www.sciencemag.org/cgi/content/full/324/5926/528/DC1

Supporting Online Material for

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

The Bovine HapMap Consortium

*To whom correspondence should be addressed. C. P. Van Tassell (curt.vantassell@ars.usda.gov), J. F.Taylor (taylorjerr@missouri.edu), and R. A. Gibbs (agibbs@bcm.tmc.edu).

Published 24 April 2009, *Science* **324**, 528 (2009)
DOI: 10.1126/science.1167936

This PDF file includes

Table of Contents

Materials and Methods

1. Bovine biological inventory database for the HapMap project

For storage and dissemination of the data generated by the Bovine HapMap Project a database with a web interface (http://www.bovinehapmap.org) was developed. The database schema (Fig. S1) and goals of the database development are described below.

Data Flow: Animals from each breed were identified by the breed champion in collaboration with the breed association. Permissions were obtained from animal owners by the breed champion and then tissue samples (semen, blood, and/or ear notches for tissue culture) were procured. The breed champion was responsible for ensuring that sufficient tissue was collected to meet the needs of the project (sufficient for 5 mg of DNA for the first 24 animals per breed or 200 µg of DNA for subsequent animals).

Extracted DNA was shipped to the central repository in the Department of Animal Science at Texas A&M University (TAMU, USA). Upon receipt at the repository, the DNA sample was assigned a project identification number and bar code. If the samples were from members of a trio, then QC genotyping was performed at the repository to verify parentage. If the sample was associated with an animal ID that had previously been received by the repository, then QC genotyping was performed to verify the sample identity. DNA concentrations were verified by PicoGreen assays (Invitrogen, Carlsbad, CA) and samples were stored at -80 ^{o}C.

Transactions: The database was designed to be dynamic with several possible permissible transactions. These transactions included:

1. Adding new biological samples to the inventory.
2. Assignment of unique IDs to track samples individually or as a batch.
3. Sending samples for genotype analysis.
4. Receiving genotype data.
5. Performing QC analysis on the genotype data.

Field Constraints: All database fields have a controlled vocabulary to avoid duplication of records and spelling mistakes. These constrained fields include tissue type, tissue units, and breed type. New database entries for these data types are allowed only from a web interface on a separate screen.

Animal Identification and Owner: Each animal was assigned a unique number by the database. The same animal having different IDs in the U.S. and in other countries was tracked through the cross reference table and was merged and assigned a new identification number. Entry of phenotypic data for an animal was optional.

***Breed*:** Each animal's breed was identified by a unique three letter ID. The breed table also included a definition field to identify breeds known to be taurine, indicine or hybrids and to identify generation numbers.

Sponsors and Breed Champions: A field was devoted to identifying individual(s) responsible for selecting animals to be genotyped, managing biological sample collection, DNA extraction and quality control. Multiple champions per breed were permitted as were multiple breeds per champion. Breed champions were able to view and edit the data pertaining to their breeds.

Pedigrees: The sire and dam information for all genotyped animals was stored in the Animal ID table. Most breeds currently have pedigree information on each animal for at least four ancestral generations. A Boolean field was included in this table for trio validation by genotyping. With flexible definitions for membership numbers, almost any relationship structure could be accommodated.

Permissions: The database was password protected. All participants who had been issued a user ID and password were able to view the data, however, data editing was limited to the respective breed champions and database administrators.

Database Queries and View Screens for Web Interface: The web interface allowed for querying the database for breed summary statistics, individual animal data, and genotyping success rates.

The Animal ID table (Fig. S2) provided a unique animal identifier and the anonymized HapMap identifier as well as critical animal specific information:

- Animal Key - Primary identifier for an animal. Most tables use this internal numeric ID to connect to animal information.
- Breed Key - Internal numeric identifier connects to the breed table to identify the standard, three character, ICAR breed code identifier.
- Num in Breed - This number along with the breed identifier was used to create the public animal ID, e.g., HOL1, ANG243, etc.
- The animal specific owner and genotyping sponsor contact identifiers that link to the contact table**.**

The Animal Process table (Fig. S3) provides:

- DNA Ship Key - pointer to the record of shipment of DNA to a genotyping lab.
- Assay Key - pointer to a combination of markers that identify an assay.
- Genotyping Batch Key - pointer to a group of genotypes produced in one batch for a breed. This field was needed for quality-control purposes, as the batch and individual animals were required to meet quality-control criteria.
- Pay Animal - on the basis of the QC data for the batch and the animal, this field was changed by the breed champion indicating that the QC criteria were met for this animal and batch.

The Animal tables (Fig. S4) define the relevant identity related information. The Pedigree table includes only relationships among animals and birth year to allow basic validation. The Breed table defines the three character breed abbreviations and provides for a full breed description. The final piece of information is a list of aliases for an animal. These aliases might include registration numbers from different countries, AI company identifiers (e.g., National Association of Animal Breeders codes), or names. A single alias was identified as the primary identifier, and this was usually the registration number from the country of origin. Aliases were only viewable by the respective breed champion and only the HapMap identifiers were publicly released.

The set of Inventory tables (Fig. S5) defines the core of the Laboratory Information Management System and identifies all locations housing tissues or DNA. A tissue table which included linked tables to constrain variables to a controlled vocabulary was included. These tables were modifiable by breed champions, as the personnel at these locations maintained tissue banks. The DNA in the inventory was expected to be maintained at a very small number of locations, such as at TAMU or the International Livestock Research Institute (Nairobi, Kenya). Finally, pending genotypic data to provide validation, a holding pen was provided for incoming DNA before it was added to the DNA inventory. Tissue samples were tracked from the tissue database through the DNA extraction process into a holding pen table, which allowed complete sample tracking.

The Assay tables (Fig. S6) defined an assay as a combination of genotypes and markers. A single SNP could be represented by multiple marker keys to allow its genotyping in different assays.

The QA/QC tables (Fig. S7) generate the information required to make payment decisions for genotypes on the basis of QC data. The Genotyped Batch table was used to define a group of genotypes within a breed. These QC parameters include percent genotype calls for all markers in the batch, percent of markers called above a threshold (e.g., 95% of the time -- this parameter was from the contract with a genotyping center), percent of animals called and percent of markers conforming to expectations of Mendelian inheritance (on the basis of trio information). In addition, to facilitate payment decisions on individual animals, the percentage of markers called was calculated and updated in a separate table. Finally, to facilitate use of the data by people not needing complete pedigree data, the trio table identified trio members. The member field was flexible to include sire, dam, offspring, maternal grandsire, etc. A table indicating that the trio had been validated was an additional QC tool.

The primary user information was maintained in a contact table (Fig. S8) that provided information for the breed champions, animal owners, and breed sponsors. The structure allowed animals within a breed to have multiple owners, sponsors, and champions, all of whom could be associated with multiple breeds. The schema stored a single Materials Transfer Agreement which was executed with an individual owner to represent all animals associated with that owner.

2. SNP discovery

***Across-Breed Detected SNP*:** Sequence reads were generated from random shotgun libraries from the Holstein, Angus, Brahman, Limousin, Jersey, and Norwegian Red breeds (Table S2). The sequence reads from these breeds were compared to release Btau20040927 (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/snp/Btau20040927/) of the bovine genome (a Hereford) with BLAST with default criteria and a 1e-50 expectation cut off. The criteria for selecting a read for SNP analysis were as follows:

a. The clone insert mate pairs mapped uniquely to a scaffold, with spacing equal to the library average insert size, and one of the BLAST hits was at least 400 bases and both reads hit with at least a 97% identity.

b. If the reads hit separate scaffolds, or if only one read of a mate pair hit the genome, then the hit length had to be 800 bases with at least a 97% identity.

c. From reads passing either of these two criteria, variants that passed the following quality criteria were selected:

i. For the query sequence, the variant base phred quality was $\geq$30 with phred quality $\geq$20 in the flanking bases.

ii. For the reference genome the variant base had phrap quality of $\geq$60 and for flanking bases $\geq$40.

iii. Not more than one additional base difference in the 20 base flanking sequence.

This method was very similar to the NQS method (*S1*) which was developed for the initial mapping of human SNP but all thresholds used here were more stringent. Table S2 shows the number of SNP detected from reads in each of the six breed groups as well as the numbers from each breed of discovery that were assayed. The SNPs are available in dbSNP (Holstein Submitter ID BCM-HGSC, Batch IDs 10002 (ss61470116 to ss61497504); Brahman Submitter ID BCM-HGSC, Batch ID 10001(ss61497505 to ss61505881); Angus Submitter ID BCM-HGSC, Batch ID 10000 (ss61505882 to ss61512060); Jersey Submitter ID BCM-HGSC, Batch ID 10003 (ss61456395 to ss61461128); Limousin Submitter ID BCM-HGSC, Batch ID 10004 (ss61569998 to ss61570804); Norwegian Red Submitter ID BCM-HGSC, Batch ID 10005 (ss61512079 to ss61570013); and mixed sets Submitter ID BCM-HGSC, Batch IDs Bos_taurus_042605 (ss38322072 to ss38337338), 001234_SS, and 856981)

The number of SNP identified per 1,000 bases from a comparison of random shotgun sequence reads to the Hereford assembly was uniform among the taurine breeds but was almost twice as great in the indicine Brahman (Fig. S9).

Other across-breed SNP discovery efforts by the consortium included:

(i) Additional sequencing of Brahman shotgun libraries discovered 40,501 putative SNP allowing a more even taurine and indicine origin of SNP representation in the study. DbSNP Submitter ID BCM-HGSC, Batch ID 856984.

(ii) Inclusion of 4,687 SNP from Holstein representing regions targeted on chromosomes 6, 14, and 25 to identify short range LD and signatures of selection. DbSNP Holstein Submitter ID BCM-HGSC, Batch ID 856982.

(iii) PCR primer design, amplification and deep sequencing of targeted ENCODE-like regions in the Holstein and Angus breeds to assess the ascertainment bias in SNP discovery and long range LD in cattle. Five regions of ~500 kb each, on chromosomes 6 and 14 were sampled with closely spaced amplicons in the middle 10kb, sample 500 bp amplicons every 5 kb out to 100kb and every 20 kb in the remainder of the 500 kb. DbSNP Submitter ID BCM-HGSC, Batch ID 856983.

(iv) Inclusion of about 500 SNP from the Interactive Bovine In Silico SNP (IBISS) database (*S2*) that were discovered from the alignment of EST sequences. DbSNP Submitter ID IBISS, Batch ID IBISS4_072005 (ss46525994 to ss46527139).

(v) Inclusion of about 500 SNP from aligning Holstein BAC-end sequences to the Hereford Assembly. DbSNP Submitter ID TAMU ANIMAL GENOMICS , Batch ID TAMU1 (ss66537302 to ss66538295).

(vi) Comparison of the parental haplotypes within the Hereford genome sequence identified putative SNP, available in dbSNP Submitter ID BCM-HGSC, Batch IDs 11000, 11001, 11002, 11003, 11004 (ss61858563 to ss62358562, ss63246193 to ss63746192, ss64408020 to ss64908019, ss65023065 to ss65523064, and ss65124342 to ss65624341).

3. Preliminary SNP validation

Tests via the Sequenom MassARRAY® (Sequenom, San Diego, CA) technology of putative SNP derived from the first set of across-breed shotgun sequence comparisons showed that approximately 80% of the SNP could be validated as real DNA sequence variants. The SNP from the IBISS (iv) and BAC-ends (v) were found to be monomorphic at rates of 9.3% and 32%, respectively. These preliminary data show an ascertainment bias due to the breed of discovery and prompted additional discovery efforts in Brahman.

4. Details of genotype QA/QC analyses

The genotypes produced in this study were collected on multiple platforms and at multiple time points. To ensure the highest possible quality data a series of filters were employed to remove lower quality markers and genotype sets for individuals. Outgroup animals were not considered during any of the QA/QC steps. The process used to identify these markers and individuals proceeded as follows:

(1) *Markers genotyped on multiple platforms*. Duplicate marker genotypes from all platforms were identified and genotypes from the platform having the highest call rate were retained.

(2) *Monomorphic markers*. Fifteen markers had no variation across all breeds and were removed.

(3) *Trio Concordances*. To evaluate the assay performance three trios (sire, dam, and offspring) per breed were genotyped. For each trio, a marker was identified as being discordant or concordant according to whether Mendelian errors were observed. Discordance counts were accumulated across breeds for each of the markers and all markers with more than two concordance errors were flagged and manually evaluated for retention in the case that the locus could actually be assigned to the X chromosome.

(4) *X - linked markers*. The cattle genome is not fully assembled and about 15% of the genome has been placed in contigs with no chromosomal assignment (identified as 'ChrUn'). Markers from the ChrUn contigs with high numbers of trio errors but for which all males were homozygous were likely to be located on the X-chromosome. All markers having an inheritance pattern consistent with X-chromosome linkage were retained.

(5) *Missing data per individual*. The total number of genotypes and missing genotypes were calculated for each individual and genotype completeness was calculated as the percentage of called genotypes of the total possible genotypes (called + missing). The resulting distribution was clearly bimodal with 10 individuals, who were removed, having genotype completeness less than 70% and the remaining 487 individuals having genotype completeness greater than 91%.

(6) *Missing data per breed*. Allele frequencies (genotype counts) and missing data for each marker by breed combination were calculated for the remaining animals to remove SNP where the genotyping data were problematic. Simple threshold filters that could be unilaterally applied across all breeds were not identified due to the intrinsic ascertainment bias in the SNP (as described in the manuscript). For example, taurine derived SNP had a higher missing data rate in indicine than in taurine animals and *vice versa*. Furthermore, there were different numbers of each breed type represented: taurine (14), indicine (3) and hybrid (2). For each marker, individual breeds that had >10% missing data were identified. Loci for which >50% of the taurine breed individuals and >50% of the indicine breed individuals had >10% missing data were excluded. Because there were only 3 indicine breeds represented in the data, this required 2 of the 3 breeds to have >10% missing data. This filter removed data for 338 loci.

(7) *Hardy-Weinberg Equilibrium (HWE)*. Next, a test for conformance to HWE within each breed was applied. Due to the large number of performed tests, an empirical threshold was determined by ordering the resulting test P-values for each breed and

identifying the threshold yielding the largest 1% of P-values. This analysis identified 3,426 loci for which at least one breed did not conform to HWE.

(8) *Locus genotyping error rate*. We next estimated the underlying genotype error rate for each individual locus for which at least one trio genotype discordance was observed as follows. For each locus, we first assumed that the probabilities of observing a genotype conditional on the true underlying genotype was a function of an unknown error parameter ε as specified in Table S3.

If we let O_M, O_F, and O_P represent the observed genotypes and T_M, T_F, and T_P represent the true genotypes of the mother, father and progeny of a trio, respectively, the probability of observing a specific trio genotype can be written:

$$P(O_M, O_F, O_P) = \sum_T P(O_M, O_F, O_P \mid T_M, T_F, T_P)\, P(T_M, T_F, T_P)$$

where the summation *T* is across all 15 genotypic combinations of true trio genotypes that are consistent with Mendelian segregation. Assuming that true genotype frequencies conform to HWE proportions and that allele frequencies are F(A) = p and F(B) = 1 – p = q we have, for example, $P(O_M = AA, O_F = AA, O_P = AB \mid T_M = AA, T_F = AA, T_P = AA) = 2(1\text{-}m)^3(1\text{-}\varepsilon)^5\varepsilon$ and $P(T_M = AA, T_F = AA, T_P = AA) = p^4$,where m is the proportion of missing data for the locus which we assume is an estimate of the probability that a randomly selected individual's genotype is missing and we use the conditional probabilities in Table S3. From the result above, the probability of observing a set of discordant trio genotypes is:

$$P_D = \sum_D P(O_M, O_F, O_P)$$

where the summation *D* is across all 16 possible discordant observed trio genotype possibilities when we allow for the possibility that one parent may have a missing genotype. Thus for a locus with N_T genotyped trios, the expected number of discordant trios is $N_T P_D$. From this, we estimated the parameter ε for each locus by equating the expected number of discordant trio genotypes to the observed number N_D of discordant trio genotypes. The estimate of ε utilizes an estimate of allele frequency and missing data rate that is averaged across breeds and also assumes HWE across breeds. These assumptions are likely violated due to different allele frequencies within breeds and the higher likelihood of missing data in certain breeds due to specific mutations that interfere with genotype calling. However, the sample size was too small to precisely estimate breed specific allele frequencies and missing data rates and thus we considered the approach to be a useful approximation to the identification of loci that were problematically scored in cattle. Finally we used the locus specific estimates of ε to produce an estimate of the locus specific genotyping error rate:

$$ER = p^2[1\text{-}(1\text{-}\varepsilon)^2]+2pq[1\text{-}(1\text{-}\varepsilon)^2\text{-}\varepsilon^2]+q^2[1\text{-}(1\text{-}\varepsilon)^2] = \varepsilon[2\text{-}\varepsilon(1+2pq)]$$

The aforementioned 3,426 loci that were found to not be in HWE were ordered based on estimated genotyping error rates, and the number of breeds found to be out of HWE. A 5% threshold for the locus genotyping error rate was chosen which means that a SNP had to have an estimated genotyping error rate >5% and at least one breed out of HWE to be excluded. This resulted in removing an additional 393 loci.

(9) *Minor allele Frequency (MAF)*. There were 264 SNP identified with a $0 < MAF < 0.04$ in all breeds which were removed.

(10) *Gender Determination*. The gender of an individual was not available for some of the genotyped animals. Several loci on the X-chromosome outside the pseudo-autosomal region were evaluated for heterozygosity in individual animals to confirm gender. Animals having low heterozygosity values (allowing for genotyping error) were identified as males and those having high heterozygosity values were identified as females. This enabled the sex of all individuals to be inferred and identified the mislabeling of gender for 7 animals from different breeds.

5. Analyses of allele frequencies: Breed differences and STRUCTURE analyses

Analysis of the entire panel of across-breed discovered SNP revealed a uniform distribution of allele frequencies by breed (Table S4) but when the individual allele frequencies were examined in groups separated by the breed of discovery there were clear and expected differences in the observed allele frequencies (Table S5).

The smallest allele frequency that could be estimated in this study was 0.045 within the Red Angus breed, however, within most other breeds the lower bound for allele frequency estimates was about 0.021. We filtered the data to remove SNP with MAF<0.04 and this should have resulted in the removal of the majority of rare SNP and haplotypes from the data.

Structure Analysis: Population structure and admixture analysis was also performed with the most commonly used software STRUCTURE (*S3*). STRUCTURE assumes a model in which there are K progenitor populations (where K can be a variable), each of which is characterized by a set of allele frequencies at each locus. The method attempts to assign individuals to populations on the basis of their genotypes, while simultaneously estimating progenitor population allele frequencies. The data were analyzed for admixture and homogeneity of breeds by varying the number of clusters (K) from 2 to 10 (Fig. S12). For this analysis, the linkage model was used assuming correlated allele frequencies among populations but varying F_{ST} for different sub-populations. The length of burn-in period, and number of MCMC replications after burn-in were each set to 1,000. The results were checked for convergence from duplicate runs.

At K = 2, the principal separation was between the taurine and indicine breeds. The hybrid breeds Sheko, Santa Gertrudis, and Beefmaster revealed their taurine and indicine admixture. Nelore and Gir cluster as pure indicine breeds, while the U.S. Brahman reveals a marginal taurine admixture. N'Dama, Charolais, and Romagnola showed marginal admixture with indicine cattle. At K > 2, the separation was primarily among the taurine cattle. At K=3, Hereford clustered with Limousin, Charolais, Piedmontese, and Romagnola while Holstein clustered with Angus, Red Angus, Norwegian Red, and Jersey. Brown Swiss and Guernsey also showed admixture from the second and third progenitor populations at K = 3. Santa Gertrudis showed admixture from all three populations. N'Dama and Jersey were identified as a single population with no admixture until K = 7. At values of K > 7 admixture was detected in these breeds.

6. Analyses performed for chromosomes with dense markers

Extended Haplotype Sharing: We analyzed the sharing of phased haplotypes extending over multiple markers across breeds for chromosomal regions with a high density of markers, including portions of chromosomes 6, 14, and 25. Haplotypes were inferred for all animals with fastPHASE version 1.2.3 (*S4*). Haplotype segments which we considered to define a single locus comprised 10 adjacent markers spanning no more than 200 kb. New loci were defined by sliding the window along the chromosome one SNP at a time, allowing haplotype loci to overlap. The number of markers per window was held constant in order to keep the range of possible alleles consistent for each window. The upper bound on the window size was imposed to exclude regions with relatively less dense markers and to reflect the hypothesis that LD in cattle extends only about 500 kb. The proportion of shared haplotypes between two populations *P1* and *P2* at locus *k* was defined as:

$$\mathrm{S}(P_1, P_2, k) = \sum_{i,j} \mathrm{S}_\mathrm{a}(i, j, k)/(2*(n_1 + n_2))$$

where *i* and *j* range over the individuals in populations P_1 and P_2, respectively, $\mathrm{S}_\mathrm{a}(i, j, k)$ is the number of shared haplotypes between individuals *i* and *j* at locus *k*, and n_1 and n_2 are the number of samples in P_1 and P_2. This raw proportion can be normalized to take into account the proportion of shared haplotypes within each of the individual populations, as follows:

$$\mathrm{S}'(P_1, P_2, k) = 2*\mathrm{S}(P_1, P_2, k)/(\mathrm{S}(P_1, P_1, k) + \mathrm{S}(P_2, P_2, k))$$

$\mathrm{S}'(P_1, P_2, k)$ has a value of 1.0 if the proportion of shared haplotypes between populations P_1 and P_2 at locus *k* is equal to the average of the proportion of shared haplotypes within the two populations P_1 and P_2. If $\mathrm{S}'(P_1, P_2, k) << 1.0$, the proportion of shared haplotypes between the two populations is much less than the average within the two populations.

We applied these measures across entire chromosomes to facilitate the identification of regions of haplotype similarity and haplotype diversity between breeds. For example, Fig. S13 shows values of S′ for haplotypes within a 5 Mb region of chromosome 6 spanning positions 36-41 Mb. Each locus in the figure is represented as a bar whose width indicates the length of the 10-marker window and whose height represents the normalized proportion S′ of shared haplotypes at that locus. The region of the chromosome in the figure exhibits a wide range of values for S′. For example, the locus near the center of the figure at 38.5 Mb represents a region spanning approximately 174 kb in which the proportion of shared haplotypes between Angus and Holstein is 0.45, not significantly different from the proportion of shared haplotypes within Angus (0.45) and within Holstein (0.46). An examination of the specific haplotypes involved shows that there are four 10-marker haplotypes that are shared between Angus and Holstein, accounting for 103/106 (97%) of the observed Holstein haplotypes and 47/54 (87%) of the observed Angus haplotypes for this segment of the chromosome. In contrast, Fig. S13 also indicates a 119 kb region at about 37.7 Mb in which the proportion of shared haplotypes between Angus and Holstein is 0.13, compared to the proportion shared within Angus (0.28) and within Holstein (0.22).

Fig. S14 shows shared haplotypes between Angus and Holstein within a 5 Mb region of BTA14 with a dense set of markers (between positions 7-12 Mb). The figure illustrates regions in which most haplotypes are shared between breeds (such as the region surrounding 9.9 Mb) and regions in which little haplotype sharing occurs between breeds (such as the 175 kb region around 7.8 Mb). As shown in Table S6, the observed mean proportion of shared haplotypes between Angus and Holstein is slightly higher on chromosome 6 (0.240) and on chromosome 14 (0.247) than on chromosome 25 (0.214). Chromosome 25 (Fig. S15) was selected as a control due to the absence of known QTL on this chromosome.

Further analysis is required to determine whether the difference in haplotype sharing on chromosome 6 is due to drift or selection. As a preliminary step, the autocorrelation coefficient was computed and the results show that the proportion of shared haplotypes exhibits significant locality, meaning that it does not randomly vary about the mean, indicating selection as a potential causal mechanism.

Clustering Breeds on the Basis of Shared Haplotypes: The proportion of shared haplotypes can be used as a distance measure for clustering breeds. The normalized distance between breeds P_1 and P_2 was calculated with:

$$\mathrm{D}'(P_1, P_2) = 1 - \sum_{\mathrm{k}} \mathrm{S}'(P_1, P_2, k)/u$$

where u is the number of loci. This is related to common measurements for genetic distance between two individuals (*S5-S7*). $\mathrm{D}'(P_1, P_2)$ has value 0 if breeds P_1 and P_2 share the same proportion of haplotypes as are shared by the individuals within each individual breed.

Figs. S16-S18 show dendrograms of cattle breeds calculated on the basis of $\mathrm{D}'(P_1, P_2)$ for all pairs of breeds over the densely genotyped chromosomes 6, 14, and 25, produced with the NEIGHBOR program and with UPGMA clustering in PHYLIP version 3.6 (*S8*). Chromosome 25 was selected on the basis of the lack of known QTL on this chromosome, thus serving as a control for artificial selection. The dendrogram derived from chromosome 25 shows several expected clusters: The taurine breeds cluster towards the upper left, followed clockwise by the African breeds (N'Dama and Sheko), the taurine × indicine hybrids (Beefmaster and Santa Gertrudis) and the indicine breeds (Nelore, Gir, and Brahman). These results are consistent with expectations on the basis of the known recent ancestries of the breeds represented in the study. Compared to the dendrogram for BTA25, some differences are apparent in the dendrograms derived from haplotypes on chromosomes 6 and 14. For example, the Beefmaster and Santa Gertrudis breeds cluster within the taurine breeds, and are most closely associated with Hereford on chromosomes 6 and 14. While the Beefmaster was originally developed as a cross among Hereford and Shorthorn with Brahman cattle, Santa Gertrudis cattle are approximately five-eighths Shorthorn and three-eighths Brahman. Thus, these breeds should cluster between the indicine and taurine breeds if the haplotype loci are neutral as seen in Fig. S16. However, if selection has occurred in favor of taurine alleles, one would expect the hybrids to cluster close to the Hereford breed as seen in Figs. S17 and S18. While Shorthorn cattle were not included in this study, they are a closely related British breed to the Hereford which may explain why the Santa Gertrudis also clustered with Hereford in these figures.

Estimation of the Distribution of Unascertained SNP MAF Frequencies: The SNP utilized in the population genetic analyses of the 19 sampled breeds were ascertained to be among the most common within the bovine genome (Table S4). To assess the extent of bias towards SNPs with high MAF, we analyzed 1,201 SNPs detected in the sequencing of 119 fragments from regions of chromosomes 6 and 14 in 18 Angus, 16 Holstein, and 5 Brahman individuals. Fig. S19 shows the expected unascertained SNPs MAF frequencies within taurine (Angus and Holstein) and indicine (Brahman) breeds estimated from these genomic regions. Despite the higher nucleotide diversity within Brahman, the estimated SNP MAF distributions are very similar which is consistent with the similar current effective population sizes for Angus ($\mathrm{N_e}$ = 136), Holstein ($\mathrm{N_e}$ = 99) and Brahman ($\mathrm{N_e}$ = 115) (Table S1). These results should be treated with some caution in view of the very limited numbers of genomic regions and animals that were surveyed. Nevertheless, these data are currently the only estimates of the unascertained SNP MAF distributions that are available for cattle.

