 April; corrected 24 April 2009



www.sciencemag.org/cgi/content/full/324/5926/522/DC1

Supporting Online Material for

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

The Bovine Genome Sequencing and Analysis Consortium*

*To whom correspondence should be addressed. E-mail: ce75@georgetown.edu
(C.G.E.), ross.tellam@csiro.au (R.L.T.), and kworley@bcm.edu (K.C.W.)

Published 24 April 2009, *Science* **324**, 522 (2009)
DOI: 10.1126/science.1169588

This PDF file includes

Materials and Methods
SOM Text
Figs. S1 to S23
Tables S1 to S14
References

Correction 24 April 2009: This file has full references for some in press and an additional sponsor. *Science* regrets the original posting in error.

Supporting Online Material

Table of Contents

Supporting Materials and Methods	3
Sequencing the Genome	3
Gene Prediction and Consensus Gene Set	3
NCBI	3
Ensembl	4
Fgenesh and Fgenesh++	4
Geneid and SGP2	4
GLEAN consensus gene set	5
Experimental Validation of Gene Set	5
Generation of cDNAs	6
Selenoproteins	6
U12 Introns	7
MicroRNAs	7
microRNA Prediction Set 1	7
microRNA Prediction Set 2	8
Merging microRNA Set 1 and microRNA Set 2	9
Bovine Official Gene Set	9
Manual Annotation	9
GC Content	10
De Novo Repeat Detection	10
Identification of Intact LINE Elements and Phylogenetic Analysis	10
Correlation of Repeat Elements	11
Simple Sequence Repeats	11
Protein Ortholog Analysis	11
Protein Phylogenetic Analysis	12
Exon Skipping	12
Homologous Synteny Blocks (HSBs) and Evolutionary Breakpoint Regions (EBRs)	12
Oxford Grid	13
Bidirectional Promoters	13
Segmental Duplications	14
Gene Ontology and Pathway Analyses for Genes Present in Segmental Duplications	14
Adaptive Evolution in the Bovine Genome	15
Annotation of genes involved in adaptive immunity	16
Metabolic Reconstruction and Identification of Metabolic Gene Losses in the Cattle Genome	16
Lysozyme Genes	16
Supporting Text	16
Experimental Validation of Gene Set	16
Selenoproteins	17
U12 Introns	17
MicroRNAs	18
GC Content	18
Phylogenetic Analysis of LINE Elements	18

Correlation of Repeat Elements	18
Simple Sequence Repeats	18
Protein Ortholog Analysis.....	19
Exon Skipping.....	19
Evolutionary Breakpoint Analysis.....	20
Oxford Grid.....	20
Bidirectional Promoters	20
Segmental Duplications	21
Gene Ontology and Pathway Analyses for Genes Present in Segmental Duplications.....	21
Adaptive Evolution	21
Adaptive Immunity	22
Metabolic Reconstruction and Gene Losses	23
Lysozyme Genes	24
Supporting Figures.....	25
Supporting Tables	48
Supplementary References and Notes	71

Supporting Materials and Methods

Sequencing the Genome

Leukocyte DNA from a Hereford cow (L1 Dominette 01449, 30% inbreeding coefficient) was used for the small insert whole genome shotgun (WGS) libraries, while the DNA for the BAC library was derived from her sire (L1 Domino 99375, 31% inbreeding coefficient). The sequencing strategy was a hybrid of the whole genome shotgun (WGS) and the hierarchical BAC clone approaches, and assembled with methods similar to the rat (S1) and sea urchin genomes (S2) [see (S3) for a full description]. Assembly used the Atlas genome assembly system (S4), and details are provided in (S3). The most recent assemblies are Btau3.1 and Btau4.0. Btau3.1 combined individual bacterial artificial chromosome (BAC) sequences assembled with overlapping WGS sequence with sequences from a previous whole genome shotgun only assembly. Sequences were placed in Btau3.1 using preliminary physical mapping data [see details in (S3)], and Btau3.1 was used for most gene specific analyses in this manuscript. Subsequently, Btau4.0 was constructed by placing the sequence using different mapping information [see details in (S3)] and adding finished sequence data, though most sequence contigs remain unchanged between Btau3.1 and Btau4.0. Btau4.0 was used for many of the global analyses presented here such as the GC content, repeat, homologous synteny block, and segmental duplication analyses. 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. The major difference in Btau4.0 compared to Btau3.1 was the positioning of sequence scaffolds onto chromosomes with a radiation hybrid physical map consisting of 3,484 markers of which 2,759 were BAC-end sequences anchored to the cattle BAC fingerprint map (S5, S6). The procedures used for mapping and ordering scaffolds on chromosomes are provided in (S3). In the Btau4.0 assembly, 90% of the total genome sequence was placed on the 29 autosomes and X chromosome and validated (S3). Of 1.04 million expressed sequence tag (EST) sequences, 95.0% were contained in the assembled contigs, giving an estimated genome size of 2.87 Gbp. The quality of the assembly was assessed (S3) by alignment to 73 finished BACs and showed that the genomic coverage was between 92.5% and 100.0% (average of 98.5%). Single nucleotide polymorphism (SNP) linkage data (S7, S8) were used to assess the order of scaffolds on three chromosomes and genome-wide (S3). In the latter instance quality was determined by comparison of SNP order on the assembly to independent genetic linkage maps constructed from 17,482 SNPs. Fewer than 0.8% of SNP were incorrectly positioned at the resolution of these maps (S3).

Gene Prediction and Consensus Gene Set

NCBI

The NCBI gene prediction process included cDNA, EST and protein alignments, with Splign and ProSplign (S9, S10). The best scoring coding sequence (CDS) was identified for all cDNA alignments with a 3-periodic fifth-order Markov model for the coding propensity score and walking Markov models for the splice signals and translation initiation and termination signals. These are the same scores used with Gnomon (S11), the NCBI *ab initio* gene prediction tool. All cDNAs with CDS scoring above a threshold were marked as coding cDNAs, and all others were marked as untranslated regions (UTRs). Some of the CDS were incomplete, meaning that they lacked translation initiation or termination signals. All protein alignments were scored the same way, and CDS that did not satisfy the threshold criterion for a valid CDS were removed. After

determining the UTR/CDS nature of each alignment, they were assembled with a modification of the Maximal Transcript Alignment algorithm (S12), taking into account not only exon-intron structure compatibility but also the compatibility of the reading frames. Two coding alignments were connected only if they both had open and compatible CDS. UTRs were connected to coding alignments only if there were appropriate translation initiation or termination signals. There were no restrictions on the connection of UTRs other than the exon-intron structure compatibility. All assemblies with a complete CDS, including the translation initiation and termination signals, were combined into alternatively spliced isoform groups. Incomplete assemblies were directed to Gnomon for extension (S11).

Ensembl

The Ensembl gene predictions were produced with an automated pipeline system with protein and cDNA evidence (S13). Bovine protein sequences were aligned to the genome with the exact alignment program Pmatch to identify their approximate genomic locations. Genewise (S14) or Exonerate (S15) was used next to create coding transcript models. Genewise transcript models were obtained on the basis of other mammal and other species protein sequences were used where no cow-specific data was available. The predictions were refined by filtering out faulty evidence and comparing to orthologous genes to identify fragmentary predictions and missed orthologs. Bovine cDNAs were used to add UTR regions to the structures. Dedicated pipeline modules were applied to identify non-coding RNA genes (ncRNA), flag potential pseudogenes in the predicted gene set and load mitochondrial genes.

Fgenesh and Fgenesh++

Two gene prediction sets were generated at Softberry Inc with Fgenesh and the Fgenesh++ pipeline (S16, S17). Fgenesh is a hidden Markov model based *ab initio* gene prediction program. Fgenesh++ is a pipeline for automatic prediction of genes, which in addition to Fgenesh, includes sequence analysis software to incorporate information from full-length cDNA alignments and similar proteins from the eukaryotic section of the NCBI NR database (S18). Both Fgenesh and Fgenesh++ used bovine-specific gene-finding parameters trained on known genes of organisms closely related to *B. taurus*. Before submitting to Fgenesh++/Fgenesh, sequences were first masked by RepeatMasker (S19) with the -cow option (which masks repeats in non-primate, non-rodent mammals). Low complexity regions and simple repeats were not masked since they can be parts of coding sequences. When using RefSeq for mapping of known mRNAs, we used only mRNAs with the status "provisional", "reviewed" or "validated", but not predicted mRNAs (status "model" or "predicted", respectively). Similarly, we excluded computationally predicted proteins in the Genbank NR database (S18) from our dataset.

Geneid and SGP2

Geneid (S20) and SGP2 (S21) were used to predict genes in the Btau3.1 genome assembly. Geneid is an *ab initio* gene finder designed with a hierarchical structure. In the first step, splice sites and start and stop codons were predicted and scored along the sequence with Position Weight Arrays. In the second step, exons were built from the sites. Exons were scored as the sum of the scores of the defining sites, plus the score derived from a Markov model for coding DNA. Finally, from the set of predicted exons, the gene structure was assembled, maximizing the sum of the scores of the assembled exons. Geneid offers some support to integrate predictions from multiple sources via external gff files. SGP2 is a comparative gene finder that

uses TBLASTX (S22) to identify regions of similarity between the target genome and any number of informant genomes. These regions were then processed and imported into Geneid, where they were used to re-score predicted exons prior to the assembly of the gene structure. The rationale is that predicted genes overlapping regions of conservation would be promoted into the final gene model.

GLEAN consensus gene set

The individual gene prediction sets were integrated with GLEAN (S23). GLEAN is a tool for creating consensus gene lists by integrating gene evidence. It uses Latent Class Analysis to estimate accuracy and error rates for each source of gene evidence, and then uses these estimates to reconstruct the consensus prediction on the basis of patterns of agreement/disagreement observed between each evidence source. GLEAN analysis labels each prediction with a confidence score reflecting the underlying support for that gene. GLEAN was run seven times with different combinations of the following gene prediction lists: NCBI, Ensembl, Fgenesh, Fgenesh++, Geneid and SGP2 as well as aligned proteins and ESTs. The proteins were from metazoan SwissProt (S24) and aligned with Exonerate (S15) with a minimum 60% percent identity and 80% alignment coverage. The ESTs were consensus dbEST (S25) assembled with TGICL (S26) with a minimum 95% identity and 90% alignment coverage. EST consensus sequences were aligned to the genome assembly with Exonerate with 98% identity and 80% alignment coverage. Coding sequences from full-length cDNA were aligned with Splign (S9), with criteria of 99% identity and 100% alignment coverage.

A "gold standard" set was created with 208 full-length coding sequences from cloned cDNAs that were not available at the time the gene lists were constructed to evaluate the seven GLEAN sets and to compare them to input gene prediction sets. These evaluations used FASTA (S27). To evaluate accuracy of intron/exon structure for each gene prediction set, the number of gold standard sequences with perfect matches to a predicted gene model was determined. A perfect match was defined as an alignment in which both sequences were completely aligned with at least 99% identity and no gaps. To evaluate completeness of a gene prediction set, the number gold standard sequences that matched a predicted gene model with at least 99% identity, not considering gaps or alignment coverage, was determined (Table S1). In addition to alignments with the gold standard, each gene prediction set was evaluated for agreement with gold standard splice sites after aligning the gold standard sequences to Btau3.1 with Splign (Table S2). A combination of the following parameters was used to select the GLEAN set as the Official Gene Set (OGS): (i) number of gene models; (ii) number of 100% identity matches to gold standard sequences; (iii) presence of matches to gold standard sequences and; (iv) agreement with splice sites of aligned cDNA (Tables S1 and S2). GLEAN5, the GLEAN set generated with NCBI, Ensembl, SGP2, and combined Fgenesh/Fgenesh++ sets, was selected as OGSv1, and is available at (S28).

Experimental Validation of Gene Set

To experimentally assess the quality of this annotation pipeline a set of 384 GLEAN5 annotations were selected and verified with RT-PCR in 12 bovine tissues. Multiple bovine organs were collected soon after sacrifice. Total RNA was prepared from frozen tissues with TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. Total RNA was converted to cDNA with Superscript III (Invitrogen) primed with random primers. For each tissue, 5 µg of total RNA was converted to cDNA. Gene models were assayed

experimentally by RT-PCR as previously described (S29-S31). Similar amounts of 12 *B. taurus* cDNAs (abomasum, cerebellum, heart, kidney, liver, lung, lymph node, muscle, spinal cord, spleen, testis and thymus, final dilution 1000x) were mixed with JumpStart REDTaq ReadyMix (Sigma, St. Louis, MO, USA) and 4 ng/ul primers (Sigma-Genosys, St. Louis, MO, USA) with a FreedomEvoBio robot (TECAN, Männedorf, Switzerland). The first ten cycles of PCR amplification were performed with the touchdown annealing temperatures decreasing from 60 to 50°C; annealing temperature of the next 30 cycles was carried out at 50°C. Amplicons were isolated on "Ready to Run" precast gels (Pfizer, New York, NY, USA) and sequenced. This procedure was used to experimentally assess sets of exon-exon junctions of the GLEAN5 and non-GLEAN gene models.

Of 26,835 GLEAN5 consensus gene models we randomly picked 3'-most exon-exon junctions of 384 genes with intervening introns longer than 1.5 kbp in length. Primers were designed with Primer3 (S32) (Tm: range 57-63°C, optimum 60°C; length, range 18-27 bp, optimum 20 bp; product length, range 200-600 bp; GC content, range 30-70, optimum 50). The non-GLEAN set was established as follows. All unique introns from the set of gene predictions used as input to GLEAN were projected onto the genome. From this set we filtered out: (i) introns supported by less than three gene prediction programs (both donor and acceptor sites predicted exactly); (ii) introns overlapping GLEAN5 models or RefSeq, Mammalian Gene Collection (S33) or mRNA alignments present in the UCSC Btau3.1 genome browser; (iii) introns from unmapped contigs, and; (iv) introns less than 1.5 kbp in length. Of the remaining introns, the 3'-most intron was selected from each of 183 gene models, defined as a set of overlapping gene predictions. About half of these (S96) were randomly chosen to be experimentally verified as described above.

Generation of cDNAs

Expressed Sequence Tags (ESTs) and full-length cDNA sequences were generated in a similar manner to other genome projects (S33, S34). Substrate mRNA was predominately harvested from tissues derived from L1 Domino 99375, L1 Dominette 01449, her female calf and an unborn male fetus. In total, 28 cDNA libraries from different tissues were used to generate 529,927 ESTs after quality filtering, which were submitted to GenBank (S35) (Table S3). Library construction utilized either a 5' cap trapping method to enrich for full-length transcripts (S36) or size fractionation to enrich for clones >1.4 kbp. From this EST dataset potentially unique full-length clones were identified either through their alignment at 50% identity for 100 bp starting at the initiating methionine to other Mammalian Gene Collection sequences (S33) or through manual inspection of alignments to the non-redundant protein database provided by NCBI. In total, 10,896 full-length cDNA clones were sequenced and subsequent analysis determined that 9,187 of these represent complete and unique coding elements (Table S3). These cDNAs were used as evidence in construction of the second Official Gene Set (OGSv2).

Selenoproteins

All human selenoproteins were mapped to the cow genome. Almost all were partially predicted by the GLEAN pipeline although one (*SelW*) had been completely missed. Of those predicted by GLEAN, only one (*GPX3*) was predicted to encode a selenocysteine (Sec)-containing protein but the predicted Sec residue was different to that obtained by manual annotation. No other selenoproteins were correctly predicted as Sec-containing proteins, either because the in-frame TGA was taken as a STOP or because the TGA-containing exon was skipped. For additional

bovine selenoproteins, we used TBLASTN to search for human-cow homologous proteins where a cysteine residue in human aligns to a TGA in the bovine genome. Of the seven sequences which passed our thresholds (E-value ≤ 0.1 , no more than one in-frame STOP, and a potential SECIS element less than 10,000 bp downstream of the end of the BLAST high scoring pair), three were human non-selenoprotein members of the GPX family, which aligned to bovine GPX selenoproteins. The remaining four were compared with the NCBI's NR database but no alignments from any other species were found to support the alignment of a cysteine residue with a TGA codon found in human versus cow.

U12 Introns

U12 introns in the bovine genome were identified with the union of several computational approaches and this set was manually refined to eliminate obvious errors: (i) human introns from U12DB (S37) were mapped to the bovine genome with GMAP (S38) with 100 bp of exonic flanking sequence; (ii) human RefSeq, cow mRNA and cow ESTs were mapped to Btau3.1 scaffolds with GMAP (S38) and the resulting introns scored and classified as U12 or U2, and; (iii) Geneid *ab initio* predictions allowing for U12 introns were made with permissive parameters and the predicted U12 introns (~2,900 total) were filtered according to alignment of flanking sequence to cattle mRNAs, cattle ESTs and the NCBI NR protein database (S18). Alignment support was considered positive if >67% of the 10 translated amino acids on either side of the predicted splice junction (14/20 in total) were aligned to the same or similar residues. The combined set of U12 introns was subjected to manual inspection to eliminate obvious mapping errors.

MicroRNAs

Bovine microRNAs (miRNAs) were independently predicted by two approaches (prediction Sets 1 and 2 below) and then combined to create a non-redundant set.

microRNA Prediction Set 1

First, a set of bovine predicted miRNAs was generated by comparison with known miRNAs as follows. Metazoan miRNA were downloaded from miRBase version 11.0 (S39) in FASTA format containing the respective start and end positions of their mature parts. WU-BLAST (S40) was used to search each of the known miRNAs with the default parameters plus a DUST filter and the *hspsepSmax 30* option for defining the maximum separating distance between two high score pairs (this allows for a varying pre-miRNA loop while still matching the better conserved 5' and 3' arms). BLASTN matches longer than 20 bp were extended at both ends to match the length of the query sequence. To remove unstable or spurious hits three filter parameters were calculated for each putative pre-miRNA. These included a minimum free energy filter (≤ -15 kcal/mol), a RANDFOLD (S41) filter estimating the stability of the folding compared to dinucleotide shuffled folded sequences (100 randomizations, p-value ≤ 0.05) and finally a RNashapes filter (S42) was used to predict the probability of the sequence folding into a simple stem-loop like shaped structure. Nevertheless, the RNashapes filter was not applied on the final predictions as some known miRNAs, like hsa-let-7a, are known to not meet criteria for stable stem-loop structures when subjected to a minimum free energy folding algorithm, such as RNAfold (S43). The putative miRNAs that passed these filters were aligned to their query miRNAs with MAFFT (S44, S45) and the conservation of the seed region was calculated by mapping the known mature miRNA region on the query miRNA to the alignment. Criteria for

bovine miRNA predictions were 100% conservation of the seed (nucleotides 2-7 of the mature miRNA) and more than 90% sequence identity over the full mature miRNA. As several miRNA (like hsa-let-7, mmu-let-7, etc) can map to the same locus, all predictions were clustered with GALAXY (S45). From a single locus the match with the highest conservation of the mature miRNA and the highest overall percent alignment identity over the entire putative pre-miRNA was used as a single representative sequence for that locus. Including 10 additional pre-miRNAs found after an update to miRBase 12.0, this approach yielded a total of 361 pre-miRNAs in the bovine genome with homology to known miRNAs in other animal genomes.

In addition to predicting bovine miRNAs on the basis of homology to known miRNAs, we used a comparative approach on the basis of a Support Vector Machine (SVM) model of hairpin-like structures followed by an orthology assignment step. This method allows prediction of novel miRNAs that do not show sequence homology to known miRNAs. The complete method is described in (S46); what follows is a brief outline of the basic principles. First, an *ab-initio* SVM model was created to score stem-loop like sequences extracted from the genomic sequence with RNAfold (S43). Second, an orthology assignment pipeline grouped putative precursors from over 40 animal species, then precursors within groups were aligned. In a third step the orthologous groups were again subjected to an SVM model designed to distinguish alignments of orthologous miRNA sequences from other ncRNA alignments or false positive predictions, taking into account typical conservation patterns in pre-miRNA sequence alignments. This approach yielded 135 putative novel bovine miRNAs that are not yet found in miRBase 12.0 with the direct homology approach. The ortholog-based and novel bovine predicted miRNAs were combined to form Set 1.

microRNA Prediction Set 2

Mature miRNAs and stem-loop precursors were downloaded from miRbase v. 10 (S39). Mature miRNAs and their respective precursors were combined into a single sequence with the mature region in lower case format. Each precursor miRNA sequence was aligned with Btau3.1 with WU-BLAST (S40). BLAST was performed by seeding only the mature region of the precursor miRNA to minimize false positives, and then allowing seed extensions outside the mature region. BLAST output was parsed and a sequence corresponding to each hit was extracted from the assembly, extending the extracted sequence to the length of the original query. A global alignment between query (precursor miRNA) and subject sequence (extracted region) was constructed with T-COFFEE (S47), and the number of substitutions was determined. The free energy of folding of the subject sequence were computed with RNAfold (S43). A PRSS analysis between the two sequences was performed with 1,000 iterations in order to assess the statistical significance of the alignment and confirm that the two sequences were homologous. PRSS is part of the Fasta sequence comparison package (S48), and works by constructing local alignments between a query and a database of shuffled subject sequences to generate a distribution of alignment scores, which is used to compute an E-value for the alignment of the query to the actual subject. In our case, the query was the precursor miRNA and the subject was the extracted region of the assembly. A RANDfold analysis of the subject sequence was performed in order to determine how likely the sequence resembled a miRNA. Most of the known miRNAs are in a structural conformation corresponding to a free energy of folding that is considerably lower than that for shuffled sequences with the same nucleotide composition, indicating a tendency in the sequence towards a stable secondary structure (S41). RANDfold was run with 1,000 iterations per sequence, and the results were tabulated. Putative miRNA homologs were

kept if they were at least 95% identical to the known miRNA or if the following conditions were met: (i) similarity score of at least 65% throughout the entire global alignment; (ii) free energy of folding ≤ -20 kcal/mol or lower; (iii) PRSS score $\leq 1e-05$, and; (iv) RANDfold score ≤ 0.015 . In many cases there were more than one miRNA per genomic locus. This was particularly true for miRNAs that are known to have several paralogs, or due to orthologous genes that are redundant in the database used (miRBase v.10). All overlapping miRNAs were clustered on the basis of genomic location and sequences within the cluster were scored based on similarity to known query miRNA. Only the sequence with the greatest similarity was used in further analysis. Putative microRNAs were analyzed with RepeatMasker to remove repetitive and transposable elements.

Merging microRNA Set 1 and microRNA Set 2

The precursor miRNA sequences from the two miRNA prediction sets were compared with WU-BLAST with default settings, except the hspsepSmax parameter was set to 30. The results were parsed for hits on the same strand with 100% sequence identity and more than 50 bp alignment length. Sequences that met these alignment criteria were considered identical miRNA loci if their start and end coordinates in the genome did not differ by more than 25 bp. This step prevents merging of paralogous loci, but allows for variation in length of precursor miRNAs predicted with different methods. Most of the predictions missed in Set 2 were located in the unassigned scaffolds and/or were new miRNAs present in miRBase version 12.0.

Bovine Official Gene Set

The OGSv1 used in global analyses and annotation was the GLEAN5 consensus set. OGSv2 was generated by: (i) rerunning GLEAN with new cDNA evidence and; (ii) incorporating manual annotations, selenoproteins and U12 intron data into the consensus gene set. Specifically, manual annotations, selenoprotein and U12 intron data were used to replace GLEAN5 gene models or add additional gene models.

Manual Annotation

Manual annotation was performed by a group of approximately 150 scientists who typically had experience with specific genes. The aims of the manual annotation effort were to confirm or correct OGSv1 automated gene models, identify genes missing from OGSv1, and identify changes in genes or gene families that comment on ruminant biology and evolution. A total of approximately 4,000 gene models were manually inspected. The initial step was to obtain the sequence of a bovine EST/ cDNA or a human or mouse protein ortholog from RefSeq (S49), Ensembl (S50), UCSC Genome Browser (S51) or Uniprot (S52). This sequence was used to search the OGSv1 translated protein database with BLASTP or BLASTX (S22). The most significant Expect values and bit scores for the bovine ortholog were generally well separated from secondary hits. Reciprocal BLAST analysis was performed to validate the ortholog. For gene families syntenic position was also used to define orthology. Genes missing from OGSv1 were identified in the assembly by comparing bovine EST/cDNA or protein homologs to the assembly with BLASTN or TBLASTN. Gene models were then annotated with one of three methods. Some participants used manual methods and web tools such as BLAST at NCBI, Ensembl and Bovine Genome Database to annotate a gene, and submitted annotations to a manual submission website at the Bovine Genome Database (S53). Other people participated in an annotation jamboree held at the Sanger Institute, and used the Otterlace annotation software

(S54) to view and edit gene models. The jamboree data were then transferred as generic feature format (gff) from the Sanger Institute to the Bovine Genome Database. The majority of the annotations were performed with the Apollo Annotation Editor (S55), which connects remotely to the Bovine Genome Database to retrieve gene prediction evidence, including OGSv1, other gene prediction sets, protein homolog alignments, and bovine EST/cDNA alignments. Regardless of the annotation method, inspections were made for completeness of the gene model, appropriate start and stop codons, untranslated regions, splicing variants and overall exon structure. Corrected or new gene models were then incorporated into OGSv2.

