

# **Title: Adaptive divergence in the bovine genome**

William Barendse<sup>a,b</sup>, Sean McWilliam<sup>a</sup>, Rowan J. Bunch<sup>a</sup> and Blair E. Harrison<sup>a</sup>

<sup>a</sup>Commonwealth Scientific and Industrial Research Organization, Queensland Bioscience Precinct, 306 Carmody Road, St. Lucia, Queensland 4067, Australia

<sup>b</sup>School of Veterinary Science, University of Queensland, Gatton, Queensland, 4343, Australia

Correspondence:

Dr. William Barendse

Email: [Bill.Barendse@csiro.au](mailto:Bill.Barendse@csiro.au)

Telephone: +61 7 3214 2444

Address: CSIRO Australia

306 Carmody Road

St. Lucia, QLD 4067

Australia

## **Abstract:**

Cattle diverged during the Pleistocene into two subspecies, one in temperate and one in tropical environments. Here we have used next generation sequencing of the indicine subspecies of cattle and compared it to the taurine subspecies. Although 23.8 million single nucleotide polymorphisms (SNP) were found, the number of fixed amino acid substitutions between the taurine and indicine subspecies was low and consistent with the Haldane predictions for adaptive selection rather than with Neutral Theory. We noted 33 regions of enhanced divergence of nonsynonymous SNP between the subspecies, which included an increased rate of deleterious variants. Signals of positive selection were found for genes associated with immunity, including the Bovine Major Histocompatibility Complex, which also showed an increased rate of deleterious amino acid variants. The genes important in sensing the environment, especially the olfactory system, showed a network wide signal of positive selection.

Key Words: adaptive, genome, evolution, cattle

## Introduction

Cattle are a rare example of adaptive divergence into temperate and tropical environments in which the subspecies show hybrid vigour on being crossed. Genome wide analyses of evolution in cattle have focussed on the expansion of gene families or on evidence of selection signatures using population genetic evidence based on genotype or haplotype frequencies, mainly in the taurine subspecies. The initial sequencing of cattle found gene families that were expanded in cattle lineages, identifying genes associated with the immune system in particular that were increased in number compared to other mammalian species (Elsik et al. 2009). In addition, analyses of genotype or haplotype frequencies have been used to identify signatures of selection and in some cases have linked these to the known locations of quantitative trait loci (Barendse et al. 2009; Flori et al. 2009; Gibbs et al. 2009; Hayes et al. 2009). Analyses of nonsynonymous and synonymous DNA variants have been performed for a limited number of amplicons for genes expected to be involved in animal production (MacEachern et al. 2009b).

The description of the genomic variation and structure of the *indicus* subspecies lags behind that of the *taurus* subspecies of cattle. The taurine subspecies is of temperate origin while the indicine subspecies is of tropical origin. Individuals of several of the breeds of taurine cattle, including Hereford, Fleckvieh, Holstein, Black Angus, and rare Japanese breeds have been sequenced (Eck et al. 2009; Elsik et al. 2009; Kawahara-Miki et al. 2011; Stothard et al. 2011; Larkin et al. 2012; Tsuda et al. 2013). There has been targeted resequencing of hundreds of individuals to investigate breed diversity for the Holstein, Fleckvieh and other breeds (Jansen et al. 2013; Daetwyler et al. 2014). These genome sequencing studies have documented variation, not only at the single nucleotide polymorphism (SNP) level, but also structural variation including copy number variation (CNV) and major structural polymorphisms (Bickhart et al. 2012). For the indicine subspecies one Brahman, one Nelore, and a handful of Gir animals have been sequenced (Barris et al. 2012; Canavez et al. 2012; Liao et al. 2013) and there is currently no published *de novo* assembly of an indicine genome.

To further our understanding of the indicine subspecies, we have resequenced 32 individuals of the Brahman breed in Australia. Animals were chosen to be as unrelated as possible and were taken to represent a wide sample of the breed in Australia, which originated in the 19<sup>th</sup> century. These individuals were compared to one another and to sequences from the taurine subspecies. Our aim was to describe the genome wide patterns in nonsynonymous and synonymous mutations to obtain evidence of adaptive selection and divergence between the subspecies.