7. Assignment of ancestral allele state

As expected, the Anoa and Water buffalo genotypes were monomorphic for the majority of filtered markers that were successfully genotyped (10,371); 48 markers were fixed for alternate alleles in the Anoa and Water Buffalo so the ancestral state for these loci could not be determined. A heterozygous genotype was produced by at least one animal for 2,497 markers. For 16 segregating markers, alternate homozygotes were observed and 1,789 markers were segregating in both species so the ancestral state for these loci also could not be assigned. The ancestral allele was assigned for 11,366 markers, of which, 10,427 were assigned to chromosomes (Table S7).

Ancestral allele determination: The latest genome assembly builds for dog (build 2), horse (build 1), and human (hg 18) were downloaded and used in a comparative genomics approach to identify the ancestral allele at each SNP locus (Fig. S20).

A total of 32,018 markers had chromosomal assignments, but of these, 1,315 were filtered because the flanking sequences altered between consecutive cattle genome builds. Orthologous sequences with high BLAST scores were used to identify the corresponding SNP base within the human, dog, and horse genome sequences. As expected from the phylogenetic tree (Fig. S20), the horse sequence generated the most hits (20,813), followed by dog (14,550), and human (12,920).

In total the two methods predicted the ancestral allele for 24,562 markers. To compare the approaches, we selected 7,864 markers for which the ancestral allele was determined by both outgroup genotyping and sequence orthology. Strict concordance (the same allele found in all 5 species, where orthology could be determined) of ancestral allele predictions was observed for 4,620 markers (58.7%). However, for 1,787 SNP (22.7%) both cattle alleles were observed in human, horse, and dog and ancestral state could not be assigned. For the remaining 1,457 SNP (18.5%) two alleles were found in human, horse, and dog, but only one was common with cattle. By excluding the more divergent human comparison, the strict concordance between the methods increased to 5,898 markers (75%). However, there were 782 SNP (9.9%) where both cattle alleles were observed in horse and dog and for the remaining 1,184 SNP (15.1%) two alleles were found in horse and dog, but only one was common with cattle.

8. Signatures of selection

Overview of methods used: The iHS statistic was calculated for each polymorphic site within a breed. The iHS statistic measures the extent of local LD, partitioned into two classes: haplotypes centered upon a SNP that carry the ancestral *versus* the derived allele (*S9*). Directional selection favoring a new mutation results in a rapid increase in the frequency of the selected allele along with the background haplotype on which the mutation arose. This phenomenon increases LD on the chromosomes which harbor the derived (selected) allele. Thus, this measure is most sensitive to a rapid increase in the frequency of the derived allele at a selected site, but the derived allele must have existed on few distinct backgrounds (haplotypes) prior to selection and have not yet reached fixation. After fixation, the iHS statistic may continue to identify regions of high LD surrounding the selected site, but may not detect selection at the selected region itself because fixation will eliminate variation at and near the selected site.

The F_{ST} statistic is a classical measure of the degree of differentiation between subpopulations and has previously been used to detect selection with genome-wide SNP data (*S10, S11*). Strong selection in one or more breeds (directional or diversifying selection) would produce differences in allele frequencies at linked neutral sites that exceed the level expected under isolation and drift alone. Conversely, stochastic variation in allele frequencies would be constrained if balancing selection were present in ancestral populations and persisted during and after breed formation. In this case, F_{ST} would indicate a lower level of differentiation between breeds than expected from isolation and drift. The distribution of F_{ST} averaged across a sliding 8 SNP window along all chromosomes and for all breeds is shown in Fig. S21. Extremely high values represent likely instances of divergent selection and extremely low values represent likely instances of balancing selection.

Finally, we employed the composite likelihood method (*S12*) to assess the unique spatial pattern of allele frequencies expected under the selective sweep model. Briefly, upon fixation, variation is eliminated at the selected site and at nearby linked regions, but more distal regions may partially or completely escape this purge if recombination allows both SNP alleles at a linked neutral site to become associated with the selected allele. The probability of this occurring increases with genetic distance from the selected site.

Assuming a selective fixation has occurred at a given chromosome position, the distribution of allele frequencies at linked neutral sites can be determined as a function of genetic distance, strength of selection, and distribution of allele frequencies prior to selection. Formulated as a composite likelihood function, the likelihood of the data under the selective sweep model is maximized with respect to chromosomal position and strength of selection. This constrained likelihood is compared to the unrestricted likelihood of a null model that assumes no selection. The logarithm of the ratio of these likelihoods is often plotted against chromosomal position to facilitate the identification of putative selective sweeps.

With all of these methods, statistical significance can only be assessed by generating an empirical null distribution from simulations that capture features of the data that are unrelated to selection. For example, an important feature to consider in a simulation of the data is the impact of ascertainment bias on allele frequency distributions. Because most of the SNP used in this study were discovered by comparing single shotgun reads from another breed to the Hereford derived reference genome, SNP with the largest differences in frequency between a discovery breed and Hereford are most likely to be polymorphic in the discovery breed. The distribution of MAF for SNP derived with this strategy, is, as a result, heavily biased toward high MAF SNP. Correction for ascertainment bias requires knowledge of the joint allele frequency distribution across all breeds. Because this distribution is unknown, our identification of genomic regions that have been subjected to recent selection should be considered tentative.

Calculation of |iHS| values: The integrated extended haplotype homozygosity score (iHS) provides a measure of recent positive selection because when an allele increases rapidly in frequency due to strong selection, it tends to be associated with high levels of haplotype homozygosity that extend further than expected under a neutral model (*S9*).

The iHS computing tool uses estimated haplotypes, estimated recombination rate and ancestral allele state to compute unstandardized iHS values:

$$\text{unstandardized iHS} = ln\left(\frac{iHH_A}{iHH_D}\right)$$

where iHH_A and iHH_D refer to the integrated extended haplotype homozygosity score (*S13*) for the ancestral and derived alleles, respectively. To adjust for the age of the SNP, the iHS values were standardized (*S9*) to obtain a final statistic with mean 0 and variance 1, regardless of SNP allele frequency:

$$\text{iHS} = \frac{ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{SD_p\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]}$$

where the expectation and standard deviation are estimated from the empirical distribution at SNP whose allele frequency *p* matches the frequency at the core SNP. These calculations were performed on a breed-by-breed basis. Large negative iHS values indicate unusually long haplotypes carrying the derived allele and large positive iHS values indicate long haplotypes carrying the ancestral allele.

Extreme positive and negative values are potentially interesting, so we plotted |iHS| against chromosome position (Fig.S22). Across breeds, the top 1% of |iHS| values exceeded 2.694. There was evidence of a sweep in progress in Limousin and Piedmontese near *MSTN* on BTA2. On BTA6 and BTA14, in regions that coincide with known QTL, there was evidence of selective sweeps in a number of breeds (Fig. S22). Table S8 contains regions demonstrating evidence of positive selection.

Population comparisons with F_{ST}: Indexes of dissimilarity such as F_{ST} between pairs of populations can be used to estimate the genetic distance between the populations. Arlequin (*S14*) was used to calculate Slatkin's genetic distance (*S15*) between the pairs of cattle breeds and fitch from PHYLIP (*S8*) was used to generate an unrooted tree.

The tree in Fig. S23 clearly separates the indicine breeds (Nelore, Gir, and Brahman) from the taurine breeds. Also the hybrid breeds (Santa Gertrudis, Beefmaster, and Sheko) are placed intermediate to the indicine and taurine breeds and the African taurine N'Dama breed is positioned distant from the taurine and indicine breeds consistent with the hypothesis of a separate African site of domestication (*S16*). Of the European breeds the British Island breeds (Hereford, Guernsey, and Jersey), and European mountain Brown Swiss are the most distinct probably reflecting their phylogeographical origin.

Multidimensional scaling plots were also generated from the F_{ST} distances and are presented in Fig. S24. In Fig. S24A in which all breeds are included in the analysis, the first dimension clearly represents an indicine *versus* taurine ancestry effect, reflecting the primary division in cattle genomes. The two pure indicine breeds are to one extreme and both the European and African taurine cattle are to the other extreme for the first dimension. Again, the hybrid Sheko, Santa Gertrudis, and Beefmaster are intermediate on this axis. It is possible that there might be some indicine ancestry in the Italian Romagnola breed - previous microsatellite work has suggested that there may be traces of indicine in Mediterranean breeds (*S17*) - and the Romagnola is the most easterly of the European breeds represented in the sample. In Fig. S24B, in which the indicine and hybrid breeds were excluded from the analysis, there is a clear separation of the West African N'Dama from the other breeds, supporting the tentative archaeological and mtDNA evidence that suggests that there may have been a separate African site of cattle domestication. Neither of the Italian Romagnola or Piedmontese breeds, which have previously been shown to have some mtDNA from Africa, show affinity with the West African N'Dama breed. When the indicine and indicine-influenced breeds and also the N'Dama were removed from the analysis, strong selection for beef *versus* dairy function separates the European breeds on the second dimension (Fig. S24C). However, there appears to be no apparent geographical separation of the southern from the northern European breeds.

9. Relatedness, parentage, and traceability

We directly compared genomic similarity estimated from SNP data with that predicted from pedigree relationship coefficients. We tracked 61 generations to the original founders of the Angus breed to calculate the additive coefficients of relationship between all pairs of genotyped individuals resulting in a highly significant (r^2 = 0.87) association between D values and additive coefficients of relationship (Fig. S25). Unrelated individuals had, on average, 78% shared alleles, although this did not affect the utility of the regression for predicting relationships in the absence of pedigree information. This is applicable to genotyping individuals to estimate relationships among the animals and using the resultant relationship matrix and performance data to predict breeding values (*S18*). This approach could also be used to manage endangered bovine populations and to minimize inbreeding in captive breeding programs.

We evaluated the number of SNP required for determining parentage and animal identity (Table S9). Optimal SNP sets were selected to minimize the total squared deviation from 0.5 for each SNP's MAF across all breeds, subject to the constraint that the selected SNP could physically be no closer than 100 kb. Exclusion or match probabilities were calculated assuming random mating and no co-ancestry (*S19*). The mean MAF calculated form 50 SNP across all breeds was 0.39±0.02 and 0.34±0.03 with 1,600 SNP. This shows that 50 well chosen SNP can generally identify parentage within breeds and establish individual identity (Fig. S26). The equivalent power for sire identification, when the dam is unknown, requires 400 SNP. These results ignore genotyping errors, LD among closely linked markers and co-ancestry. However, it is clear that sufficient high utility markers exist within the set characterized here to design panels for parentage and identity in the majority of popular breeds.

SOM Text{ TC "**SOM Text**" \f A \l "1" **}**

Additional acknowledgments. The Bovine Hapmap Consortium gratefully acknowledges the logistical support provided by the Alliance for Animal Genome Research, Direction National de l'Elevage, Guinea and Ethiopian Institute of Agricultural Research, Ethiopia. National Research Initiative Grants from the USDA CSREES supported in part JG, CAG, VLM, and JC (2007-35604-17870) and JFT and RDS (2005-35205-15448, 2005-35604-15615, 2006-35205-16701, 2006-35616-16697 and 2008-35205-04687). Science Foundation Ireland 02/IN.1/B256 supported in part LPLL and DJL. Jonathon Pritchard and William Wen kindly provided the iHS computing tool.

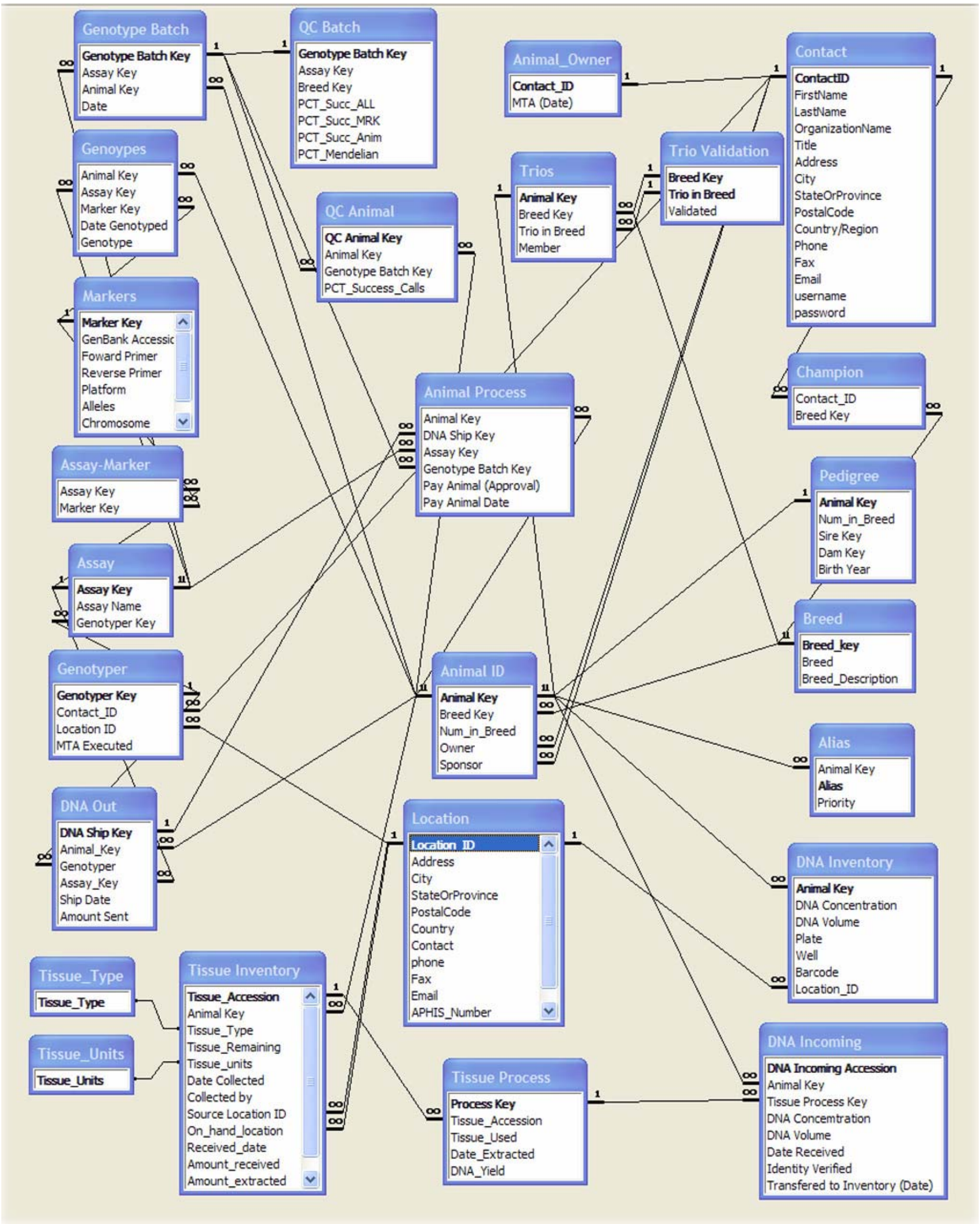
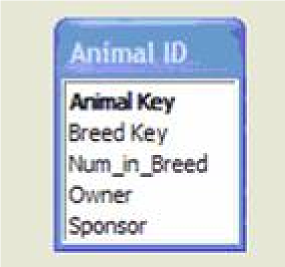


Fig. S1. Database Schema.



Animal ID

Animal Key

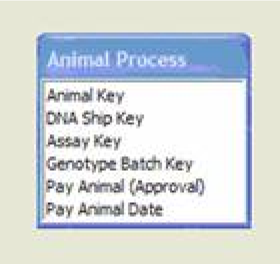
Breed Key

Num_in_Breed

Owner

Sponsor

Fig. S2 Screen shot of the Animal ID table.



Animal Process

Animal Key
DNA Ship Key
Assay Key
Genotype Batch Key
Pay Animal (Approval)
Pay Animal Date

Fig. S3. Screen shot of the Animal Process table.

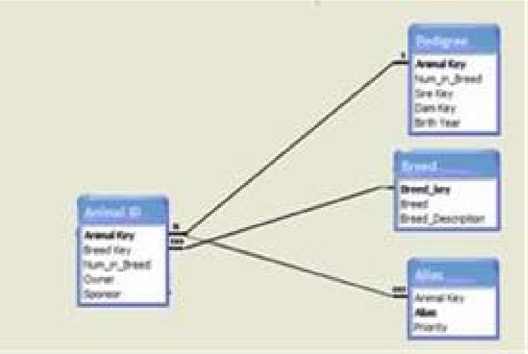


Fig. S4. Screen shot of the Animal tables.

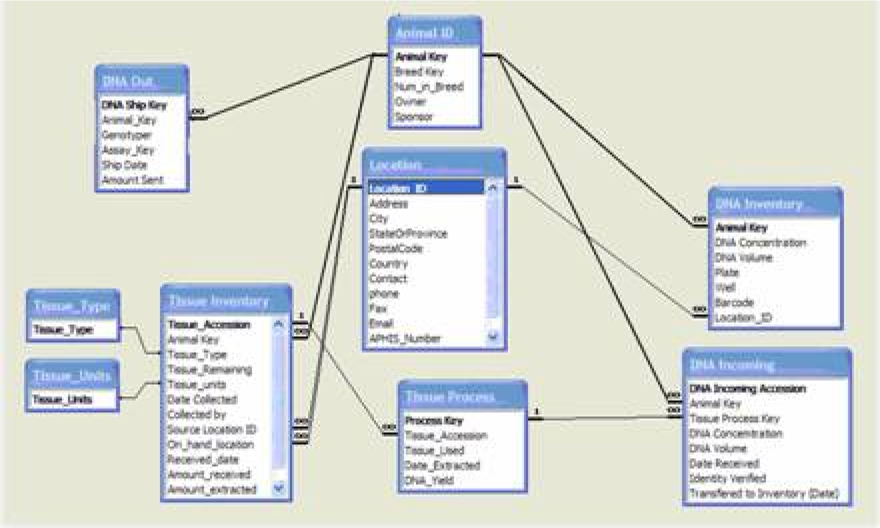
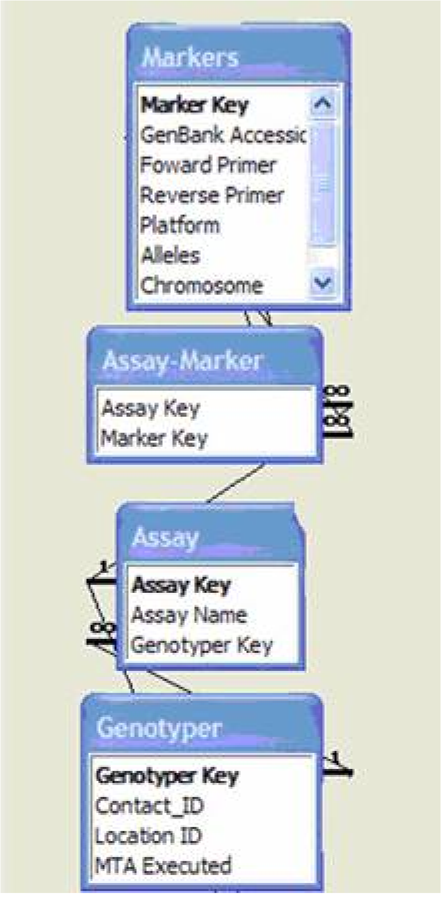


Fig. S5. Screen shot of the Inventory tables.



Markers

Marker Key
GenBank Accessic
Foward Primer
Reverse Primer
Platform
Alleles
Chromosome

Assay-Marker

Assay Key
Marker Key

Assay

Assay Key
Assay Name
Genotyper Key

Genotyper

Genotyper Key
Contact_ID
Location ID
MTA Executed

Fig. S6. Screen shot of the Assay tables.

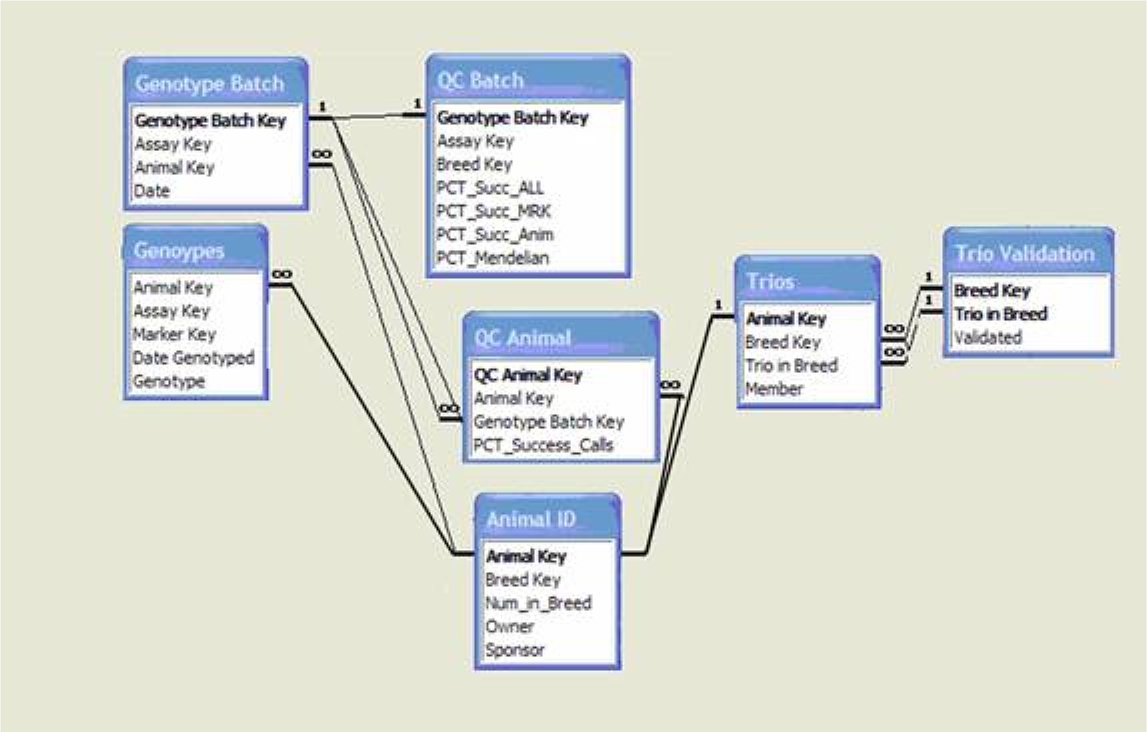


Fig. S7. Screen shot of the QA/QC Tables.

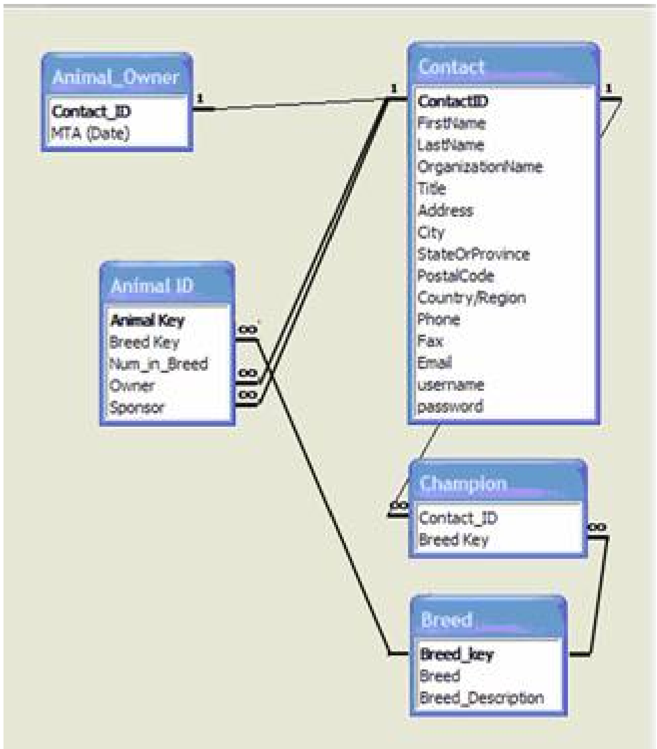


Fig. S8. Screen shot of the User tables.

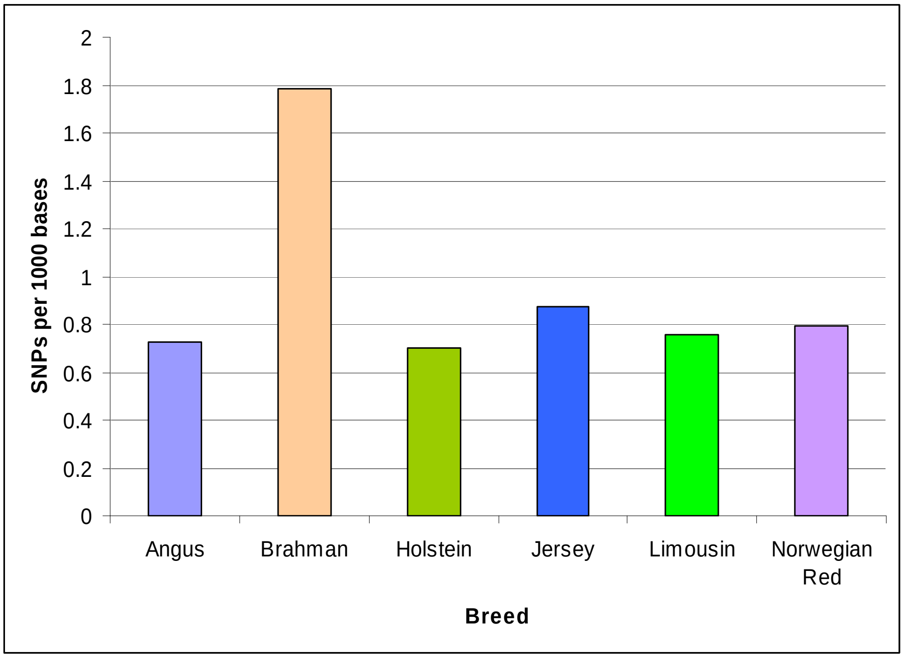


Fig. S9. Frequency of SNP identified by comparing random shotgun reads to the Hereford assembly.