GC Content

Genomic sequences were partitioned into segments by the binary recursive segmentation procedure, *D_{JS}*, proposed by (S56). In this procedure the chromosomes are recursively segmented by maximizing the difference in GC content between adjacent subsequences. The process of segmentation was terminated when the difference in GC content between two neighboring segments was no longer statistically significant (S57). We compared the distribution of GC-content lengths among *B. taurus* (Btau4.0), *H. sapiens* (NCBI Build 36.3), and *M. musculus* (NCBI Build M37.1), and used a G-test goodness-of-fit test to determine that none of the distributions of segment length differed from any other.

De Novo Repeat Detection

The Btau4.0 assembly was used for *de novo* repeat identification. Repeats were identified with PALS/PILER (S58) and RepeatScout (S59). Output from both approaches were combined and clustered to create globally alignable clusters with BLASTCLUST (NCBI BLAST distribution). Clusters were then aligned with MUSCLE (S60) and consensus sequences for each cluster generated with PILER (S58). Repeat consensus sequences were compared to SwissProt (S24) to identify protein coding gene families inappropriately included in the repeat set. Consensus sequences identified as similar to protein coding sequences but not similar to retroposon or endogenous retrovirus protein coding sequences were removed from the consensus set. This process was performed twice to ensure that no protein coding genes were overlooked. A consensus set of 23,725 repeats was produced. Consensus sequences were subsequently aligned to the OGSv1 gene models and tRNA sequences. A set of 1,233 gene models was identified which contained repeat sequences on this basis along with a set of tRNA genes. Consensus repeat sequences, in conjunction with RepeatMasker mammalian consensus repeats, were then used to mask the Btau4.0 assembly for use by other groups in evolutionary breakpoint and segmental duplication (SD) analyses. Known repeat sequences within the consensus repeat set were identified with RepeatMasker and also aligned to a custom transposable element protein database. This allowed the partitioning of the repeats into three broad classes: (i) identical to or containing a single subsequence identical to a single previously known repeat class; (ii) similar to one or more previously known repeat classes, and; (iii) unknown, or new repeats.

Identification of Intact LINE Elements and Phylogenetic Analysis

Intact L1 (L1_BT) and BovB were identified with PALS (S58), by aligning consensus L1 and BovB sequences to Btau4.0, with criteria of $\geq 70\%$ identity over $\geq 90\%$ of the query sequence. Sequences were globally aligned with MUSCLE (S60) and the alignments used to create maximum likelihood trees with RAxML (S61) with the GTRCAT substitution model, and an initial 200 bootstraps followed by a maximum likelihood search. Because L1 repeats occur more

often in high GC regions, while BovB repeats occur more often in low GC regions (indicated by their positive and negative correlations with GC, see below), we identified a subset of these repeats located in 162 1.5 Mbp bins that contained equal numbers of BovB and L1_BT intact repeats in order to minimize the effect of surrounding genomic sequence composition. Jukes-Cantor substitution rates (S62) were calculated for this sub-set, omitting all positions containing gaps and missing data. Standard error estimates were obtained by a bootstrap procedure (500 repeats). Calculation of the Jukes-Cantor substitution rates and editing the trees for appearance were performed with MEGA v4 (S63). Potentially active elements were identified by scanning for ORF of the appropriate length with CLC Sequence Viewer 5 (S64).

Correlation of Repeat Elements

To calculate correlations of repeat element classes with each other, each chromosome was divided into 1.5 Mbp segments (bins) beginning at the 5' end. For each bin we calculated: (i) the number of repeats from each repeat group on the basis of both RepeatMasker output from our custom library, and SSR coordinates from Phobos v. 3.3.2 (S65) output that were entirely within the bin; (ii) the number of OGSv1 gene models that started in the bin (gene density); (iii) the GC content, and; (iv) the number of SDs identified by both WSSD and WGAC (both defined below) and located 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, GC content and SD.

Simple Sequence Repeats

The Btau3.1 assembly, ovine BAC end sequences, human (NCBI Build 36) and dog (CanFam 2.0) genome assemblies were processed through the SPUTNIK program (S66) with the options: '-s 10 -F 300' to search for simple sequence repeats (SSRs). SSRs had to exceed a score of 10 i.e. more than 10 bp of repeat sequence. Three hundred bp segments were retrieved on either side of each SSR. These data were further filtered to remove 'SSRs' with large poly N tracts in the flanking sequences. Outputs were categorized into di-, tri-, tetra- and pentanucleotide repeats and subsequent statistical analyses were conducted. In addition, resultant bovine SSRs were used in a BLASTN search against cARTY SINE, with parameters set to >85% identity and an overlap >100 bp. The same parameters were used in a BLASTN (S67) search of the resultant human SSRs against RepBase9.05 (S68).

Protein Ortholog Analysis

Orthologous relationships between genes of cattle (OGSv1), human, mouse, rat, dog, platypus and opossum (Ensembl v45, which included the Ensembl August 2006 human gene-build) were inferred through all-against-all protein sequence similarity searches with the Smith-Waterman algorithm and retaining only the longest predicted transcript per locus. The orthologous groups were then formed by: (i) grouping recently duplicated sequences with >97% identity within genomes to be treated subsequently as single sequences; (ii) forming triangles and tuples of the mutually reciprocal best hits between genomes, and; (iii) expanding the seed orthologous groups by inclusion of co-orthologous sequences that are more similar to the orthologous gene than to any other gene in any other genome, requiring also that all members of the group have matches overlapping by at least 20 bp.

Protein Phylogenetic Analysis

Multiple protein sequence alignments of strictly defined single-copy orthologs across all species were produced with MUSCLE (S60). The well-aligned regions of these alignments were extracted with GBlocks (S69) with default parameters and then concatenated into a superalignments. Phylogenetic analysis of superalignments was performed with the maximum-likelihood method as implemented in PHYML (S70) with the JTT model for amino acid substitutions, a gamma correction with four discrete classes, an estimated alpha parameter and proportion of invariable sites.

Exon Skipping

Genome-wide exon skipping events were analyzed across four placental mammals: human, mouse, dog and cow. Each event was defined as a triplet of exons with evidence of skipping for the middle exon. With comparative methods we mapped the documented human events onto the three other species and classified them according to whether the skipped exon was conserved or not as assessed by direct cDNA and/or EST sequencing of the transcripts sampled from a broad range of tissues.

The initial data set represented 1,930 exon triplets representing exon-skipping events in human extracted from Ensembl (S50) and EST sequences (S25). These triplets were aligned to mouse, dog and cow with Exonerate (S15). In each species the triplets were classified according to whether the three exons were aligned, whether all exons except the middle one were aligned, or any other possibility. Furthermore, we annotated the aligned triples that had evidence of skipping of the middle exon in mouse from EST data, with dbEST (S25).

To uncover whether exon-skipping events were evolutionary conserved, we experimentally assessed whether it was possible to amplify by RT-PCR both corresponding isoforms in bovine tissues, as previously described (S30, S31, S71, S72). Duos of RT-PCR reactions were performed in 16 different bovine tissues (abomasum, adrenal gland, amygdala, bladder, cerebellum, heart, kidney, liver, lung, mesenteric lymph node, muscle, spinal cord, spleen, testis, thymus, thyroid) with pairs of primers mapping in the first and third exon and in the first and second exon of a triplet of exons with evidence of skipping for the middle exon in human. Tissues were collected at the slaughterhouse and were immediately frozen on dry ice. Total RNA was extracted with TRIzol reagent (100 mg tissue for 1 ml Trizol, Invitrogen), passed across an RNeasy column (Qiagen, Valencia, CA, USA), and used as a template for complementary DNA (cDNA) synthesis with Superscript III (Invitrogen) according to the manufacturer's protocols. Similar amounts of cow cDNAs were mixed with REDTaq ReadyMix (Sigma) and 4 ng/μl of primers (Sigma-Genosys) with a TECAN FreedomEvo robotic workstation in 25 μl final volume. The ten first cycles of the PCR amplification were performed with a touchdown annealing temperatures decreasing from 60 to 50°C; annealing temperature of the next 30 cycles was carried out at 50°C. Amplicons were separated on "Ready to Run" precast gels (Pharmacia) and sequenced.

Gene Ontology (GO) (S73) annotations were downloaded from the Ensembl BioMart (S74) set and GO terms assigned to each exon-triplet. We then compared the entire GO trees of each subset of triplets, obtained according to the conservation in other species, with the GO annotations of all genes with cassette exons in human, using the hypergeometric test.

Homologous Synteny Blocks (HSBs) and Evolutionary Breakpoint Regions (EBRs)

A set of 280,000 cattle BAC-end sequences (BESs) from the CHORI-240 library (S6) were

repeat-masked and compared for similarity to human genome build 36, mouse genome build 37, dog genome build 2, macaque genome build 2 with BLASTN (S67). Unique BLASTN hits with E-values $<1 \times 10^{-10}$ and hit length ≥ 100 bp were selected for the construction of pair-wise comparative maps of the cattle, human, macaque, mouse, dog genomes. The same set of cattle BESs was aligned to Btau4.0 with BLAT with the fastMap option (S75). The genomic sequence coordinates of unique hits ≥ 100 bp and $\geq 98\%$ identity were taken as positions of corresponding BESs in cattle chromosomes. For the pig-cattle comparison, a set of 350,000 pig BESs was compared with Btau4.0 with BLASTN. Single-hit BESs (E-value $<1 \times 10^{-10}$; hit length ≥ 100 bp) were integrated with the porcine physical map downloaded from the Sanger Institute's website (S76). Pig BAC clones with known coordinates on pig chromosomes and with a unique BLASTN match in the cattle genome were used to build a cattle-pig comparative map. Homologous synteny blocks were defined for each of the cattle pair-wise comparative maps. The minimum HSB size used for this analysis was 500 kbp in cattle genome sequence coordinates. In addition, sets of one-to-one orthologs downloaded from Ensembl (S50) were used to define pair-wise HSBs between human, chimp, macaque, dog, mouse, mouse, rat and opossum genomes with the same criteria for HSB definition as described above. The Evolution Highway comparative chromosome browser (S77) with recent upgrades (S78) was used to visualize pair-wise HSBs on cattle and human chromosomes. An automated option of Evolution Highway was used to identify cattle-specific EBRs, artiodactyl EBRs (shared between cattle and pig), and ferungulate EBRs (shared between cattle, pig, dog) in cattle and human sequence coordinates according to rules defined in (S77). A t-test with unequal variances was used to identify repeat families that were unequally distributed in EBRs when compared to the rest of the genome. The same test was applied to compare densities of segmental duplications in EBRs and the rest of the genome. Benjamini-Hochberg critical values were calculated to control for false positive discovery rate (S79).

Oxford Grid

An Oxford Grid (S80) was constructed to show the extent of chromosomal conservation between cattle (BTA) and humans (HSA). It was drawn from 46,947 cattle BAC-end sequences showing homology with sequence from build 36 of the human genome. A similar Oxford Grid was constructed comparing chromosomal conservation between cattle (BTA) and mouse (MMU).

Bidirectional Promoters

Bidirectional promoters represent a shared regulatory sequence, or the amalgamation of two promoter regions of oppositely oriented genes, which are separated by no more than 1,000 bp. These promoters occur frequently in the human genome and regulate genes with essential functions (S81, S82). We mapped bidirectional promoters in the cattle genome using a multi-tiered approach. The first step was the identification of candidate bidirectional promoters in three separate bovine datasets (spliced bovine EST alignments, bovine RefSeq annotations (S49) available at the UCSC Cow Genome Browser and bovine OGSv1). The high stringency set of bidirectional promoter candidates met the criterion that the transcriptional start sites of the oppositely oriented genes were no more than 1,000 bp apart.

Given the frequency of incomplete annotations at the 5' ends of genes in most species, we also identified a low-stringency set of oppositely oriented gene pairs without imposing a limit on their maximal intervening distance. Bovine ESTs were aligned to these OGSv1 (coding) annotations to gain support that 5'UTR regions could be mapped onto this low-stringency set

producing more reliable bidirectional promoter predictions from the computationally derived gene set. A subsequent step was to map bovine candidate bidirectional gene pairs onto their human orthologs. This approach was useful to confirm predictions in the cow genome, to assess the completeness of annotations in the human genome and to identify potential rearrangements between genomes. Only those candidate bovine bidirectional promoters for which the gene pairs could be mapped to orthologous, oppositely oriented gene pairs in the human genome (while meeting the 1,000 bp maximal distance criterion) were considered to be confirmed as bidirectional promoters in the cow genome.

Segmental Duplications

The segmental duplication (SD) content of the cow genome was assessed with two different methods; one dependent on the assembly (WGAC) and one from an assessment of excess depth-of-coverage of whole-genome shotgun sequence data (WSSD) against the Btau4.0 genome assembly. The assembly-dependent method, BLAST-based whole genome assembly comparison (WGAC) (S83), was used to identify a total of 129,555 pairwise alignments representing putative SDs of size greater than 1 kbp and with >90% identity. High-copy repeat sequences were initially removed with RepeatMasker (S19) and a newly constructed library of cow repeats described above. Initial seed alignments that were >250 bp and >88% identity, with repeats subsequently reintroduced, were used to create local alignments and optimal global alignments that were >1 kbp and had >90% sequence identity. Duplication intervals that were >94% identity and >10 kbp in size (after chaining across gap regions in the cow genome) and not supported by WSSD (see below), were excluded from the genome-wide calculation of SDs. As larger, high-identity duplications (> 94%) are frequently collapsed within working draft sequence assemblies (S84) or may represent artifactual duplications within an assembly (S83), we compared these assembly-based results to whole genome shotgun sequence detection (WSSD) database of cow SDs. WSSD identified regions >10 kbp in length with a significant excess of high-quality WGS reads (S85) within overlapping 5 kbp windows. We established thresholds on the basis of the alignment of WGS reads against 96 unique cow BACs. The analysis was from a comparison of 23,971,214 *B. taurus* WGS reads against 400 kbp segments of the Btau4.0 assembly. 13,523,039 reads were remapped to the assembly on the basis of the following criteria: >94% sequence identity; >200 bp non-repeat-masked sequence and at least 200 bp of PhredQ >30 bp. We excluded regions with repeats with < 10% divergence from their consensus and all bovine-specific repeat sequences.

Gene Ontology and Pathway Analyses for Genes Present in Segmental Duplications

To determine whether SDs are associated with gene functional categories, we tested for enrichment of Gene Ontology (GO) terms (S73) and InnateDB pathways (S86) among genes located in SD intervals identified by both WGAC and WSSD methods (S28). GO terms and InnateDB pathways were assigned to bovine OGSv1 gene models by transferring annotations from human orthologs, identified as described in *Protein Ortholog Analysis* methods. Many of the bovine genes were found to have one-to-many orthology relationships with human genes. To overcome the potential bias of assigning multiple similar GO and pathway annotations to the same bovine gene due to these cases, five ortholog sets were created. Each of the five sets included all bovine OGSv1 genes for which there was at least one human ortholog. For each bovine gene that had multiple human orthologs, a single human gene was randomly selected for annotation transfer. This process was repeated to generate five unbiased datasets. The Gene

Ontology and Pathway Analysis tools in InnateDB (S86) were then used to investigate which GO terms and pathways were significantly over-represented in each of the five sets with the hypergeometric distribution. P-values were corrected for multiple testing with the Benjamini-Hochberg correction for the FDR.

Adaptive Evolution in the Bovine Genome

To identify genes on the bovine lineage that have evidence of adaptive evolution we compared 10,519 bovine genes to their putative orthologs (where present) in the human, mouse, rat, dog, opossum and platypus genomes. Putative orthology was assigned as described above (see Protein *Ortholog Analysis* methods). Of the 10,519 orthologous groups, 1,531 orthologs were missing in the human genome, 940 in mouse, 1,352 in rat, 1,895 in dog, 2,563 in opossum and 4,999 in platypus.

Most methods that are commonly implemented to investigate evidence of positive selection require the provision of a phylogenetic tree for each orthologous set of genes. Genome projects to date have tended to investigate evidence of positive selection either in the analysis of pairwise alignments or in datasets of strict 1:1 orthologs in a small number of comparison species, in which the ortholog was present in all examined species. These analyses could use a single simple species tree to represent the phylogenetic relationship of each orthologous gene set. In this analysis, we took advantage of the increasing number of mammalian genome sequences and analyzed orthologs from seven species for evidence of adaptive evolution. Increasing the number of orthologous sequences in each alignment is expected to increase the power to detect positive selection. Because not all orthologs were found in each of the seven examined genomes, our orthologous gene datasets contained a variable number of sequences. It was therefore necessary to individually reconstruct a phylogenetic tree for each dataset. To do this, neighbor-joining phylogenetic trees were reconstructed from coding sequence alignments of each orthologous gene dataset with the neighbor algorithm implemented in PHYLIP (S87, S88). The ratios of nonsynonymous substitutions per nonsynonymous site (d_N) to synonymous substitutions per synonymous site (d_S), indicated by d_N/d_S , were estimated by maximum likelihood (ML) for each gene from coding sequence alignments of each of the 10,519 orthologous groups with the codeml program from PAMLv4 (S89). Ratios greater than one indicate positive selection (S90). Two models were implemented to test the statistical significance of variable selective pressure specifically on the bovine lineage. In the one-ratio model, which acts as the null model ($\text{NSSites} = 0$, $\text{model} = 0$), each lineage was modeled to have the same d_N/d_S ratio. The ratio is constrained between 0 and 1, and thus does not allow for the presence of positive selection. The second model is a model of bovine-specific evolution, where the bovine lineage was selected as the ‘foreground’ lineage and d_N/d_S was specifically allowed to vary unconstrained on this lineage only (model 2). The two models were compared with the Likelihood Ratio Test (LRT), calculated from the log likelihood ($\ln L$) values of both models. Twice the difference between $\ln L_{\text{model2}}$ and $\ln L_{\text{one-ratio}}$ (i.e. $2\delta l$) was compared to a χ^2 distribution to obtain the p-values.

To determine if any particular functional categories were significantly associated with genes subject to positive selection we mapped via orthology, bovine genes to human molecular function and biological process terms from the GO (S73) and PANTHER (S91) databases. For each ontology term, the distribution of log likelihood ratios associated with genes mapped to the term was compared, with a one-sided Mann-Whitney U (MWU) test, to the distribution of all log likelihood ratios, similarly to the method previously described (S92). This approach has the

potential to identify categories of genes that have a tendency towards being subject to positive selection despite the majority of genes not having stringent evidence of positive selection. To investigate which pathways were represented in genes that had evidence of positive selection ($d_N/d_S > 1$), bovine genes were mapped to pathway annotation via their human orthologs with InnateDB (S86). InnateDB is a platform to facilitate systems level analysis, which integrates pathway and molecular interaction data from the major publicly available databases.

Annotation of genes involved in adaptive immunity

Several gene families important in acquired immunity were manually annotated with Btau3.1 and in some cases Btau4.0 in combination with OGSv1 with the procedures described above (see *Manual Annotation methods*).

Metabolic Reconstruction and Identification of Metabolic Gene Losses in the Cattle Genome

The metabolic pathways encoded in the cattle genome were reconstructed with the OGSv1 (of the basis of Btau3.1). For this purpose, OGSv1 gene models were matched against those in CattleCyc, which contains a curated cattle-specific pathway genome database (S93). Briefly, CattleCyc was created by comparing cattle metabolic genes with those in a human-specific metabolic pathway/genome database reconstructed with the human genome (build 36) and Pathway Tools (S93). The OGSv1 and CattleCyc gene models were assumed to be identical if they shared a common chromosomal location and were encoded on the same strand. In addition to the OGSv1 gene models, six genes from Build 2.1 (not present in Btau3.1) and 11 mitochondrial genes were used to reconstruct cattle-specific metabolic pathways. Missing cattle metabolic genes were identified on the basis of gene orthology with human metabolic genes and assessed by sequence similarity and synteny of selected mammalian genomes (cattle, human, dog, mouse and chimpanzee). To confirm the absence of cattle genes, cattle orthologs were also searched for in other assemblies (Btau2.1 and Btau4.0). Mammalian phylogenomic foot-printing was also conducted; human protein sequences were compared with trace whole genome sequences of all of the whole genome sequenced mammals including 2X coverages with TBLASTN (S22).

Lysozyme Genes

BLAST (S22) with an E-value threshold of 0.000001 was used to identify protein sequences similar to human lysozyme C, in human, mouse, and bovine RefSeq (S49) protein sets, and in bovine OGSv1. Protein sequences arising from distinct genes were aligned with CLUSTALW (S94), and the resulting alignment was used to build a neighbor-joining tree with the PHYLIP programs PROTDIST, SEQBOOT, NEIGHBOR, and CONSENSE (S87). The frequency of lysozyme C sequences in EST libraries in GenBank was determined by matching sequences to entries in NCBI's UniGene database (S18), followed by analysis with UniGene's EST ProfileViewer.

Supporting Text

Experimental Validation of Gene Set

Eighty-two percent of the 384 GLEAN5 consensus gene set models that were selected for verification by RT-PCR in 12 bovine tissues were confirmed, a result similar to the 85%

confirmation rate obtained by combining Ensembl, Twinscan and SGP2 in the annotation of the chicken genome (S29, S95). There was strong correlation between RT-PCR success rate and the level of support for the predictors. As expected, cDNA and EST were the best indicators of successful RT-PCR, but neither was very sensitive, while Ensembl showed a good balance between sensitivity (74%), specificity (84%) and positive predictive value (95%). Completeness of the annotation was monitored through experimental verification of exon-exon junction of gene structures predicted by at least three gene prediction methods and not included in the consensus GLEAN5 gene set. Only 18% of (17 out of 96) exon-exon junctions of these latter gene models could be experimentally confirmed by RT-PCR in at least one of the set of 12 bovine tissues.

Selenoproteins

The bovine selenoproteins were annotated as described above. SelW was found to have two Sec residues in *B. taurus*, with an arginine residue (in human) converted to a Sec residue in the bovine protein. In addition, the cow SelP ortholog has twelve Sec residues whereas the human sequence has ten. Seven sequences were identified which were potential bovine expansions of selenoprotein families. Three were members of the GPX family, and aligned to human non-selenoprotein members of the GPX family. The remaining four were compared with the NCBI's NR database but not supported by alignments from other species. Finally, we checked whether there has been an increase of the gene copy numbers of individual selenoproteins in the cow genome and indeed, one (*SelI*) was found to be present in more copies in cow (two) than in other mammals (one).

U12 Introns

The minor splicing pathway has been well-conserved among vertebrates and the number of U12-type introns found in different mammalian species does not show appreciable variation (S37, S96, S97). However, these studies have generally relied on searching for homologs of human U12 introns in other vertebrates, making estimates of intron gain and loss difficult: losses may be due to incomplete genome sequence and gains could be missed due to reliance on homology searches. U12 introns, an atypical class of splicesomal introns, have AT-AC at their 5' and 3' ends, respectively, while canonical U2-type introns have GT-AG, respectively. U12 introns have historically posed a challenge to accurate gene annotation due to their rarity in mammalian genomes (<0.5% of introns are U12), however, their highly conserved donor and branch point sequences permit their accurate identification. We have included them in the transcript annotations and thus enhanced the quality and utility of the genome annotation for subsequent analyses. The U12 complement of the bovine genome is comparable in size to that of other investigated mammals (S37, S96, S97).

In total, 798 U12 introns were identified in the cattle genome that either exhibit evolutionary conservation or were supported by alignment to expressed sequence. Only 515 (65%) were included in the GLEAN5 annotation, which justified the more exhaustive annotation (Fig. S1). Only three introns contained in the GLEAN5 set that could be classified as U12-type were missing from the U12 annotations, suggesting that the approach was nearly exhaustive. A breakdown of terminal dinucleotide combinations (GT-AG, AT-AC, AT-AA, etc) was performed according to whether or not they were predicted by GLEAN5. We found a bias toward GT-AG and canonical introns in general in the GLEAN5 set. Their splice site scores are on average greater (Geneid splice site scores of 5.29 vs. 4.77 for donors and 6.79 vs. 5.50 for acceptors).

MicroRNAs

We identified 361 bovine miRNA genes with homology to experimentally verified microRNAs in the miRBase 12.0 and 135 novel microRNAs with a comparative genomic approach (S46). The 496 bovine miRNAs were grouped into 298 homolog families. About half of the bovine miRNA occur in 60 genomic miRNA clusters, in which 2 to 7 miRNA genes are separated by less than 10 kbp (Fig. S2). A notable exception was a 43 kbp cluster on BTA21 harboring approximately 40 contiguous microRNA genes that is orthologous to a large cluster on human 14q32.31 (S98). This region is imprinted in the mouse (S99).

GC Content

Animal genomes are not uniform in their long-range sequence composition, but are composed of a mosaic of sequence stretches of variable lengths that differ widely in their GC compositions. Whether these stretches meet the criteria of isochores [*sensu* (S100)], or should better be referred to as GC-content domains (S101) is a matter of debate (S57, S102-S104). In animal genome sequences studied to date, the distribution of GC-content domain lengths (plotted on a log-log scale) was found to follow a heavy-tail distribution with power-law decay exponents ranging from -1.12 to -1.15 . The genome of the *B. taurus* genome is no exception and the compositionally homogeneous segments in its genome, as in all other genomes studied so far, do not have a characteristic length; rather, there is an abundance of short segments and only a small number of longer segments.

A comparison of the distributions of GC-content lengths among *B. taurus* (Btau4.0), *H. sapiens* (NCBI Build 36.3), and *M. musculus* (NCBI Build M37.1) is shown in Fig. S3. Interestingly, the bovine has the lowest abundance of small size GC-content domains (<2 kbp) relative to the other three genomes. The GC contents of their small domains span from 7% to 82%. In contrast, the mid- and long-size GC-content domains (3 kbp - 1 Mbp) in *B. taurus* are more frequent than in human but the long size domains are less frequent than in mouse. Only a small fraction (3%) of the homogeneous domains are longer than 300 kbp, however their mean GC content (39.6%) is significantly lower than the mean GC content for the entire genome (41.7%).