## Results

We identified 23.8 million single nucleotide polymorphisms (SNPs) different to the Hereford reference sequence across the 32 bulls with an average yield of 5.9 million SNP per individual (Table S3) after collecting 670 Gb of genomic sequence. Of these, 0.62% were coding sequence SNP (cSNP) and 0.25% were nonsynonymous SNP (nsSNP) (Table 1). When normalized for the number of SNP identified in each bull, the proportion of SNP of each type was relatively similar between individuals. However, there was seven times greater variation in the proportion of nsSNP between individuals than for intergenic SNP

(Table 1). The degree of similarity between animals and amounts of sequence obtained per individual are contained in the Supplementary Information and Tables S1 and S2. The FASTQ file for each bull was submitted to the 1000 Bulls database (1000Bulls.org) and the VCF file for each bull was submitted to dbSNP under the handle WBARENDSE (BRA2 – SNP and BRA2 – indel) and BioSample object IDs SAMN03166172 to SAMN03166203.

The sequencing yielded on average approximately 560,000 small indels per individual (Table S3), of which 0.028% were nonsynonymous coding indels, which was approximately 10 times less than nsSNP (Table 1). The proportion of intergenic indels per individual and the variability between individuals was similar to that found for SNPs. Furthermore, the variability for nonsynonymous indels as a proportion of the total was greater than the corresponding variability for nsSNPs, indicating that individuals differed more substantially from each other in coding sequence indels than in cSNPs. Indels in coding sequence are more likely than SNPs to represent mutations that have strong effects on protein structure and processing.

To characterise the SNP discovery distribution, we plotted the number of Brahman bulls that shared one or more SNP that are different to the cattle reference sequence. Of the 23.8 million SNP, 6.33 million (26.5%) showed either one or two alleles but occurred only in a single animal (Fig 1A). At the other extreme, 12,231 (0.05%) of the SNP were common to at least 27 of the sires, representing SNP that were at high frequency in Brahman animals compared to the Hereford reference sequence. The distribution in Fig 1A appears to reach an asymptote. To determine the decay of the curve we plotted the log number of SNP by log number of bulls (Fig 1B) and found that it was not linear across the entire range, showing that the asymptote decays more quickly than expected. This plot shows a greater deficit of numbers of SNP common to all sires than would be expected given the initial decline in sharing of alleles. To determine whether this might be due to low levels of sequence coverage, the data from the 21 bulls with the most sequence were plotted (Fig S2). This showed much greater linearity of the log-log plot across the range, and in combination with the plot of the rate of change in SNPs shared between bulls (Fig S3) suggests that the accelerated decay is partly due to low genome coverage for some bulls. This lack of similarity would also have a contribution from introgression of some taurine alleles into Brahman cattle (Sanders 1980), which would reduce the number of fixed substitutions between taurine and indicine animals. Extrapolation of the log-log plot in its linear phase suggests approximately  $10^5$  fixed differences between taurine and indicine cattle across the genome. The low level of fixed differences between taurine and indicine cattle is unlikely to be due to lack of discovery of SNPs, because not only were 23.8 million SNPs identified, the method of comparison is insensitive to low copy number coverage of a particular polymorphic site, and most importantly, each additional bull added progressively fewer new SNPs and an asymptote was reached. Since it is the rare SNPs that are discovered last, it is unlikely that there is a large number of fixed SNP differences between taurine and indicine breeds that is yet to be discovered.

Surprisingly, there were more small indels common to all bulls than SNPs common to all bulls (Fig 1B), even though there were fewer indels in total and fewer indels that were found only in a single bull or common to a small number of bulls. The cross-over point was at 23 bulls. There were more than 10 times as many small indels common to all 32

bulls than shared SNPs. The distribution for indels was linear over a larger range of bulls than SNPs, and the decline in number of common small indels was rapid only at the end of the distribution. However, contrary to this pattern of more fixed small indels than SNP, there were no indels in coding sequence that were common to all 32 bulls, and only 5 such indels that were found in more than 27 of the bulls.

Neutral theory predicts that neutral and near neutral mutations, whether slightly advantageous or deleterious, have similar mean times to fixation or extinction (Kimura and Ohta 1971). Therefore, if most nsSNP are neutral or near neutral they should have effectively the same distribution as synonymous SNPs (sSNP). However, we found that there were substantially fewer nsSNP than sSNP in the middle frequency range than at the extremes of the distribution (Fig 1C & D). The log-log plot (Fig 1D) clearly shows that there are more nsSNPs possessed by all 32 bulls than sSNPs even though there are fewer nsSNPs in total. Furthermore, at the other extreme, of SNPs found in only a single individual, the nsSNPs almost equalled the sSNPs. This suggests that a substantial number of the nsSNPs are not neutral or near neutral in their effects. Extrapolation from the linear part of the log-log plot suggests that the number of fixed nonsynonymous substitutions between taurine and indicine cattle is of the order of 900 to 1,000 (Fig 1D, Fig S2, Fig S3). The total number of nsSNPs, irrespective of the number of bulls in which they occur, was 79,309.