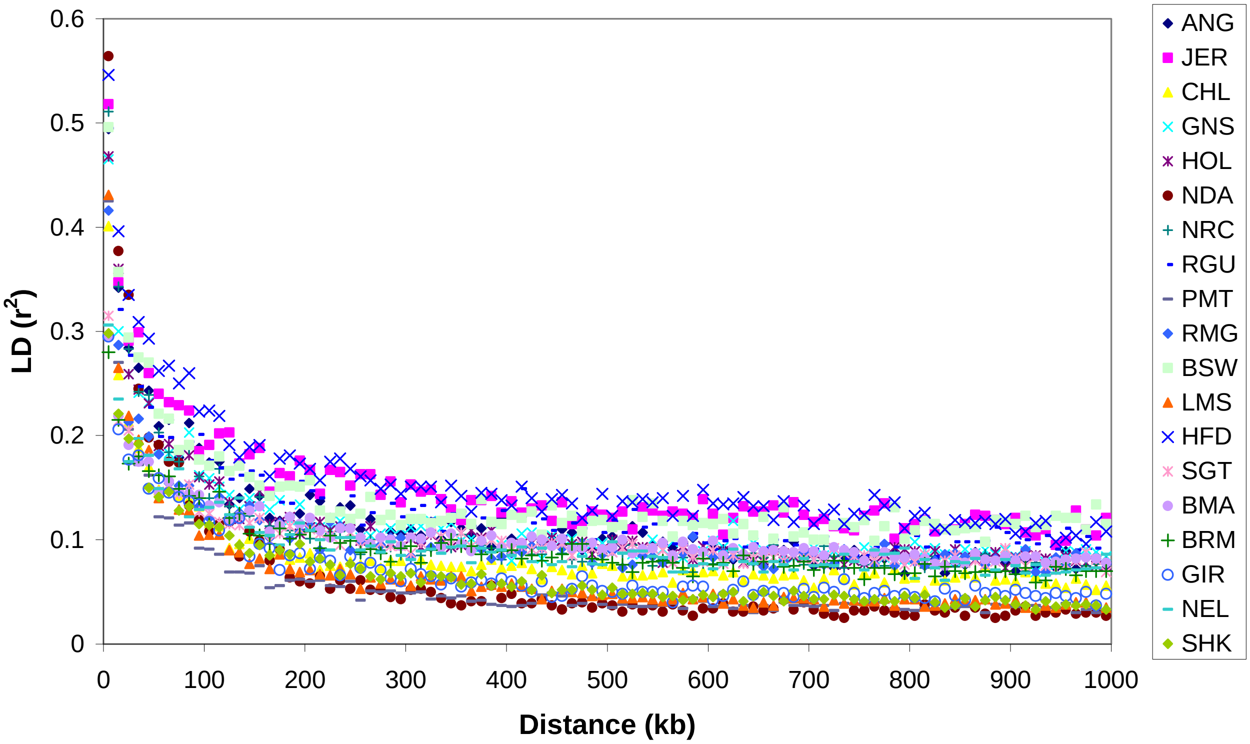


Fig. S10. Change in linkage disequilibrium (r^2) between marker pairs with increasing distance for all breeds. Values are the genome-wide average r^2 values within 10-kb bins.
ANG = Angus, JER = Jersey, CHL = Charolais, GNS = Guernsey, HOL = Holstein, NDA = N'Dama, NRC = Norwegian Red, RGU = Red Angus, PMT = Piedmontese, RMG = Romagnola, BSW = Brown Swiss, LMS = Limousin, HFD = Hereford, SGT = Santa Gertrudis, BMA = Beefmaster, BRM = Brahman, GIR = Gir, NEL = Nelore, SHK = Sheko.

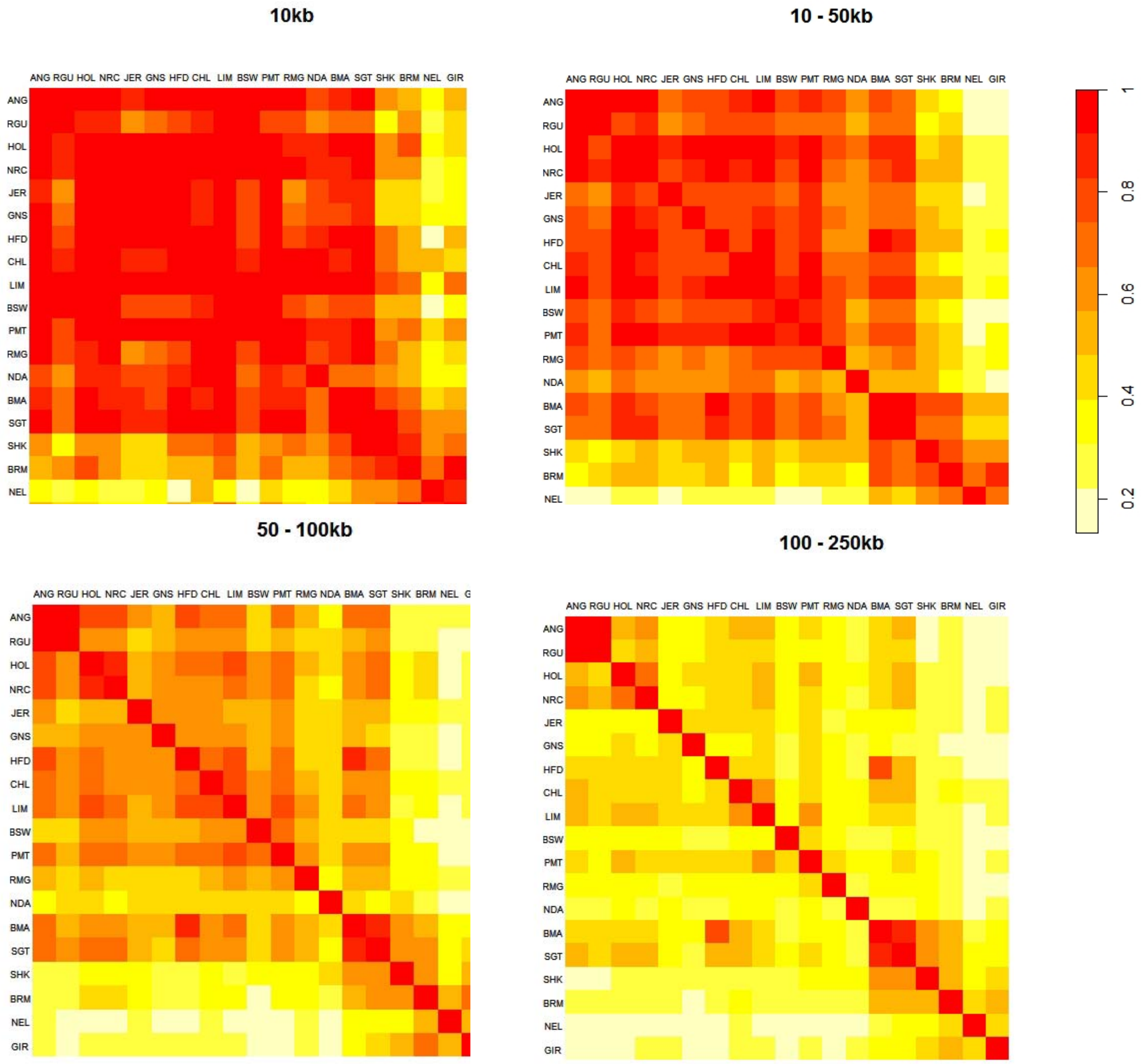
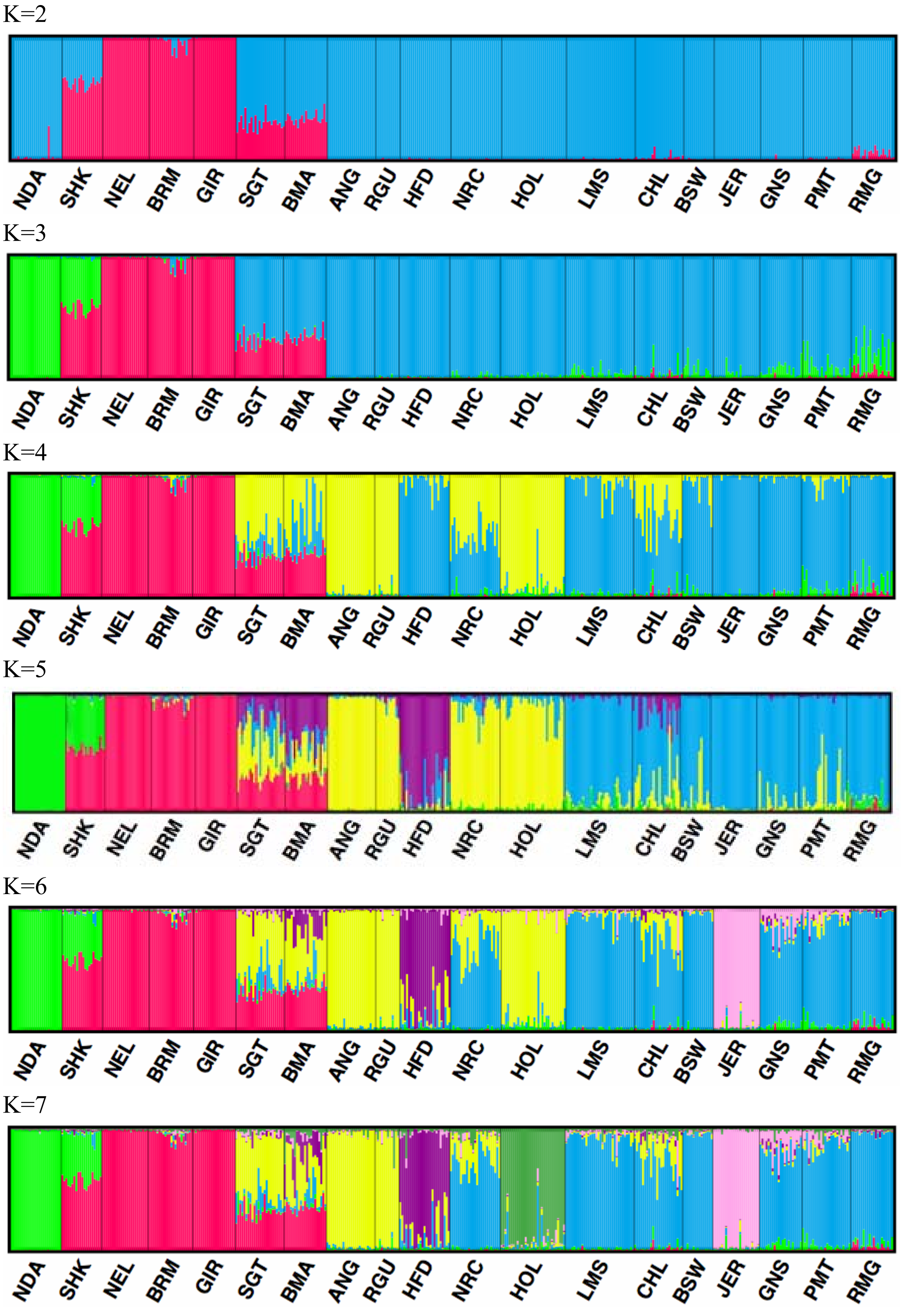


Fig. S11. Correlation between r values between breeds for marker pairs. ANG = Angus, RGU = Red Angus, HOL = Holstein, NRC = Norwegian Red, JER = Jersey, GNS = Guernsey, HFD = Hereford, CHL = Charolais, LMS = Limousin, BSW = Brown Swiss, PMT = Piedmontese, RMG = Romagnola, NDA = N'Dama, BMA = Beefmaster, SGT = Santa Gertrudis, SHK = Sheko, BRM = Brahman, NEL = Nelore, GIR = Gir.


K=2
NDA SHK NEL BRM GIR SGT BMA ANG RGU HFD NRC HOL LMS CHL BSW JER GNS PMT RMG
K=3
NDA SHK NEL BRM GIR SGT BMA ANG RGU HFD NRC HOL LMS CHL BSW JER GNS PMT RMG
K=4
NDA SHK NEL BRM GIR SGT BMA ANG RGU HFD NRC HOL LMS CHL BSW JER GNS PMT RMG
K=5
NDA SHK NEL BRM GIR SGT BMA ANG RGU HFD NRC HOL LMS CHL BSW JER GNS PMT RMG
K=6
NDA SHK NEL BRM GIR SGT BMA ANG RGU HFD NRC HOL LMS CHL BSW JER GNS PMT RMG
K=7
NDA SHK NEL BRM GIR SGT BMA ANG RGU HFD NRC HOL LMS CHL BSW JER GNS PMT RMG

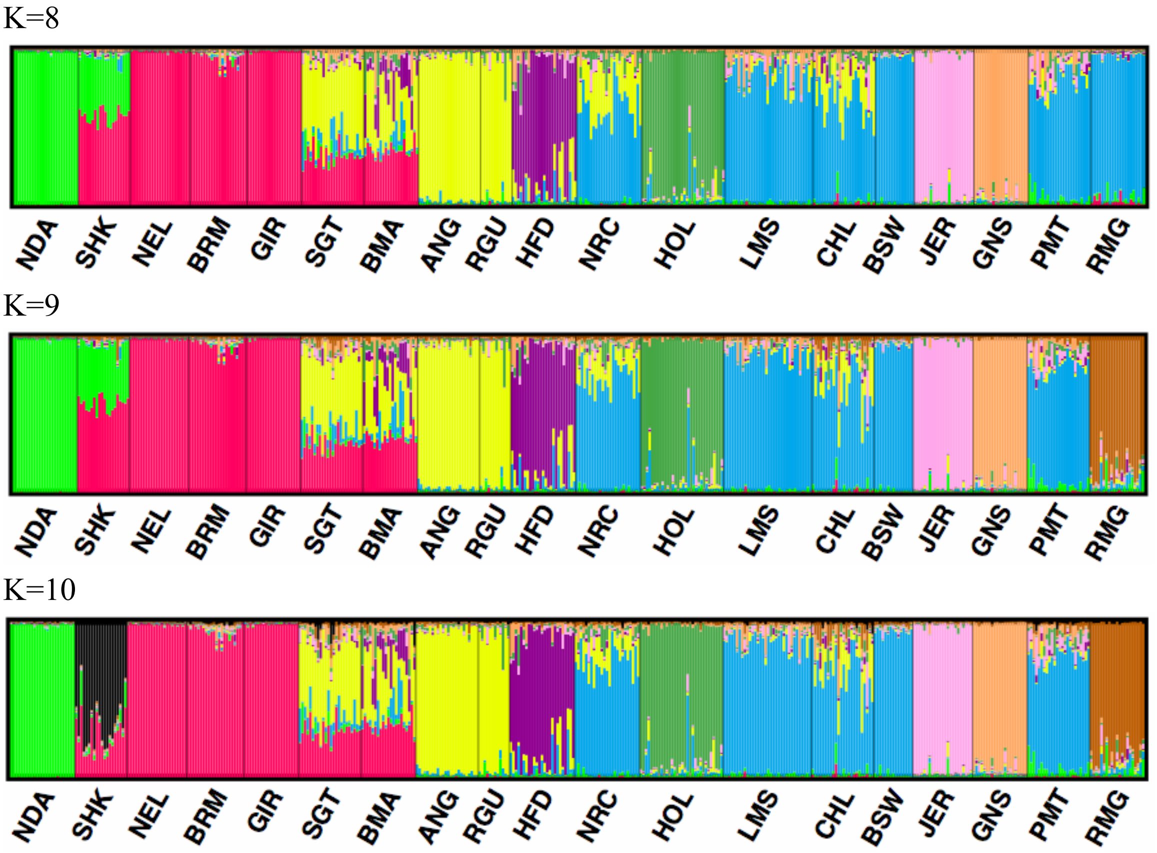


Fig. S12. Population classification across 19 breeds based on InSTRUCT assuming values of K (the number of ancestral populations) from 2 to 10. Each individual is represented by a vertical bar, often partitioned into colored segments representing the proportion of an individual's genome derived from each progenitor population. Breeds are separated by black lines, with breed names indicated below each panel. 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.

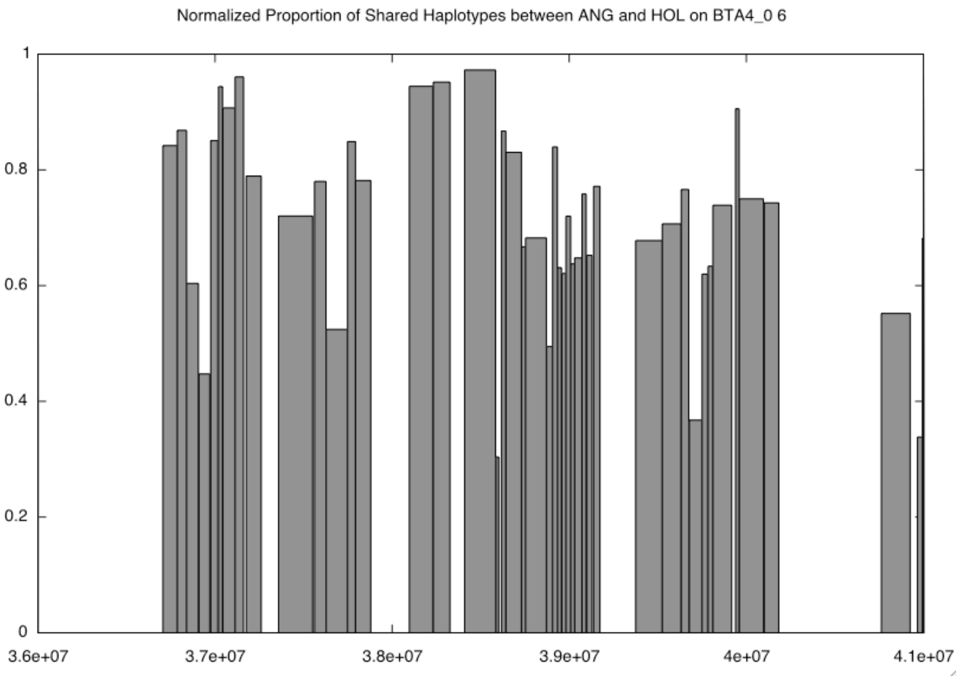


Fig. S13. Normalized proportion of shared multi-marker haplotypes between Angus and Holstein within a region of chromosome 6.

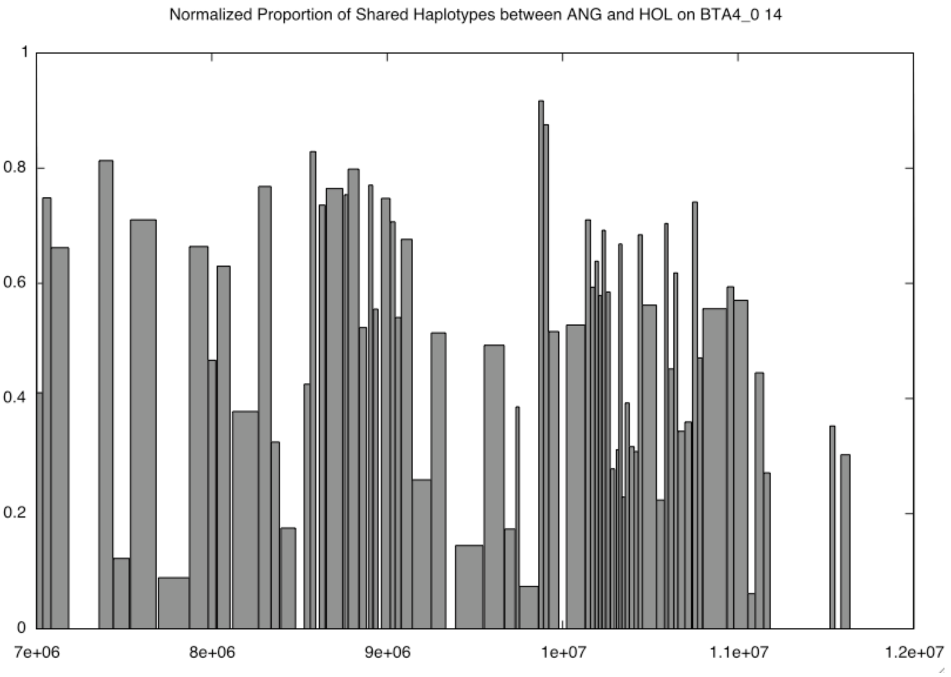


Fig. S14. Normalized proportion of shared multi-marker haplotypes between Angus and Holstein within a region of chromosome 14.

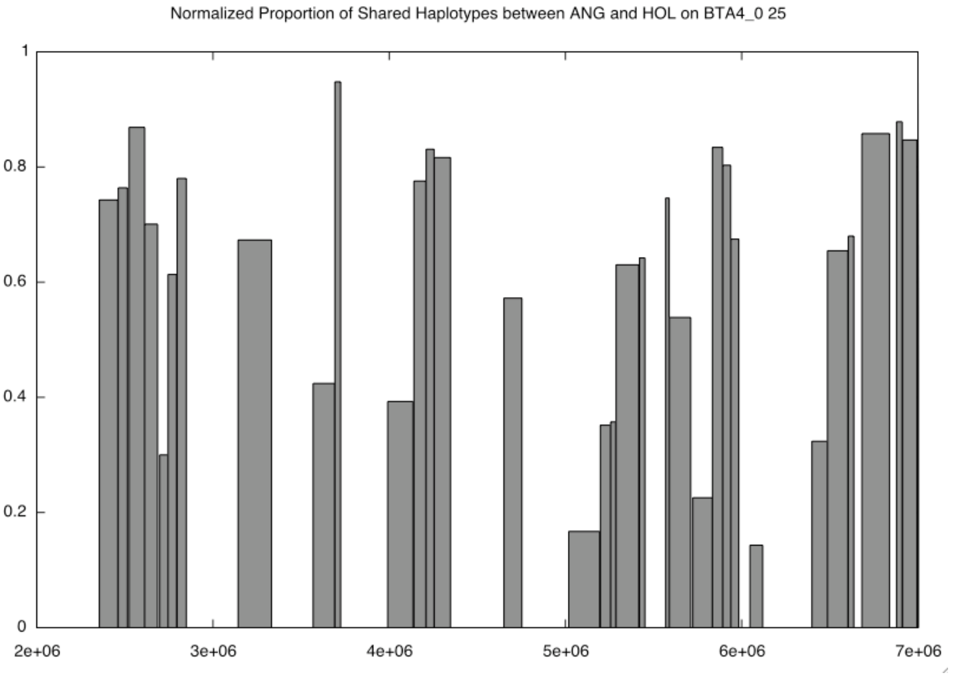


Fig. S15. Normalized proportion of shared multi-marker haplotypes between Angus and Holstein within a region of chromosome 25.

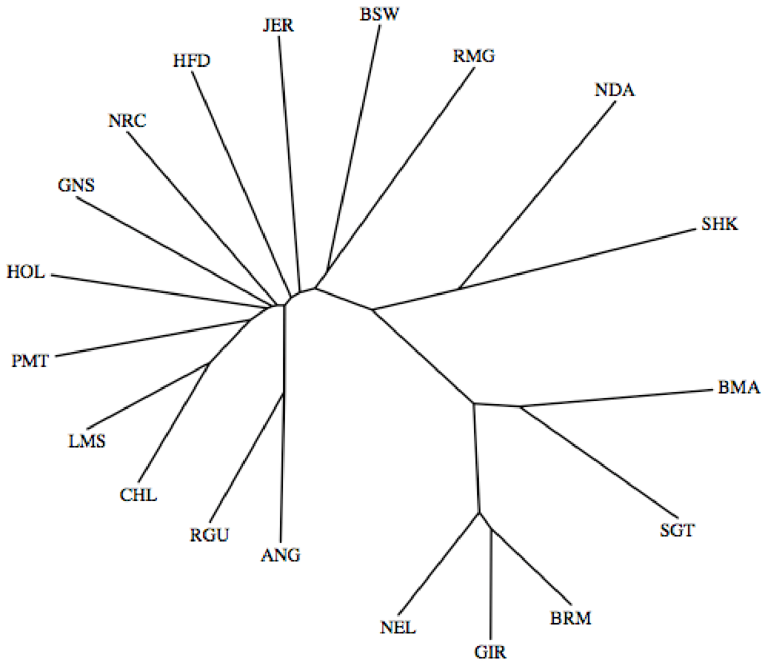


Fig. S16. Dendrogram of cattle breeds from the proportion of shared haplotypes on chromosome 25.

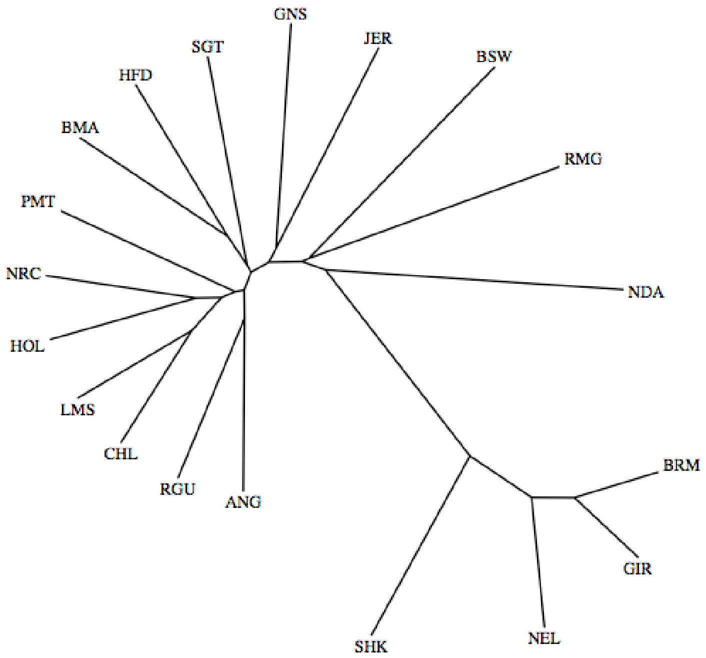


Fig. S17. Dendrogram of cattle breeds from the proportion of shared haplotypes on chromosome 6.

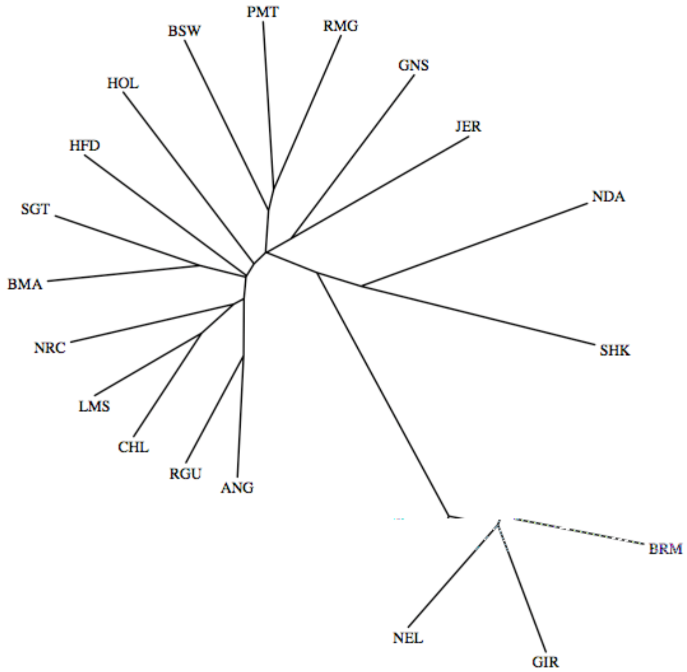


Fig. S18. Dendrogram of cattle breeds from the proportion of shared haplotypes on chromosome 14.