Phylogenetic Analysis of LINE Elements

The maximum likelihood tree of BovB elements, with 11 terminal clusters with branch lengths less than 0.02, indicates that a number of recent retrotransposition events of BovB have occurred, which is evidence for continued activity of BovB retrotransposons (Fig. S4). A similar analysis for L1_BT repeats is shown in Fig. S5.

Correlation of Repeat Elements

Figure S6 shows correlations among the repeat groups, gene density, GC content, and SD. A chromosome map of high and low density ancient repeats, L2/MIR (a LINE/SINE pair) and BovB, and more recent repeats, BovB/Art2A (BovB derived SINE pair) is shown in Fig. S7.

Simple Sequence Repeats

Figures S8 and S9 show the relative frequencies of different dinucleotide and trinucleotide repeats, respectively. Comparative frequencies of trinucleotide SSRs in the human, canine, bovine and ovine sequences are shown in figure S10. The latter information was obtained from

ovine BAC end sequence data. The structure of Bov-A2, which consists of an AGC repeat tail, duplicated regions with 16 bp palindromes and AC(T)nC repeats, makes it a suitable predecessor for the emergence of new SSRs. Indeed, of the SSR identified, 12% of AAGT, 20% of AAGTG, 12% of AAAGT and 41% of AGCAT were associated with Bov-A2.

Protein Ortholog Analysis

All orthologous classifications and the corresponding species copy-number distribution are available from (S105). Some of the cattle unique genes may be spurious gene predictions, while the paucity of unique genes in dog and platypus may be due to highly conservative gene predictions (Fig. 1A).

About 1,000 orthologs shared between rodents and laurasiatherians appear to be missing in human (Fig. 1B). The majority of these encode G-protein-coupled receptors, including olfactory receptors, which may represent actual human lineage gene losses, as the human genome is the most completely sequenced and is the best annotated. However, since this analysis was performed, a newer human Ensembl release has incorporated over 300 new genes similar to those that appear to have been missing, suggesting that some of the lost orthologs in human may be due to incomplete annotation of the human genome. The genome-wide phylogenetic analysis with the single-copy orthologs clearly supports the accepted phylogeny in which humans and rodents are sister lineages to the exclusion of laurasiatherians (Fig. 1D). The maximum likelihood approach allowed us to estimate the relative rates of molecular evolution along each of the branches, and attributes the higher level of divergence between human and rodents to an elevated rate of the rodent evolution.

Exon Skipping

A total of 277 cases with different conservation patterns in human and mouse were examined in 16 different cow tissues by RT-PCR. It was assumed that this comprehensive set of bovine tissues encompassed the tissues where these exon-skipping events were initially discovered. The 277 cases were divided into: (i) a 'conserved' set, which included 163 cases in which exon skipping occurred in both human and mouse, and; (ii) a 'non-conserved' set, which included 114 cases in which exon skipping occurred in human but not mouse.

Of the 277 cases, we could detect expression for 188 in cow (Table S5). More specifically, expression was detected for 122 (75%) of the 'conserved' set, whereas only 66 (58%) showed expression from the 'non-conserved' set. The results confirm that the majority of the exon-skipping with EST evidence in human and mouse also have regulated exon skipping in cattle. In particular, 71 (58%) of the 'conserved' set showed evidence of exon skipping in cattle, while only 20 (17.5%) were confirmed in the 'non-conserved' set. The vast majority (80%; 57 out of 71) of the cow substantiated "conserved" set of exon-skipping events show expression in the identical set of tissues for both transcript isoforms, i.e. the one including all three exons of the triplet (long form) and the one skipping its middle exon. This percentage falls to 45% (9 out of 20) for the "non-conserved set", because for about one third (35%, 7 out of 20) of these triplets the isoform skipping the middle exon was amplified only in a subset of the tissues positives for the longer isoform. This analysis also showed specific exon losses or loss of alternative splicing in one or the other lineages. For instance, the middle exon is constitutive in cattle for 27 (22%) of the 'conserved' set and for 37 (56%) of the 'non-conserved' set. For the remainder of each set (24 in the 'conserved' set and 9 in the 'non-conserved' set), the middle exon was not detected in any isoform in cattle. Finally, 62% of all cases (117 out of 188) showed lack

of conservation of the skipping event in mouse or cattle, or both, from which we estimate that approximately 40% of the exon skipping is conserved across mammals, which agrees with the upper bound obtained from previous analyses involving human and rodents (S106-S110).

After performing a functional analysis of the differential conservation in cattle with the GO categories associated to the human events, we verified that among the 91 cases with a evolutionary conserved exon-skipping event there is an over-representation of the GO terms *regulation of transcription* ($p = 0.0018$) and *development* ($p = 0.0071$). The 64 cases for which the alternative splicing is apparently lost, (i.e., the alternative exon is constitutive in cattle) have over-representation of the GO term *catalytic role in protein modification* ($p = 0.0003$). Finally, the 36 cases that lost the regulated exon in cattle are over-represented in the GO terms *phosphorylation* ($p = 0.0005$) and *kinase activity* ($p = 0.0028$). However, after adjustment for multiple testing with the Bonferonni correction none of these GO terms was significant.

Evolutionary Breakpoint Analysis

Examples of cattle-specific EBRs, artiodactyl EBRs (shared between cattle and pig), and ferungulate EBRs (shared between cattle, pig, dog) in cattle and human sequence coordinates are shown Fig. S11 and Fig. 2, respectively. A full chromosome-by-chromosome display of human-cattle (and other genomes) homologous synteny blocks and genome organization is provided at the Evolution Highway website (S111). Repeat families that were unequally distributed in EBRs when compared to the rest of the genome are listed in Tables S6 and S7.

Oxford Grid

The Oxford Grid shows that most cattle chromosomes correspond primarily to part or all of one human chromosome (Fig. S12). The main difference between each cattle chromosome and its corresponding human chromosome is multiple rearrangements. For around one-third of cattle chromosomes, the largest segment conserved with a human chromosome comprises more than half of the cattle chromosome. In contrast, the Oxford Grid showing a comparison of the cattle and mouse chromosomes constructed in a similar manner shows much less correspondence. An expandable, zoomable, hyperlinked version of this grid is available at (S112). Similar grids for cattle versus pig, dog, macaque and mouse are also available.

Bidirectional Promoters

A total of 1,574 bidirectional promoters were identified in the cattle genome with the three datasets (Fig. S13) (see *Supporting Materials and Methods*). A total of 5,156 low-stringency candidate bidirectional promoters were predicted with data from OGSv1 genes. This result is close to the number predicted from analyses in the human genome (S82). Aligned ESTs concurred at the 5' UTRs for 220 of OGSv1 genes, and extended the 5' ends of 225 of them. All of these predictions met the 1,000 bp maximal intervening distance criterion (see *Supporting Materials and Methods*). The remaining low-stringency predictions require further experimental or computational evidence. As an example, the set of human bidirectional promoters controlling protein-coding genes was compared to the cattle genome. Of the 1,369 promoters examined, 85% were verified with the same regulatory structure in the cattle genome.

An example of a bovine candidate bidirectional promoter that did not validate in human involved the gene *CYB5R4* (Fig. S14), which has been implicated in diabetes (S113). The orthologous gene in the cow genome has evidence for a bidirectional promoter with the partner gene containing minimal coding potential and strong RNA secondary structure (Fig. S15). The

chromatogram from the original sequencing of the EST (DV834581) that supports the novel bovine gene shows an intact cloning structure at the 5' end and poly A tail at the 3' end thereby validating the orientation of this gene with respect to *CYB5R4*. No detectable homology exists for this partner gene in the human genome.

Segmental Duplications

An estimated 3.1% (94.4 Mbp) of the cattle genome consists of SDs (Figs. S16 and S17). This estimate includes 1,020 duplication intervals identified by WGAC and WSSD, as well as those that were detected only by WSSD, which likely represent collapsed duplications within the assembly (S28). There were also a number of examples of tandem SDs (Figs. S17 and S18). 47% (45.2/94.4 Mbp) of the SDs were found on scaffolds that have not been assigned to a chromosome (Fig. S17).

In addition to the estimated 94.4 Mbp of SDs in the bovine genome, we identified SDs in Btau4.0 that appear to be artifacts of the assembly process. A total of 1860 pairwise alignments (>20 kbp, >94% identity) corresponding to 92.45 Mbp of apparent duplicated sequence in Btau4.0 could not be substantiated by WSSD (i.e. WGAC+/WSSD-). These are predominantly intrachromosomal in origin. Excluding the unassigned scaffolds (present in ChrUn), there are a total of 364/402 (91%) pairwise alignments that map within 1 Mbp of one another, suggesting that these may represent local errors in the assembly. As expected, there was enrichment of SDs in ChrUn, which contains unassigned sequence.

Gene Ontology and Pathway Analyses for Genes Present in Segmental Duplications

GO terms and Pathways with statistically significant enrichments ($p < 0.05$) in all five sets are shown in Tables S8 and S9.

Adaptive Evolution

A total of 2,210 genes were identified that have evidence for variable selective pressure on the bovine lineage and 71 bovine genes were identified with $d_N/d_S > 1$ under model 2 (Table S10). Of these, 40 were also significant with the LRT and have statistically significant evidence of adaptive evolution on the bovine lineage. The bovine specific model described above is conservative in that it assumes that there is variable selective pressure only on the specified bovine lineage. It may be an unrealistic assumption that orthologous genes from the other divergent mammalian species are subject to similar selective pressure. To overcome this assumption, we also compared the null one-ratio model to another model, the free-ratios model, which allows variable selective pressure on all the lineages. An additional 16 of the 71 genes with $d_N/d_S > 1$ on the bovine lineage were found where the free-ratios model was significantly favored ($p < 0.05$). It should be noted that the overall analysis can be susceptible to the draft nature of some of the genome sequences employed, the alignment quality for each gene across such divergent species and issues associated with incorrect gene predictions. Therefore, we manually extracted and aligned the sequences of ten genes with $d_N/d_S > 1$. These genes included *IFNAR2*, *IFNG*, *CD34*, *TREM1*, *TREML1*, *FCER1A*, *IL23R*, *IL24*, *IL15* and *LEAP2*. Of these, six (*IFNAR2*, *CD34*, *TREM1*, *FCER1A*, *IL24* and *IL15*) were confirmed as significant in at least one analysis model. Candidate genes with evidence of positive selection from this and other genome-wide analyses require additional future analyses to confirm these signatures.

Genes that have $d_N/d_S > 1$ on the bovine lineage and are significant under either model 2 or the free-ratios model were found to be associated with a range of GO biological processes including

immune response (IL24, IL15, IL23R, LEAP2, TREM1), cell adhesion (CD34), transcription (CBX7, HIST1H1C, ZNF771), and lipid metabolism (FABP6, LIPE, PNPLA4). 108 GO Molecular Function and 130 GO Biological Process terms had significant MWU p-values ($p < 0.05$), however, only five molecular function terms were significant ($FDR < 0.1$) after correction for multiple testing with the Benjamini and Hochberg correction for the false discovery rate (FDR). The latter terms included: *extracellular-glutamate-gated ion channel activity; ionotropic glutamate receptor activity; sodium channel activity; diacylglycerol binding; kainite selective glutamate receptor activity.* Only two PANTHER ontology terms (*Glutamate receptor* and *Cation transport*) were significant after p-value correction.

No over-representation of any InnateDB pathway was apparent in this dataset. InnateDB was also used to investigate and visualize the molecular interaction networks of the genes that had evidence of positive selection and their interacting partners (Fig. S19). Two different pairs of genes, with $d_N/d_S > 1$ on the bovine lineage, were observed to interact with each other. *SNRPD1* and *SNRPD2*, which are both small nuclear ribonucleoproteins involved in spliceosome assembly, were found to be interacting partners. An interaction between glycosylphosphatidylinositol anchor attachment protein 1 (*GPA1*) and the eukaryotic translation initiation factor (*EIF3E*) was also observed but it should be noted that this interaction was only supported by a single yeast 2-hybrid experiment.

Adaptive Immunity

Annotation of the T-cell receptor (TCR) alpha and beta loci was problematic due to poor scaffolding in the regions of interest. The total annotated genes for TCR alpha in the Btau4.0 assembly is 71 functional variable regions, 38 joining regions and a single constant region. This compares with 54 TCR alpha variable regions in humans and 98 in mice including some pseudogenes. Both humans and mice have 61 joining regions. The 38 joining region genes identified in cattle are most likely an incomplete list.

The repertoire of TCR beta gene segments in cattle has been expanded by extensive duplication. A total of 133 variable genes were identified with subgroups VB9 and VB6 comprising 35 and 40 members, respectively. 79 of the VB segments appear to be functional. In addition, there are three clusters of D, J and C genes, which total 3 DB, 17 JB and 3 CB genes. The bovine TCR beta locus shares many similarities with human and mouse TCR beta loci but may differ in organization and it contains substantially more gene segments. Phylogenetic analysis suggests that the expansions of certain VB subgroups are distinct between humans and cattle raising interesting questions about the evolutionary pressures influencing this immunologically important locus.

In contrast to the genomic organization of human and mouse TCR gamma genes, in cattle these genes are found at two loci on BTA4 with each locus containing three Variable-Constant cassettes. The bovine TCR delta genes are found within the TCR alpha locus like in human and mice. In addition, bovine *TRDV4* was found downstream of the *TRDC* gene and in an inverted orientation, similar to orthologous human and mouse genes. In contrast to the single *TRDV1* gene found in humans and mice, an expansion to 52 genes was observed in cattle for sequences belonging to the *TRDV1* family.

Workshop Cluster 1 (*WC1*) genes encode a family of scavenger receptor cysteine-rich (SRCR) proteins found exclusively on $\gamma\delta$ T cells in cattle, sheep and swine but not humans or mice. There are at least 13 *WC1* genes distributed within two regions on BTA5.

Analysis of the Btau3.1 assembly identified 63 immunoglobulin lambda variable (*IGLV*)

and 22 immunoglobulin kappa variable (*IGKV*) genes. 33 genes (25 *IGLV* and 8 *IGKV*) are apparently functional. This is significantly lower than the number of functional light chain variable genes in human (33 lambda and 44 kappa genes totaling 77 genes) or mouse genomes [8 lambda and 97 kappa genes totaling 105 genes, data from IMGT database (*S114*)]. The heavy chain locus was not annotated as most of it was missing from Btau3.1. However, the available data on the light chain genes suggest that post-recombinatorial mechanisms might contribute to generation of the bovine pre-immune antibody repertoire.

The organizational features of the bovine Major Histocompatibility Complex (MHC) (called BoLA in the cattle) and the MHCs of other ruminants are unique in that genes of the class II region occur in two segments, called IIa and IIb, about 20 Mbp apart on BTA23. The BoLA region in the assembly spans scaffolds 10, 11, 13, 30, 31 and 39 and contains a total of 154 predicted genes, which compares well with MHC regions in other mammalian species. Annotation revealed that there are 60 genes within the BoLA class I region, 38 genes within the BoLA class IIa and IIb region, and 56 genes within the BoLA class III region. The haplotype organization of the BoLA classical Class I genes is very different from that of the human and mouse with different numbers of classical class I genes depending on haplotype (*S115-S117*). This resulted in great difficulty in accurately assigning genes to gene models and only three classical class I loci were identified on the basis of only homology with known bovine class I cDNA sequences (*S118*). Two of the genes identified in the assembly encoded novel proteins homologous to MHC class I heavy chains. Three *MIC* genes were identified in the assembly although there is independent evidence for four *MIC* genes (*S119*), suggesting that one *MIC* gene may be missing from the assembly. A cluster of three non-classical class I genes were identified among the *MIC* genes, a distribution quite different from human (*S119*).

The Class IIb region is similar to previously published sequence data (*S120*), with the inclusion of splice variants within genes: *TAP2* (*S121*) and novel splice variants for *PSMB8* and *RXRb*. There are multiple duplications in the class IIa region making it less straightforward to annotate with human and mouse cDNA and ESTs. The order of classical class II genes is: *BoLA DQA2*, *BoLA DQA2-1*, *BoLA DQB*, *BoLA DQA*, *BoLA DRB3*, *BoLA DRA*. Thus, it is clear that the reference animal L1 Dominette 01449 had a haplotype with duplicated *DQ* genes and may even have three *DQA* loci, although it is more likely that the reference animal was heterozygous. The single *DRB3* sequence was identified as *DRB3*1002* (*S118*).

Metabolic Reconstruction and Gene Losses

The results of metabolic reconstruction demonstrate a strong degree of conservation among the comprehensive set of genes involved in core mammalian metabolism. A total of 116 pathway holes (“missing enzymes”), or 14% of the total reactions in pathways, were identified (Tables S11, S12 and S13). The fraction of complete pathways and of missing enzymes is similar to that obtained for the human and mouse genomes and is a reflection of the quality of both the assembly and genome annotation (*S122*).

Evidence for the loss of *PLA2G4C* (EC 3.1.1.4) is as follows (Fig. S22). Flanking genes in the human genome are located in a single cattle contig (AAFC03078605) that has high quality sequence coverage. The interval between genes that flank *PLA2G4C* (phospholipase A2, group IVC; cytosolic and calcium-independent) is shorter in the cattle, dog and horse genomes (3.6 kbp on average) than in rodent, primate and platypus genomes (77.8 kbp on average). Phylogenetic foot-printing with 25 high and low coverage mammal genome sequences (Table S13) with TBLASTN of the human *PLA2G4C* protein sequence against WGS reads revealed strong

homology scores in Euarchontoglires: Primate (human, chimpanzee, macaque, orangutan, rhesus, but not lemur), Rodentia (mouse, rat guinea pig), Lagomorpha (pika but not rabbit) and Scandentia (treeshrew). In Afrotheria, the gene was found in Afrosoricida (tenrec) but not Proboscidea (elephant). *PLA2G4C* was found in the basal Didelphimorpha (opossum) and Monotremata (platypus) genomes, but not in a Xenarthran (armadillo). In contrast, none of the Laurasiatherian clades, i.e., the Ferungulates (cattle, dog, cat, horse), Chiroptera (bat) or Eulipotyphla (hedgehog) had any evidence for the gene. On the basis of these results, we can conclude that *PLA2G4C* was present in the ancestor of all mammals given its presence in both opossum and platypus. The gene was deleted after the divergence of Laurasiatheria and Euarchontoglires from the Boreoeutherian ancestor and may also have been deleted independently in Xenarthra and in some lineages of Afrotheria, Lagomorpha and Primates. Although several species have only low coverage sequence, which may obfuscate the results, the phylogenetic distribution pattern of *PLA2G4C* in mammals, particularly those with high coverage sequence, is consistent with its deletion approximately 87-97 Mya in the Laurasiatherian lineages. These results account for the absence of the gene in the cattle genome.

Lysozyme Genes

Bovine C-type lysozyme genes are listed in Table S14, and a neighbor joining tree is shown in Fig. S23.

Supporting Figures

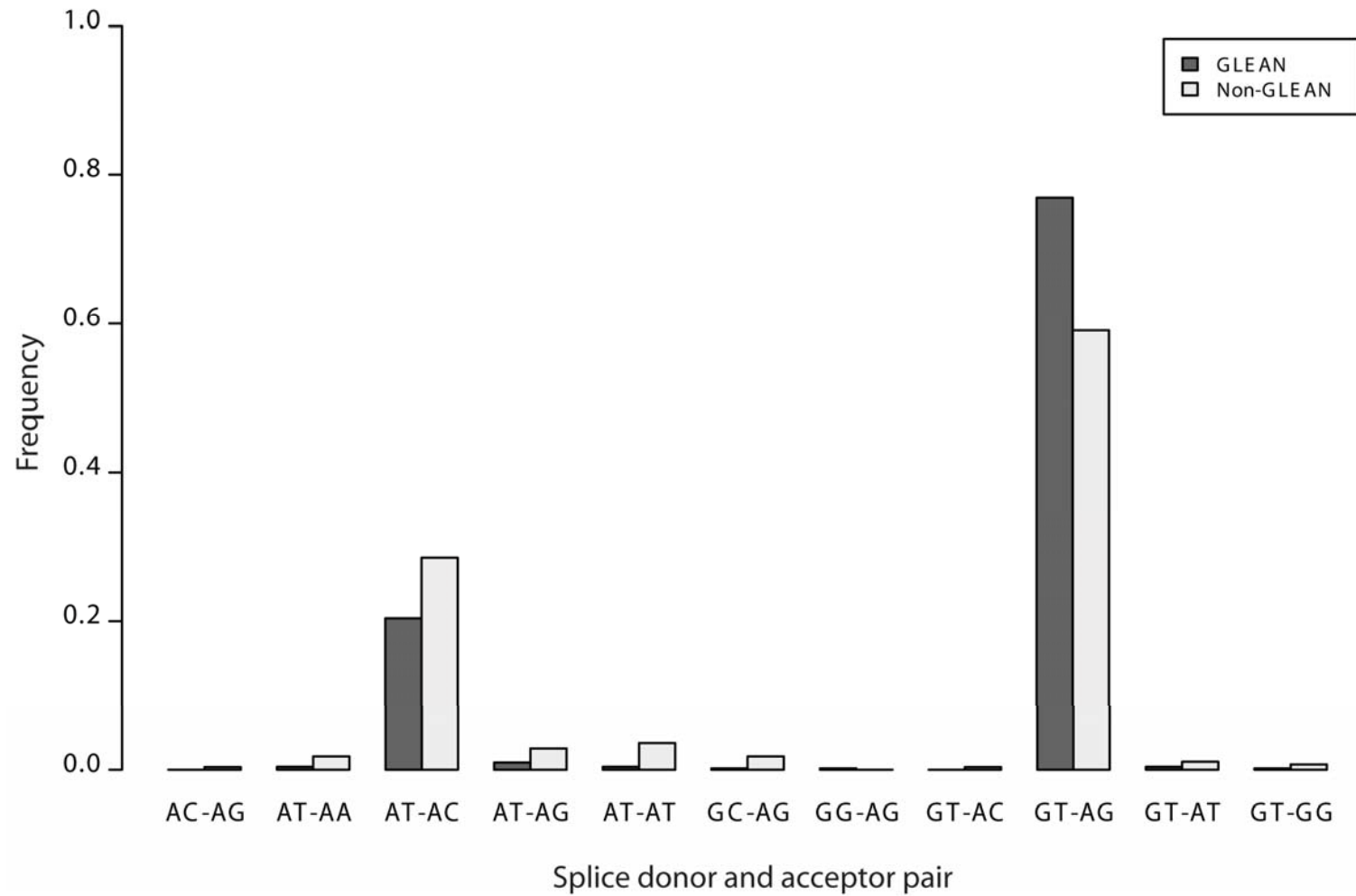


Fig. S1. Frequencies of intron donor and acceptor combinations according to whether or not they were predicted by GLEAN5.

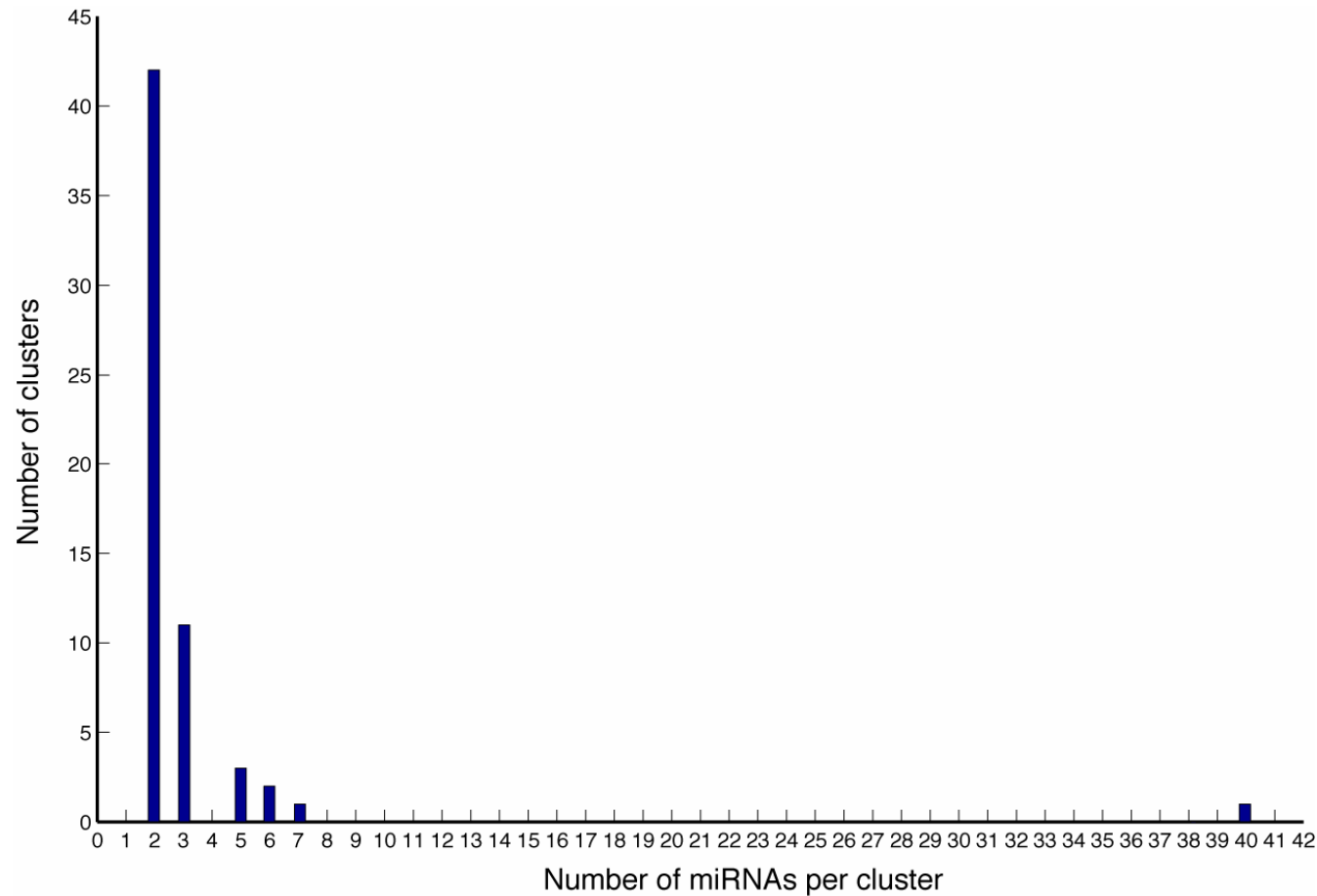


Fig. S2. Number of clusters of miRNA containing two or more miRNA. A total of 496 bovine miRNAs were grouped into 298 homolog families. miRNAs were considered to form a cluster if they were separated by less than 10 kbp. About half of the bovine miRNA occur in genomic clusters containing 2 to 7 miRNA genes with the exception of one large 43 kbp long cluster on BTA21:59,594,412-59,637,311 (Btau3.1) containing 40 miRNAs.