To ascertain whether SNP calling, especially the ratio of nsSNP to sSNP, was affected by the draft status of the bovine genome or stringency of SNP calling, we took several approaches to check for errors including examining the location of copy number variants, transcript length and quality threshold for calling of SNPs (Supplementary Information). None of these showed effects that would materially influence the ratio of nsSNP to sSNP in this data set, although we found CNV to have a higher ratio of nsSNP to sSNP than the rest of the genome. We did find that the number of SNPs was strongly correlated to transcript length (Fig 2A, Fig S1), and most genes with large numbers of SNPs have long coding sequences. Furthermore, the number of SNPs per bp was low except for some genes with very short transcript lengths, especially those in the range 50 to 200 bp (Fig 2B). Importantly, we found 33 regions with striking excesses of nsSNP to sSNP (Fig S4, Table S5), and one of these regions included the Bovine Major Histocompatibility Complex (MHC) (Fig S5).

Positive selection is often termed relaxed selection, implying that genes with high mutation rates are more likely to show high levels of nsSNPs. Here we show (Fig 3A) that high numbers of SNP per coding sequence base pair are more likely to be found in values of  $-0.5\log(\omega)$  around zero. A contrast between genes under positive and negative selection (Table 2) shows that although there were significantly ( $G_{adj} = 736.3$ , 2df,  $P < 0.00001$ ) more genes under negative selection with low rates of SNP per bp, there were substantial numbers of genes under positive selection that had a low rate of SNP per bp. The average  $-0.5\log(\omega)$  for genes with numbers of SNP per bp  $\geq 0.01$  (Table S6) was markedly less negative than that for genes with numbers of SNP per bp  $< 0.01$ . Although genes with positive selection were more likely to have higher numbers of SNP, there were a large number of genes with evidence of positive selection that had low numbers of SNP per bp and a large number of genes with evidence of negative selection that had a large number of SNP per bp.

Many of the very short genes have large numbers of SNP per bp (Fig 2B) so to examine the impact of numbers of SNP per transcript on estimates of positive selection, the length of transcript was plotted against evidence for positive selection (Fig 3B). Although there is an obvious shift towards shorter transcript lengths for genes with positive selection, this is not driven by transcripts with very high numbers of SNP per bp, such as those with lengths below 500 bp (Fig 3A, Table 3). Removing the transcripts below 500 bp changes the relative length of transcripts under negative selection compared to transcripts under positive selection from 1.74 to 1.65. Shorter transcript length of genes with positive selection appears to be a general pattern in the data and not due to artefacts.

To determine the impact of nsSNP, we characterised the distribution of SIFT predictions. Firstly, fixed nonsynonymous substitutions, and nsSNPs near fixation for opposite alleles, showed SIFT predictions that were very much less damaging than nsSNPs in general. For example, comparing nsSNPs found in 28 or more of the bulls compared to the rest, the average SIFT score = 0.66 S.E.M = 0.017 for those near fixation while the average SIFT score = 0.41 S.E.M. = 0.001 for the rest, yielding  $t = 37.03$ ,  $P < 0.00001$ . Using the SIFT codes *deleterious* and *tolerated*, the same comparison yielded an odds ratio of 4.3 for *tolerated* substitutions for those near fixation and  $G_{adj} = 91.17$ , 1df,  $P < 0.00001$ . Secondly, splitting the data into those transcripts with number of SNP per bp  $> 0.01$  compared to those with SNP per bp  $\leq 0.01$ , irrespective of the number of bulls the SNP was found in, the average SIFT score = 0.40 S.E.M. = 0.002 for SNP per bp  $> 0.01$  while the average SIFT score = 0.42 S.E.M. = 0.003 for SNP per bp  $\leq 0.01$ , yielding  $t = 5.89$ ,  $P < 0.00001$ . Using the SIFT codes *deleterious* and *tolerated*, the same comparison yielded an odds ratio of 1.4 for *tolerated* substitutions for those with SNP per bp  $\leq 0.01$  and  $G_{adj} = 222.8$ , 1df,  $P < 0.00001$ . This comparatively small difference is highly significant due to the very large number of elements. Thirdly, the 33 regions with extremely large numbers of nsSNP compared to sSNP (Fig S4A, Table S5) had genes with 970 deleterious to 3,076 tolerated nsSNP compared to 16,982 deleterious and 58,281 tolerated nsSNP for the rest of the genome, an odds ratio of 1.15 ( $G_{adj} = 12.95$ , 1df,  $P = 0.0003$ ). Finally, the distribution of deleterious nsSNPs in transcripts with high SNP per bp was compared to the Poisson distribution and there were substantially more transcripts with either very high numbers or very low numbers of deleterious mutations than expected (Table S7), resulting in a goodness of fit chisquare  $\chi^2_{10} = 7242.2$ ,  $P < 0.00001$ . This striking divergence from random was found whether transcripts were segregated on number of tolerated nsSNPs or not (Table S8). A high excess of deleterious nsSNP are consistent with relaxed selection but an excess of zero deleterious nsSNP but high numbers of tolerated nsSNP are implausible under relaxed selection. This suggests other modes of selection, such as balancing selection.