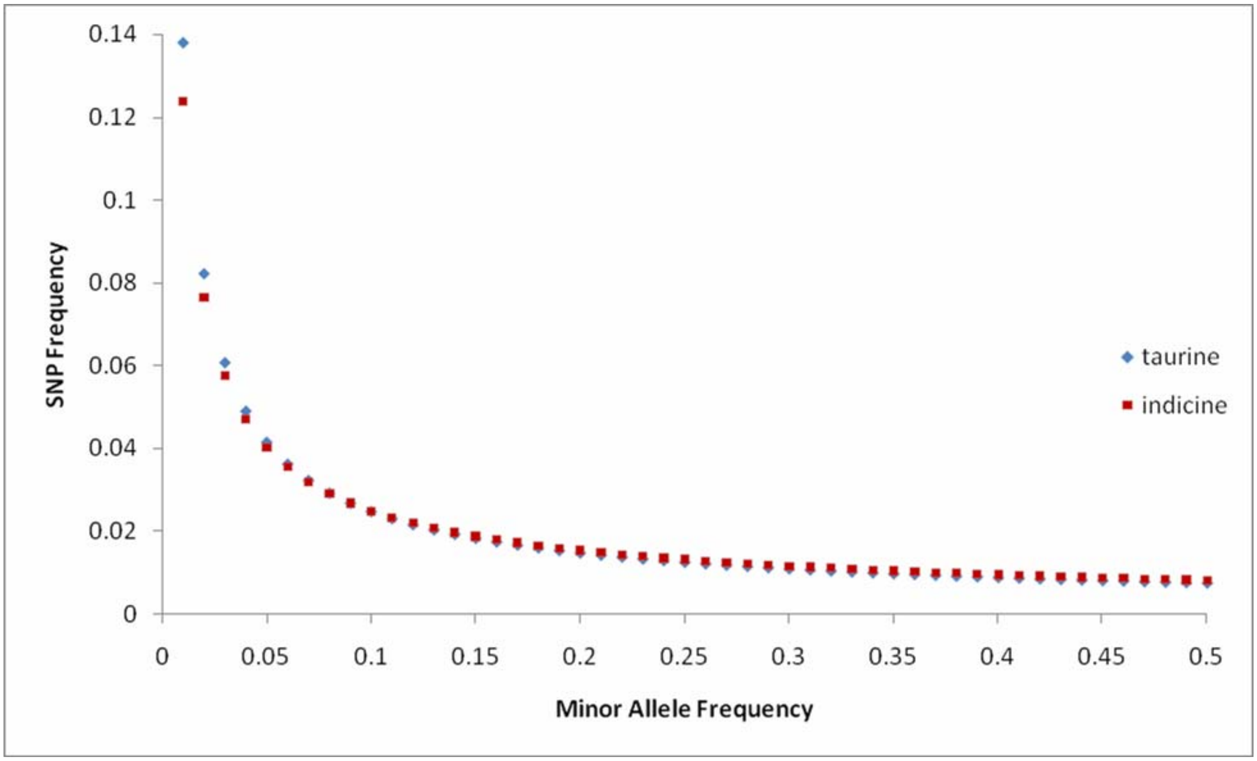


Fig. S19. Estimated distributions of unascertained SNP minor allele frequency distributions in taurine (Angus and Holstein) and indicine (Brahman) cattle produced from deep sequencing 199 regions on chromosomes 6 and 14.

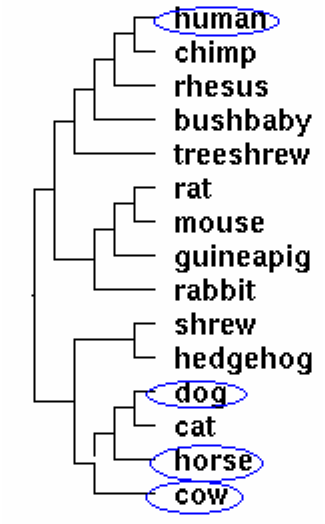


Fig. S20. Evolutionary relationship between species for which genome sequence assemblies were used to establish the ancestral allele at each bovine SNP by comparative analysis.

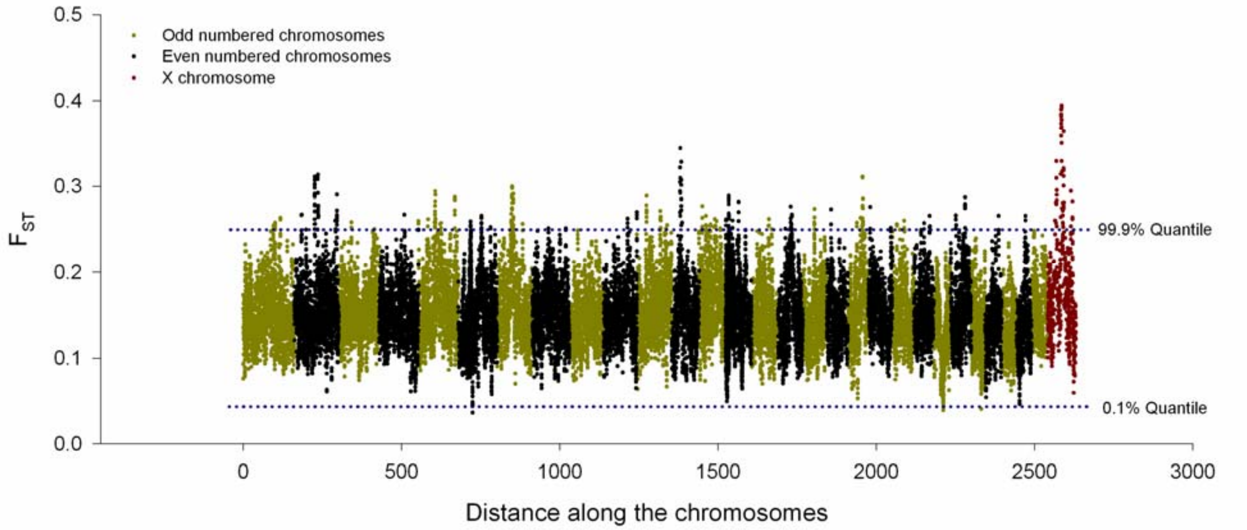


Fig. S21. Genome-wide scan for positive selection. The distribution of F_{ST} averaged across a sliding 8 SNP window is shown for all breeds. Dashed lines represent the 0.1% and 99.9% quantiles for the genome-wide F_{ST} values.

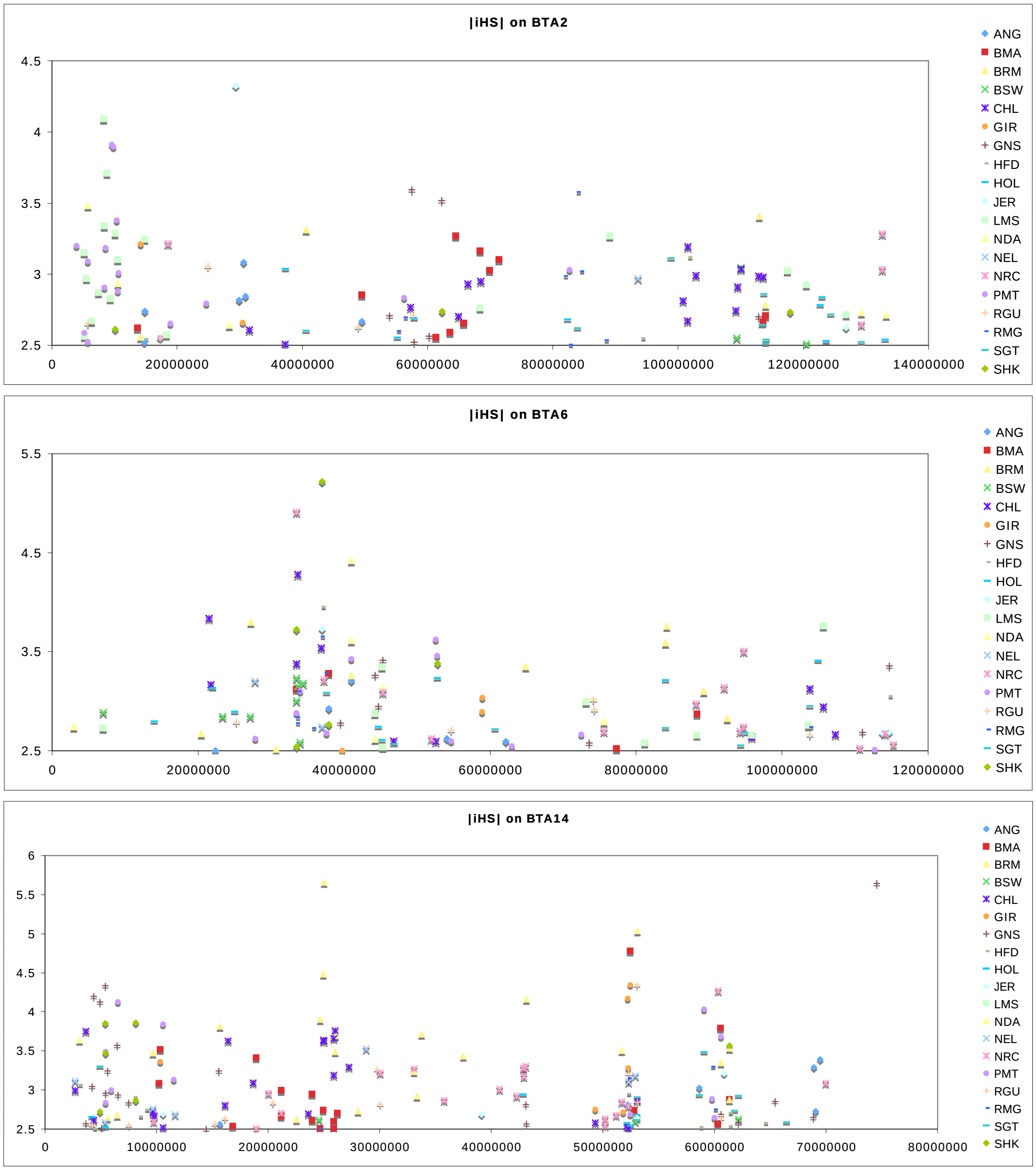


Fig. S22. Plots of SNP from chromosomes 2, 6, and 14 with extreme iHS values. SNP with |iHS| >2.5 are plotted.

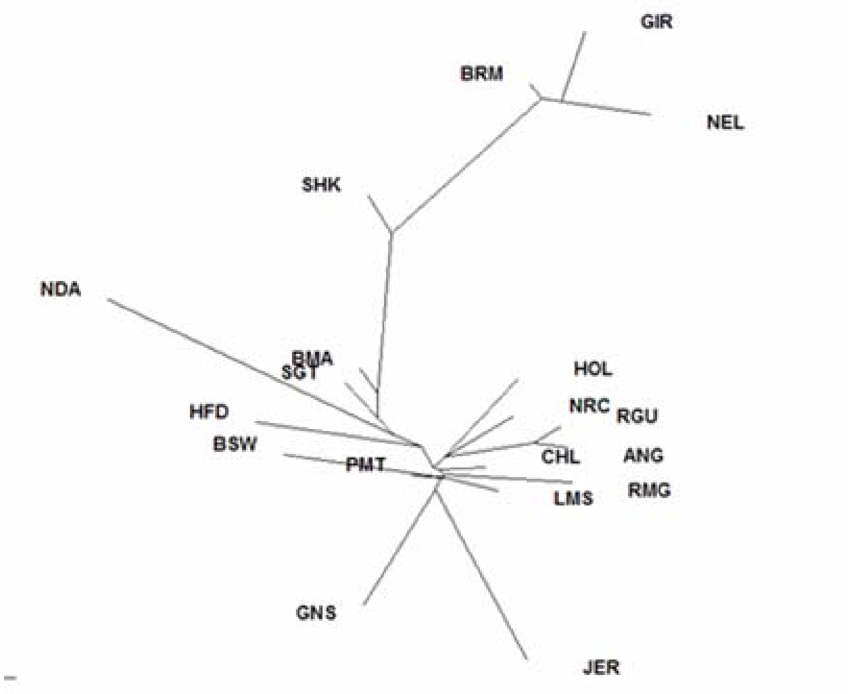


Fig. S23. Unrooted tree from F_{ST} distances computed between pairs of breeds. ANG = Angus, BMA = Beefmaster, BRM = Brahman, BSW = Brown Swiss, CHL = Charolais, GIR = Gir, GNS = Guernsey, HFD = Hereford, HOL = Holstein, JER = Jersey, LMS = Limousin, NDA = N'Dama, NEL = Nelore, NRC = Norwegian Red, PMT = Piedmontese, RGU = Red Angus, RMG = Romagnola, SGT = Santa Gertrudis, SHK = Sheko.

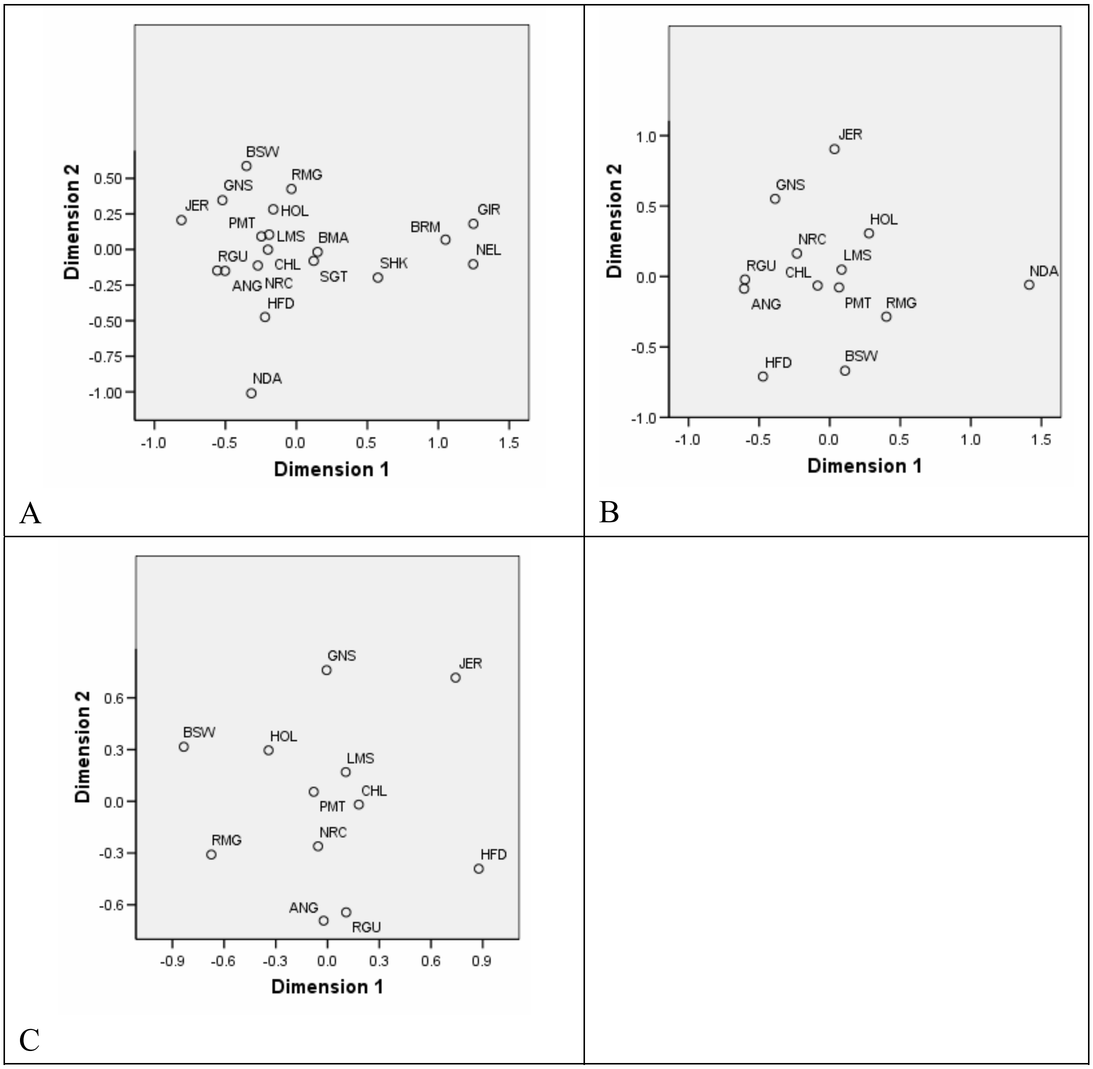


Fig. S24. Multi-dimensional scaling plots with the genetic distances from pairwise F_{ST} estimates between the breeds; (A) All 19 breeds included in the analysis; (B) Analysis performed after excluding indicine and indicine influenced breeds; and (C) Analysis performed after excluding indicine, indicine influenced and African breeds. ANG = Angus, BMA = Beefmaster, BRM = Brahman, BSW = Brown Swiss, CHL = Charolais, GIR = Gir, GNS = Guernsey, HFD = Hereford, HOL = Holstein, JER = Jersey, LMS = Limousin, NDA = N'Dama, NEL = Nelore, NRC = Norwegian Red, PMT = Piedmontese, RGU = Red Angus, RMG = Romagnola, SGT = Santa Gertrudis, SHK = Sheko.

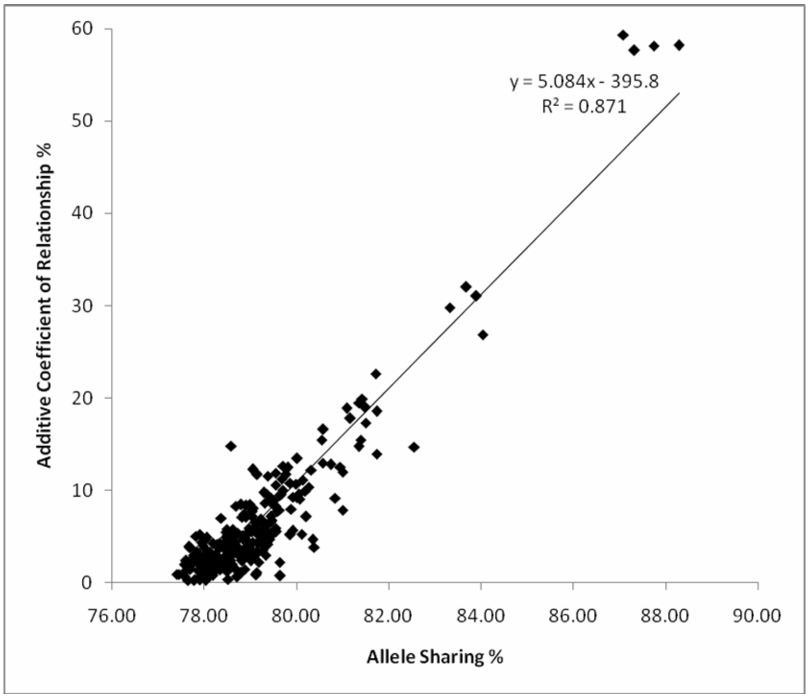


Fig. S25. Relationship between average allele sharing at 31,303 loci and additive coefficient of relationship estimated from a 61 generation pedigree among all 253 pairs derived from 23 registered Angus individuals.

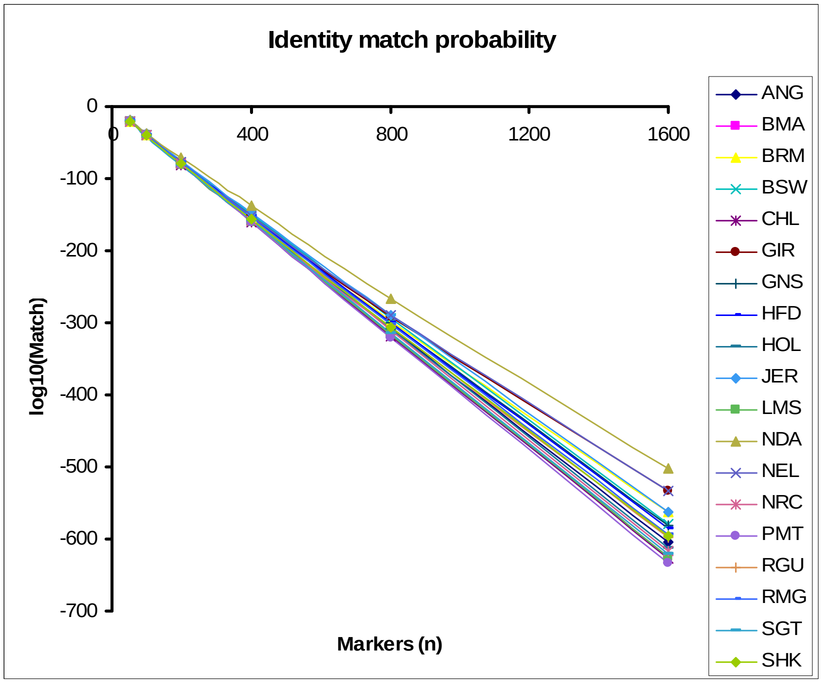


Fig. S26. Plot of $\log_{10}$(match probability) versus SNP marker number by breed for the establishment of individual identity.

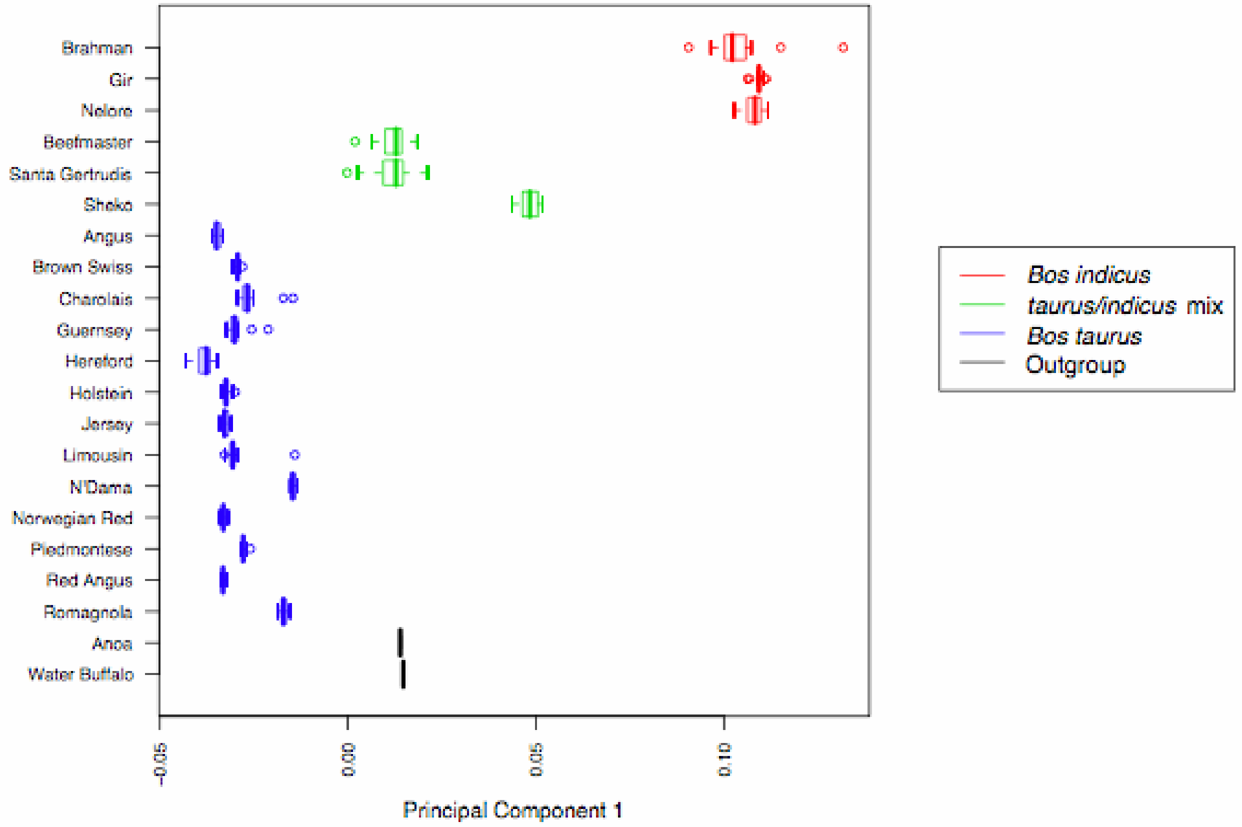


Fig. S27. Principal component 1 based on all ascertained markers. Taurine breeds (blue) remain separated from indicine breeds (red), and the admixed breeds (green) are intermediate in the analysis.

Table S1 Summary of sampled bovine populations

Breed and abbreviation	Number genotyped	Country of sampling	Land of origin	Current primary geographical distribution[1]	Estimated global populatio n size	Current effective population size	Primary purpose	Breed Champion	Particular characteristics
Taurine									
Angus ANG	27	USA and NZ	Scotland	Global	>10 M[2]	136	Beef	J.C.M., R.D.G	Black coat Meat quality
Brown Swiss BSW	24	USA	Switzerland	Alpine Europe and Americas	7 M	61	Dairy	C.P.V.T., T.S.S.	Brown coat Rugged appearance
Charolais CHL	24	US	France	France, NA, Brazil and South Africa	>12 M	110	Beef	J.L.W.	White to cream coat Large body size
Guernsey GNS	21	USA and UK	Channel Islands	NA, UK, Oceania and South Africa	75,000	76	Dairy	J.L.W., C.P.V.T.	Tan and white coat Refined structure
Hereford HFD	27	USA and NZ	UK	Global	>5 M[3]	97	Beef	R.D.G, J.C.M.	Red coat with white face
Holstein HOL	53	USA and NZ	Netherlands	Global	>65 M	99	Dairy	J.C.M., C.P.V.T.	Black and white coat High milk yield
Jersey JER	28	USA and NZ	Channel Islands	Global	>2.5 M	73	Dairy	J.C.M., C.P.V.T.	Small size Milk quality
Limousin LMS	42	USA and France	France	France, UK and NA	>4 M[4]	174	Beef	R.D.G, S.S.	Red coat Muscularity
N'dama NDA	25	Guinea	West Africa	West Africa	7 M	228	Multi-purpose	O.H.	Fawn coat, small size Trypanosome resistance
Norwegian Red NRC	25	Norway	Norway	Norway	260,000	106	Dairy/ Dual purpose	S.L.	Red and white coat
Piedmontese PMT	24	Italy	Italy	Italy	400,000	167	Beef/ Dual purpose	P.M.	Gray coat, dark skin Muscularity
Red Angus RGU	12	USA and Canada	Scotland	NA, Australia	>300,000[5]	85	Beef	R.D.G.	Red coat Meat quality, docility
Romagnola RMG	24	Italy	Italy	Italy, USA, and Australia	30,000	92	Beef	P.A.	Ivory to gray coat, black skin Muscularity
Sheko SHK	20	Ethiopia	Ethiopia	East Africa	<5,000	145	Multi-purpose	P.J.B., O.H.,	Small size, Brown coat of variable shades

								J.F.G.	Trypanosome resistance
Indicine									
Brahman BRM	25	USA and Australia	USA	Australia, USA, Tropics	>4 M	115	Beef	W.B., C.A.G.	Gray coat, humped Heat and disease tolerance
Gir GIR	24	Brazil	India	Asia and South America	>3 M	133	Dairy, Multi-purpose	A.R.C.	Mottled red and white coat Humped, Heat and pest tolerant
Nelore NEL	24	Brazil	India	South America	>100 M	86	Beef	A.R.C.	White to gray coat, humped Heat and pest resistance
Hybrid									
Beefmaster BMA	24	USA	USA	Americas	>1 M[6]	106	Beef	L.C.S.	Variable, predominantly red coat Robustness and adaptability
Santa Gertrudis SGT	24	USA	USA	USA, Brazil, Australia	~5 M[7]	107	Beef	R.D.G	Red coat, Drought, heat, insect & disease tolerance Robustness and adaptability
Outgroup									
Anoa	2	Indonesia	Indonesia	Indonesia	<5000	Unknown	None	J.L.W.	Small, black to brown coat, straight and rear-pointing horns
Mediterranean Buffalo	2	Italy	Italy and Romania	India	~400,000	Unknown	Dairy, Dual purpose	P.M.	Brown to black coat, compact form

Table S2 Number of shotgun reads from six different breeds used for SNP discovery.

				SNP	SNP Assayed
Breed	**Reads**	**Mapped**	**q20 Bases**	**Discovered**	**(%)**
Angus	26,170	12,785	10,999,492	7,971	2,230 (6.0)
Brahman	10,995	5,685	4,947,910	8,817	6,554 (17.5)
Holstein	143,498	58,600	51,232,556	35,922	16,145 (43.1)
Jersey	10,400	6,431	5,478,967	4,795	591 (1.6)
Limousin	3,548	1,743	1,537,014	1,164	412 (1.0)
Norwegian Red	154,347	87,684	75,258,669	59,580	8,393 (22.4)
Unknown					3,145 (8.4)
Total	348,958	172,928	149,454,608	118,249	37,470 (100)

Table S3 Conditional probabilities of observing genotypes given the true underlying genotype.