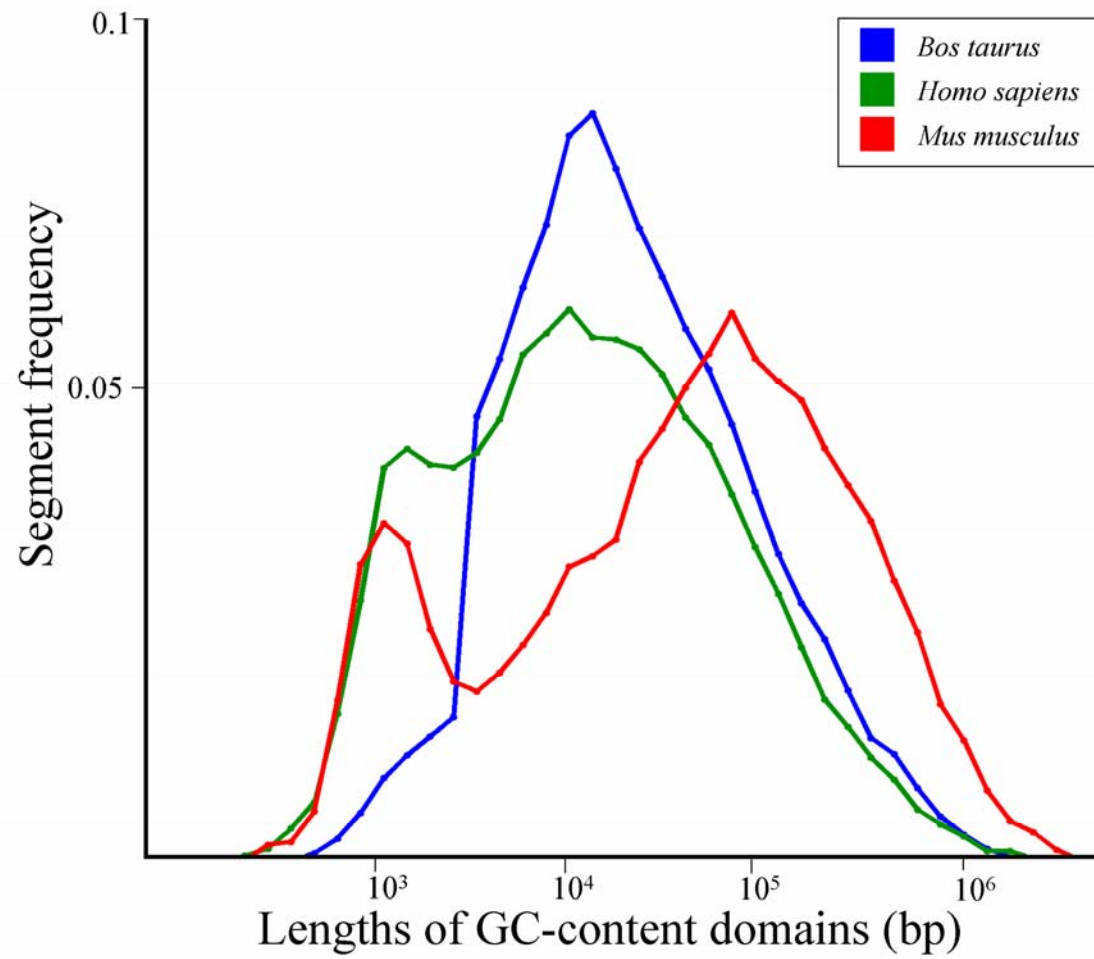


Fig. S3. The frequency of GC-content domain segments in cattle, human, and mouse.

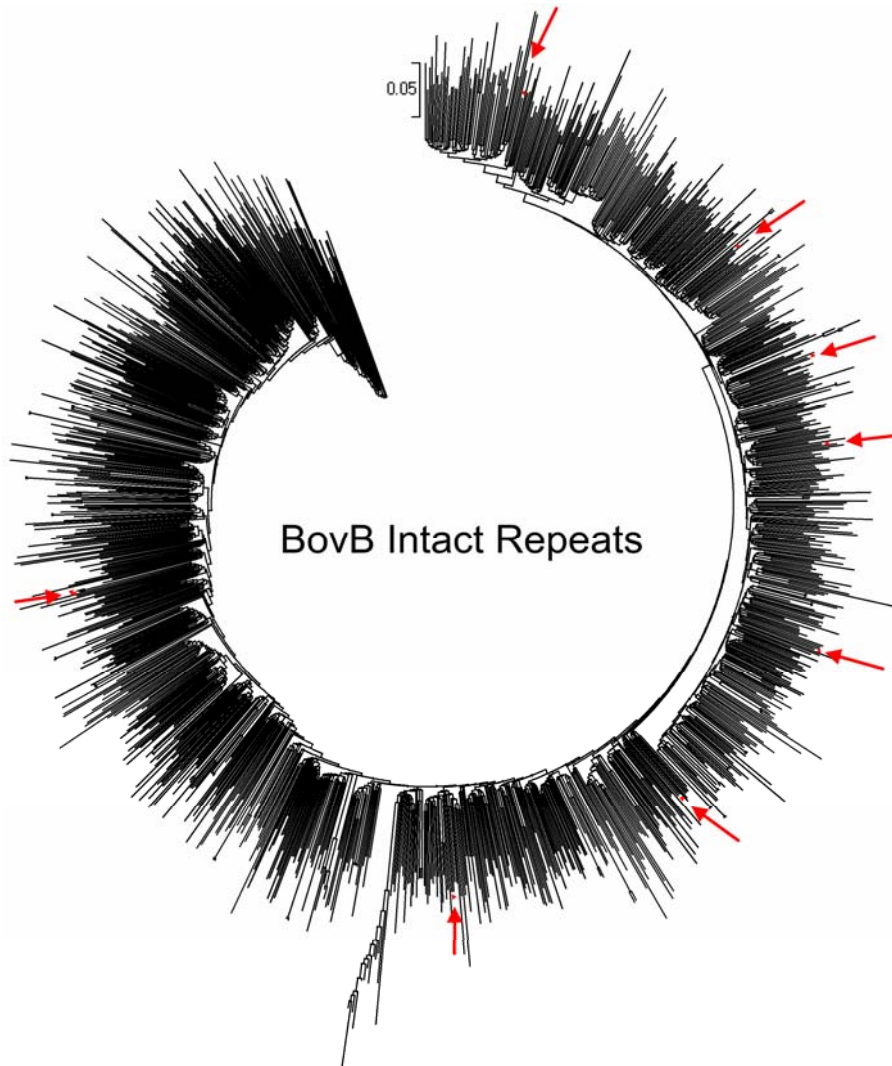


Fig. S4. Maximum likelihood tree derived from global alignments of intact/full length BovB intact repeats. Red arrows/triangles indicate potentially active LINEs of the basis of their intact ORF content.

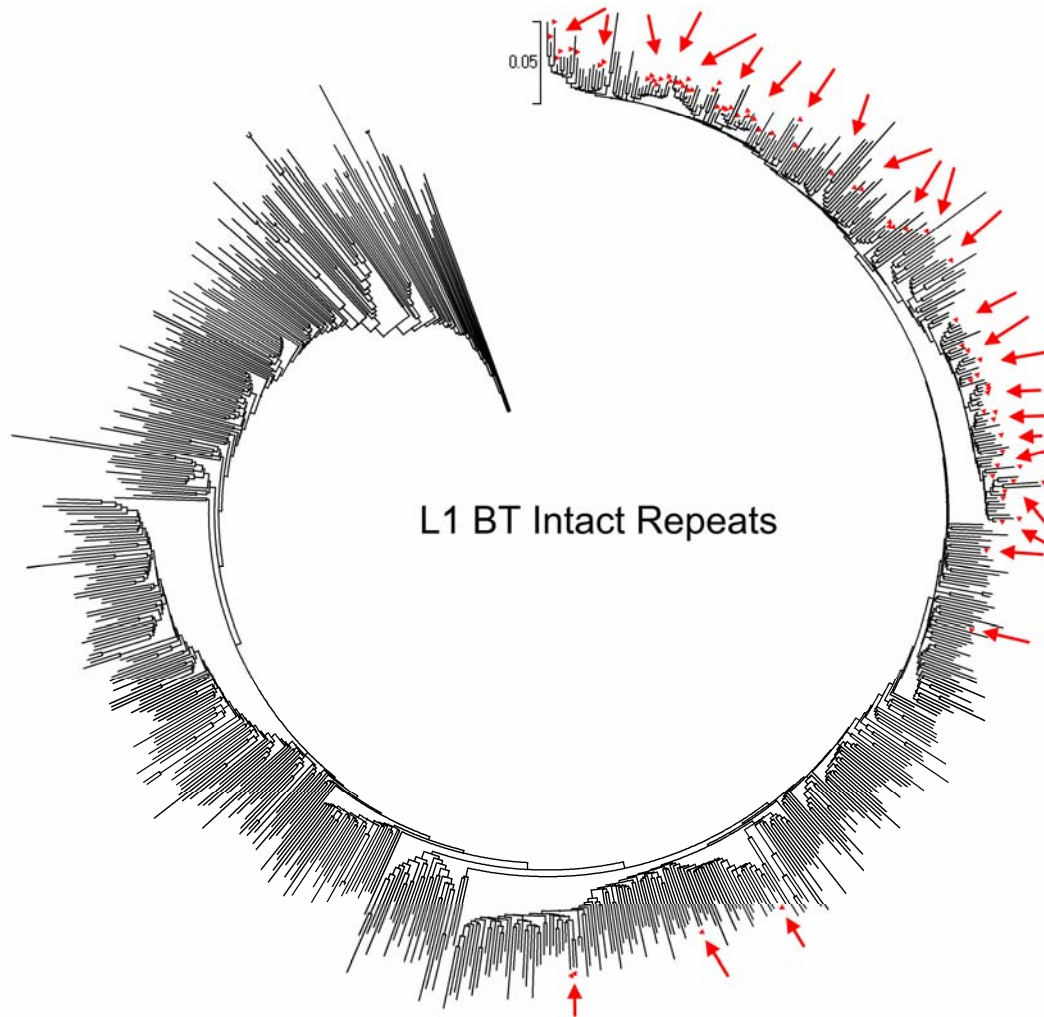


Fig. S5. Maximum likelihood tree derived from global alignments of intact/full length L1_BT intact repeats. Red arrows/triangles indicate potentially active LINEs on the basis of their intact ORF content.

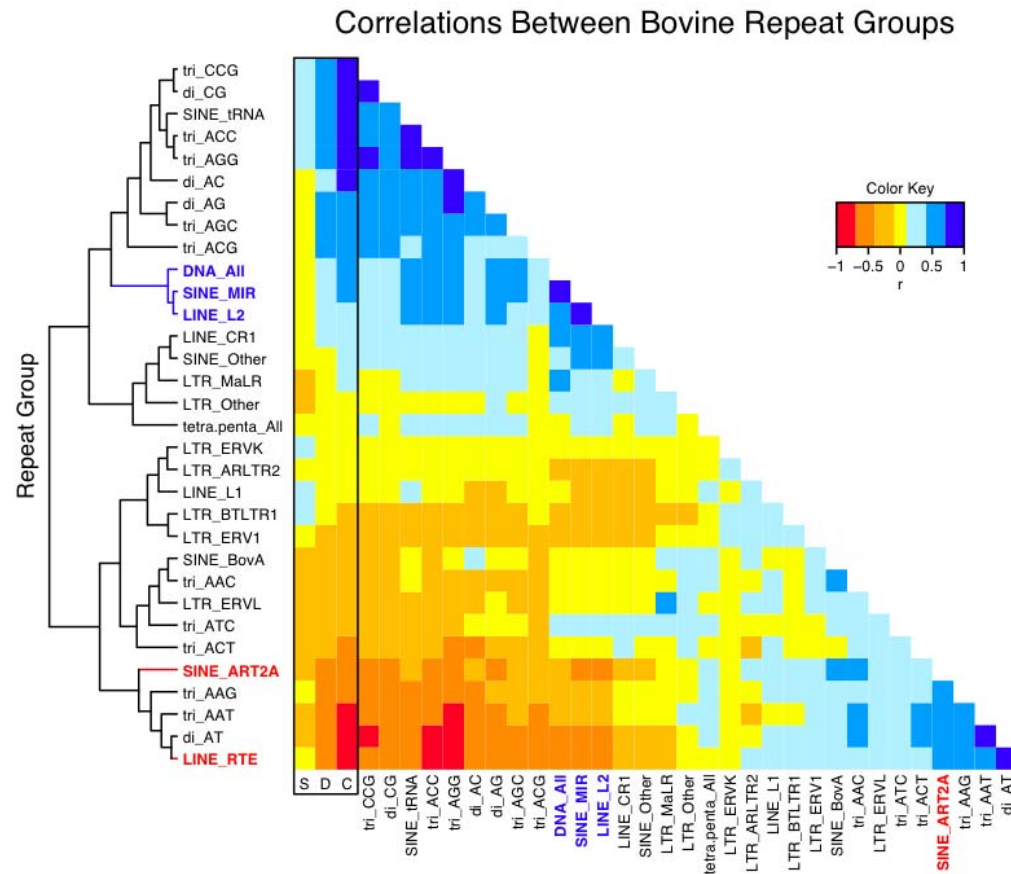


Fig. S6. Correlation analysis of repeat groups. Pairwise correlations among the repeat groups and between the repeat groups and segmental duplication (column S), gene density (column D), and GC content (column C). Repeat groups were clustered on the basis of all their correlations. Yellow cells have non-significant correlations ($>5\%$ 2-tailed test after Bonferroni correction). Blue cells indicate significant positive correlations, and the orange/red cells indicate significant negative correlations.

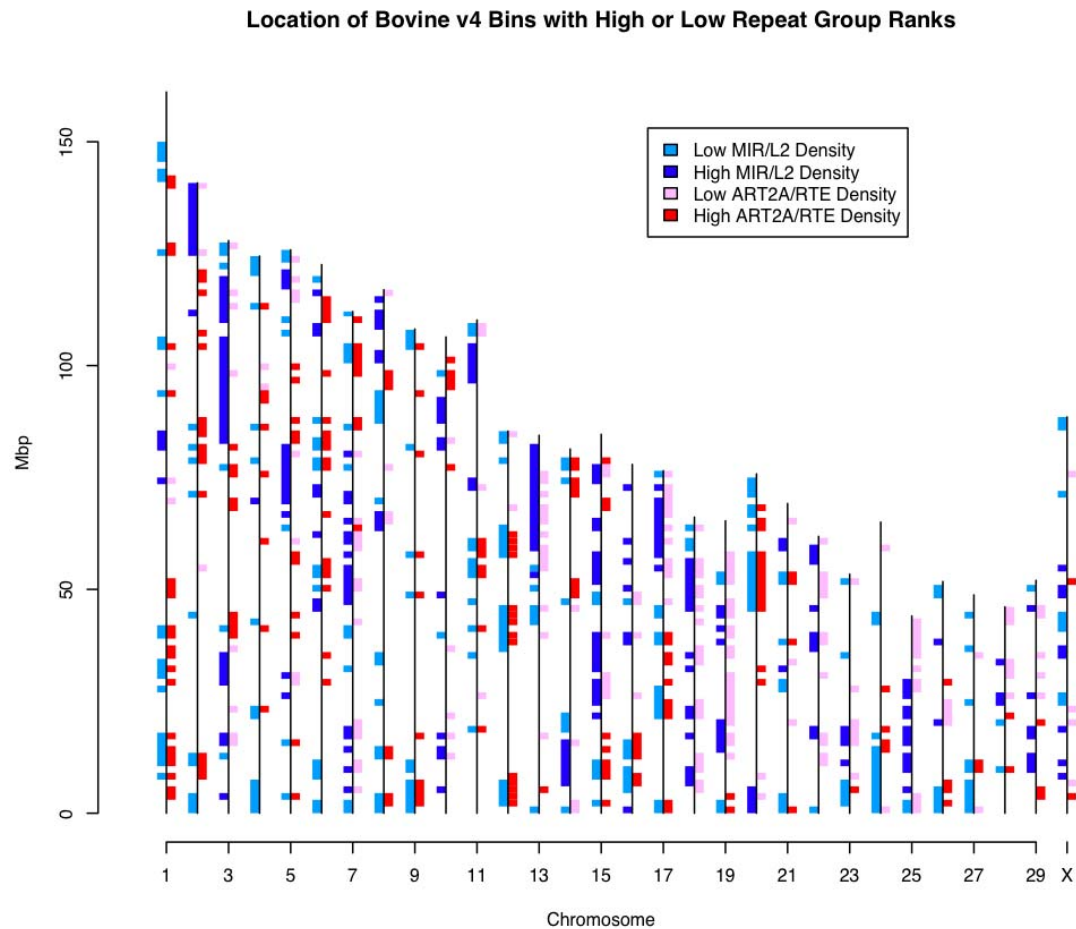


Fig. S7. Spatial distribution of ancient and new extreme density bins for repeats. The top and bottom 5% tails of the bins for L2/MIR and BovB/Art2A (Art2A/RTE) correlations were identified based on an expected random distribution in Btau4.0. High density L2/MIR regions often form contiguous blocks and these regions never overlap with high density Art2A/RTE regions. High density L2/MIR blocks may correspond to ancestral mammalian genome domains which have not been invaded by new, BovB derived repeats.

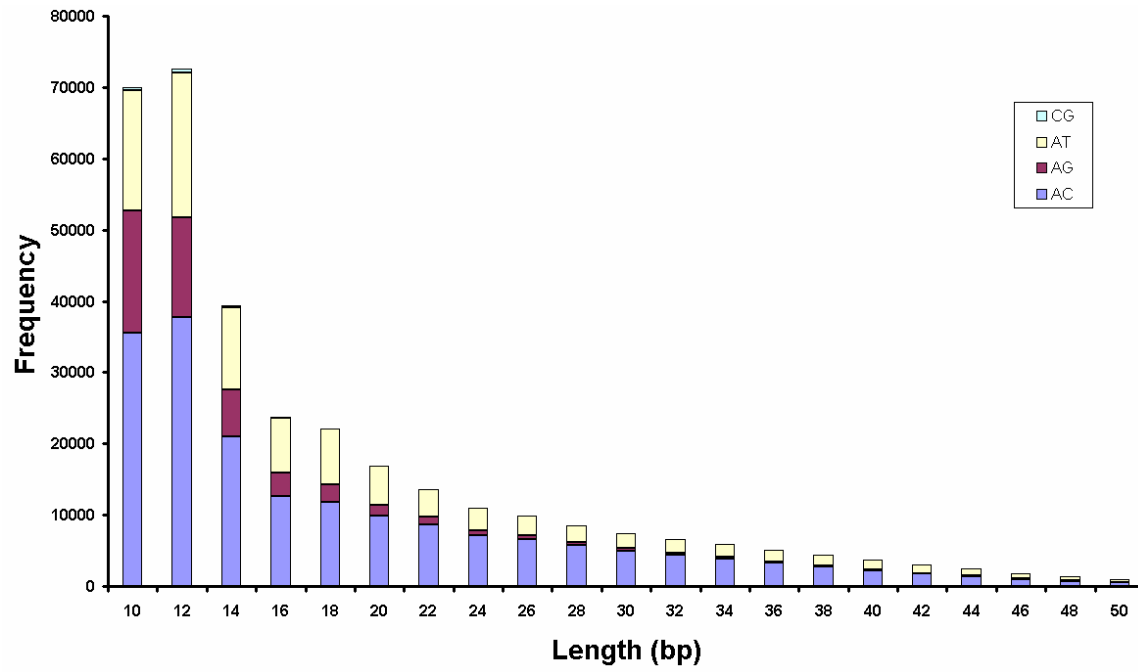


Fig. S8. Frequency of dinucleotide simple sequence repeats of varying lengths in the bovine genome.

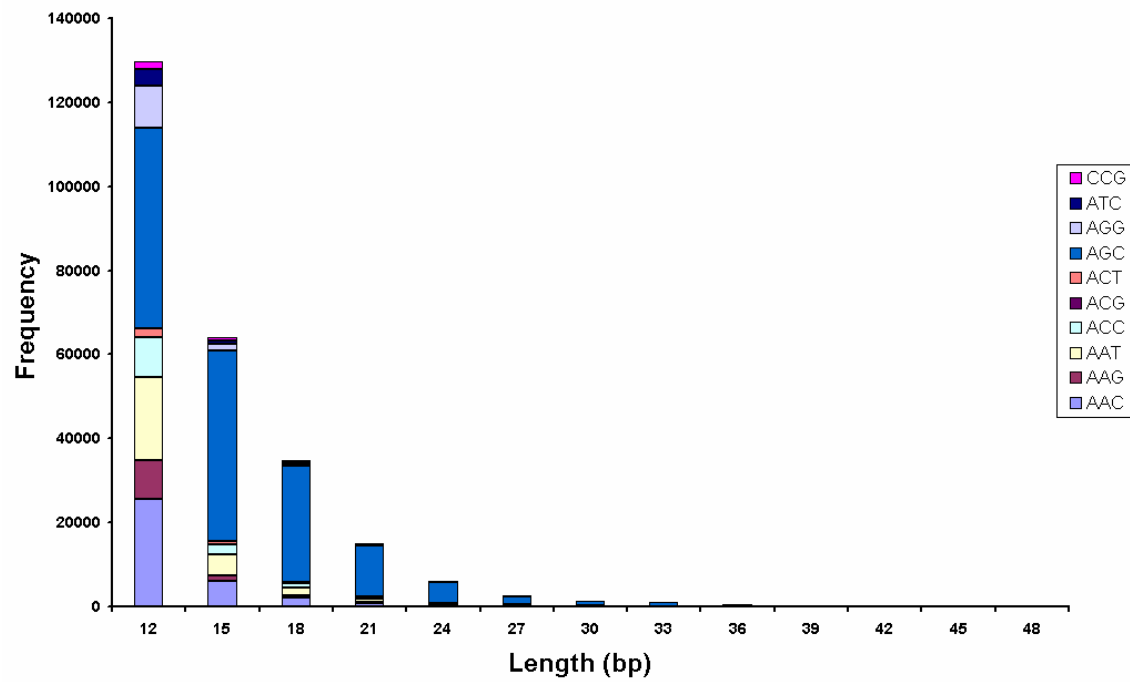


Fig. S9. Frequency of trinucleotide simple sequence repeats of varying lengths in the bovine genome.

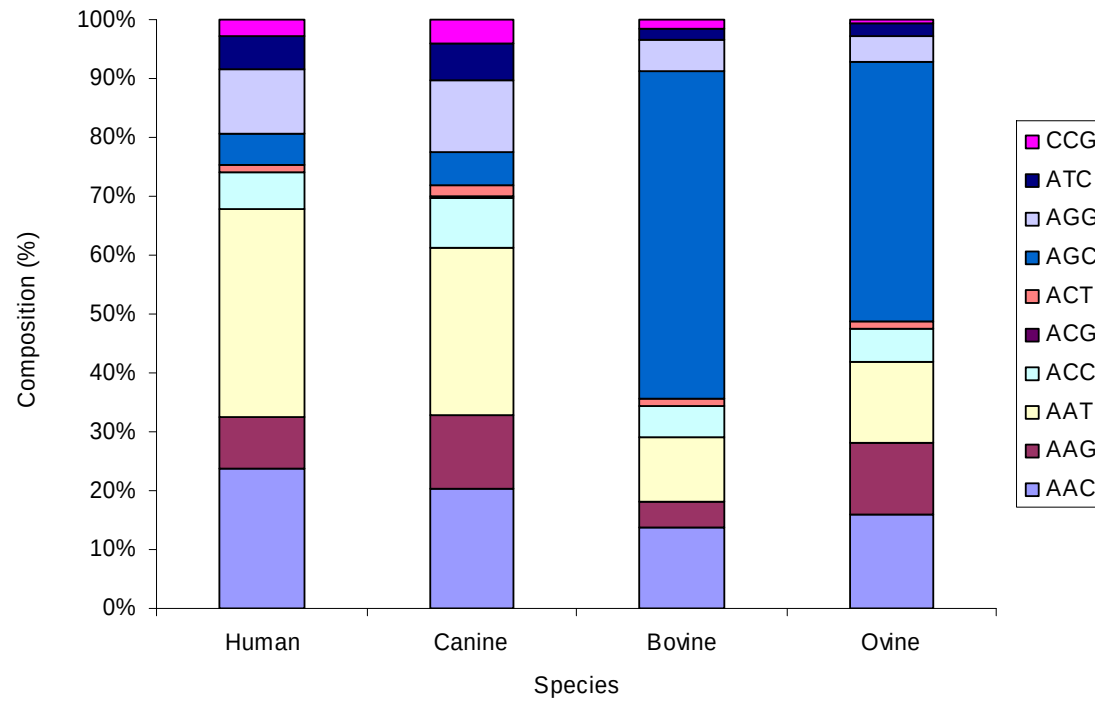


Fig. S10. Comparative compositions of trinucleotide simple sequence repeats (SSRs) in four mammalian species.