		Observed Genotype			
		AA	AB	BB	Σ
True Genotype	AA	$(1-\varepsilon)^2$	$2(1-\varepsilon)\varepsilon$	ε^2	1
	AB	$(1-\varepsilon)\varepsilon$	$(1-\varepsilon)^2+\varepsilon^2$	$(1-\varepsilon)\varepsilon$	1
	BB	ε^2	$2(1-\varepsilon)\varepsilon$	$(1-\varepsilon)^2$	1

Table S4 Distributions of SNP minor allele frequencies within and across cattle breeds and for outgroup species. Overall distribution of minor allele frequencies across all animals included in the taurine, indicine, composite, and African breeds of cattle.

		Average MAF[1]							
	Num SNP	**Including Monomorphic**	**Excluding Monomorphic**	**Monomorphic**	**0< MAF ≤0.1**	**0.1< MAF ≤0.2**	**0.2< MAF ≤0.3**	**0.3< MAF ≤0.4**	**0.4≤ MAF ≤0.5**
Angus	33326	0.190	0.241	0.21	0.19	0.15	0.17	0.14	0.15
Brown Swiss	31072	0.175	0.239	0.27	0.17	0.15	0.15	0.14	0.13
Charolais	33326	0.206	0.239	0.14	0.21	0.17	0.15	0.16	0.17
Guernsey	33326	0.184	0.238	0.23	0.20	0.15	0.14	0.14	0.15
Hereford	33326	0.215	0.261	0.18	0.15	0.14	0.18	0.17	0.18
Holstein	33326	0.208	0.248	0.16	0.18	0.17	0.16	0.18	0.16
Jersey	33326	0.171	0.233	0.27	0.19	0.13	0.15	0.12	0.14
Limousin	33326	0.201	0.238	0.16	0.21	0.17	0.16	0.16	0.16
Norwegian Red	33326	0.204	0.253	0.20	0.16	0.15	0.17	0.15	0.17
Piedmontese	31072	0.202	0.245	0.18	0.16	0.18	0.18	0.16	0.14
Red Angus	31072	0.186	0.252	0.26	0.15	0.13	0.19	0.12	0.15
Romagnola	31072	0.186	0.236	0.21	0.18	0.17	0.15	0.15	0.13
Taurine	33326	0.194	0.244	0.03	0.25	0.18	0.18	0.18	0.18
Beefmaster	33326	0.225	0.247	0.09	0.17	0.20	0.19	0.18	0.16
Santa Gertrudis	33326	0.216	0.240	0.10	0.22	0.19	0.16	0.18	0.15
Composite	33326	0.220	0.243	0.04	0.21	0.21	0.19	0.18	0.17
N'Dama	33326	0.137	0.230	0.41	0.14	0.13	0.12	0.11	0.10
Sheko	31072	0.187	0.243	0.23	0.19	0.16	0.13	0.16	0.13
African	33326	0.161	0.237	0.20	0.21	0.18	0.16	0.14	0.12
Brahman	33326	0.153	0.195	0.22	0.30	0.17	0.11	0.12	0.09
Gir	33326	0.132	0.201	0.34	0.21	0.15	0.12	0.09	0.08
Nelore	33326	0.134	0.202	0.34	0.22	0.15	0.10	0.10	0.09
Indicine	33326	0.140	0.199	0.15	0.36	0.17	0.12	0.10	0.09
Anoa	20834	0.042	0.371	0.89			0.06		0.05
Water Buffalo	20834	0.035	0.386	0.91			0.04		0.05
Outgroups	20834	0.038	0.378	0.82		0.05	0.06	0.02	0.04
Overall	33326	0.185	0.237	0.00	0.22	0.23	0.19	0.18	0.18

[1] Average MAF values calculated

Table S5 Distribution of SNP by minor allele frequencies across all breeds when SNP were discovered by comparison of sequences produced from the indicated breed to the Hereford assembly.

(A) Breed of SNP Discovery: Angus

		Minor Allele Frequencies		Fraction of All SNP					
	Num SNP	Include Monomorphic	Exclude Monomorphic	Monomorphic	0 < MAF≤ 0.1	0.1 < MAF≤ 0.2	0.2 < MAF≤ 0.3	0.3 < MAF≤ 0.4	0.4 < MAF≤ 0.5
Angus	2148	0.238	0.258	0.08	0.17	0.17	0.21	0.18	0.20
Anoa	1945	0.036	0.371	0.90	0.00	0.00	0.05	0.00	0.05
Beefmaster	2148	0.234	0.256	0.08	0.16	0.19	0.19	0.18	0.19
Brahman	2148	0.134	0.179	0.25	0.33	0.16	0.10	0.09	0.08
Brown Swiss	1949	0.195	0.241	0.19	0.19	0.17	0.15	0.15	0.15
Buffalo	1945	0.035	0.378	0.91	0.00	0.00	0.04	0.00	0.05
Charolais	2148	0.229	0.252	0.09	0.18	0.18	0.17	0.18	0.19
Gir	2148	0.115	0.187	0.38	0.23	0.15	0.09	0.09	0.06
Guernsey	2148	0.200	0.246	0.19	0.19	0.15	0.15	0.16	0.16
Hereford	2148	0.244	0.265	0.08	0.16	0.16	0.20	0.19	0.21
Holstein	2148	0.219	0.241	0.09	0.20	0.18	0.18	0.18	0.16
Jersey	2148	0.192	0.233	0.18	0.21	0.15	0.17	0.14	0.16
Limousin	2148	0.224	0.241	0.07	0.21	0.19	0.18	0.16	0.18
Ndama	2148	0.153	0.228	0.33	0.17	0.14	0.14	0.12	0.10
Nelore	2148	0.120	0.193	0.38	0.22	0.15	0.08	0.09	0.08
Norwegian Red	2148	0.220	0.252	0.12	0.18	0.15	0.20	0.16	0.18
Peidmontese	1949	0.229	0.251	0.09	0.18	0.18	0.19	0.19	0.17
Red Angus	1949	0.227	0.263	0.13	0.16	0.14	0.21	0.15	0.19
Romangola	1949	0.209	0.244	0.14	0.18	0.16	0.19	0.18	0.14
Santa Gertrudis	2148	0.226	0.248	0.09	0.21	0.18	0.16	0.19	0.17
Sheko	1949	0.178	0.236	0.25	0.19	0.16	0.13	0.15	0.12

(B) Breed of SNP Discovery: Brahman

		Minor Allele Frequencies		Fraction of All SNP					
	Num SNP	Include	Exclude						
		Monomorphic	Monomorphic	Monomorphic	0 < MAF ≤ 0.1	0.1 < MAF ≤ 0.2	0.2 < MAF≤ 0.3	0.3 < MAF≤ 0.4	0.4 < MAF≤ 0.5
Angus	6406	0.095	0.214	0.55	0.14	0.08	0.09	0.06	0.07
Anoa	177	0.028	0.417	0.93	0.00	0.00	0.02	0.00	0.05
Beefmaster	6406	0.203	0.222	0.08	0.21	0.24	0.19	0.16	0.12
Brahman	6406	0.236	0.252	0.06	0.19	0.21	0.18	0.20	0.17
Brown Swiss	6378	0.091	0.213	0.57	0.13	0.09	0.08	0.07	0.06
Buffalo	177	0.024	0.386	0.94	0.00	0.00	0.03	0.00	0.03
Charolais	6406	0.118	0.174	0.32	0.31	0.12	0.08	0.09	0.08
Gir	6406	0.198	0.235	0.16	0.19	0.18	0.18	0.15	0.14
Guernsey	6406	0.103	0.192	0.46	0.23	0.09	0.08	0.07	0.08
Hereford	6406	0.108	0.233	0.54	0.13	0.08	0.09	0.08	0.09
Holstein	6406	0.101	0.213	0.53	0.15	0.09	0.08	0.08	0.07
Jersey	6406	0.089	0.224	0.60	0.12	0.07	0.07	0.06	0.07
Limousin	6406	0.107	0.205	0.48	0.19	0.10	0.07	0.09	0.07
Ndama	6406	0.082	0.212	0.61	0.12	0.08	0.07	0.06	0.06
Nelore	6406	0.200	0.238	0.16	0.19	0.18	0.17	0.16	0.15
Norwegian Red	6406	0.100	0.234	0.57	0.12	0.08	0.07	0.07	0.08
Peidmontese	6378	0.109	0.212	0.49	0.16	0.10	0.09	0.08	0.08
Red Angus	6378	0.095	0.236	0.60	0.10	0.07	0.10	0.06	0.07
Romangola	6378	0.120	0.203	0.41	0.19	0.13	0.10	0.09	0.08
Santa Gertrudis	6406	0.199	0.221	0.10	0.24	0.21	0.18	0.14	0.13
Sheko	6378	0.226	0.258	0.13	0.19	0.17	0.15	0.20	0.16

(C) Breed of SNP Discovery: Holstein

		Minor Allele Frequencies			Fraction of All SNP				
		Include	Exclude						
	Num SNP	Monomorphic	Monomorphic	Monomorphic	0 < MAF ≤ 0.1	0.1 < MAF ≤ 0.2	0.2 < MAF≤ 0.3	0.3 < MAF≤ 0.4	0.4 < MAF≤ 0.5
Angus	15687	0.210	0.241	0.13	0.20	0.16	0.19	0.16	0.17
Anoa	9685	0.039	0.372	0.89	0.00	0.00	0.05	0.00	0.05
Beefmaster	15687	0.230	0.254	0.10	0.16	0.19	0.19	0.19	0.18
Brahman	15687	0.136	0.182	0.25	0.31	0.16	0.10	0.10	0.08
Brown Swiss	13708	0.198	0.247	0.20	0.17	0.15	0.17	0.16	0.15
Buffalo	9685	0.032	0.383	0.92	0.00	0.00	0.04	0.00	0.04
Charolais	15687	0.228	0.251	0.09	0.19	0.18	0.17	0.18	0.19
Gir	15687	0.118	0.192	0.39	0.21	0.15	0.11	0.08	0.07
Guernsey	15687	0.204	0.244	0.17	0.20	0.16	0.15	0.15	0.17
Hereford	15687	0.240	0.266	0.10	0.16	0.15	0.20	0.19	0.20
Holstein	15687	0.246	0.260	0.05	0.16	0.19	0.18	0.21	0.20
Jersey	15687	0.194	0.237	0.18	0.21	0.14	0.17	0.14	0.16
Limousin	15687	0.224	0.244	0.08	0.21	0.18	0.17	0.18	0.18
Ndama	15687	0.149	0.234	0.36	0.15	0.14	0.12	0.12	0.11
Nelore	15687	0.121	0.191	0.37	0.23	0.14	0.09	0.09	0.08
Norwegian Red	15687	0.221	0.252	0.12	0.18	0.17	0.18	0.16	0.18
Peidmontese	13708	0.226	0.250	0.10	0.16	0.19	0.21	0.18	0.16
Red Angus	13708	0.207	0.254	0.18	0.16	0.15	0.21	0.14	0.17
Romangola	13708	0.201	0.241	0.17	0.18	0.18	0.16	0.17	0.15
Santa Gertrudis	15687	0.221	0.246	0.10	0.21	0.18	0.16	0.18	0.16
Sheko	13708	0.178	0.240	0.26	0.19	0.15	0.13	0.16	0.12

(D) Breed of SNP Discovery: Jersey

		Minor Allele Frequencies		Fraction of All SNP					
	Num SNP	Include Monomorphic	Exclude Monomorphic	Monomorphic	0 < MAF ≤ 0.1	0.1 < MAF ≤ 0.2	0.2 < MAF≤ 0.3	0.3 < MAF≤ 0.4	0.4 < MAF≤ 0.5
Angus	576	0.199	0.238	0.16	0.20	0.15	0.19	0.14	0.15
Anoa	576	0.040	0.363	0.89	0.00	0.00	0.06	0.00	0.05
Beefmaster	576	0.221	0.241	0.08	0.19	0.21	0.17	0.21	0.14
Brahman	576	0.133	0.172	0.23	0.35	0.18	0.07	0.08	0.09
Brown Swiss	576	0.188	0.234	0.20	0.20	0.16	0.15	0.15	0.14
Buffalo	576	0.032	0.394	0.92	0.00	0.00	0.03	0.00	0.05
Charolais	576	0.225	0.253	0.11	0.17	0.19	0.18	0.14	0.20
Gir	576	0.115	0.184	0.37	0.22	0.15	0.12	0.07	0.06
Guernsey	576	0.200	0.245	0.19	0.18	0.16	0.15	0.14	0.18
Hereford	576	0.241	0.268	0.10	0.15	0.16	0.19	0.19	0.22
Holstein	576	0.203	0.229	0.12	0.25	0.17	0.14	0.18	0.15
Jersey	576	0.221	0.255	0.13	0.17	0.17	0.18	0.13	0.21
Limousin	576	0.215	0.236	0.09	0.22	0.17	0.18	0.18	0.16
Ndama	576	0.151	0.230	0.35	0.15	0.14	0.15	0.11	0.10
Nelore	576	0.120	0.200	0.40	0.20	0.15	0.08	0.10	0.07
Norwegian Red	576	0.207	0.249	0.17	0.18	0.15	0.16	0.18	0.16
Peidmontese	576	0.226	0.252	0.10	0.16	0.20	0.16	0.19	0.18
Red Angus	576	0.196	0.252	0.22	0.16	0.14	0.19	0.12	0.17
Romangola	576	0.201	0.239	0.16	0.17	0.20	0.18	0.16	0.13
Santa Gertrudis	576	0.204	0.232	0.12	0.24	0.16	0.17	0.16	0.14
Sheko	576	0.175	0.234	0.25	0.21	0.13	0.14	0.16	0.10

(E) Breed of SNP Discovery: Limousin

		Minor Allele Frequencies		Fraction of All SNP					
	Num SNP	Include Monomorphic	Exclude Monomorphic	Monomorphic	0 < MAF ≤ 0.1	0.1 < MAF ≤ 0.2	0.2 < MAF≤ 0.3	0.3 < MAF≤ 0.4	0.4 < MAF≤ 0.5
Angus	576	0.199	0.238	0.16	0.20	0.15	0.19	0.14	0.15
Anoa	576	0.040	0.363	0.89	0.00	0.00	0.06	0.00	0.05
Beefmaster	576	0.221	0.241	0.08	0.19	0.21	0.17	0.21	0.14
Brahman	576	0.133	0.172	0.23	0.35	0.18	0.07	0.08	0.09
Brown Swiss	576	0.188	0.234	0.20	0.20	0.16	0.15	0.15	0.14
Buffalo	576	0.032	0.394	0.92	0.00	0.00	0.03	0.00	0.05
Charolais	576	0.225	0.253	0.11	0.17	0.19	0.18	0.14	0.20
Gir	576	0.115	0.184	0.37	0.22	0.15	0.12	0.07	0.06
Guernsey	576	0.200	0.245	0.19	0.18	0.16	0.15	0.14	0.18
Hereford	576	0.241	0.268	0.10	0.15	0.16	0.19	0.19	0.22
Holstein	576	0.203	0.229	0.12	0.25	0.17	0.14	0.18	0.15
Jersey	576	0.221	0.255	0.13	0.17	0.17	0.18	0.13	0.21
Limousin	576	0.215	0.236	0.09	0.22	0.17	0.18	0.18	0.16
Ndama	576	0.151	0.230	0.35	0.15	0.14	0.15	0.11	0.10
Nelore	576	0.120	0.200	0.40	0.20	0.15	0.08	0.10	0.07
Norwegian Red	576	0.207	0.249	0.17	0.18	0.15	0.16	0.18	0.16
Peidmontese	576	0.226	0.252	0.10	0.16	0.20	0.16	0.19	0.18
Red Angus	576	0.196	0.252	0.22	0.16	0.14	0.19	0.12	0.17
Romangola	576	0.201	0.239	0.16	0.17	0.20	0.18	0.16	0.13
Santa Gertrudis	576	0.204	0.232	0.12	0.24	0.16	0.17	0.16	0.14
Sheko	576	0.175	0.234	0.25	0.21	0.13	0.14	0.16	0.10

(F) Breed of SNP Discovery: Norwegian Red

		Minor Allele Frequencies		Fraction of All SNP					
	Num SNP	Include Monomorphic	Exclude Monomorphic	Monomorphic	0 < MAF ≤ 0.1	0.1 < MAF ≤ 0.2	0.2 < MAF≤ 0.3	0.3 < MAF≤ 0.4	0.4 < MAF≤ 0.5
Angus	8106	0.211	0.245	0.14	0.19	0.16	0.18	0.16	0.17
Anoa	8097	0.047	0.369	0.87	0.00	0.00	0.07	0.00	0.06
Beefmaster	8106	0.228	0.249	0.08	0.17	0.19	0.19	0.19	0.17
Brahman	8106	0.129	0.173	0.26	0.34	0.15	0.10	0.09	0.07
Brown Swiss	8106	0.195	0.237	0.18	0.19	0.17	0.16	0.16	0.14
Buffalo	8097	0.039	0.391	0.90	0.00	0.00	0.04	0.00	0.06
Charolais	8106	0.226	0.249	0.09	0.19	0.18	0.18	0.17	0.19
Gir	8106	0.113	0.187	0.40	0.22	0.14	0.10	0.08	0.07
Guernsey	8106	0.202	0.246	0.18	0.19	0.16	0.15	0.15	0.18
Hereford	8106	0.239	0.262	0.09	0.17	0.16	0.21	0.18	0.20
Holstein	8106	0.218	0.240	0.09	0.21	0.18	0.17	0.18	0.17
Jersey	8106	0.183	0.227	0.20	0.22	0.15	0.16	0.14	0.14
Limousin	8106	0.221	0.241	0.08	0.21	0.18	0.17	0.17	0.18
Ndama	8106	0.148	0.234	0.37	0.15	0.14	0.13	0.11	0.11
Nelore	8106	0.112	0.188	0.40	0.23	0.14	0.08	0.09	0.07
Norwegian Red	8106	0.246	0.263	0.07	0.16	0.17	0.22	0.18	0.21
Piedmontese	8106	0.224	0.247	0.09	0.17	0.20	0.19	0.19	0.16
Red Angus	8106	0.209	0.253	0.18	0.17	0.15	0.21	0.13	0.18
Romangola	8106	0.204	0.242	0.16	0.18	0.18	0.16	0.17	0.15
Santa Gertrudis	8106	0.216	0.241	0.10	0.23	0.18	0.15	0.19	0.15
Sheko	8106	0.173	0.235	0.27	0.20	0.15	0.12	0.15	0.12

Table S6 Mean and standard deviation for observed proportion of shared haplotypes between Angus and Holstein.

Chromosome	Loci	mean S	s.d. S	mean S′	s.d. S′
BTA6	132	0.634	0.207	0.240	0.117
BTA14	201	0.581	0.228	0.247	0.134
BTA25	85	0.586	0.221	0.214	0.133

Table S7 Assignment of ancestral SNP allele by chromosome and spacing (bp) between markers
Includes markers from the 4.5K and 7.5K sets that were not genotyped in the outgroup species

BTA	# Valid Markers[1]	Ave. Spacing	Min. Spacing	Max. Spacing	# Ancestral Known	Ancestral Ave. Spacing	Ancestral Min Spacing	Ancestral Max Spacing
1	1489	96692	7	941423	598	241000	25	2438107
2	1464	89633	8	730260	580	225420	12	1587630
3	1284	92737	3	712246	521	226764	3	1823586
4	1284	91340	6	904416	524	223121	9	1733402
5	1217	95435	20	756584	470	245975	21	1726378
6	2420	48138	11	903871	536	217229	23	1977165
7	1058	97809	4	986266	398	257240	21	1576673
8	1179	92399	9	1019449	459	236783	29	1411567
9	976	104787	2	940400	386	261034	21	2091272
10	1103	93008	7	1104187	454	226258	20	2052176
11	1205	89337	5	875942	483	223156	28	1680469
12	864	92536	1	946290	336	238253	26	1459213
13	968	84357	5	890466	346	234274	7	1558751
14	2722	29578	5	623416	370	217768	5	1454769
15	818	97649	2	1944406	291	264361	23	2074585
16	858	86291	8	867493	333	221505	8	1769846
17	803	88922	1	569866	326	219432	36	1256942
18	640	95311	15	1019891	273	223550	17	2328223
19	665	93815	4	972841	252	246760	16	1366682
20	860	87167	7	1421140	334	222701	36	1920696
21	650	99322	15	799431	231	276120	25	1740887
22	678	88235	4	1123980	264	220842	12	2059685
23	567	86520	8	580792	211	231122	8	1424476
24	675	89824	21	680079	265	229322	15	1272328
25	1184	34673	1	879814	169	242198	50	1041640
26	581	84460	16	1075463	238	205827	25	1462344
27	480	92917	17	881457	177	252882	26	1620904
28	501	88724	20	670421	192	232262	19	1632138
29	467	101204	10	1189809	176	267392	6	2373272
X	565	159461	20	2255471	234	382459	2	2535130
UN	2802				939			
Total	33027	89076	1	2255471	11366	240434	2	2535130

Table S8 Regions with evidence of recent positive selection detected by iHS. The summary includes those regions where multiple SNP separated by <1 Mb had |iHS| > 2.694 (top 1%). ANG = Angus, BMA = Beefmaster, BRM = Brahman, BSW = Brown Swiss, CHL = Charolais, GIR = Gir, GNS = Guernsey, HFD = Hereford, HOL = Holstein, JER = Jersey, LMS = Limousin, NDA = N'Dama, NEL = Nelore, NRC = Norwegian Red, PMT = Piedmontese, RGU = Red Angus, RMG = Romagnola, SGT = Santa Gertrudis, SHK = Sheko.

BTA	Position (Mb)	Breeds[1]
1	68.5-70.2	BRM, CHL, NDA, PMT
	83.4-84.5	LMS, SGT, SHK
	109.7-110.6	NRC, PMT, SGT
2	5.1-10.7	LMS, PMT, NDA
	112.9-114.1	BMA, BRM, CHL, GNS, SGT
3	79.0-79.4	GNS, RGU, SGT
	96.5-98.8	ANG, HFD, SGT
	102.1-104.7	ANG, HOL, LMS, RGU, RMG
4	95.1-96.6	BMA, PMT, RGU
5	32.7-33.8	LMS, NRC, SGT
	54.1-55.6	SGT
	79.4-80.6	LMS, NRC, RMG, SGT
	104.5-105.6	HOL, LMS, PMT
6	33.4-34.4	BMA, BSW, CHL, PMT, RMG, SHK
	36.8-37.8	ANG, BMA, HFD, JER, NEL, NRC, RMG, SGT, SHK
	44.2-45.3	GNS, HOL, LMS, NDA, NRC
7	14.6-15.3	JER
	37.4-39.1	JER, RMG, SGT
	45.4-47.1	SGT
8	57.1-57.7	BSW, CHL, GIR, JER
9	55.1-55.7	GIR, HFD, PMT
10	13.6-15.2	PMT, RMG
	53.3-53.4	CHL, HOL, LMS, NDA, SGT
11	27.7-28.1	NRC
	61.0-63.9	GNS, HFD, NDA, PMT
	68.9-71.6	GNS, LMS, RGU, SGT
13	9.6-10.5	CHL, NEL, SHK
	18.7-19.9	CHL, HFD, HOL, PMT
	25.0-27.0	BMA, BSW, GIR, GNS, NEL, NRC, RMG, RGU, SGT
	41.8-42.8	BSW, GNS, NEL, NRC, RGU
	71.5-73.8	BSW, HOL, NRC, PMT, SHK
14	2.6-10.4	BMA, BRM, CHL, GIR, GNS, HFD, LMS, NDA, NEL, PMT, SGT
	15.5-16.9	BRM, CHL, GNS, LMS
	23.9-30.1	BMA, BRM, CHL, LMS, NEL, NRC, RGU
	42.3-43.2	BRM, GNS, HOL, LMS, NRC
	51.6-53.3	ANG, BMA, BRM, CHL, GIR, GNS, HFD, LMS, NDA, NEL, NRC, RGU, RMG, SGT
	58.6-62.2	ANG, BMA, BRM, HFD, JER, LMS, NRC, PMT, RMG, SGT, SHK
15	11.2-17.0	BMA, GIR, NEL
16	3.0-3.1	CHL, BMA
	43.9-44.7	ANG, GIR, HFD, NDA, RGU
	66.2-66.3	LMS, NEL, PMT

17	29.0-29.8	BMA, RGU
	63.0-63.5	ANG, BSW, GNS, HFD, SHK
18	14.9-15.0	GIR, RGU, RMG
19	22.3-22.5	HFD
	24.8-25.4	BMA, GNS,SHK
20	18.2-19.6	BSW, GNS, LMS
	22.7-24.1	BRM, BSW, GIR, HOL, JER, RMG
	32.0-33.1	HOL, RMG
	44.3-44.7	HOL, NEL, SHK
22	29.1-29.2	BMA, BRM, LMS, PMT, SGT
23	30.6-33.3	NDA, NEL
25	6.9-7.8	CHL, HFD, HOL, JER, LMS, SHK
	12.9-13.3	BMA, JER, NEL, PMT, RGU
28	26.6-28.4	CHL, HFD, NRC, SGT, SHK

Table S9 Minimum number of SNP required for four generalized traceability and parentage scenarios. Identity and paternity analyses require the ability to identify any animal (or sire) in the world (1.3 Billion animals) with <1 chance in a million of a false match. Parentage analysis requires the ability to identify parentage in a herd of 80,000 cows and 4000 sires with <1 chance in a million of a false match

Comparison	Threshold (Match probability)	Number of SNP required
Identity[1]	$1.3\mathrm{x}10^{-15}$	>50
Paternity dam unknown[1]	$1.3\mathrm{x}10^{-15}$	>400
Paternity dam known[1]	$3.2\mathrm{x}10^{-14}$	>100
Parentage[2]	$3.2\mathrm{x}10^{-14}$	>100

Supplementary References

S1. D. Altshuler *et al.*, Nature **407,** 513 (2000).

S2. R. J. Hawken, W. C. Barris, S. M. McWilliam, B. P. Dalrymple, *Mamm. Genome* **15,** 819 (2004).

S3. D. Falush, M. Stephens, J. K. Pritchard, *Genetics* **164,** 1567 (2003).

S4. P. Scheet, M. Stephens, *Am. J. Hum. Genet.* **78,** 629 (2006).

S5. A. M. Bowcock *et al.*, *Nature* **368,** 455 (1994).

S6. J. L. Mountain, L. L. Cavalli-Sforza, *Am. J. Human Genet.***61,** 705 (1997).

S7. D. J. Witherspoon *et al.*, *Genetics* **176,** 351 (2007).

S8. J. Felsenstein, *Cladistics* **5,** 164 (1989).

S9. B. F. Voight *et al.*, *PLoS Biol* **4,** e72 (2006).

S10. J. M. Akey *et al.*, *Genome Res.* **12,** 1805 (2002).

S11. L. L. Cavalli-Sforza, *Proc. R. Soc. Lond. B Biol. Sci.* **164,** 362 (1966).

S12. R. Nielsen *et al.*, *Genome Res.* **15,** 1566 (2005).

S13. P. C. Sabeti *et al.*, *Nature* **449,** 913 (2007).

S14. L. Excoffier, G. Laval, S. Schneider, *Evolutionary Bioinformatics* 1, 47 (2005).

S15. M. Slatkin, *Genetics* **139,** 457 (1995).

S16. O. Hanotte *et al.*, *Science* **296,** 336 (2002).

S17. T. Cymbron *et al.*, *Proc. Biol. Sci.* **272,** 1837 (2005).

S18. C. R. Henderson. *J Dairy Sci*. **57**, 963 (1974).

S19. K. G. Dodds, M. L. Tate, J. A. Sise, *J. Anim. Sci.* **83**, 2271 (2005).

Supporting Online Material

Table of Contents

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 50ºC; annealing temperature of the next 30 cycles was carried out at 50ºC. 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^{o}C, optimum 60^{o}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 (*S*96) 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 where 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-intio* 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 ≤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 50ºC; annealing temperature of the next 30 cycles was carried out at 50ºC. 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 ≥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 (NSsites = 0, 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_{model2} and ln $L_{one\text{-}ratio}$ (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*; *ionotrophic 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 (*GPAA1*) 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 γδ 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 1 cDNA sequences (*S118*). Two of the genes identified in the assembly encoded novel proteins homologous to MHC class 1 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 Proboscidia (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

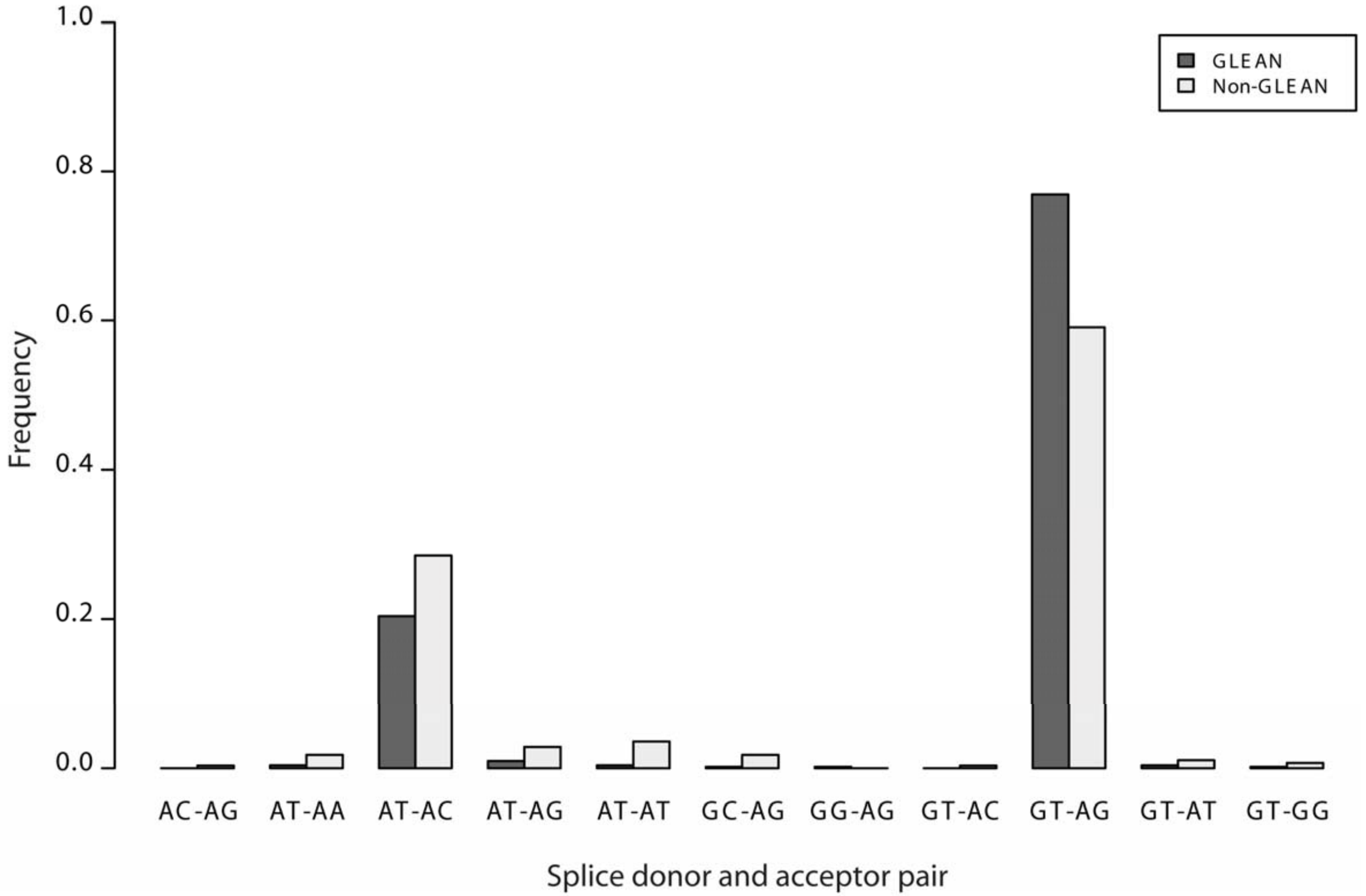


Fig. S1. Frequencies of intron donor and acceptor combinations according to whether or not they were predicted by GLEAN5.

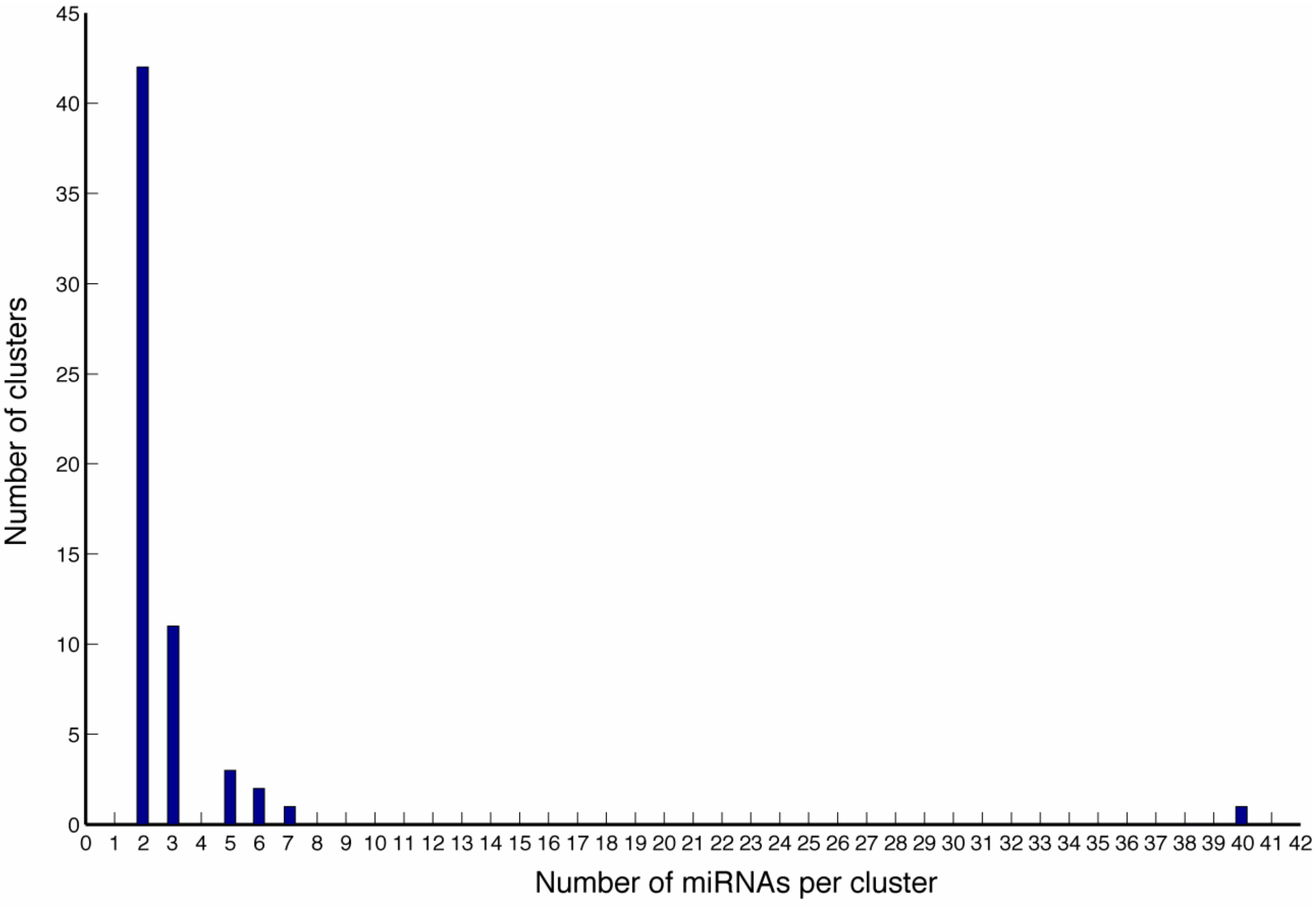


Fig. S2. Number of clusters of miRNA containing two or more miRNA. A total of 496 bovine miRNAs were grouped into 298 homolog families. miRNAs were considered to form a cluster if they were separated by less than 10 kbp. About half of the bovine miRNA occur in genomic clusters containing 2 to 7 miRNA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

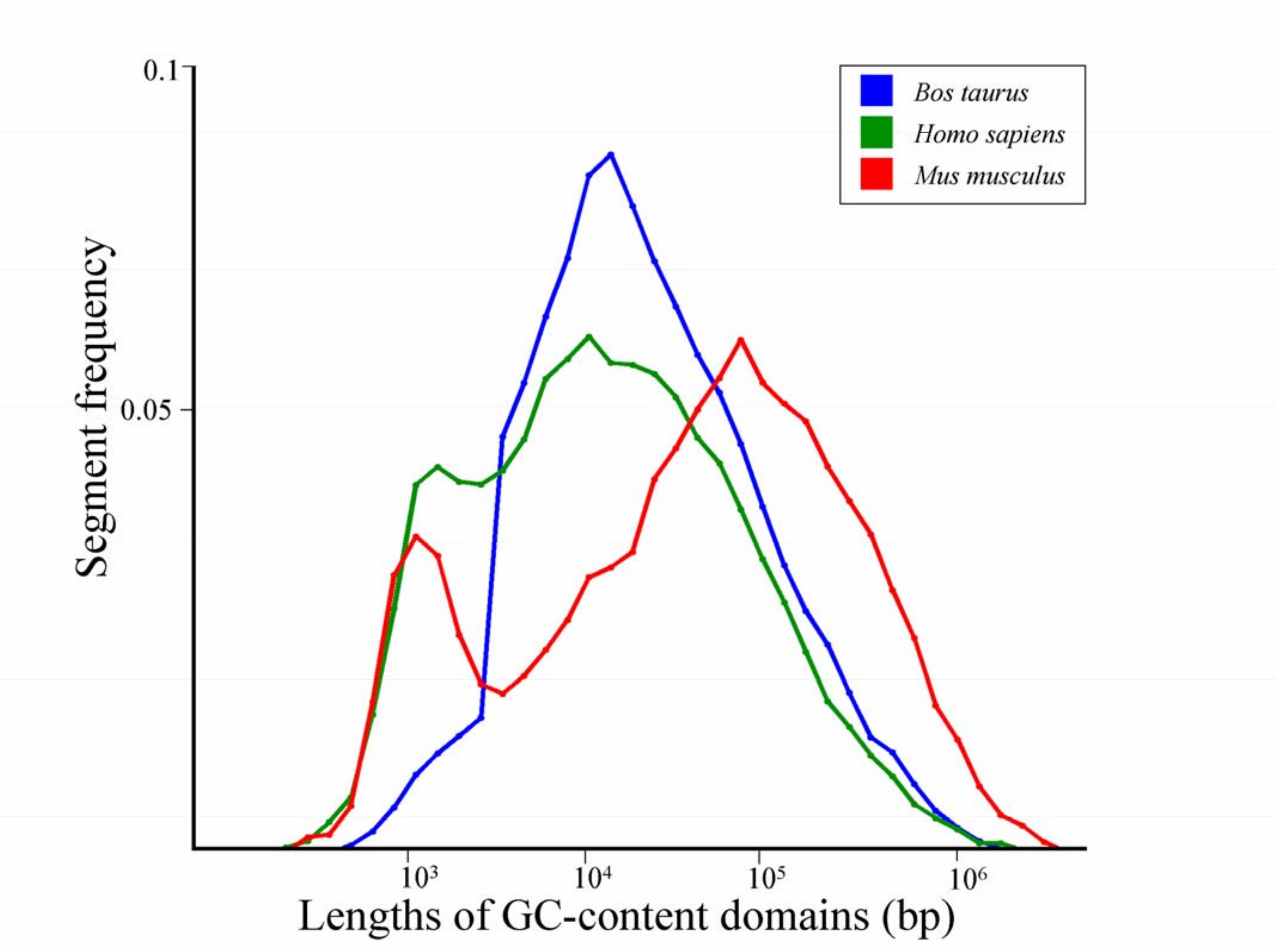


Fig. S3. The frequency of GC-content domain segments in cattle, human, and mouse.

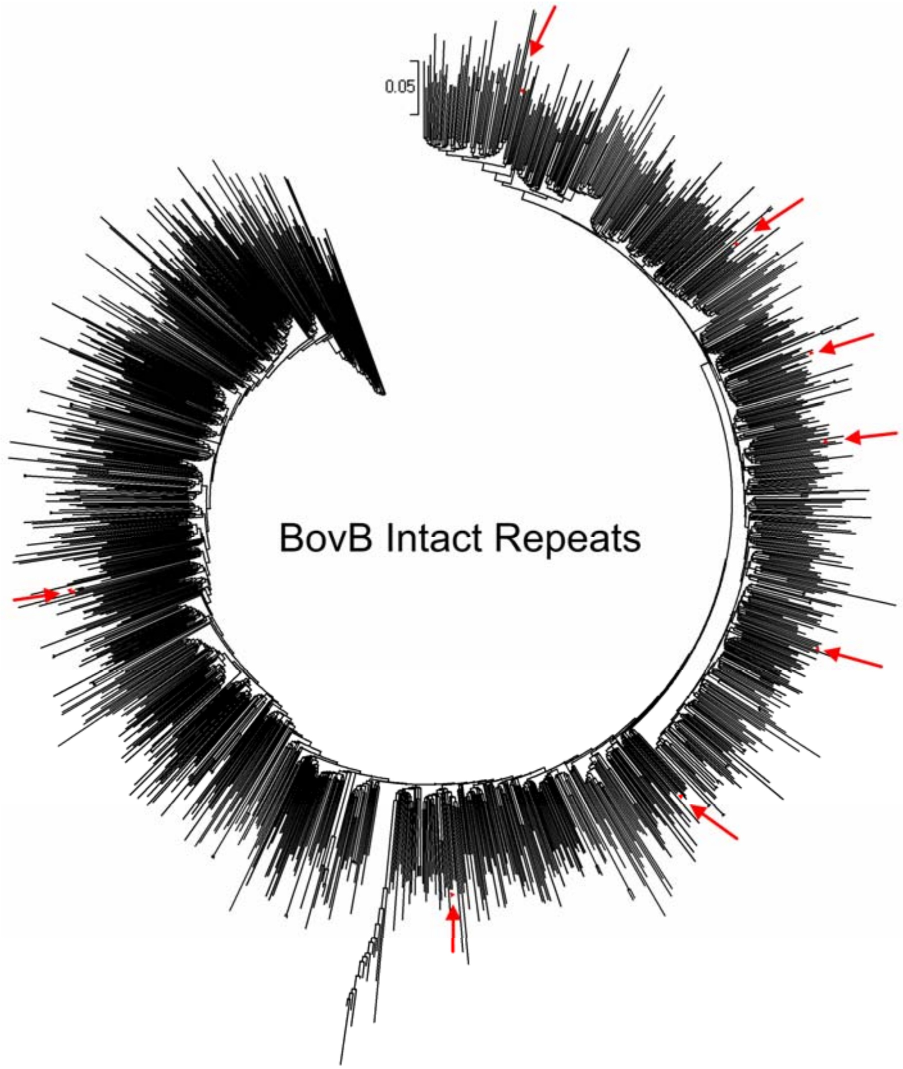


Fig. S4. Maximum likelihood tree derived from global alignments of intact/full length BovB intact repeats. Red arrows/triangles indicate potentially active LINEs of the basis of their intact ORF content.

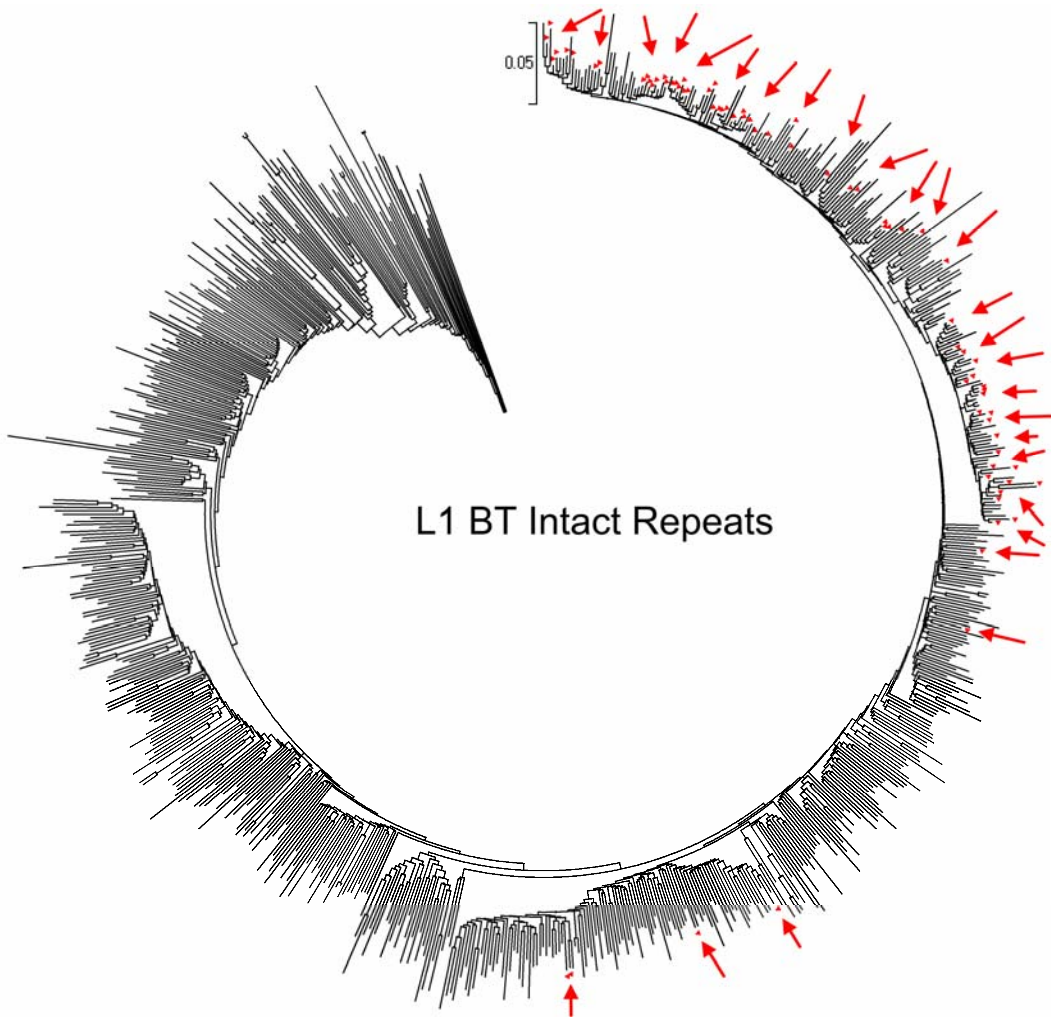


Fig. S5. Maximum likelihood tree derived from global alignments of intact/full length L1_BT intact repeats. Red arrows/triangles indicate potentially active LINEs on the basis of their intact ORF content.

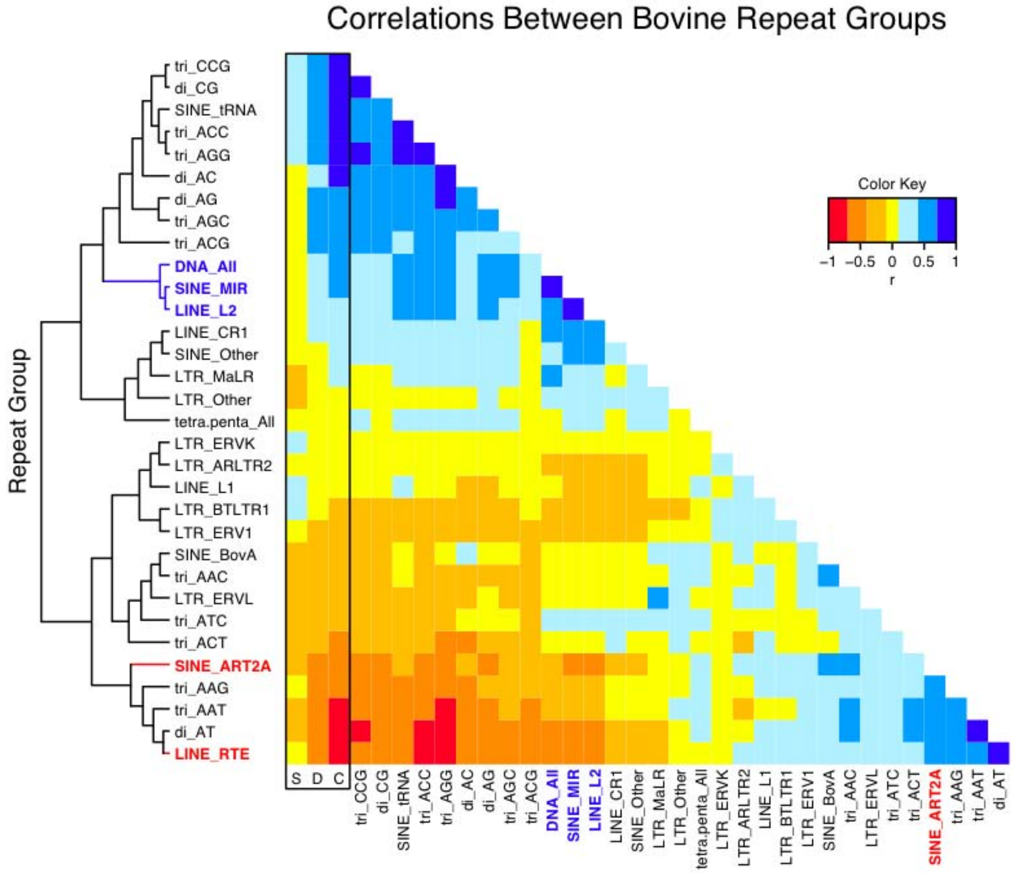


Fig. S6. Correlation analysis of repeat groups. Pairwise correlations among the repeat groups and between the repeat groups and segmental duplication (column S), gene density (column D), and GC content (column C). Repeat groups were clustered of the basis of all their correlations. Yellow cells have 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.

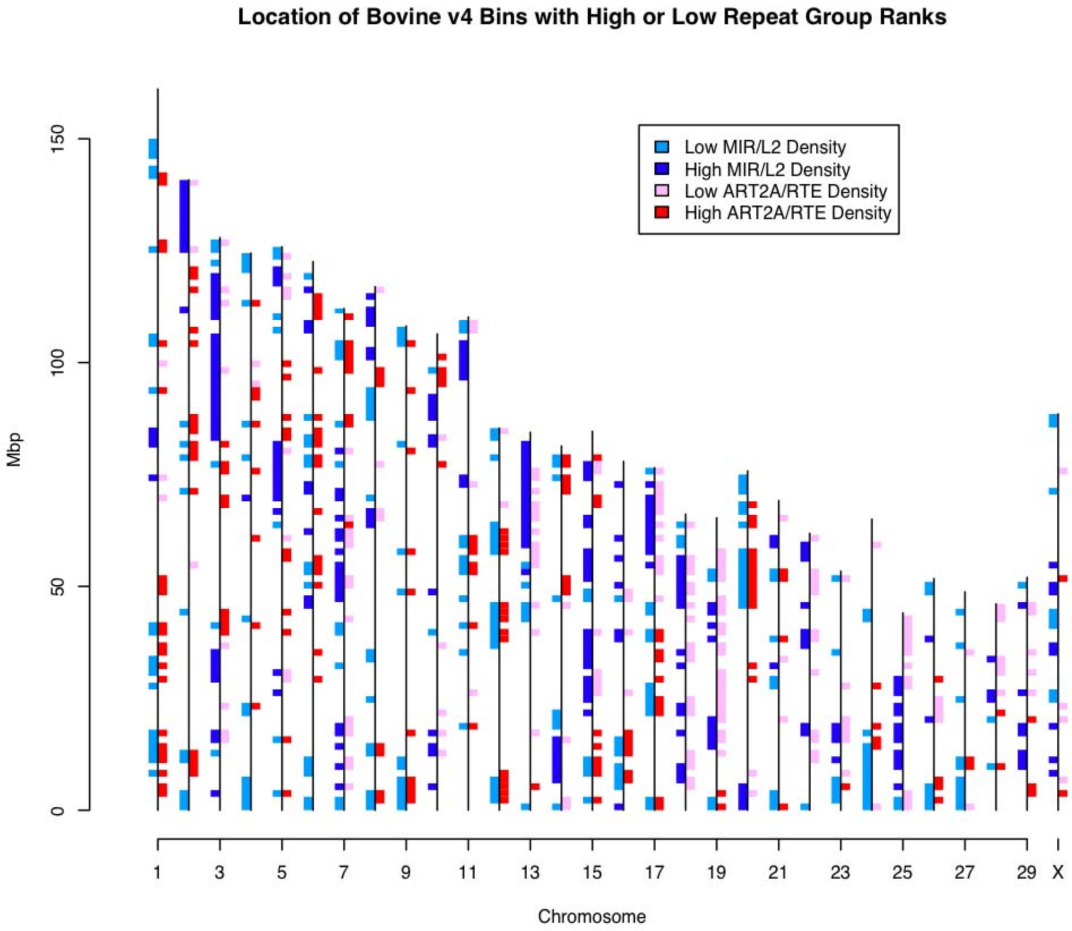


Fig. S7. Spatial distribution of ancient and new extreme density bins for repeats. The top and bottom 5% tails of the bins for L2/MIR and BovB/Art2A (Art2A/RTE) correlations were identified based on an expected random distribution in Btau4.0. High density L2/MIR regions often form contiguous blocks and 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.