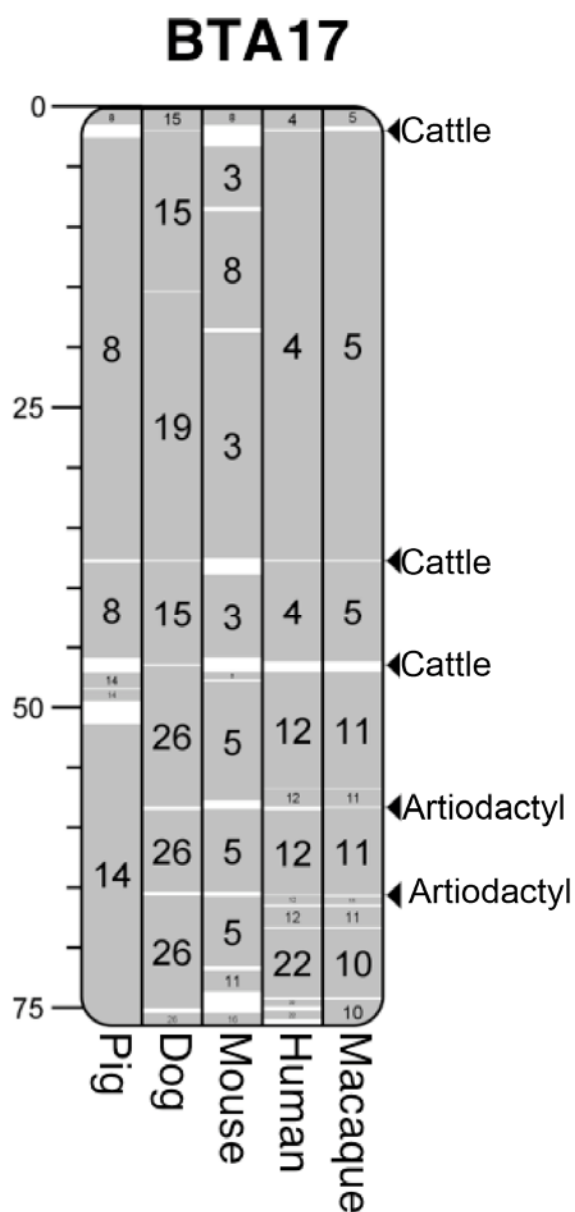


Fig. S11. Cattle-specific and artiodactyl breakpoint regions in BTA17. Homologous syntenic blocks (HSBs) defined using BTA17 (Btau4.0) as the reference genome are indicated by grey shading. Macaque (RheMac2, build 2), human (HSA36, build 36), dog (CFA2, build 2), mouse (MMU9, build 37) and pig (physical map) were used for pair-wise comparisons. White areas correspond to evolutionary breakpoint regions. Arrows to the right of the chromosome ideogram indicate positions of cattle-specific and artiodactyl breakpoint regions. Cattle-specific rearrangements are identified as breakpoints that appear common to chromosomes of all other species when overlaid on the cattle genome. Artiodactyl-specific rearrangements are specific to the chromosomes of pigs and cattle, i.e., they appear at the same location in the genomes of all other species except pigs when aligned to the cattle genome. The alignments were visualized using the Evolution Highway comparative chromosome browser (*S111*).

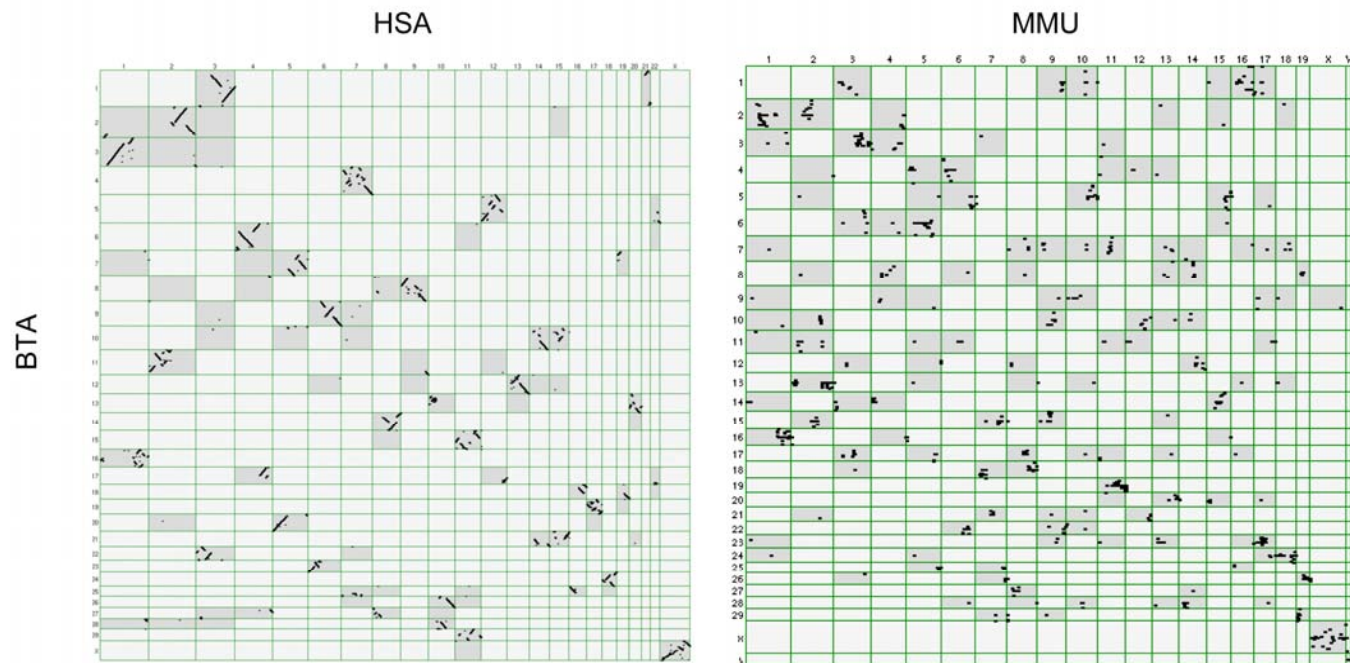


Fig. S12. Oxford grids showing chromosome comparisons between species. An Oxford grid showing the extent of chromosomal conservation between cattle (BTA) and humans (HSA) was drawn from 46,947 cattle BAC-end sequences showing high homology with sequence from build 36 of the human genome. An Oxford grid also shows the chromosomal conservation between the cattle and mouse genome sequences. Expandable, zoomable, hyperlinked versions of these grids and similar grids for cattle versus pig, dog, macaque and mouse are available at (S112).

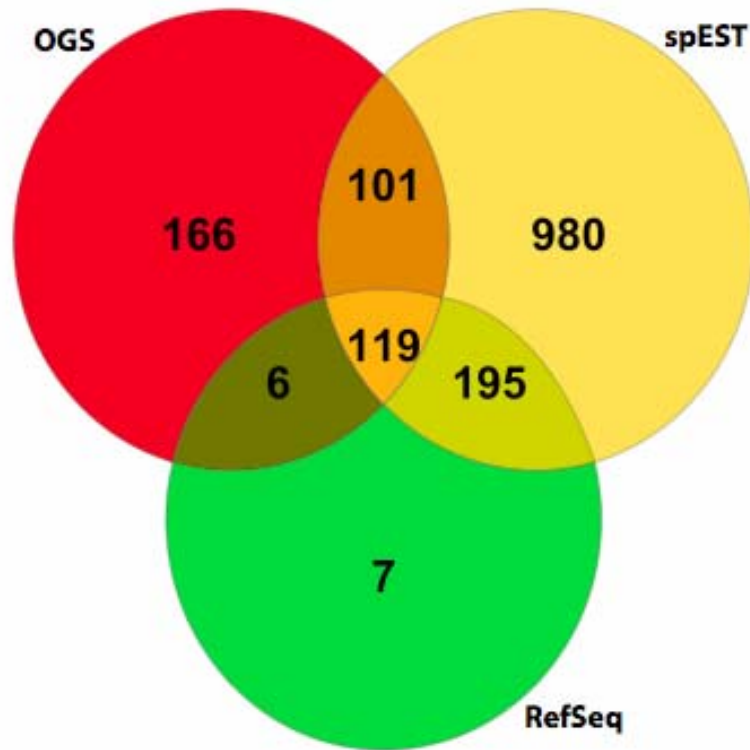


Fig. S13. Venn diagram showing bidirectional promoters mapped in the cattle genome. Annotation sources were OGSv1 (red), spliced ESTs (yellow), and RefSeq genes (green).

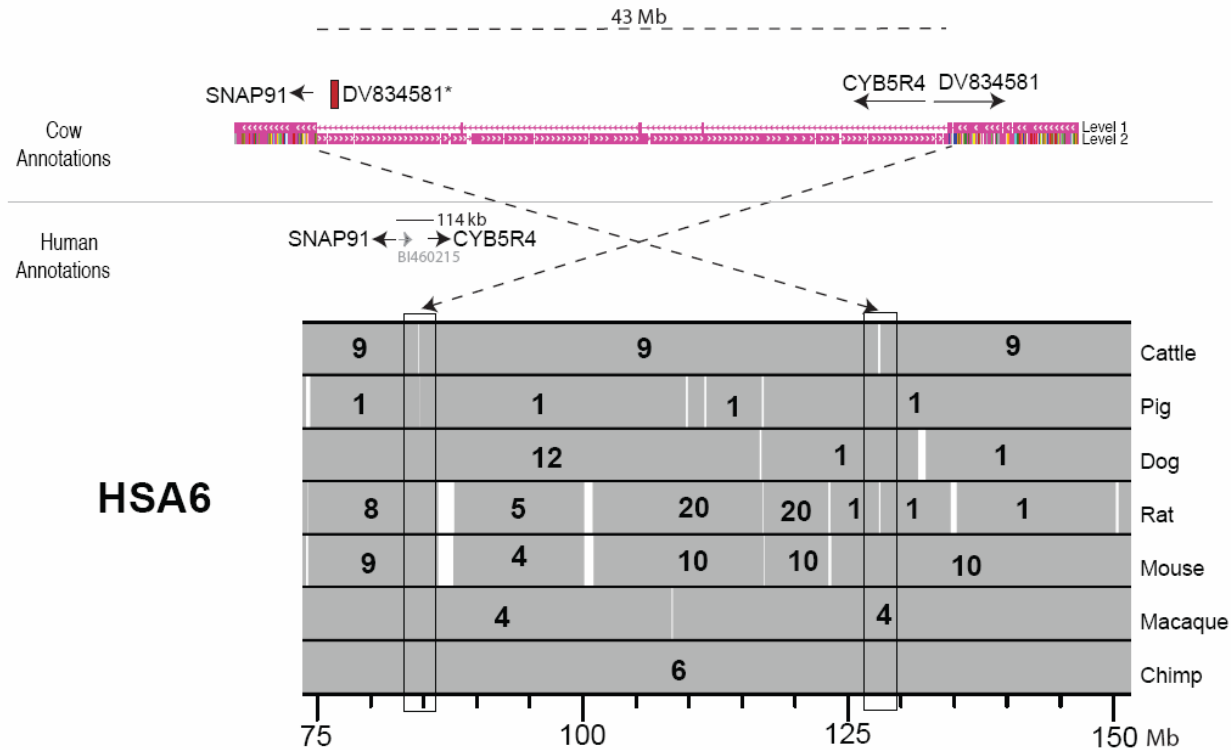


Fig. S14. An inversion of 43 Mbp in BTA9 is shown relative to human chromosome 6 (HSA6), which maintains an ancestral organization of the region as seen from the comparison of HSA6 to the orthologous regions of pig, dog, rat, mouse, macaque, and chimp chromosomes (lower panel). The inversion in BTA9 is illustrated with the “human net” track downloaded from the UCSC Genome Browser (upper panel) (S51). The *SNAP91* gene flanks the region of interest in both the cattle and human genomes. In human, *SNAP91* is separated from *CYB5R4* by only 114,000 bp. In cattle, an upstream, alternative promoter of *CYB5R4* acts as a bidirectional promoter for a putative RNA-gene, *DV834581*. This gene is a cattle-specific gene, similar to a region of homology near the *SNAP91* gene (*DV834581**). The rearrangement was identified by mapping the *CYB5R4* bidirectional promoter of cattle to the human genome, which revealed that the orthologous gene pair did not exist in human.

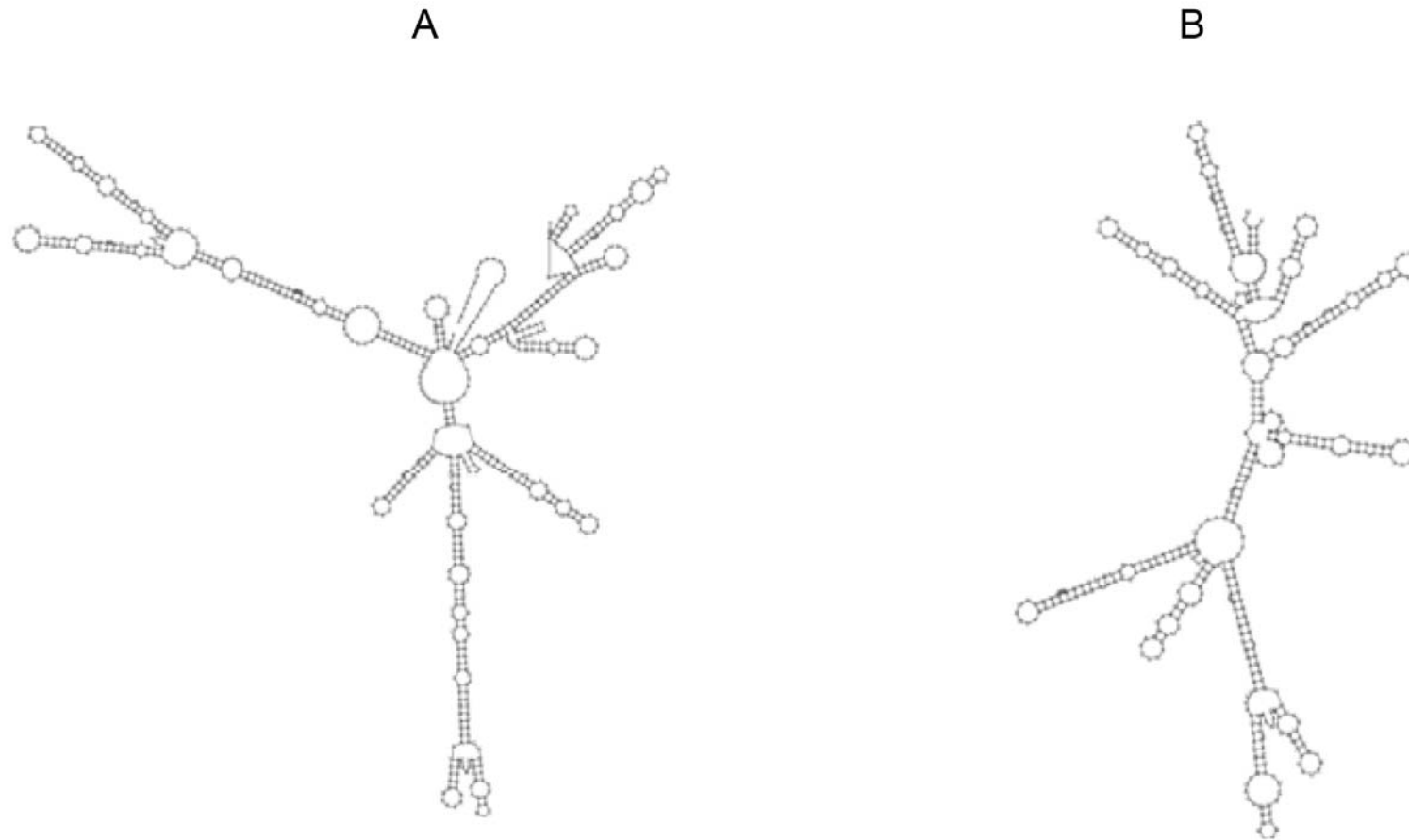


Fig. S15. RNA secondary structure of the putative non-coding gene partner to *CYB5R4* in cattle. Two isoforms (A and B) of this non-coding gene are produced by alternative splicing and both have been subjected to analysis of RNA secondary structure with RNAfold (S43). No orthologous gene was found in primates or rodents.

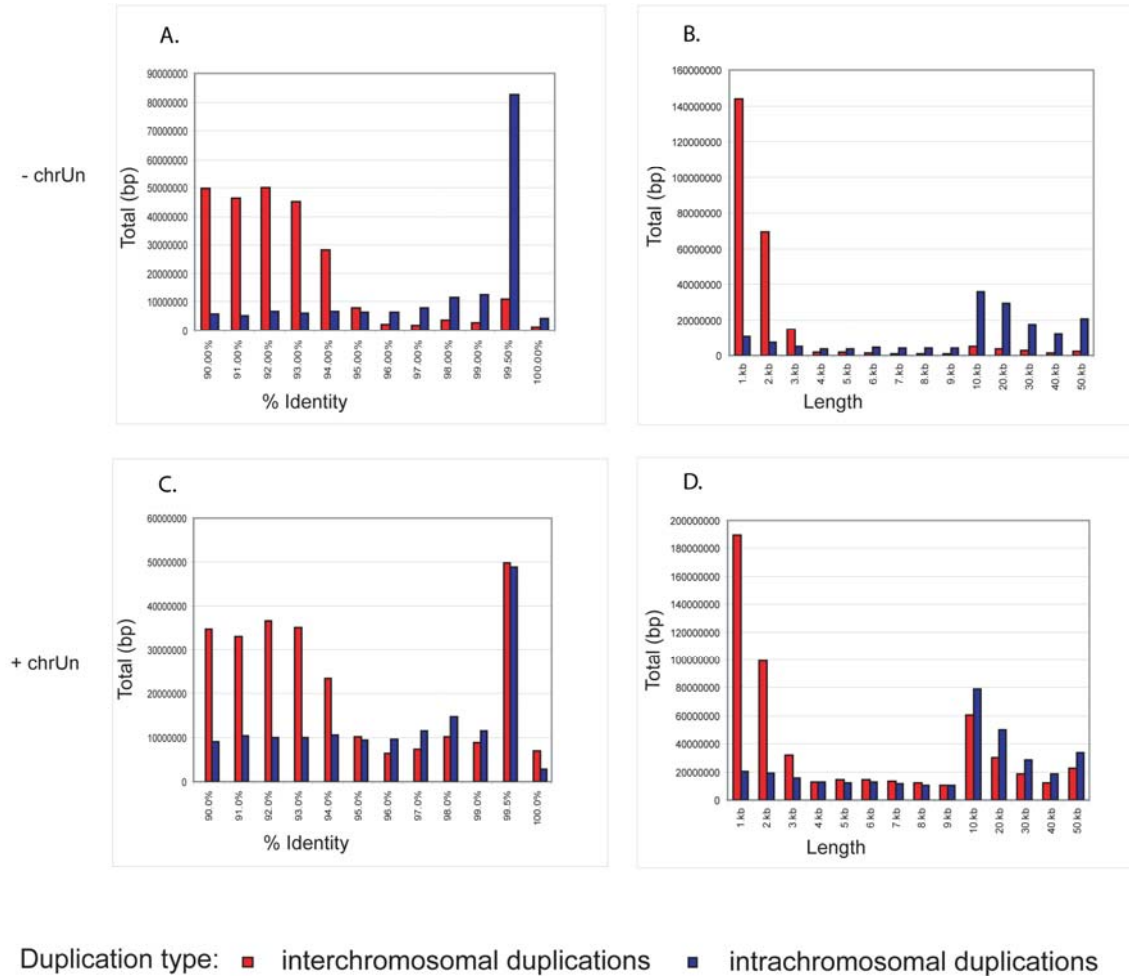


Fig. S16. Percent identity and length distribution of cattle segmental duplications. The total number of aligned bp for interchromosomal and intrachromosomal segmental duplications (SDs) was plotted as a function of the percent identity and length as identified by the whole-genome analysis comparison method (WGAC) of the Btau4.0 assembly. (A) Percent identity distribution without consideration of sequence that could not be mapped to a chromosome (chrUn). (B) Length distribution without chrUn. (C) Percent identity distribution including chrUn. (D) Length distribution including chrUn.

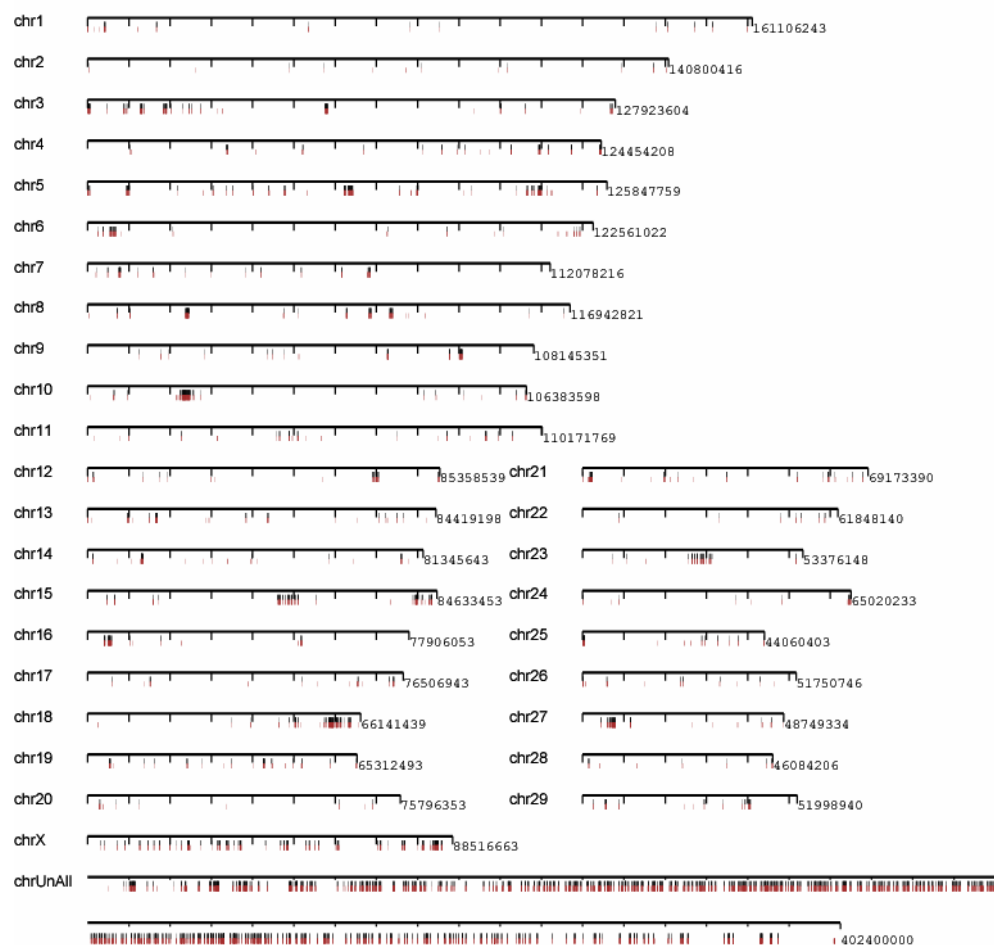


Fig. S17. Map of cattle segmental duplications. The locations of large (>10 kbp) SDs defined by WGAC (black) and WSSD (brown) methods are shown mapped onto bovine chromosomes. Each tick mark represents 10 Mbp. Large clusters of SDs are clearly identifiable within the genome although in some cases, the precise sequence structure of these regions has not yet been resolved.