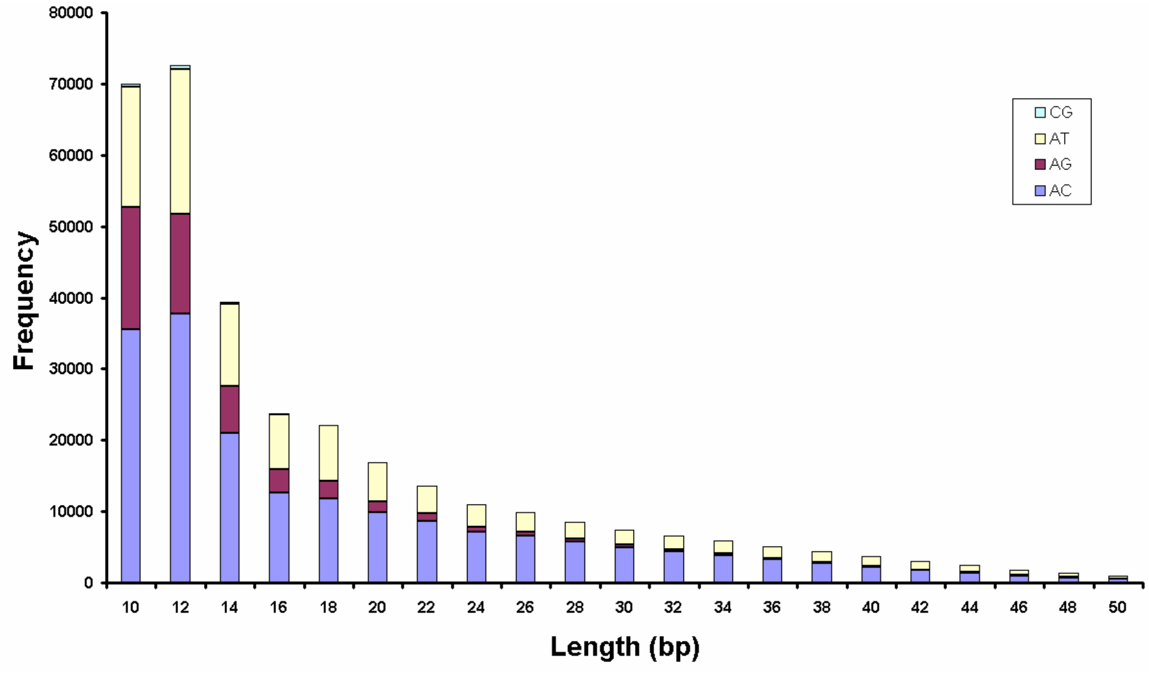


Fig. S8. Frequency of dinucleotide simple sequence repeats of varying lengths in the bovine genome.

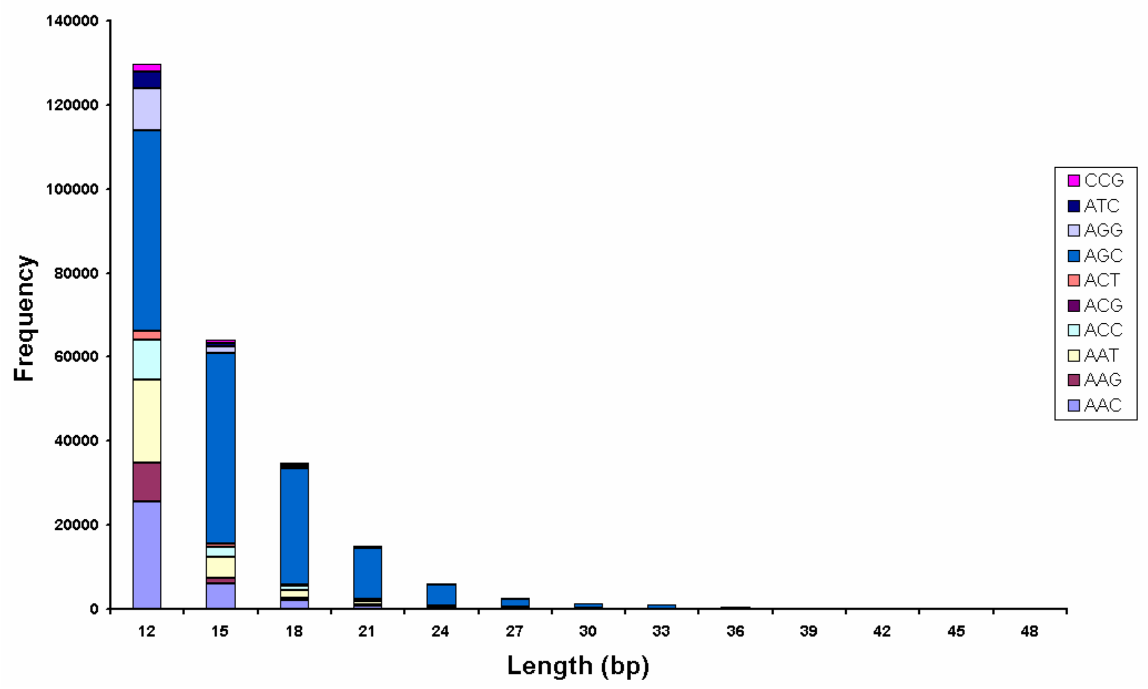


Fig. S9. Frequency of trinucleotide simple sequence repeats of varying lengths in the bovine genome.

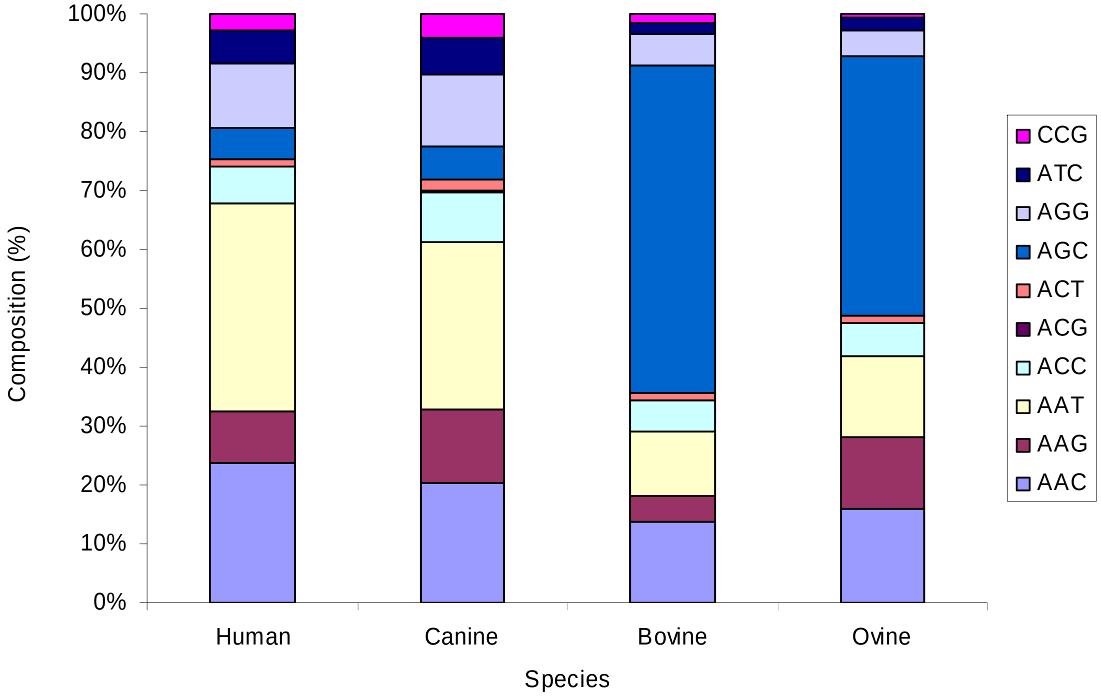


Fig. S10. Comparative compositions of trinucleotide simple sequence repeats (SSRs) in four mammalian species.

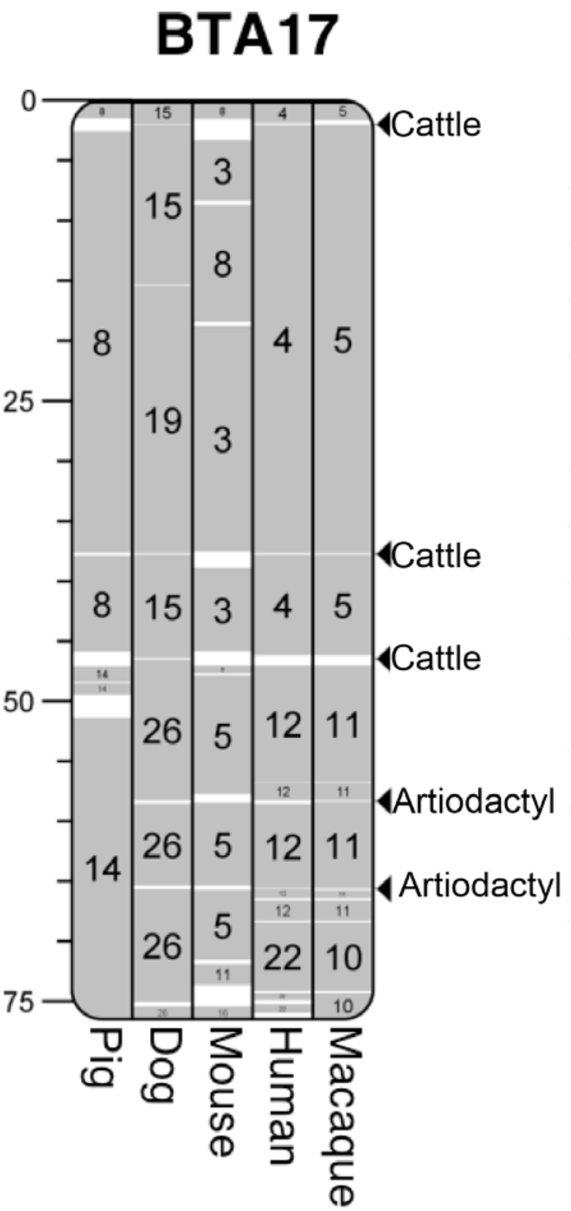


Fig. S11. Cattle-specific and artiodactyl breakpoint regions in BTA17. Homologous synteny blocks (HSBs) defined using BTA17 (Btau4.0) as the reference genome are indicated by grey shading. Macaque (RheMac2, build 2), human (HSA36, build 36), dog (CFA2, build 2), mouse (MMU9, build 37) and pig (physical 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*).

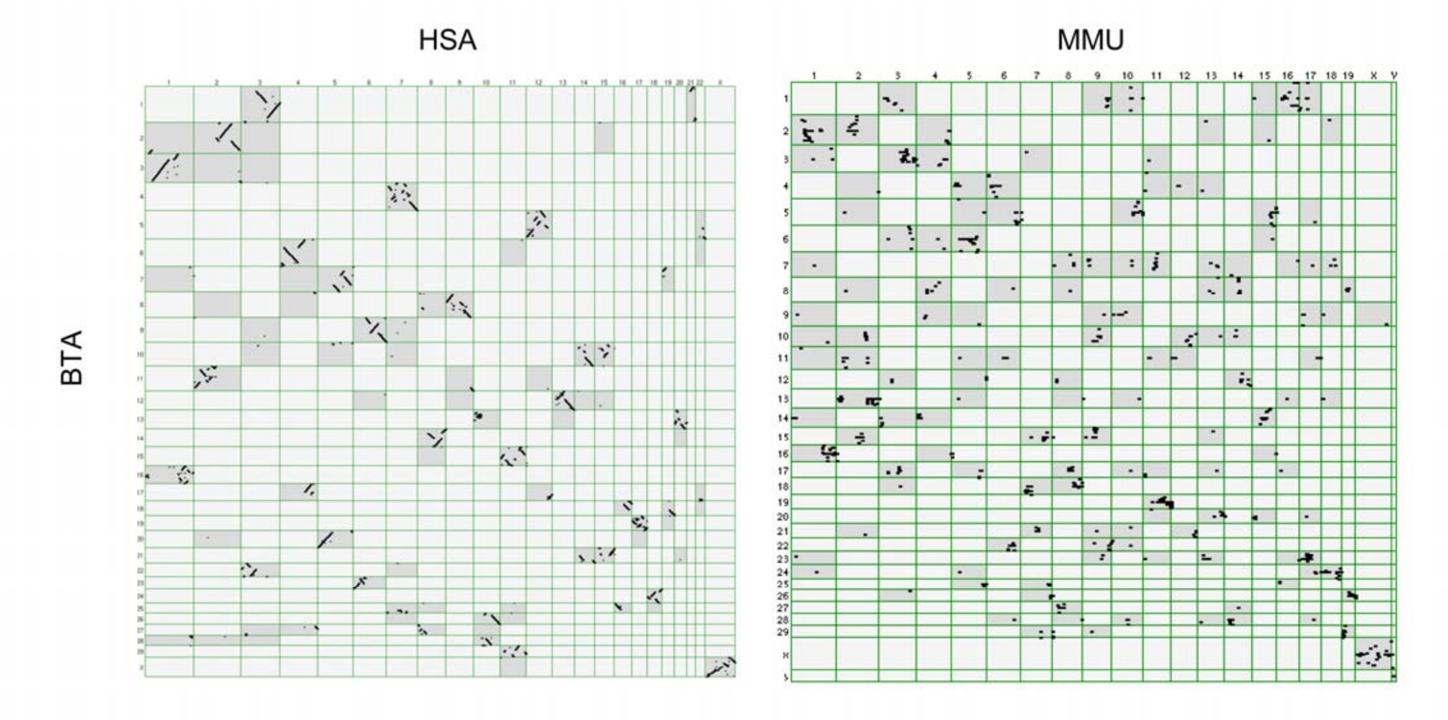


Fig. S12. Oxford grids showing chromosome comparisons between species. An Oxford grid showing the extent of chromosomal conservation between cattle (BTA) and humans (HSA) was drawn from 46,947 cattle BAC-end sequences showing high homology with sequence from build 36 of the human genome. An Oxford grid 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*).

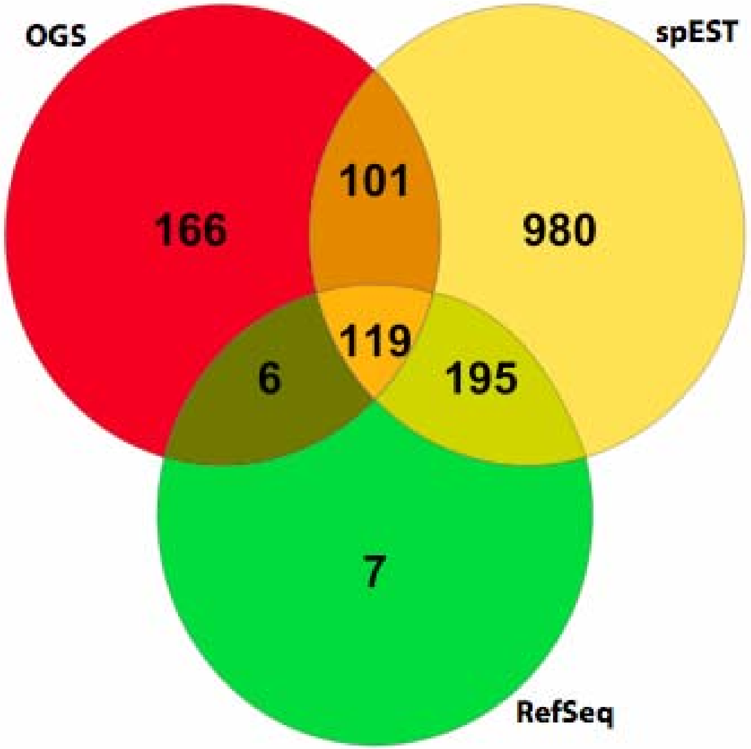


Fig. S13. Venn diagram showing bidirectional promoters mapped in the cattle genome. Annotation sources were OGSv1 (red),

spliced ESTs (yellow), and RefSeq genes (green).

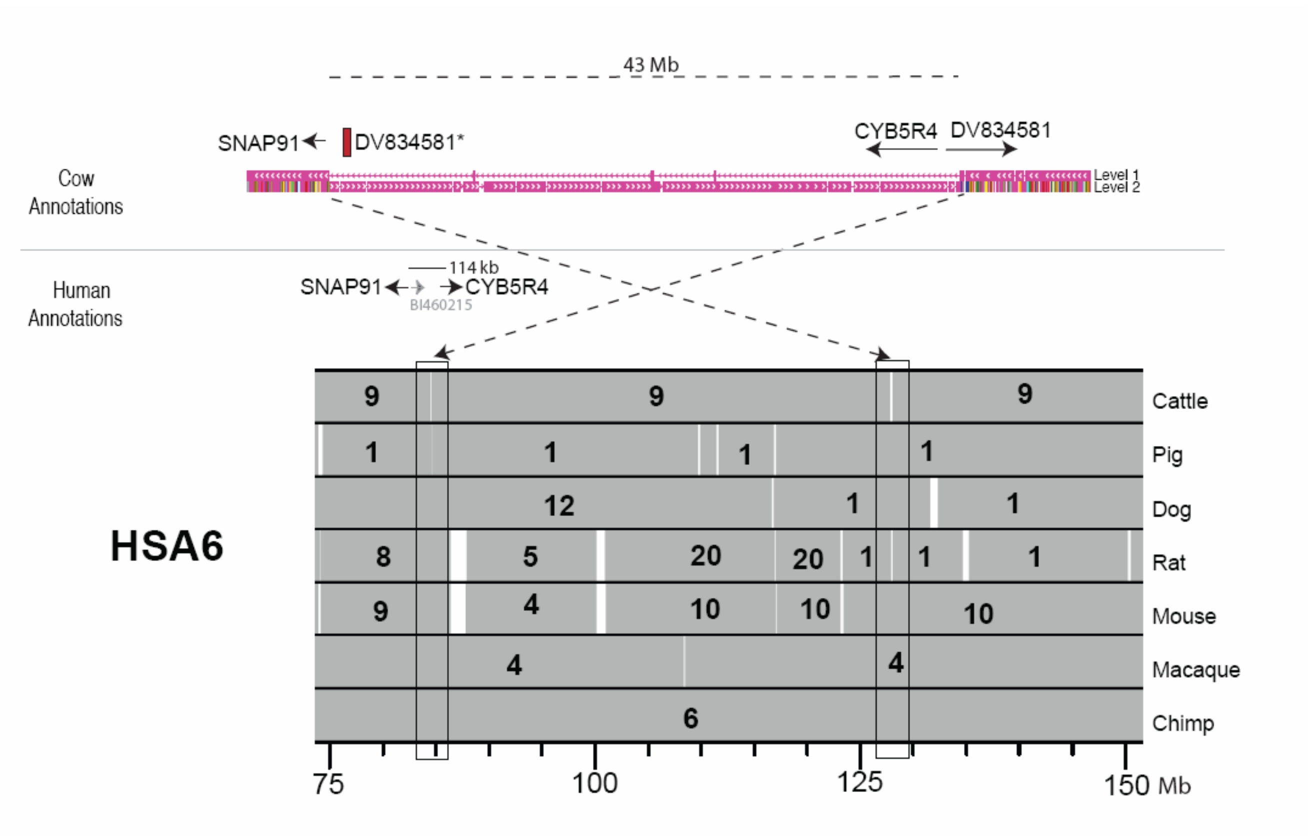


Fig. S14. An inversion of 43 Mbp in BTA9 is shown relative to human chromosome 6 (HSA6), which maintains an ancestral organization of the region as seen from the comparison of HSA6 to the othologous regions of pig, dog, rat, mouse, macaque, and chimp chromosomes (lower panel). The inversion in BTA9 is 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.

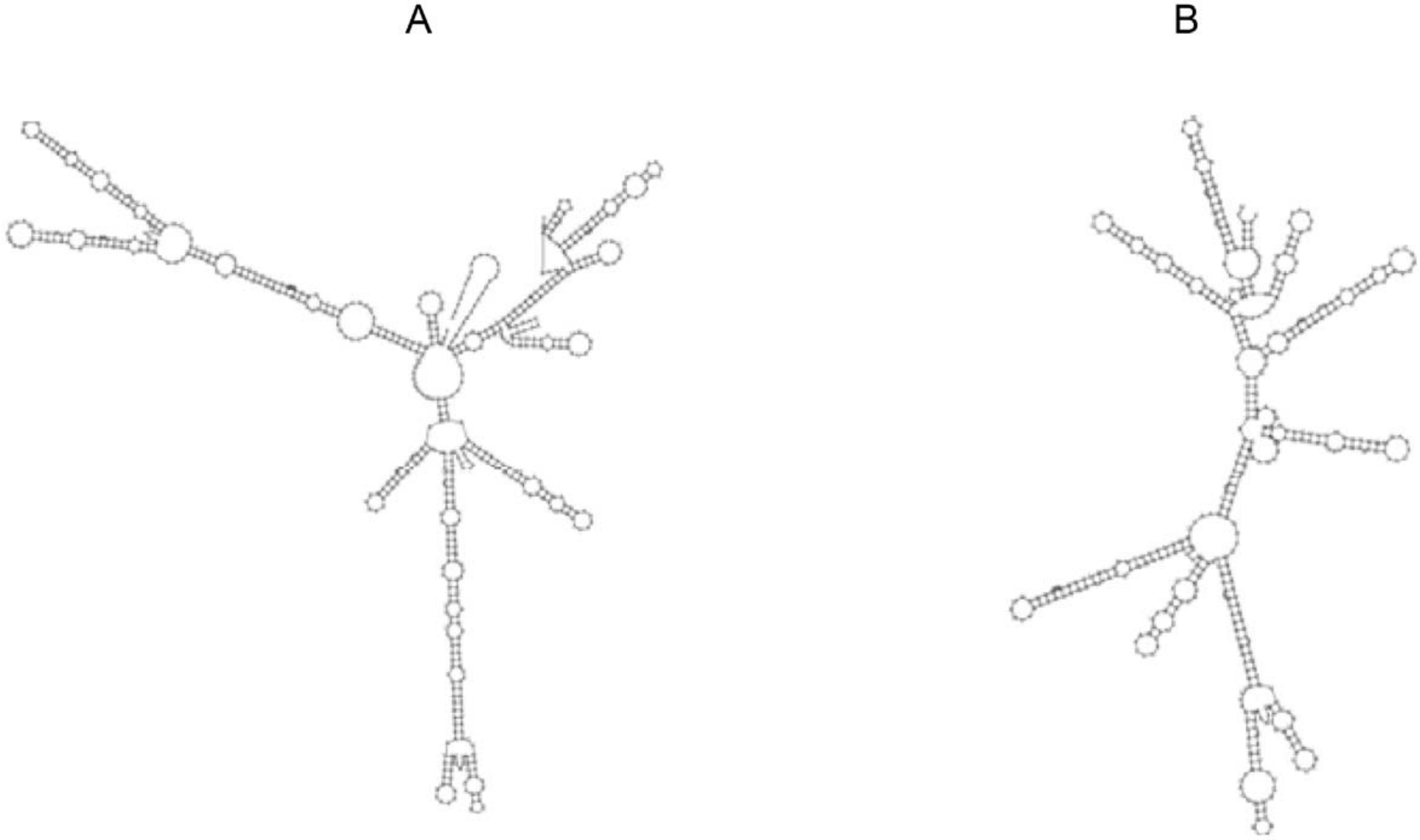


Fig. S15. RNA secondary structure of the putative non-coding gene partner to *CYB5R4* in cattle. Two isoforms (A and B) of this non-coding gene are produced by alternative splicing and both have been subjected to analysis of RNA secondary structure with RNAfold (*S43*). No orthologous gene was found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

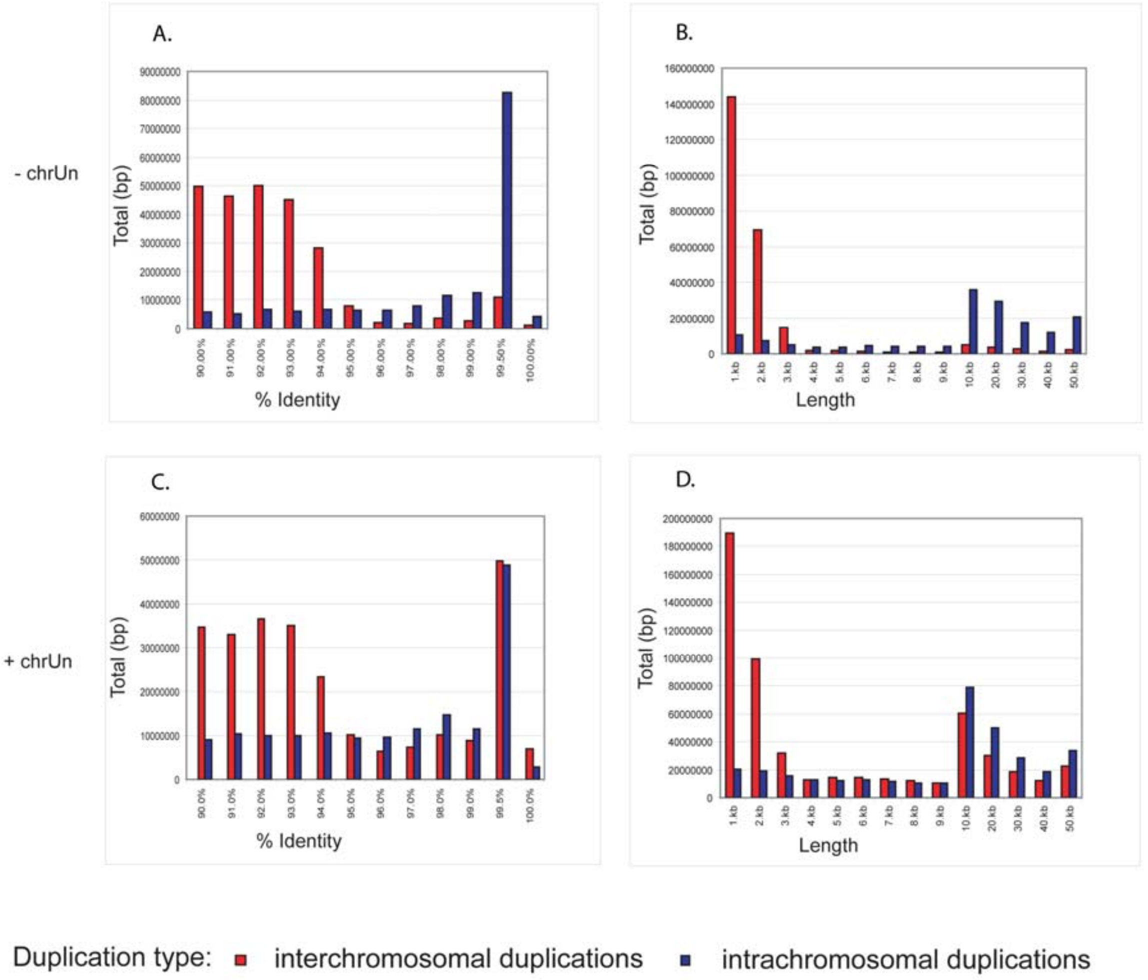


Fig. S16. Percent identity and length distribution of cattle segmental duplications. The total number of aligned bp for interchromosomal and intrachromosomal segmental duplications (SDs) was plotted as a function of the percent identity and length as identified by the whole-genome analysis comparison 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.