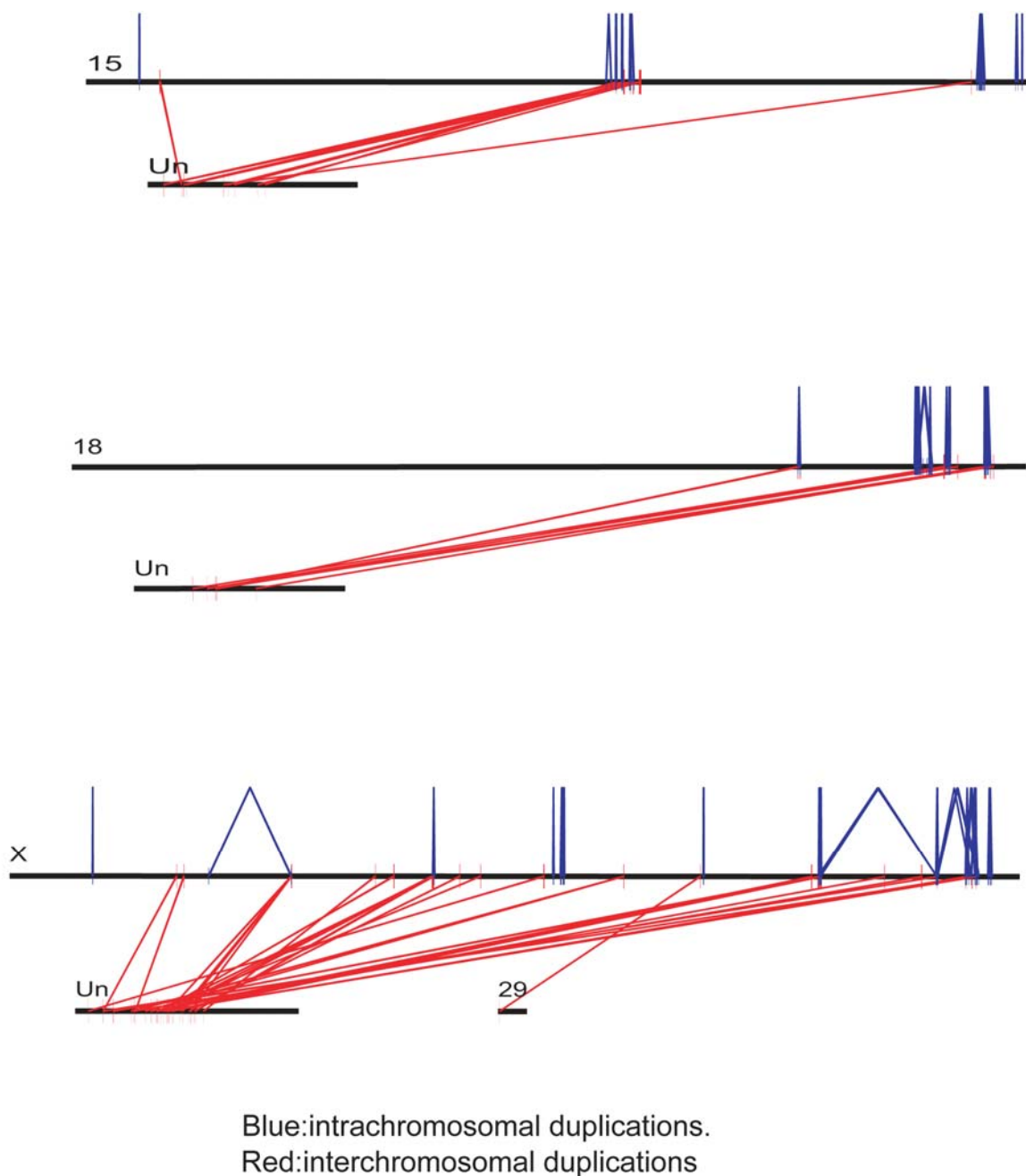


Fig. S18. Intrachromosomal tandem segmental duplications. Large (>20 kbp) interchromosomal and intrachromosomal duplication pairwise alignments (identified by both WSSD and WGAC) are depicted for three chromosomes (BTA18, BTA15 and BTAX). A preponderance of tandem duplications as opposed to interspersed intrachromosomal duplications is shown. Many of the pairwise alignments yet unassigned to a chromosome may reflect missing tandem duplications for these regions. Red, interchromosomal duplications; blue, intrachromosomal duplications. Un, chromosome unknown (unassigned scaffolds).

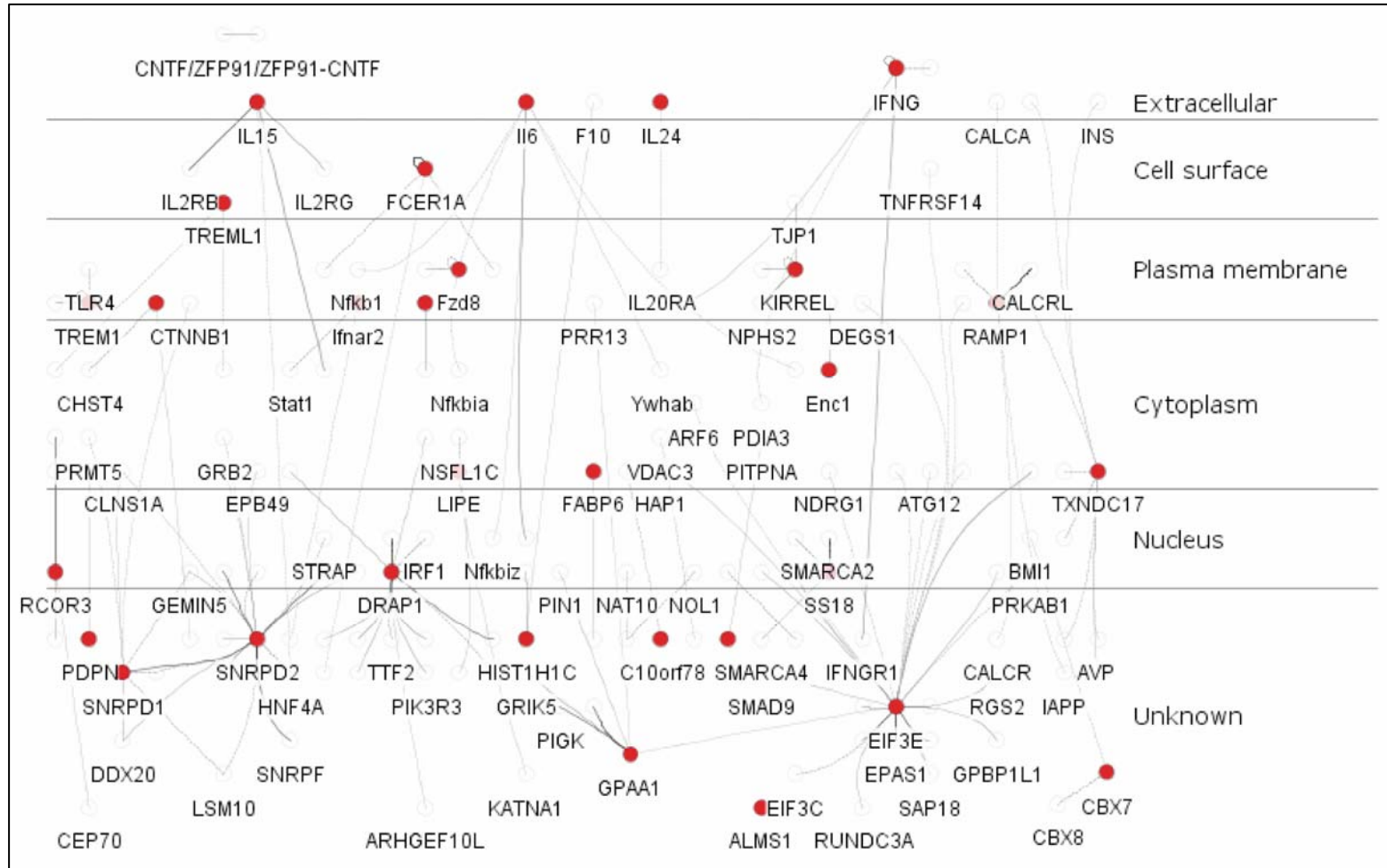


Fig. S19. Molecular interaction networks of the genes that had evidence of positive selection and their interacting partners. InnateDB (S86) was used to investigate and visualize the molecular interaction networks of the genes which had evidence of positive selection and their interacting partners.

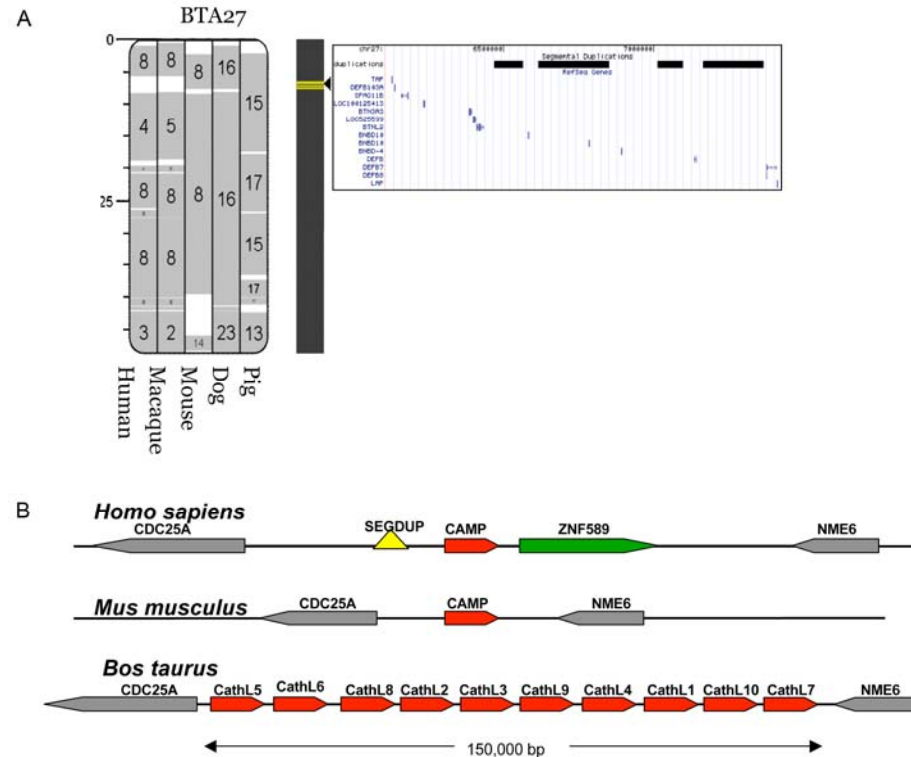


Fig. S20. Examples of changes in the comparative organizations of innate immune related genes. (A) A cattle-specific evolutionary breakpoint region (EBR) in BTA27 showing co-location of segmental duplications (SDs) with a cluster of β -defensin genes. Homologous synteny blocks (HSBs) defined using BTA27 (Btau4.0) as the reference genome are indicated by grey shading. Macaque (RheMac2, build 2), human (HSA36, build 36), dog (CFA2, build 2), mouse (MMU9, build 37) and pig (physical map) were used for pair-wise comparisons. White areas correspond to EBRs. The arrow to the right of the chromosome box showing SDs on BTA27 indicates correspondence between this region and a cattle-specific EBR. The alignments were visualized using the Evolution Highway comparative chromosome browser (*S111*). The SD was defined by both assembly-dependent (WGAC) and assembly-independent (WSSD) methods. A cluster of β -defensin genes is shown in the right hand panel, which is derived from the UCSC cattle genome browser (*S51*). (B) Comparative organization of the cathelicidin gene locus in each of the human, mouse and cattle genomes. Arrows represent genes and their orientations. Introns are included within each gene representation. Human and mouse cathelicidin (*CAMP*) and cattle cathelicidins (*CATHL1* to *CATHL10*) are shown in red while a human SD (SEGDU) is shown by a yellow triangle).

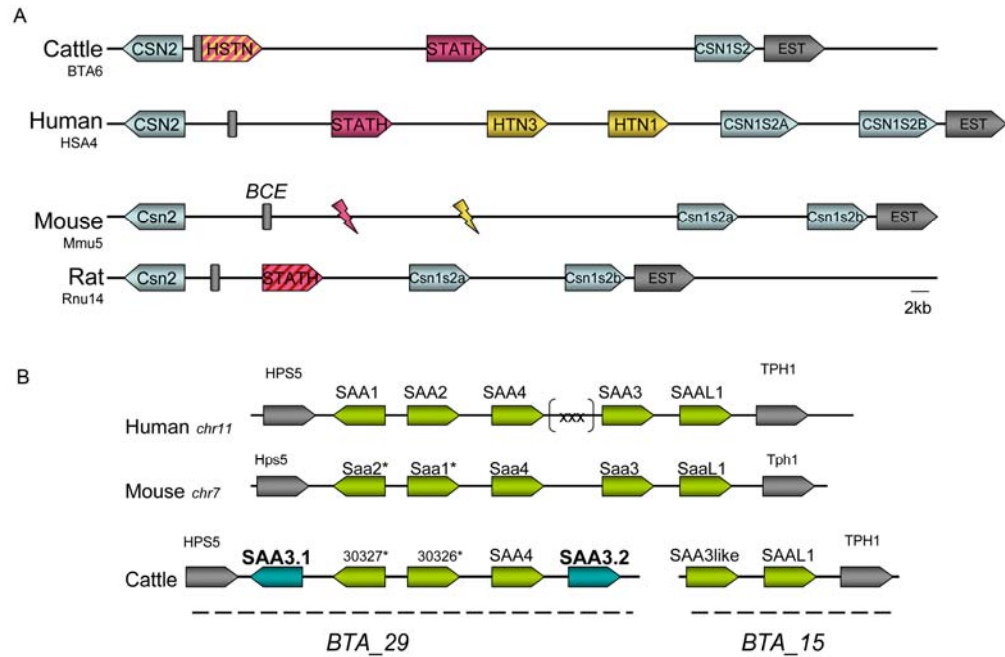


Fig. S21. Comparative representations of the organizations of the serum amyloid A and casein gene clusters in mammalian genomes. (A) Diagrammatic representation of the organization of the casein gene cluster in cattle, human, mouse and rat highlighting the histatin-statherin region. The β -casein gene (*CSN2*), Histatherin gene (*HSTN*), Statherin gene (*STATH*), histatin genes (*HNT1*, *HNT3*), bovine α -S2 casein gene (*CSN1S2*) and two alpha-S2 casein paralogs (*CSN2S2A* and *CSN2S2B*) are colored according to their gene families while the β -casein up-stream enhancer (*BCE*) is colored grey. The broken arrows in the mouse map indicate the locations of remnants of statherin- and histatin-like genes. (B) Serum amyloid A (*SAA*) gene clusters in the human, murine and bovine genomes. Duplication on BTA 29 resulted in *SAA3.1* and *SAA3.2*, which are located about 80 kbp apart and encode protein isoforms with 96% amino acid sequence identity. The duplication is linked to a bovine specific chromosomal breakpoint (BTA29/BTA15) with a *SAA3*-like pseudo gene on BTA15. These rearrangements have also led to the deletion of *SAAL1* and two flanking genes (*THP1* and *SERGEF*) from BTA29, while BTA15 gained *SAAL1* and the flanking genes. The dog and horse *SAA* gene clusters reside in an uninterrupted syntenic region. Three non *SAA*-genes (*MRGPRX3*, *MRGPRX4* and *LOC494141*) are present in the gap (xxx) represented on the human diagram between *SAA4* and *SAA3*. Two new bovine paralogs are represented by 30327* (ENSBTA00000030327) and 30326* (ENSBTA00000030326). The arrow associated with each gene shows its orientation and includes all introns.

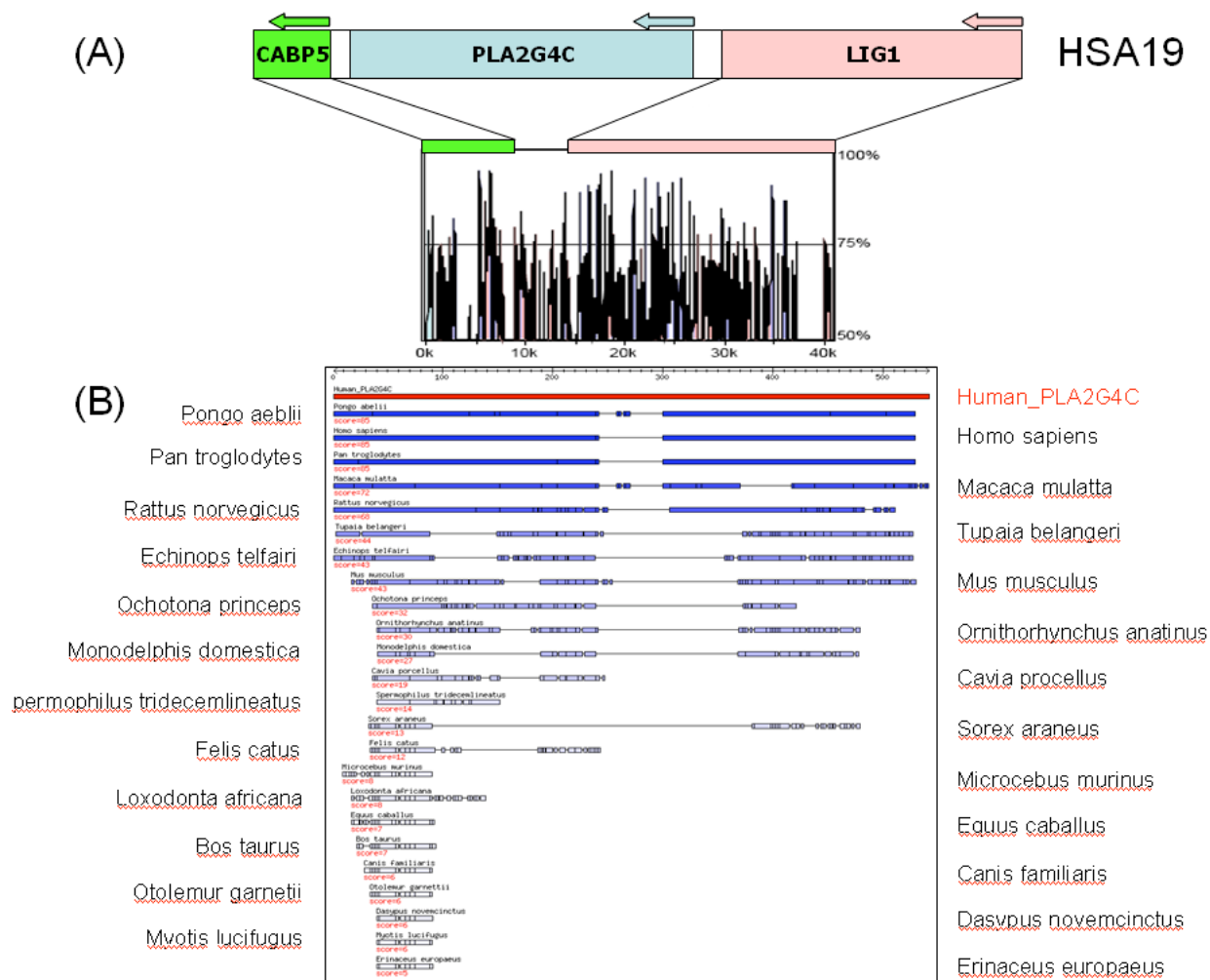


Fig. S22. Deletion of *PLA2G4C* in the cattle genome. (A) The upper panel shows orientation of *PLA2G4C* and flanking genes, *CABP5* and *LIG1*, on HSA19. The lower panel shows a DNA sequence identity plot in 100 bp sliding windows of the cattle genome. The flanking orthologs of human *PLA2G4C*, *CABP5* and *LIG1* are present in the cattle genome on BTA18. (B) Approximately 68 kbp corresponding to *PLA2G4C* are missing in the cattle genome (and dog and other Laurasiatherians), despite very high quality assembly of the sequence in this region.

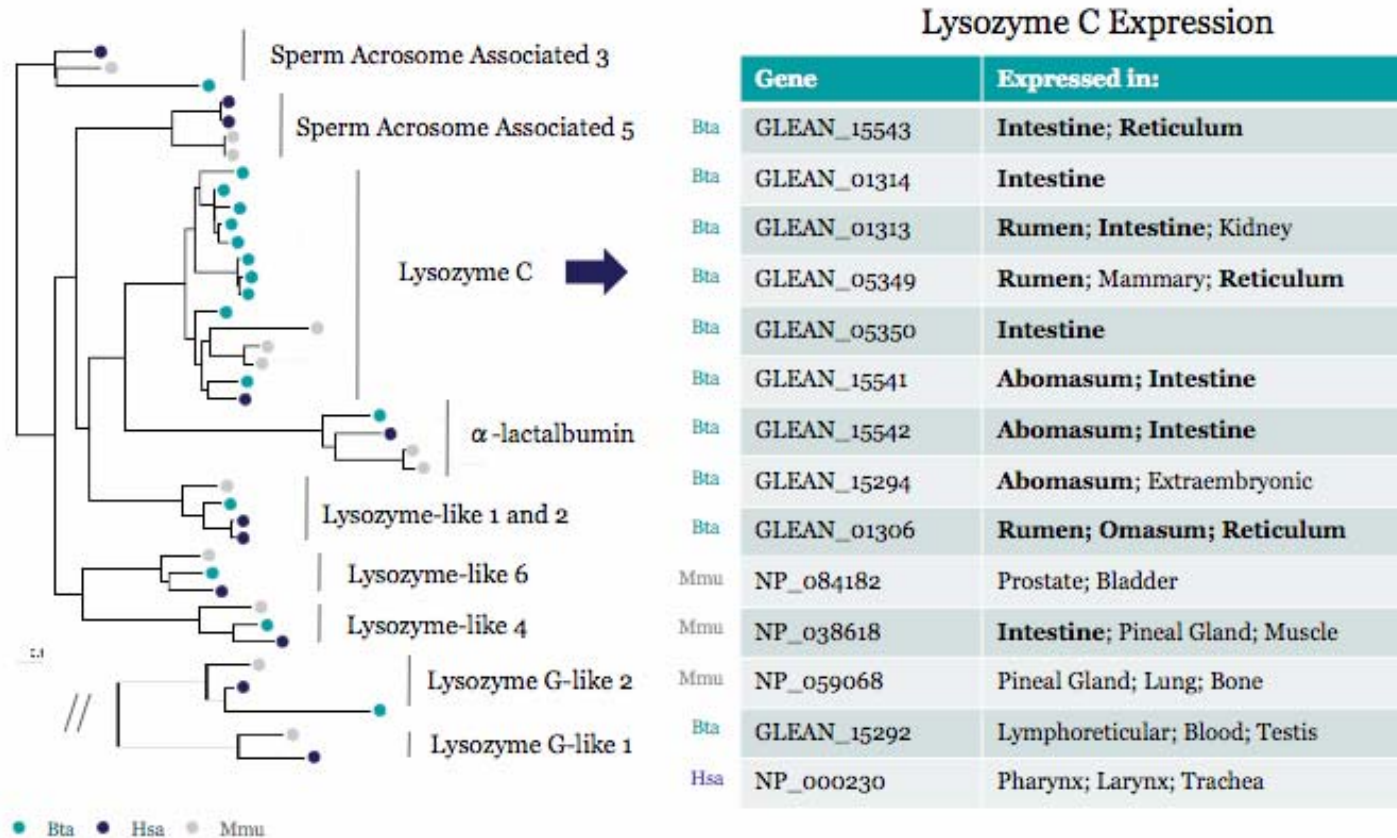


Fig. S23. Unrooted neighbor-joining tree depicting evolutionary relationships among lysozyme-related proteins from bovine, human, and mouse. Note the expansion of the lysozyme C family in cattle (Bta) (*S123*). The lysozyme C group shown in the tree was observed in 98 out of 100 bootstrap replicates. The scale bar represents 0.1 changes per site. The lysozyme G-like sequences were analyzed separately due to their extensive divergence from the other sequences. For each lysozyme C sequence the tissues showing highest expression are listed (up to three), on the basis of the frequency of sequences in EST libraries available in GenBank. Tissues particularly important for digestion or nutrient absorption are shown in bold.

Supporting Tables

Table S1. Evaluation of gene prediction sets with full-length cDNAs

Gene Set	Prediction Type	Number of Gene Loci¹	Number of Gene Models²	Perfect Matches to Full-length CDS³	High Identity Matches to Full-length CDS³	Input Gene Prediction Sets to GLEAN⁴
Ensembl	evidence	22,790	28,222	96	188	
NCBI	evidence	23,311	24,378	102	181	
Fgenesh++	evidence	48,914	49,237	88	185	
SGP2	ab initio	37,891	37,891	37	173	
Fgenesh	ab initio	67,464	67,464	25	166	
Geneid	ab initio	40,557	40,557	24	154	
GLEAN1	consensus	44,528	44,528	92	177	Ensembl, NCBI, Fgenesh++, Fgenesh, SGP2, Geneid
GLEAN2	consensus	22,460	22,460	107	183	Ensembl, NCBI, Fgenesh++
GLEAN3	consensus	31,193	31,193	100	191	Ensembl, NCBI, Combined (Fgenesh++, Fgenesh), SGP, Geneid
GLEAN4	consensus	28,332	28,332	102	189	Ensembl, NCBI, Combined (Fgenesh++, Fgenesh), Geneid
GLEAN5	consensus	26,835	26,835	105	189	Ensembl, NCBI, Combined (Fgenesh++, Fgenesh), SGP2
GLEAN6	consensus	25,497	25,497	107	187	Ensembl, NCBI, Fgenesh++, SGP2

GLEAN7	consensus	25,937	25,937	106	187	Ensembl, NCBI, Fgenesh++, Geneid
--------	-----------	--------	--------	-----	-----	----------------------------------

¹The number of gene loci was computed after grouping gene models with overlapping CDS.

²Gene model number may be different than the original dataset because gene models that crossed scaffolds were split into multiple gene models. Gene models that were not able to be translated were removed. Ensembl gene prediction was performed on chromosome assemblies and individual unassigned scaffolds. All other gene predictions were performed entirely on individual scaffolds.

³208 full-length coding sequences (CDS) from cloned cDNAs, with start and stop codons, that were not available in GenBank before 12/1/2006, were compared to the gene predictions with FASTA. Perfect match was defined as $\geq 99\%$ identity over the entire length of both sequences. High identity match was defined as 99% identity with no length limitation.

⁴In addition to gene prediction sets, alignments of bovine ESTs and SwissProt metazoan protein homologs were used in all GLEAN sets. ESTs were assembled into contigs, and then aligned to the assembly with Exonerate. Criteria for including an EST alignment in the GLEAN analysis were $\geq 98\%$ identity and alignment over $\geq 80\%$ of the contig length. SwissProt metazoan homologs were aligned to the assembly with Exonerate. Criteria used were $\geq 70\%$ identity and alignment $\geq 80\%$ of the protein length. In some cases the Fgenesh and Fgenesh++ sets were combined into one set for assignment of false positive and false negative parameters by GLEAN.

Table S2. Analysis of splice sites

Gene Set	Unique Predicted Donor/Acceptor Sites	Internal cDNA Donor/Acceptor Sites¹	Perfect Matches to cDNA Donor/Acceptor Site	Perfect Matches per Internal cDNA Donor/Acceptor Site	Perfect Matches per Predicted Donor/Acceptor Site	Donor Matches	Acceptor Matches
Ensembl	181,901	2,939	2,850	0.97	0.0162	2,889	2,903
NCBI	159,090	2,898	2,838	0.98	0.0182	2,880	2,878
Fgenesh++	258,985	2,945	2,716	0.92	0.0114	2,836	2,841
SGP	179,761	2,827	2,307	0.82	0.0157	2,561	2,521
Fgenesh	309,984	2,788	2,048	0.73	0.0090	2,480	2,509
Geneid	163,969	2,617	1,939	0.74	0.0160	2,231	2,227
GLEAN1	237,869	2,753	2,636	0.96	0.0116	2,737	2,730
GLEAN2	165,209	3,009	2,939	0.98	0.0182	2,984	2,991
GLEAN3	186,887	3,026	2,937	0.97	0.0162	3,006	3,002
GLEAN4	180,607	3,023	2,939	0.97	0.0167	3,000	3,004
GLEAN5	178,726	3,027	2,954	0.98	0.0169	3,005	3,004
GLEAN6	175,024	3,026	2,956	0.98	0.0173	3,003	3,003
GLEAN7	175,395	3,026	2,946	0.97	0.0173	3,000	3,006

¹ Internal cDNA donor/acceptor sites are the number of cDNA splice junctions that align between the start and stop codons of gene models. The total number of cDNA splice junctions is 3,073.

Table S3. Summary of DNA sequencing from 28 cDNA libraries

Name	Tissue	Developmental Stage	Total Number of ESTs Sequenced	Clones Selected for Full-Length Sequencing
LB001	Peyer's Patch	female, 6 months old ¹	1,536	0
LB002	Ileum	female, 6 months old ¹	21,888	1,109
LB003	Mammary Gland	female, 4 years old ²	1,536	14
LB004	Liver	female, 6 months old ¹	67,200	1,549
LB005	Testis	male, 7 years old ³	21,504	485
LB011	Hypothalamus	female, 8.5 months old ⁴	68,736	934
LB012	Heart ventricular	female, 8.5 months old ⁴	21,504	411
LB013	Thymus	female, 8.5 months old ⁴	41,856	872
LB014	Fetal Liver	male, 6 months old/ fetal ⁵	26,880	324
LB016	Uterus	female 8.5 months old ⁴	24,192	598
LB017	Fetal Ascending Colon	male, 6 months old / fetal ⁵	24,960	580
LB018	Fetal cerebral cortex	male, 6 months old / fetal ⁵	21,888	122
LB019	Fetal Pons	male, 6 months old / fetal ⁵	20,736	376
LB020	fetal medulla	male, 6 months old / fetal ⁵	19,584	100
LB021	Fetal spinal column	male, 6 months old / fetal ⁵	12,672	101
LB022	fetal cerebellum	male, 6 months old / fetal ⁵	25,728	262
LB023	Fetal Lung	male, 6 months old / fetal ⁵	26,496	150
LB024	Fetal pancreas	male, 6 months old /fetal ⁵	36,864	34
LB025	calf hippocampus	female, 8.5 months old ⁴	21,504	318
LB026	calf thalamus	female, 8.5 months old ⁴	11,518	289
LB027	calf basal ganglia	female, 8.5 months old ⁴	20,736	349
LB028	calf cerebral cortex	female, 8.5 months old ⁴	11,520	222
LB029	Fetal skin	female, 6 months old ¹	37,248	673

LB030	Fetal muscle	female, 6 months old ¹	38,784	402
LB032	Placenta	male, 6 months old / fetal ⁵	24,576	79
LB033	Ovary	female, 8.5 months old ⁴	32640	20
LB034	kidney	female, 8.5 months old ⁴	43008	55
LB035	Rumen	female, 8.5 months old ⁴	35328	0

¹ Angus female calf.

² Hereford cow (L1 Dominette 01449).

³ Hereford bull, (L1 Domino 99375).

⁴ Hereford, female calf, progeny of L1 Dominette 01449

⁵ Hereford, fetal, male, progeny of L1 Dominette 01449

Table S4. Repeat composition of the bovine genome

Group	Number	Total bp	Percent Coverage of Genome
LINE_L1	616,259	328,664,804	11.26352
LINE_RTE	376,067	313,409,818	10.74072
LINE_L2	132,485	34,553,185	1.18416
LINE_CR1	14,524	3,083,954	0.10569
	1,139,335	679,711,761	23.29409
SINE_BovA	1,839,497	294,459,617	10.09129
SINE_ART2A	348,768	121,997,595	4.18092
SINE_tRNA	388,920	57,981,206	1.98705
SINE_MIR	301,335	40,569,445	1.39034
SINE_Other	4,322	432,334	0.01482
	2,882,842	515,440,197	17.66441
LTR_MaLR	135,536	42,285,673	1.44915
LTR_ERVL	69,540	25,833,994	0.88534
LTR_ERV1	68,518	23,706,917	0.81245
LTR_ERVK	4,038	1,536,800	0.05267
	277,632	93,363,384	3.19961
DNA_All	244,174	57,157,641	1.95882
LTR_BTLTR1	11,338	6,494,236	0.22256
LTR_ARLTR2	14,358	4,127,734	0.14146
LTR_Other	8,656	1,773,440	0.06078
	34,352	12,395,410	0.42480

di_AC	539,678	6,835,776	0.23427
di_AT	440,644	4,957,167	0.16988
di_AG	375,243	3,537,184	0.12122
di_CG	9,081	85,400	0.00293
	1,364,646	15,415,527	0.52830
tri_AGC	285,325	3,910,867	0.13403
tri_AAT	231,133	2,361,839	0.08094
tri_AGG	199,279	1,945,722	0.06668
tri_AAG	194,774	1,894,234	0.06492
tri_AAC	163,282	1,793,155	0.06145
tri_ACC	100,462	1,043,099	0.03575
tri_ATC	81,511	799,361	0.02739
tri_ACT	32,644	334,552	0.01147
tri_CCG	19,735	219,746	0.00753
tri_ACG	1,769	17,524	0.00060
	1,309,914	14,320,099	0.49075
tetra.penta_All	2,979,000	36,540,043	1.25225
Interspersed Repeat Total	4,578,335	1,358,068,393	46.54174
SSR Total	5,653,560	66,275,669	2.27130

Table S5. Exon skipping

	Shared Cases¹	Non-shared Cases²
Total number of cases tested by RT-PCR in cow	163	114
Expression in cow demonstrated by RT-PCR	122	66
Middle exon exists and is regulated in cow	71	20
Middle exon exists and is constitutive in cow	27	37
Middle exon does not exist in cow	24	9

¹Exon skipping occurring in human and mouse.

²Exon skipping occurring in human but not mouse.

Table S6. Densities of repeat families found to differ significantly in cattle-, artiodactyl- and ferungulate-specific evolutionary breakpoint regions (EBRs)

Repeats	Cattle EBRs	Other intervals	Artiodactyl EBRs	Other intervals	Ferungulate EBRs	Other intervals
Number of 10 kbp intervals	2,025	261,406	451	262,980	174	263,257
LTR-MaRL	115.9*	148.9	131.6	148.6	97.9*	148.7
SINE-tRNA-Glu	97.7*	108.2	131.8*	108.1	82.9	108.1
SINE-BovA	582.8*	671.2	576.2*	670.7	544.5*	670.6
LINE-RTE	903.3*	655.8	772.2*	657.5	880.4*	657.5
LTR-ERV1	88.5*	42.5	44.3	42.8	42.7	42.8
LTR-ERVL	93.7	95.9	150.0*	95.8	117.0	96.0
LINE-L1	1614.4*	1153.3	1468.2*	1156.3	1755.6*	1156.5
LINE-L2	146.0*	244.4	153.5*	243.8	121.0*	243.7
SINE-MIR	126.4*	224.5	131.3*	224.0	108.7*	223.9

*FDR<0.05

Table S7. Density of LINE-L1 elements found to differ significantly in cattle-, artiodactyl-, and ferungulate-specific evolutionary breakpoint regions (EBRs)

Repeats	Cattle EBRs	Other Intervals	Artiodactyl EBRs	Other Intervals	Ferungulate EBRs	Other Intervals
number of 10 kbp intervals	2,025	261,406	451	262,980	174	263,257
HAL1	14.8*	22.5	12.7	22.5	6.7	22.4
L1ME4a	12.1*	20.3	8.1	20.3	14.5	20.3
L1ME1	13.6*	30.5	31.0	30.4	8.7	30.4
L1M4c	39.4*	13.6	26.5	18.5	14.6	13.8
L1M3	61.9*	23.8	49.0	24.0	32.9	24.1
L1MA9	106.4*	76.3	110.5	76.5	123.6	76.5
L1M2	38.5*	15.5	42.5	15.7	38.1	15.7
L1_BT	650.2*	330.2	518.8	332.4	974.1*	332.3

*FDR < 0.05

Table S8. Gene ontology analysis of genes in segmental duplications^{1,2}

GO Acc	GO term	Set 1	Set 2	Set 3	Set 4	Set 5
GO:0004984	olfactory receptor activity [molecular_function]	3.22E-32	2.82E-33	5.35E-35	1.65E-34	1.51E-30
GO:0007608	sensory perception of smell [biological_process]	3.55E-25	6.95E-25	4.22E-28	9.87E-28	1.07E-24
GO:0001584	rhodopsin-like receptor activity [molecular_function]	2.26E-22	3.10E-24	3.31E-25	1.58E-25	1.63E-22
GO:0007186	G-protein coupled receptor protein signaling pathway [biological_process]	5.77E-20	7.92E-22	4.78E-23	8.49E-22	6.10E-20
GO:0004872	receptor activity [molecular_function]	8.28E-20	1.11E-19	4.11E-20	3.14E-20	1.97E-18
GO:0050896	response to stimulus [biological_process]	2.80E-18	3.46E-18	1.48E-20	3.25E-20	7.71E-18
GO:0004977	melanocortin receptor activity [molecular_function]	5.05E-18	1.37E-19	5.97E-19	1.51E-16	1.24E-16
GO:0006955	immune response [biological_process]	5.60E-14	1.30E-11	9.38E-14	1.77E-13	1.37E-13
GO:0006952	defense response [biological_process]	7.99E-11	3.63E-11	8.68E-11	1.60E-11	1.37E-12
GO:0007165	signal transduction [biological_process]	9.33E-10	3.66E-11	8.39E-11	2.61E-11	3.06E-09
GO:0042612	MHC class I protein complex [cellular_component]	3.91E-09	8.13E-07	8.38E-08	1.74E-06	9.35E-08
GO:0005126	hematopoietin/interferon-class	3.54E-08	6.38E-10	2.72E-11	2.71E-11	2.74E-11

	(D200-domain) cytokine receptor binding [molecular_function]					
GO:0019882	antigen processing and presentation [biological_process]	3.16E-07	1.44E-05	2.69E-06	2.59E-05	3.45E-06
GO:0016021	integral to membrane [cellular_component]	1.14E-05	5.44E-07	4.17E-07	1.59E-05	5.66E-06
GO:0042742	defense response to bacterium [biological_process]	5.31E-05	1.59E-03	3.61E-04	5.79E-05	6.92E-06
GO:0047115	"trans-1,2-dihydrobenzene-1,2-diol dehydrogenase activity [molecular_function]"	1.01E-04	8.28E-05	1.09E-02	9.13E-05	8.82E-05
GO:0009615	response to virus [biological_process]	1.01E-04	8.73E-06	1.43E-06	1.89E-06	1.73E-06
GO:0016020	membrane [cellular_component]	9.62E-05	2.86E-05	2.13E-05	5.56E-04	7.39E-05
GO:0005132	interferon-alpha/beta receptor binding [molecular_function]	1.48E-04	1.31E-04	2.81E-06	3.58E-06	3.41E-06
GO:0032393	MHC class I receptor activity [molecular_function]	2.89E-04	2.34E-04	4.76E-03	4.49E-03	4.32E-03
GO:0004888	transmembrane receptor activity [molecular_function]	5.77E-04	2.29E-03	1.16E-04	7.48E-04	7.28E-04
GO:0005576	extracellular region [cellular_component]	7.53E-04	1.44E-04	2.57E-06	4.08E-06	8.74E-06
GO:0047026	3-alpha-hydroxysteroid dehydrogenase (A-specific) activity [molecular_function]	3.91E-03	3.70E-03	4.56E-03	4.14E-03	4.03E-03
GO:0005615	extracellular space [cellular_component]	3.96E-03	2.77E-03	5.18E-03	9.34E-04	8.68E-04

GO:0003823	antigen binding [molecular_function]	7.78E-03	6.83E-03	8.69E-03	9.67E-06	1.12E-03
GO:0042157	lipoprotein metabolic process [biological_process]	7.54E-03	6.82E-03	8.91E-03	8.76E-03	8.42E-03
GO:0004949	cannabinoid receptor activity [molecular_function]	1.36E-02	1.19E-02	1.52E-02	1.72E-02	1.60E-02
GO:0030101	natural killer cell activation [biological_process]	1.78E-02	1.71E-02	1.93E-02	2.20E-02	2.00E-02
GO:0003956	NAD(P)+-protein-arginine ADP- ribosyltransferase activity [molecular_function]	2.66E-02	2.77E-02	2.83E-02	3.23E-02	2.91E-02
GO:0045089	positive regulation of innate immune response [biological_process]	3.43E-02	3.61E-02	3.73E-02	4.11E-02	3.74E-02
GO:0030162	regulation of proteolysis [biological_process]	3.43E-02	3.61E-02	3.73E-02	4.11E-02	3.74E-02

¹Many of the bovine genes were found to have one-to-many orthology relationships with human genes. To overcome the potential bias of assigning multiple similar GO and pathway annotations to the same bovine gene due to these cases, five ortholog sets were created. Each of the five sets included all bovine OGSv1 genes for which there was at least one human ortholog. For each bovine gene that had multiple human orthologs, a single human gene was randomly selected for annotation transfer. This process was repeated to generate five unbiased datasets.

²Probabilities associated with each GO term are shown.

Table S9. Pathway analysis of genes in segmental duplications^{1,2}

Pathway ID	Pathway Name	Set 1	Set 2	Set 3	Set 4	Set 5
1460	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	3.88E-12	9.57E-13	1.95E-15	1.67E-11	5.92E-12
578	Natural killer cell mediated cytotoxicity	2.78E-12	2.76E-12	1.64E-15	5.87E-15	1.67E-14
493	Antigen processing and presentation	1.09E-10	1.79E-10	5.19E-13	1.94E-12	7.63E-12
386	IFN alpha signaling pathway (JAK1 TYK2 STAT1)	4.79E-06	5.94E-06	2.69E-09	2.20E-07	2.21E-09
400	IFN alpha signaling pathway (JAK1 TYK2 STAT3)	4.79E-06	5.94E-06	2.69E-09	2.20E-07	2.21E-09
380	IFN alpha signaling pathway (JAK1 TYK2 STAT1 STAT3)	5.91E-06	7.33E-06	3.99E-09	2.96E-07	3.28E-09
431	Regulation of autophagy	1.23E-04	8.65E-06	3.43E-07	5.38E-07	2.83E-07
392	IFN alpha signaling pathway (JAK1 TYK2 STAT1 STAT2)	7.32E-06	7.94E-06	5.83E-09	3.96E-07	4.79E-09
1201	Classical antibody-mediated complement activation	1.88E-03	2.57E-04	2.26E-04	3.29E-04	1.93E-04
515	Cytokine-cytokine receptor interaction	1.80E-03	6.05E-04	2.98E-05	7.12E-05	1.10E-04
1148	Downstream signaling in naïve CD8+ T cells	5.54E-04	5.72E-04	5.61E-06	9.65E-05	4.55E-06
568	Jak-STAT signaling pathway	6.30E-03	2.07E-03	4.16E-04	7.08E-04	3.31E-04
564	Toll-like receptor signaling pathway	1.00E-02	2.73E-03	4.39E-04	6.91E-04	3.61E-04

¹Many of the bovine genes were found to have one-to-many orthology relationships with human genes. To overcome the potential bias

of assigning multiple similar GO and pathway annotations to the same bovine gene due to these cases, five ortholog sets were created. Each of the five sets included all bovine OGSv1 genes for which there was at least one human ortholog. For each bovine gene that had multiple human orthologs, a single human gene was randomly selected for annotation transfer. This process was repeated to generate five unbiased datasets.

²Probabilities associated with each Pathway ID are shown.

Table S10. Genes subject to positive selection

Bovine Gene ID	Human Ensembl ID	Mouse Ensembl ID	Gene Symbol	Gene Name	d _N /d _S	LRT Value	Model2 p-value	Model1 p-value
GLEAN_25382	ENSG00000164304	ENSMUSG00000044566	CAGE1	cancer antigen 1	1.06	16.52	0.001***	
GLEAN_15922		ENSMUSG00000028189	Ctbs	chitinase, di-N-acetyl- chromosome X open reading frame 30	1.04	33.05	0.001***	
GLEAN_10454	ENSG00000205081	ENSMUSG00000071735	CXorf30	ectodermal-neural cortex 1	1.17	19.75	0.001***	
GLEAN_07078		ENSMUSG00000041773	Enc1	ecotropic viral integration site 2A	37.95	62.25	0.001***	
GLEAN_25881	ENSG00000126860		EVI2A	fatty acid binding protein 6, ileal (gastrotropin)	4.24	17.59	0.001***	
GLEAN_11540	ENSG00000170231	ENSMUSG00000020405	FABP6	leucine zipper protein 2	1.07	14.05	0.001***	
GLEAN_00966	ENSG00000187398	ENSMUSG00000063297	LUZP2		1.07	11.90	0.001***	
GLEAN_02457				RIKEN cDNA 2310040G07 gene	1.73	14.88	0.001***	
GLEAN_22866		ENSMUSG00000072852	2310040G07Rik	chromosome 10 open reading frame 78	1.48	8.01	0.01**	
GLEAN_25210	ENSG00000156384	ENSMUSG00000025066	C10orf78	coiled-coil domain containing 121	1.05	7.77	0.01**	
GLEAN_18851	ENSG00000176714	ENSMUSG00000029138	CCDC121	eukaryotic translation initiation factor 3, subunit E	1.04	7.52	0.01**	
GLEAN_13996	ENSG00000104408	ENSMUSG00000022336	EIF3E	transcript expressed during hematopoiesis 2	999	7.00	0.01**	
GLEAN_20989	ENSG00000154035	ENSMUSG00000018931	IDBG-35547	interferon (alpha and beta) receptor 2	999	8.21	0.01**	
GLEAN_07910		ENSMUSG00000022971	Ifnar2	interleukin 15	1.51	7.13	0.01**	
GLEAN_07098	ENSG00000164136	ENSMUSG00000031712	IL15	KIAA1274	1.79	8.65	0.01**	
GLEAN_16592	ENSG00000107719	ENSMUSG00000020092	KIAA1274	podoplanin	1.54	7.61	0.01**	
GLEAN_25910	ENSG00000162493	ENSMUSG00000028583	PDPN	phospholipase A2, group X patatin-like phospholipase domain containing 4	1.44	7.87	0.01**	
GLEAN_00418		ENSMUSG00000022683	Pla2g10	RAB24, member RAS	999	7.35	0.01**	
GLEAN_12455	ENSG00000006757		PNPLA4	oncogene family small nuclear ribonucleoprotein D1	8.16	6.67	0.01**	
GLEAN_01453	ENSG00000169228	ENSMUSG00000034789	RAB24	polypeptide 16kDa	999	7.87	0.01**	
GLEAN_07723	ENSG00000167088	ENSMUSG00000002477	SNRPD1	transmembrane protein 139	999	9.29	0.01**	
GLEAN_21922	ENSG00000178826		TMEM139		1.06	7.63	0.01**	

GLEAN_18597	ENSG00000129235	ENSMUSG00000020803	TXNDC17	thioredoxin domain containing 17	2.82	8.84	0.01**	
GLEAN_02453	ENSG00000116127	ENSMUSG00000063810	ALMS1	Alstrom syndrome 1	999	5.05	0.05*	
GLEAN_26048	ENSG00000161929	ENSMUSG00000057135	C17orf87	chromosome 17 open reading frame 87	1.50	5.63	0.05*	
GLEAN_20249	ENSG00000100307	ENSMUSG00000053411	CBX7	chromobox homolog 7	999	4.11	0.05*	
GLEAN_20841	ENSG00000174059	ENSMUSG00000016494	CD34	CD34 molecule	1.04	6.26	0.05*	
GLEAN_05820	ENSG00000173401	ENSMUSG00000020213	GLIPR1L1	GLI pathogenesis-related 1 like 1	1.61	6.56	0.05*	
GLEAN_19967	ENSG00000187837		HIST1H1C	histone cluster 1, H1c	999	4.32	0.05*	
GLEAN_22648	ENSG00000183208		IDBG-29982	CDNA FLJ27017 fis, clone SLV05746.	999	4.23	0.05*	
GLEAN_04862	ENSG00000137463	ENSMUSG00000037161	IDBG-38342	ovary-specific acidic protein	1.73	5.11	0.05*	
GLEAN_12673	ENSG00000187942	ENSMUSG00000070666	LDLRAD2	low density lipoprotein receptor class A domain containing 2	1.49	4.16	0.05*	
GLEAN_15535	ENSG00000079435		LIPE	lipase, hormone-sensitive	999	5.71	0.05*	
GLEAN_05536		ENSMUSG00000034041	Ly1l	lymphoblastic leukemia mixed lineage kinase domain-like	1.69	5.90	0.05*	
GLEAN_09957	ENSG00000168404	ENSMUSG00000012519	MLKL	receptor (G protein-coupled)	1.30	5.09	0.05*	
GLEAN_10081	ENSG00000132329		RAMP1	activity modifying protein 1	999	4.05	0.05*	
GLEAN_10642	ENSG00000125743		SNRPD2	small nuclear ribonucleoprotein D2	3.15	4.70	0.05*	
GLEAN_19853	ENSG00000124731	ENSMUSG00000042265	TREM1	polypeptide 16.5kDa triggering receptor expressed on myeloid cells 1	1.19	4.11	0.05*	
GLEAN_19856	ENSG00000161911	ENSMUSG00000023993	TREML1	triggering receptor expressed on myeloid cells-like 1	1.13	5.30	0.05*	
GLEAN_00853	ENSG00000179965	ENSMUSG00000054716	ZNF771	zinc finger protein 771	999	4.61	0.05*	
GLEAN_25096		ENSMUSG00000032068	1600029D21Rik	RIKEN cDNA 1600029D21 gene	1.17	1.65	ns	0.05*
GLEAN_16995	ENSG00000156170	ENSMUSG00000050323	C8orf38	chromosome 8 open reading frame 38	1.21	1.60	ns	
GLEAN_11591	ENSG00000160345		C9orf116	chromosome 9 open reading frame 116	999	2.63	ns	0.01**
GLEAN_19036	ENSG00000175550	ENSMUSG00000024914	DRAP1	DR1-associated protein 1 (negative cofactor 2 alpha)	999	1.56	ns	
GLEAN_02180	ENSG00000198842	ENSMUSG00000026564	DUSP27	dual specificity phosphatase	3.44	0.001	ns	

GLEAN_21713	ENSG00000203666		EFCAB2	27 (putative) EF-hand calcium binding domain 2	999	2.27	ns	
GLEAN_08191	ENSG00000185942	ENSMUSG00000055761	FAM77D	Na+/K+ transporting ATPase interacting 3	999	1.97	ns	0.01**
GLEAN_00283	ENSG00000179639	ENSMUSG00000005339	FCER1A	Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide	1.01	1.68	ns	0.001**
GLEAN_06658		ENSMUSG00000036904	Fzd8	frizzled homolog 8	5.86	1.95	ns	*
GLEAN_13471		ENSMUSG00000030050	Gkn1	gastroke 1	1.47	2.33	ns	
GLEAN_03300	ENSG00000197858	ENSMUSG00000022561	GPAA1	glycosylphosphatidylinositol anchor attachment protein 1 homolog (yeast)	999	1.43	ns	0.01**
GLEAN_06498	ENSG00000177875	ENSMUSG00000029875	IDBG-29906	CDNA FLJ32221 fis, clone PLACE6004005.	2.80	2.64	ns	0.001**
GLEAN_24551	ENSG00000178385	ENSMUSG00000051344	IDBG-79678	pleckstrin homology domain containing, family M, member 1-like	999	1.11	ns	0.05*
GLEAN_10120	ENSG00000111537	ENSMUSG00000055170	IFNG	interferon, gamma	1.57	1.53	ns	0.001**
GLEAN_23729		ENSMUSG00000049093	IL23r	interleukin 23 receptor	3.04	0.76	ns	*
GLEAN_25653	ENSG00000162892	ENSMUSG00000026420	IL24	interleukin 24	1.19	2.56	ns	0.01**
GLEAN_08602		ENSMUSG00000025746	IL6	interleukin 6	1.22	3.59	ns	
GLEAN_26229	ENSG00000183853	ENSMUSG00000041734	KIRREL	kin of IRRE like (Drosophila)	2.53	0.92	ns	
GLEAN_20621		ENSMUSG00000036216	Leap2	liver-expressed antimicrobial peptide 2	999	0.73	ns	0.01**
GLEAN_19128	ENSG00000188906	ENSMUSG00000036273	LRRK2	leucine-rich repeat kinase 2	999	0.46	ns	0.001**
GLEAN_10270	ENSG00000130045	ENSMUSG00000021396	NXNL2	nucleoredoxin-like 2	755.65	1.16	ns	*
GLEAN_14741	ENSG00000181355	ENSMUSG00000047094	OFCC1	orofacial cleft 1 candidate 1	1.40	0.61	ns	0.001**
GLEAN_22171	ENSG00000117625	ENSMUSG00000037395	RCOR3	REST corepressor 3	1.63	0.08	ns	*
GLEAN_25591		ENSMUSG00000046110	Serinc4	serine incorporator 4	1.60	2.51	ns	
GLEAN_08455	ENSG00000141380	ENSMUSG00000037013	SS18	synovial sarcoma translocation, chromosome 18	999	2.78	ns	0.01**
GLEAN_09242	ENSG00000127362		TAS2R3	taste receptor, type 2, member 3	1.18	3.74	ns	0.05*
GLEAN_14444	ENSG00000198270	ENSMUSG00000029452	TMEM116	transmembrane protein 116	1.50	2.49	ns	
GLEAN_18052	ENSG00000121895	ENSMUSG00000037913	TMEM156	transmembrane protein 156	1.03	2.35	ns	

GLEAN_15561	ENSMUSG00000052906	Ubx ^{d6}	UBX domain containing 6	1.40	2.03	ns	0.01**
GLEAN_23649	ENSMUSG00000055633	Zfp580	zinc finger protein 580	999	2.20	ns	0.05*
GLEAN_12950				1.22	0.08	ns	

Table S11. The number of unique, matched cattle metabolic pathway genes (N=736) in the GLEAN gene models and in CattleCyc.