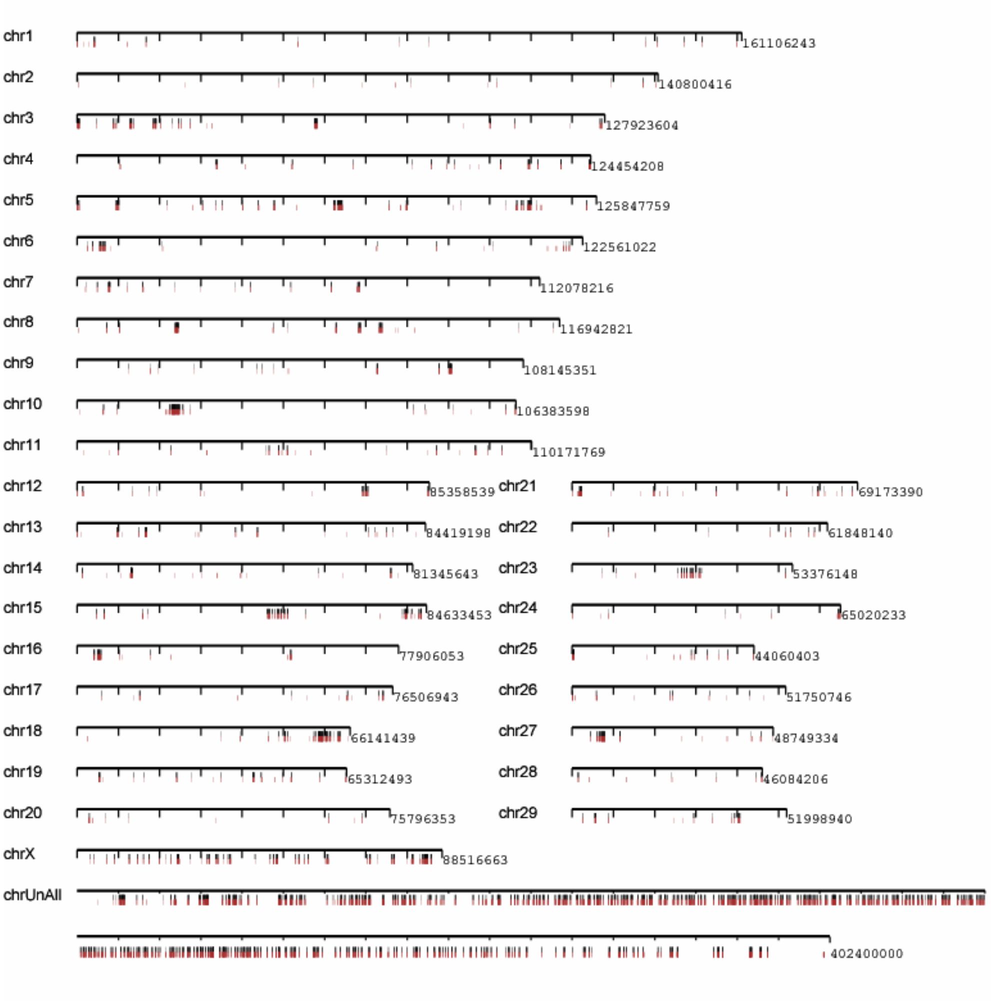


Fig. S17. Map of cattle segmental duplications. The locations of large (>10 kbp) SDs defined by WGAC (black) and WSSD (brown) methods are shown mapped onto bovine chromosomes. Each tick mark represents 10 Mbp. Large clusters of SDs are clearly identifiable within the genome although in some cases, the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

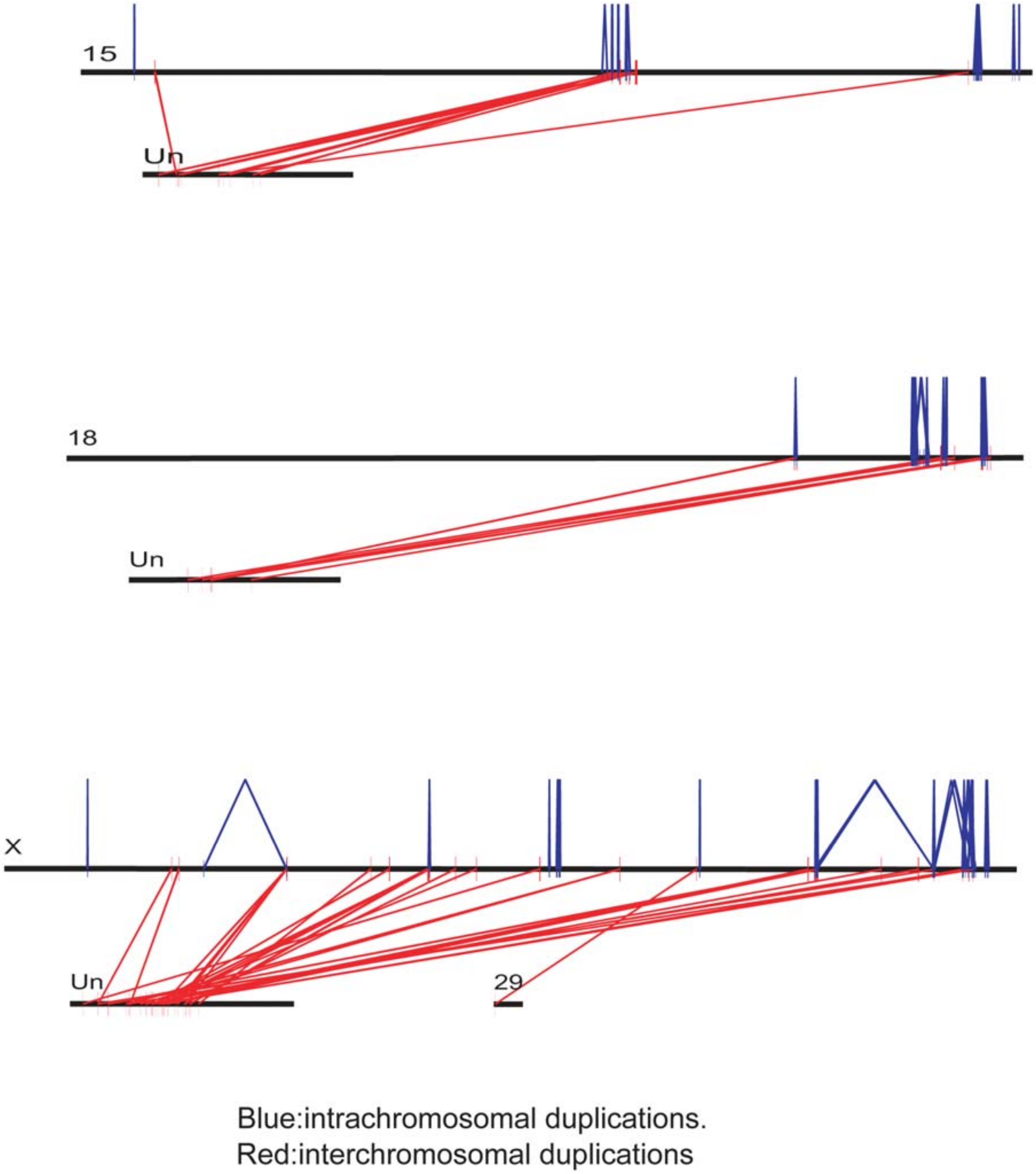

15
Un
18
Un
X
Un
29
Blue:intrachromosomal duplications.
Red:interchromosomal duplications

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).

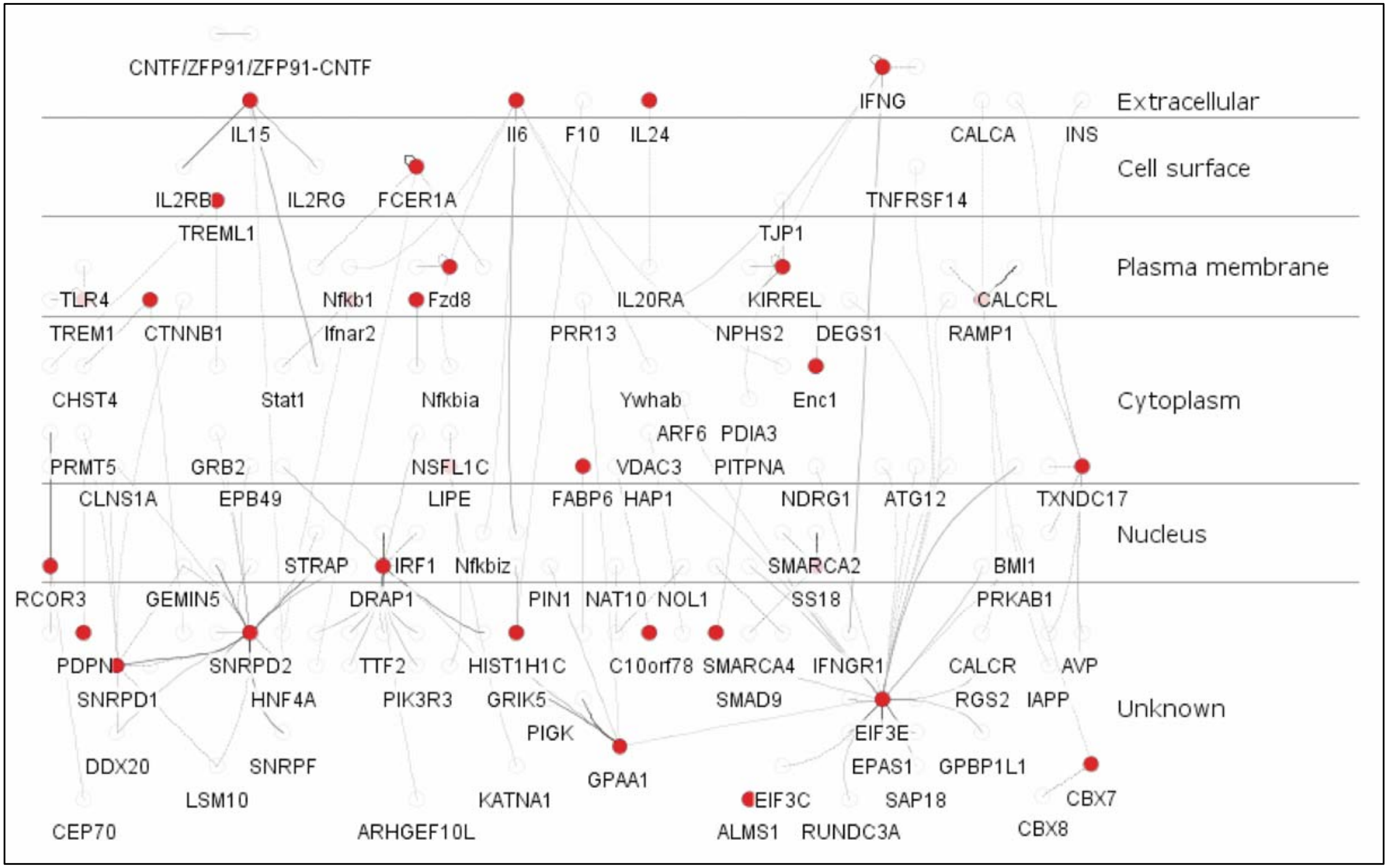


Fig. S19. Molecular interaction networks of the genes that had evidence of positive selection and their interacting partners. InnateDB (*S86*) was used to investigate and visualize the molecular interaction networks of the genes which had evidence of positive selection and their interacting partners.

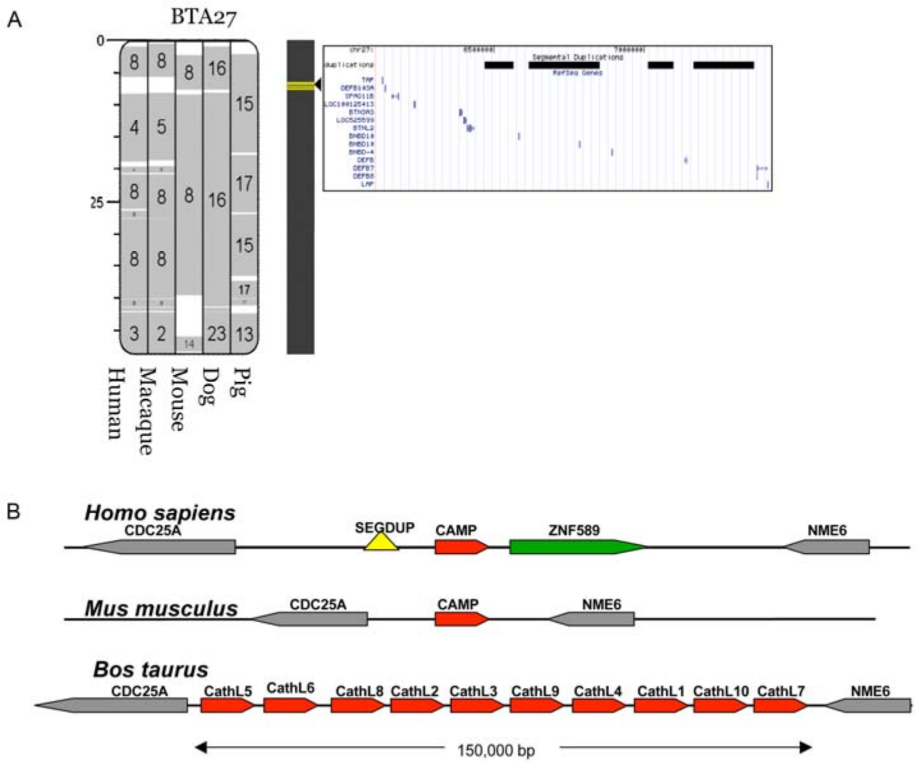


Fig. S20. Examples of changes in the comparative organizations of innate immune related genes. (A) A cattle-specific evolutionary breakpoint region (EBR) in BTA27 showing co-location of segmental duplications (SDs) with a cluster of β-defensin genes. Homologous synteny blocks (HSBs) defined using BTA27 (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 (SEGDUP) is shown by a yellow triangle).

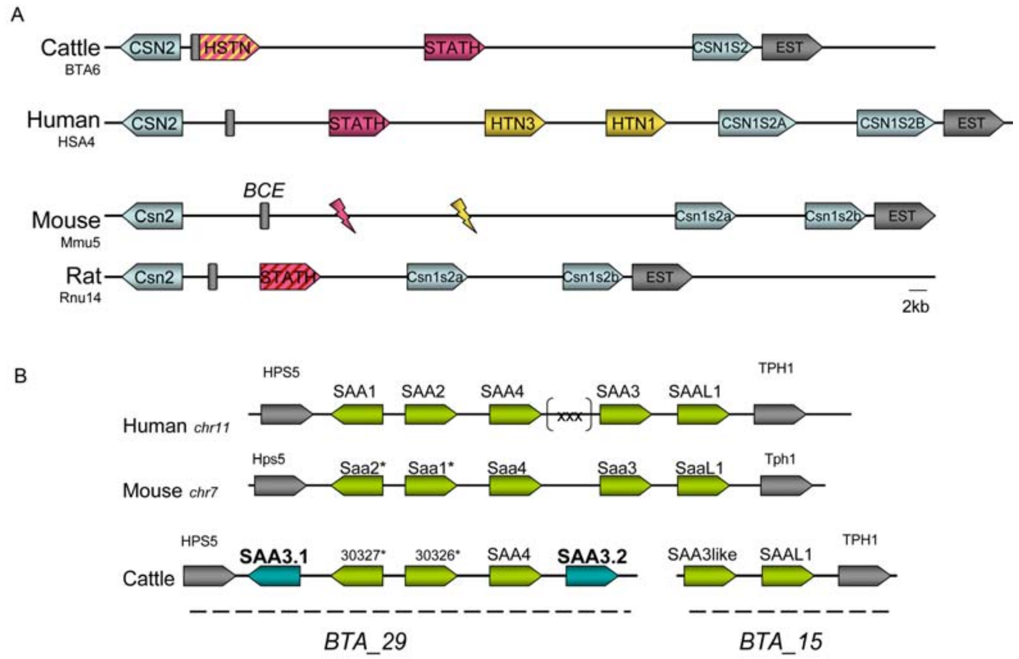


Fig. S21. Comparative representations of the organizations of the serum amyloid A and casein gene clusters in mammalian genomes. (A) Diagrammatic representation of the organization of the casein gene cluster in cattle, human, mouse and rat highlighting the histatin-statherin region. The β-casein gene (*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.

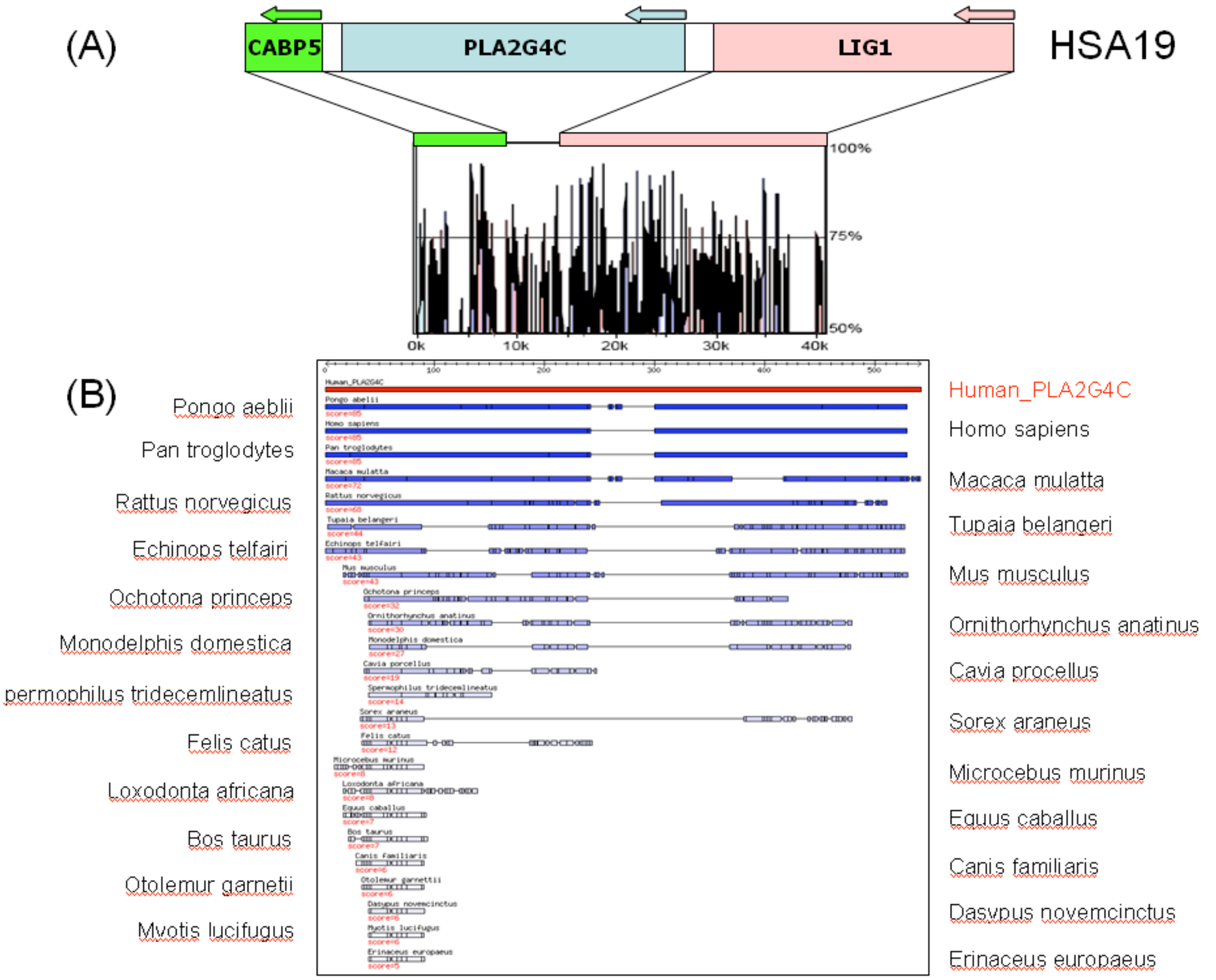
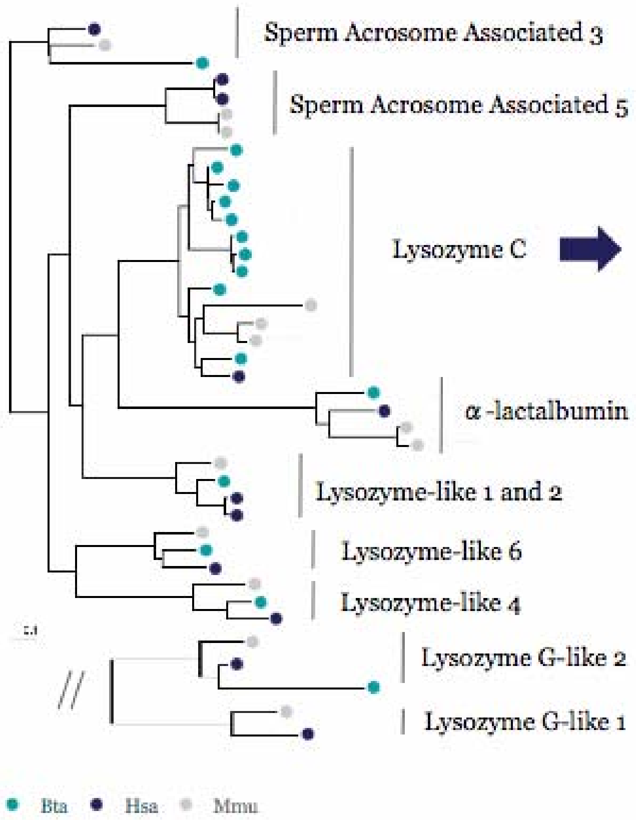


Fig. S22. Deletion of *PLA2G4C* in the cattle genome. (A) The upper panel shows orientation of *PLA2G4C* and flanking genes, CABP5 and LIG1, on HSA19. The lower panel shows a DNA sequence identity plot in 100 bp sliding windows of the cattle genome. The flanking orthologs of human *PLA2G4C*, *CABP5* 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.



Lysozyme C Expression

	Gene	Expressed in:
Bta	GLEAN_15543	**Intestine; Reticulum**
Bta	GLEAN_01314	**Intestine**
Bta	GLEAN_01313	**Rumen; Intestine;** Kidney
Bta	GLEAN_05349	**Rumen**; Mammary; **Reticulum**
Bta	GLEAN_05350	**Intestine**
Bta	GLEAN_15541	**Abomasum; Intestine**
Bta	GLEAN_15542	**Abomasum; Intestine**
Bta	GLEAN_15294	**Abomasum;** Extraembryonic
Bta	GLEAN_01306	**Rumen; Omasum; Reticulum**
Mmu	NP_084182	Prostate; Bladder
Mmu	NP_038618	**Intestine**; Pineal Gland; Muscle
Mmu	NP_059068	Pineal Gland; Lung; Bone
Bta	GLEAN_15292	Lymphoreticular; Blood; Testis
Hsa	NP_000230	Pharynx; Larynx; Trachea

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 Loci[1]	Number of Gene Models[2]	Perfect Matches to Full-length CDS[3]	High Identity Matches to Full-length CDS[3]	Input Gene Prediction Sets to GLEAN[4]
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

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

[2] Gene 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.

[3] 208 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 ≥99% identity over the entire length of both sequences. High identity match was defined as 99% identity with no length limitation.

[4] In 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 ≥98% identity and alignment over ≥80% of the contig length. SwissProt metazoan homologs were aligned to the assembly with Exonerate. Criteria used were ≥70% identity and alignment ≥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 Sites[1]	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	Hypothalmus	female, 8.5 months old[4]	68,736	934
LB012	Heart ventrical	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 Cases[1]	**Non-shared Cases**[2.]
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

[1] Exon skipping occurring in human and mouse.

[2] Exon 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

[1] 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.

[2] Probabilities 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

[1] 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.

[2] Probabilities 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	chitobiase, di-N-acetyl-	1.04	33.05	0.001***	
GLEAN_10454	ENSG00000205081	ENSMUSG00000071735	CXorf30	chromosome X open reading frame 30	1.17	19.75	0.001***	
GLEAN_07078		ENSMUSG00000041773	Enc1	ectodermal-neural cortex 1	37.95	62.25	0.001***	
GLEAN_25881	ENSG00000126860		EVI2A	ecotropic viral integration site 2A	4.24	17.59	0.001***	
GLEAN_11540	ENSG00000170231	ENSMUSG00000020405	FABP6	fatty acid binding protein 6, ileal (gastrotropin)	1.07	14.05	0.001***	
GLEAN_00966	ENSG00000187398	ENSMUSG00000063297	LUZP2	leucine zipper protein 2	1.07	11.90	0.001***	
GLEAN_02457					1.73	14.88	0.001***	
GLEAN_22866		ENSMUSG00000072852	2310040G07Rik	RIKEN cDNA 2310040G07 gene	1.48	8.01	0.01**	
GLEAN_25210	ENSG00000156384	ENSMUSG00000025066	C10orf78	chromosome 10 open reading frame 78	1.05	7.77	0.01**	
GLEAN_18851	ENSG00000176714	ENSMUSG00000029138	CCDC121	coiled-coil domain containing 121	1.04	7.52	0.01**	
GLEAN_13996	ENSG00000104408	ENSMUSG00000022336	EIF3E	eukaryotic translation initiation factor 3, subunit E	999	7.00	0.01**	
GLEAN_20989	ENSG00000154035	ENSMUSG00000018931	IDBG-35547	transcript expressed during hematopoiesis 2	999	8.21	0.01**	
GLEAN_07910		ENSMUSG00000022971	Ifnar2	interferon (alpha and beta) receptor 2	1.51	7.13	0.01**	
GLEAN_07098	ENSG00000164136	ENSMUSG00000031712	IL15	interleukin 15	1.79	8.65	0.01**	
GLEAN_16592	ENSG00000107719	ENSMUSG00000020092	KIAA1274	KIAA1274	1.54	7.61	0.01**	
GLEAN_25910	ENSG00000162493	ENSMUSG00000028583	PDPN	podoplanin	1.44	7.87	0.01**	
GLEAN_00418		ENSMUSG00000022683	Pla2g10	phospholipase A2, group X	999	7.35	0.01**	
GLEAN_12455	ENSG00000006757		PNPLA4	patatin-like phospholipase domain containing 4	8.16	6.67	0.01**	
GLEAN_01453	ENSG00000169228	ENSMUSG00000034789	RAB24	RAB24, member RAS oncogene family	999	7.87	0.01**	
GLEAN_07723	ENSG00000167088	ENSMUSG00000002477	SNRPD1	small nuclear ribonucleoprotein D1 polypeptide 16kDa	999	9.29	0.01**	
GLEAN_21922	ENSG00000178826		TMEM139	transmembrane protein 139	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	Lyl1	lymphoblastomic leukemia	1.69	5.90	0.05*	
GLEAN_09957	ENSG00000168404	ENSMUSG00000012519	MLKL	mixed lineage kinase domain-like	1.30	5.09	0.05*	
GLEAN_10081	ENSG00000132329		RAMP1	receptor (G protein-coupled) activity modifying protein 1	999	4.05	0.05*	
GLEAN_10642	ENSG00000125743		SNRPD2	small nuclear ribonucleoprotein D2 polypeptide 16.5kDa	3.15	4.70	0.05*	
GLEAN_19853	ENSG00000124731	ENSMUSG00000042265	TREM1	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	

				27 (putative)				
GLEAN_21713	ENSG00000203666		EFCAB2	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	gastrokine 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	
GLEAN_23729		ENSMUSG00000049093	Il23r	interleukin 23 receptor	3.04	0.76	ns	0.001***
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	Ubxd6	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	CattleCyc[1]/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.

[1] Amalgamated 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 models[1]

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

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

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

Gene Identifier (source species)	Calculated pI[1]	No. of Adaptive Sequence Changes for Stomach Function[2]	Most Prevalent Sites of Expression[3]
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)

[1] The 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.

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

[3] Expression 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* (In Press).

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* **Chapter 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* **Chapter 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. Vejnar, 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*, (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* (In Press).

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* (In Press).

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. Rijnkels: USDA Agricultural Research Service Cooperative Agreement 6250-51000-048; 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 Hogues, 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.