Number of Gene Models	CattleCyc¹/GLEAN	GLEAN/CattleCyc
Matched Genes	706	729
One-to-one	683	683
One CattleCyc-to-two GLEAN	23	46
Unmatched Genes	27	N.A.
Btau3.1	10	N.A.
Btau2.1	6	N.A.
Mitochondria	11	N.A.

¹Amalgamated cattle genome database built from Btau3.1, unincorporated scaffolds from Btau2.1 and the mitochondrial genome

Table S12. Statistics for cattle-specific pathway genome database reconstructed from the GLEAN5 gene models¹

Database Statistics	N
Metabolic pathways	217
Enzymatic reactions	1,409
Enzymes	1,492
Pathway holes (missing enzymes)	
Number of pathway holes	116
Pathway holes (% of total reactions in pathways)	14%
Pathways with no holes	162
Pathways with 1 hole	30
Pathways with 2 holes	11
Pathways with 3 holes	2
Pathways with 4 holes	6
Pathways with 5 holes	6
Total number of pathways with holes	55

¹In addition to the GLEAN5 gene models, 6 and 11 additional genes from Btau2.1 and the mitochondrial genome, respectively, were used to reconstruct cattle-specific metabolic pathways.

Table S13. Genome assemblies used for metabolic reconstruction analyses

Species	Genome Assembly
Cattle (<i>Bos taurus</i>)	Btau3.1/Btau4.0
Horse (<i>Equus caballus</i>)	EquCab2
Dog (<i>Canis familiaris</i>)	CanFam2.0
Cat (<i>Felis catus</i>)	CAT
Microbat (<i>Myotis lucifugus</i>)	myoLuc1
Hedgehog (<i>Erinaceus europaeus</i>)	eriEur1
Shrew (<i>Sorex araneus</i>)	sorAra1
Mouse (<i>Mus musculus</i>)	NCBI m37
Rat (<i>Rattus norvegicus</i>)	RGSC 3.4
Guinea pig (<i>Cavia porcellus</i>)	cavPor2
Squirrel (<i>Spermophilus tridecemlineatus</i>)	speTri1
Rabbit (<i>Oryctolagus cuniculus</i>)	RABBIT
Pika (<i>Ochotona princeps</i>)	OchPri2.0
Tree shrew (<i>Tupaia belangeri</i>)	tupBel1
Lemur (<i>Microcebus murinus</i>)	micMur1
Bushbaby (<i>Otolemur garnettii</i>)	otoGar1
Macaque (<i>Macaca mulatta</i>)	MMUL 1.0
Orangutan (<i>Pongo abelii</i>)	PPYG2
Chimpanzee (<i>Pan troglodytes</i>)	PanTro 2.1
Human (<i>Homo sapiens</i>)	NCBI 36
Armadillo (<i>Dasypus novemcinctus</i>)	ARMA
Elephant (<i>Loxodonta africana</i>)	BROAD E1
Tenrec (<i>Echinops telfairi</i>)	TENREC
Opossum (<i>Monodelphis domestica</i>)	monDom5
Platypus (<i>Ornithorhynchus anatinus</i>)	Oana-5.0
Chicken (<i>Gallus gallus</i>)	WASHUC2

Table S14. Sequence and expression characteristics of lysozyme C-like proteins in bovine, mouse and human

Gene Identifier (source species)	Calculated pI¹	No. of Adaptive Sequence Changes for Stomach Function²	Most Prevalent Sites of Expression³
GLEAN_15541 (cow)	6.32	7	Abomasum (4156); Intestine (29)
GLEAN_15294 (cow)	7.03	6	Abomasum (2944); Extraembryonic tissue (29)
GLEAN_15542 (cow)	6.46	7	Abomasum (14763); Intestine (204); Lymphoreticular (53)
GLEAN_01306 (cow)	9.87	1	Rumen (9173); Omasum (3171); Reticulum (454)
GLEAN_15292 (cow)	7.72	0	Lymphoreticular (345); Blood (235); Testis (70)
GLEAN_15543 (cow)	9.59	2	Intestine (1273); Reticulum (100)
GLEAN_01314 (cow)	9.20	2	Intestine (29)
GLEAN_01313 (cow)	9.35	2	Rumen (408); Intestine (175); Kidney (26)
GLEAN_05349 (cow)	9.55	3	Rumen (2338); Mammary gland (126); Reticulum (100)
GLEAN_05350 (cow)	9.46	3	Intestine (307)
NP_000230 (human)	9.28	0	Pharynx (2844); Larynx (2332); Trachea (2059);
NP_084182 (mouse)	9.63	0	Prostate (1499); Bladder (595)
NP_059068 (mouse)	8.99	0	Pineal gland (1018); Lung (753); Bone (743)
NP_038618 (mouse)	9.47	0	Intestine (337); Pineal gland (245); Muscle (72)

¹The adaptation of lysozyme C class genes to acidic environments has been associated with a decrease in isoelectric point (S123). Calculation of pI is from the mature protein sequence.

²Of the seven adaptive sequence changes described in (S123).

³Expression measurements are given in transcripts per million, which were obtained from NCBI's "UniGene's EST ProfileViewer".

Supplementary References and Notes

- S1. Rat Genome Sequencing Consortium, *Nature* **428**, 493-521 (2004).
- S2. Sea Urchin Genome Sequencing Consortium, *Science* **314**, 941-952 (2006).
- S3. Y. Liu *et al.*, *BMC Genomics* **10**, 180 (2009)
- S4. P. Havlak *et al.*, *Genome Res* **14**, 721-732 (2004).
- S5. A. Everts-van der Wind *et al.*, *Proc Natl Acad Sci U S A* **102**, 18526-18531 (2005).
- S6. W. M. Snelling *et al.*, *Genome Biol* **8**, R165 (2007).
- S7. H. Nilsen *et al.*, *Anim Genet* **39**, 97-104 (2008).
- S8. A. Prasad *et al.*, *BMC Genomics* **8**, 310 (2007).
- S9. Y. Kapustin, A. Souvorov, T. Tatusova, D. Lipman, *Biol Direct* **3**, 20 (2008).
- S10. B. Kiryutin, A. Souvorov, *ISMB*, (2005).
- S11. A. Souvorov, T. Tatusova, D. Lipman, *ISMB*, (2004).
- S12. B. J. Haas *et al.*, *Nucleic Acids Res* **31**, 5654-5666 (2003).
- S13. V. Curwen *et al.*, *Genome Res* **14**, 942-950 (2004).
- S14. E. Birney, M. Clamp, R. Durbin, *Genome Res* **14**, 988-995 (2004).
- S15. G. S. Slater, E. Birney, *BMC Bioinformatics* **6**, 31 (2005).
- S16. A. A. Salamov, V. V. Solovyev, *Genome Res* **10**, 516-522 (2000).
- S17. V. Solovyev, in *Handbook of Statistical Genetics*, D. J. Balding, M. Bishop, C. Cannings, Eds. (John Wiley & Sons, Chichester, England ; Hoboken, NJ, 2007), pp. 97-159.
- S18. E. W. Sayers *et al.*, *Nucleic Acids Res*, (2008).
- S19. <http://www.repeatmasker.org>.
- S20. E. Blanco, G. Parra, R. Guigo, *Curr Protoc Bioinformatics* **chap. 4**, Unit 4 3 (2007).
- S21. G. Parra *et al.*, *Genome Res* **13**, 108-117 (Jan, 2003).
- S22. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389-3402 (1997).
- S23. C. G. Elsik *et al.*, *Genome Biol* **8**, R13 (2007).
- S24. A. Bairoch, B. Boeckmann, S. Ferro, E. Gasteiger, *Brief Bioinform* **5**, 39-55 (2004).
- S25. M. S. Boguski, T. M. Lowe, C. M. Tolstoshev, *Nat Genet* **4**, 332-333 (1993).
- S26. G. Pertea *et al.*, *Bioinformatics* **19**, 651-652 (Mar 22, 2003).
- S27. W. R. Pearson, D. J. Lipman, *Proc Natl Acad Sci U S A* **85**, 2444-2448 (1988).
- S28. http://bovinegenome.org/bovine_genome_consortium/datasets.html.
- S29. Chicken Genome Sequencing Consortium, *Nature* **432**, 695-716 (2004).
- S30. R. Guigo *et al.*, *Proc Natl Acad Sci U S A* **100**, 1140-1145 (2003).
- S31. A. Reymond *et al.*, *Genomics* **79**, 824-832 (2002).
- S32. S. Rozen, H. J. Skaletsky, in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, S. Krawetz, S. Misener, Eds. (Humana Press, Totowa, NJ, 2000), pp. 365-386.
- S33. D. S. Gerhard *et al.*, *Genome Res* **14**, 2121-2127 (2004).
- S34. R. L. Strausberg *et al.*, *Proc Natl Acad Sci U S A* **99**, 16899-16903 (2002).
- S35. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, D. L. Wheeler, *Nucleic Acids Res* **36**, D25-30 (2008).
- S36. P. Carninci *et al.*, *Genome Res* **10**, 1617-1630 (2000).
- S37. T. S. Alioto, *Nucleic Acids Res* **35**, D110-115 (2007).
- S38. T. D. Wu, C. K. Watanabe, *Bioinformatics* **21**, 1859-1875 (2005).
- S39. S. Griffiths-Jones, H. K. Saini, S. van Dongen, A. J. Enright, *Nucleic Acids Res* **36**, D154-158 (2008).

40. <http://blast.wustl.edu>.
- S41. E. Bonnet, J. Wuyts, P. Rouze, Y. Van de Peer, *Bioinformatics* **20**, 2911-2917 (2004).
- S42. P. Steffen, B. Voss, M. Rehmsmeier, J. Reeder, R. Giegerich, *Bioinformatics* **22**, 500-503 (2006).
- S43. I. L. Hofacker, *Curr Protoc Bioinformatics* **chap. 12**, Unit 12.2 (2004).
- S44. K. Katoh, H. Toh, *BMC Bioinformatics* **9**, 212 (2008).
- S45. B. Giardine *et al.*, *Genome Res* **15**, 1451-1455 (2005).
- S46. D. Gerlach, E. V. Kriventseva, N. Rahman, C. E. Vejnár, E. M. Zdobnov, *Nucleic Acids Res*, (Oct 15, 2008).
- S47. C. Notredame, D. G. Higgins, J. Heringa, *J Mol Biol* **302**, 205-217 (2000).
- S48. W. R. Pearson, *Methods Enzymol* **183**, 63-98 (1990).
- S49. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res* **35**, D61-65 (2007).
- S50. P. Flicek *et al.*, *Nucleic Acids Res* **36**, D707-714 (2008).
- S51. D. Karolchik *et al.*, *Nucleic Acids Res* **36**, D773-779 (2008).
- S52. UniProt Consortium, *Nucleic Acids Res* **37**, D169 (2008).
- S53. <http://BovineGenome.org>.
- S54. S. M. Searle, J. Gilbert, V. Iyer, M. Clamp, *Genome Res* **14**, 963-970 (2004).
- S55. S. E. Lewis *et al.*, *Genome Biol* **3**, RESEARCH0082 (2002).
- S56. P. Bernaola-Galvan, R. Roman-Roldan, J. L. Oliver, *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **53**, 5181-5189 (1996).
- S57. N. Cohen, T. Dagan, L. Stone, D. Graur, *Mol Biol Evol* **22**, 1260-1272 (2005).
- S58. R. C. Edgar, E. W. Myers, *Bioinformatics* **21 Suppl 1**, i152-158 (2005).
- S59. A. L. Price, N. C. Jones, P. A. Pevzner, *Bioinformatics* **21 Suppl 1**, i351-358 (2005).
- S60. R. C. Edgar, *Nucleic Acids Res* **32**, 1792-1797 (2004).
- S61. A. Stamatakis, T. Ludwig, H. Meier, *Bioinformatics* **21**, 456-463 (2005).
- S62. T. H. Jukes, C. R. Cantor, in *Mammalian Protein Evolution*, H. N. Munro, Ed. (Academic Press, New York, 1969), pp. 21-123.
- S63. K. Tamura, J. Dudley, M. Nei, S. Kumar, *Mol Biol Evol* **24**, 1596-1599 (2007).
- S64. <http://www.clcbio.com>.
- S65. L. Kraemer *et al.*, *BMC Bioinformatics* **10**, 41 (2009).
- S66. <http://espressosoftware.com/pages/sputnik.jsp>.
- S67. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**, 403-410 (1990).
- S68. J. Jurka *et al.*, *Cytogenet Genome Res* **110**, 462-467 (2005).
- S69. G. Talavera, J. Castresana, *Syst Biol* **56**, 564-577 (2007).
- S70. S. Guindon, O. Gascuel, *Syst Biol* **52**, 696-704 (2003).
- S71. Mouse Genome Sequencing Consortium, *Nature* **420**, 520-562 (2002).
- S72. A. Reymond *et al.*, *Genomics* **78**, 46-54 (2001).
- S73. M. Ashburner *et al.*, *Nat Genet* **25**, 25-29 (2000).
- S74. A. Kasprzyk *et al.*, *Genome Res* **14**, 160-169 (2004).
- S75. W. J. Kent, *Genome Res* **12**, 656-664 (2002).
- S76. S. J. Humphray *et al.*, *Genome Biol* **8**, R139 (2007).
- S77. W. J. Murphy *et al.*, *Science* **309**, 613-617 (2005).
- S78. D. M. Larkin *et al.*, *Genome Research* **19**, 770-777 (2009).
- S79. D. Thissen, L. Steinberg, D. Kuang, *Journal of Educational and Behavioral Statistics* **27**, 77-83 (2002).

- S80. J. H. Edwards, *Ann Hum Genet* **55**, 17-31 (1991).
- S81. N. D. Trinklein *et al.*, *Genome Res* **14**, 62-66 (2004).
- S82. M. Q. Yang, L. M. Koehly, L. L. Elnitski, *PLoS Comput Biol* **3**, e72 (2007).
- S83. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res* **11**, 1005-1017 (2001).
- S84. X. She, Z. Cheng, S. Zollner, D. M. Church, E. E. Eichler, *Nat Genet* **40**, 909-914 (2008).
- S85. J. A. Bailey *et al.*, *Science* **297**, 1003-1007 (2002).
- S86. D. J. Lynn *et al.*, *Mol Syst Biol* **4**, 218 (2008).
- S87. J. Felsenstein. (Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005).
- S88. <http://evolution.genetics.washington.edu/phylip.html>.
- S89. Z. Yang, *Mol Biol Evol* **24**, 1586-1591 (2007).
- S90. L. D. Hurst, *Trends Genet* **18**, 486 (2002).
- S91. H. Mi, N. Guo, A. Kejariwal, P. D. Thomas, *Nucleic Acids Res* **35**, D247-252 (2007).
- S92. Rhesus Macaque Genome Sequencing and Analysis Consortium, *Science* **316**, 222-234 (2007).
- S93. S. Seo, H. A. Lewin, *BMC Systems Biology* **3**, 33 (2009).
- S94. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res* **22**, 4673-4680 (1994).
- S95. E. Eyras *et al.*, *BMC Bioinformatics* **6**, 131 (2005).
- S96. A. Levine, R. Durbin, *Nucleic Acids Res* **29**, 4006-4013 (2001).
- S97. N. Sheth *et al.*, *Nucleic Acids Res* **34**, 3955-3967 (2006).
- S98. E. A. Glazov, S. McWilliam, W. C. Barris, B. P. Dalrymple, *Mol Biol Evol* **25**, 939-948 (May, 2008).
- S99. H. Seitz *et al.*, *Genome Res* **14**, 1741-1748 (2004).
- S100. G. Bernardi, *Gene* **241**, 3-17 (2000).
- S101. E. S. Lander *et al.*, *Nature* **409**, 860-921 (2001).
- S102. O. Clay, G. Bernardi, *Mol Biol Evol* **22**, 2315-2317 (2005).
- S103. W. Li, *Gene* **300**, 129-139 (Oct 30, 2002).
- S104. W. Li, P. Bernaola-Galvan, P. Carpena, J. L. Oliver, *Comput Biol Chem* **27**, 5-10 (2003).
- S105. http://cegg.unige.ch/cow_genome.
- S106. Z. Kan, P. W. Garrett-Engele, J. M. Johnson, J. C. Castle, *Nucleic Acids Res* **33**, 5659-5666 (2005).
- S107. B. Modrek, C. J. Lee, *Nat Genet* **34**, 177-180 (2003).
- S108. Q. Pan *et al.*, *Trends Genet* **21**, 73-77 (2005).
- S109. R. Sorek, R. Shamir, G. Ast, *Trends Genet* **20**, 68-71 (2004).
- S110. T. A. Thanaraj, F. Clark, J. Muilu, *Nucleic Acids Res* **31**, 2544-2552 (2003).
- S111. <http://evolutionhighway.ncsa.uiuc.edu>.
- S112. <http://oxgrid.angis.org.au/cattle/>.
- S113. J. Xie *et al.*, *Proc Natl Acad Sci U S A* **101**, 10750-10755 (2004).
- S114. <http://imgt.cines.fr>.
- S115. S. A. Ellis, K. T. Ballingall, *Immunol Rev* **167**, 159-168 (1999).
- S116. S. A. Ellis *et al.*, *Immunogenetics* **50**, 319-328 (1999).
- S117. J. Birch, G. Codner, E. Guzman, S. A. Ellis, *Immunogenetics* **60**, 267-273 (2008).
- S118. <http://www.ebi.ac.uk/ipd/mhc/bola/>.
- S119. J. Birch, C. De Juan Sanjuan, E. Guzman, S. A. Ellis, *Immunogenetics* **60**, 477-483 (2008).

- S120. C. P. Childers *et al.*, *Anim Genet* **37**, 121-129 (2006).
- S121. A. P. Ambagala, Z. Feng, R. G. Barletta, S. Srikumaran, *Immunogenetics* **54**, 30-38 (2002).
- S122. P. D. Karp *et al.*, *Nucleic Acids Res* **33**, 6083-6089 (2005).
- S123. D. M. Irwin, *J Mol Evol* **41**, 299-312 (1995).
- S124. We thank funding officers for their ongoing efforts: J. Peterson (Project Officer), C. Bennet, A. Felsenfeld, M. Guyer, J. Malone, L. Wang, and K. Wetterstrand of NHGRI; R. D. Green (Project Officer) and S. M. Kappes of USDA ARS; D. Hamernik (Project Officer) of USDA CSREES; C. Bell (Genome Canada); R. Baker and A. Crawford (AgResearch Ltd.); B. Church (ASRA); E. Dressler of the National Beef Council; K. A. Eversole of Eversole & Associates; S. Moore; W. Roberts (State of Texas); R. Tellam (Project Officer) of CSIRO; and R. Wortham (Texas Beef Council).
- S125. Analysis of the bovine genome was supported by the following sources of funding for individual consortium members. C. Elsik: USDA National Research Initiative (NRI) grant 2007-35616-17882, Kleberg Foundation, Georgetown University, the Texas Agricultural Experiment Station; H. Lewin: USDA Cooperative State Research Education and Extension Service, Livestock Genome Sequencing Initiative (AG 2005-34480-15939); E. Eichler: NIH grant HG002385; M. Rijkels: USDA Agricultural Research Service Cooperative Agreement 6250-51000-048; E. Zdobnov and S. Antonarakis: The Swiss National Science Foundation; D. Lynn, M. Whiteside, and F. Brinkman: Genome Canada, Genome BC and the Michael Smith Foundation for Health Research; D. Bradley and L. Lau: Science Foundation Ireland; F. Brinkman: Canadian Institutes of Health Research; E. Memili: Mississippi Agricultural and Forestry Experiment Station; A. Iivanainen: The Academy of Finland (122540/2007) and Research Funds of The University of Helsinki (914/51/2006); C. Gill: USDA NRI grant 2007-35604-17870; S. Hiendleder and C. Fitzsimmons: JS Davies Bequest; J. Taylor and R. Schnabel: USDA NRI grants 2005-35205-15448, 2005-35604-15615, 2006-35205-16701 and 2006-35616-16697. This research was also supported in part by the Intramural Research Program of the NIH National Library of Medicine, EADGENE (EU Contract No. FOOD-CT-2004-506416), and Institute Strategic Grant funding from the Biotechnology and Biological Sciences Research Council.
- S126. We thank the following individuals for contributions to this research: Jonathan Usmar for preparing the Oxford Grid; Shelia Alexander, Cal Davison, R.J. Hubbard, Whisper Kelly, Vicki Leesburg, Kathy Meidinger, Ryan Rienstra, Brooke Shipp, and Heidi Stroh for collection of tissue samples for cDNA sequencing. We thank the following individuals for contributions to sequence production: Carlana Clashette Allen, Ugonna Sharon Anosike, Ashton Vashawn Bell, Carla Bickman, Veronica Cardena, Kelvin Carter, Alejandra Chavez, Dean Chavez, Hau-Seng Chu, Raynard Cockrell, Mary Louise Davila, Latarsha Davy-Carroll, Shawn Denson, Victor E. Ebong, Sonia Fernandez, Pushpa Ranjani Fernando, Courtney Sherell Francis, Jason Garner, Ricardo M. Garcia III, Tiffany Evette Garrett, Brandy A. Harbes, Ebere Sylvia Onyirioha Hawkins, Marilyn Hogue, Barbara Hollins, LaToya Howell, Bennie Johnson, Laquisha Monique King, Haika Kisamo, Liza Alvarez Lago, Chuan-Yar Lai, Fremiet Lara, Fitzherbert Henderson Legall III, Thanh-Kim Thi Le, Dhammika Liyanage, Pamela London, Lorna M. Lorensuhewa, Renita C. Madu, Evangelina Martinez, Tittu Mathew, Christian Mercado, Iracema Cavazos Mercado, Mala Munidasa, Dinh Ngoc Ngo Phong Quoc Nguyen,

Ogechi O. Nwaokelemeh, Melissa Obregon, Evelyn Odeh, Chibueze G. Onwere, Andrea Alexandra Parra, Heidie A. Paul, Agapito Perez, Yolanda Yaneth Perez, Eltrick L. Primus, Maria Puazo, Juana B. Quiroz, Dina Rabata, Moazzam Sana, Brian W. Schneider, Ida Sisson, Richard P. Sorelle, Rosenie Thelus, Nicole Thomas, Rachel Diane Thorn, Reshaunda Devon Thornton, Zulma Y. Trejos, Kamran Usmani, Courtney Sherell White, Aneisa C. Williams. We thank the following individuals for contributions to sequence finishing: Guan Chen, Marcus D. Coyle, Alicia C. Hawes, Laronda R. Jackson, Ziad Mohid Khan, Zhangwan Li, Wen Liu, Lesette M. Perez, Hua Shen, Suzhen Wang, Qiaoyan Wang, Jennifer Eunyoung Watt, Jianling Zhou.

- S127. Mention of trade names or commercial products is solely for the purpose of providing information and does not imply recommendation, endorsement or exclusion of other suitable products by any of the institutional participants.