One may predict that genes that are well known, conserved and annotated between species might be subject to more purifying selection. In ENSEMBL, all genes with transcripts have an ENSEMBL name but in some cases these transcripts cannot be linked to known gene identifiers. Of genes with positive selection, 780 had conserved gene names but 816 did not, whereas of genes with negative selection 8,092 had conserved gene names but 1,454 did not. This gives an odds ratio of 5.31 and a  $G_{adj} = 807.7$ , 1df,  $P < 0.00001$ .



To determine whether specific pathways or sets of genes were under selection, genes under positive selection were compared to those under negative selection. Using GO terms and GOrilla (Eden et al. 2009), we found that genes in the immune system and in detection of chemical stimulus involved in sensory perception were significantly represented among the genes under positive selection (Supplementary Information Data File 1). Plots in the region of the Bovine MHC showed elevated numbers of nsSNPs (Fig S5). An examination of the numbers of tolerated and deleterious nsSNP for MHC coding sequences showed much higher than average tolerated (27.75 vs 9.43,  $t = 5.70$   $P < 0.00001$ ) nsSNPs and slightly higher numbers of deleterious nsSNPs (7.25 vs 3.19,  $t = 2.66$   $P = 0.008$ ) than other transcripts with large numbers of SNPs per bp.

There was no clear evidence of differences in numbers of genes with excess positive or negative selection at the level of whole chromosomes (Fig 4). One possibility could be the X chromosome, due to initial formation of the Brahman breed using some taurine dams in the 19<sup>th</sup> century (Sanders 1980), followed by extensive backcrossing using indicine bulls and then the addition of purebred indicine cattle of both sexes to generate an indicine breed. This would predict that the X chromosome would show fewer genes under positive selection, because the comparison is to the taurine reference sequence. On the contrary, we found that the X chromosome had some of the strongest evidence for genes under positive selection. The average  $-0.5\log(\omega)$  across all chromosomes = -0.287 (n=30), S.E.M. 0.007, range -0.354 to -0.163. The value for the X chromosome = -0.163 for n=216 genes, making it the chromosome with the highest level of positive selection, with BTA15 and BTA9 the next highest values (Table S9). It had the 7<sup>th</sup> lowest number of genes with sufficient cSNP to report values for  $-0.5\log(\omega)$ , which was inconsistent with its length and may be a signature from the use of taurine dams.

## Discussion

The analysis of whole genome sequence data from 32 bulls showed that specific classes of genes have evidence for positive selection but surprisingly small numbers of SNPs or indels were fixed between taurine and indicine cattle. We show that shorter genes are more likely to show evidence of positive selection, that signatures of both relaxed and balancing selection were found, and that genes that are under positive selection are approximately 5 times more likely to be poorly annotated than genes under negative selection. We found that genes in the immune system and systems that sense the environment are more likely to show evidence of positive selection. As these subspecies were adapted to temperate versus tropical environments before domestication, these differences likely represent natural selection prior to domestication. Although genes under positive selection were distributed across all chromosomes, there were 33 regions with striking excesses of nsSNPs, which might be regions of focus for divergence between indicine and taurine cattle, and these regions include a slight increase in deleterious nsSNP. Finally, the interindividual variability for intergenic SNPs or indels was low compared to the interindividual variability for nsSNPs and coding sequence indels. As these latter two types of variant are more likely under selection this implies increased variability between individuals for genetic material that is under selective scrutiny. Further research might indicate ways in which this information can be used in prediction of the phenotype.

Surprisingly, only a small proportion of the differences, whether coding or noncoding, SNP or indel, showed fixed differences between Brahman (indicine) cattle and the reference (taurine) sequence. While some of the low number can be attributed to lower levels of genome coverage in some individuals or perhaps to past introgression from taurine germ plasm, if one took as the estimate the lowest number reached in the linear section of the log-log plot this would still only represent  $10^5$  fixed differences, much less than 1% of the genetic variants identified in these cattle.

The number of fixed amino acid substitutions between these two subspecies is small and is consistent with adaptive evolution. Using the extrapolations from the log-log plots as noted above, we would expect between 900 and 1,000 fixed nonsynonymous substitutions. The divergence between the taurine and indicine subspecies occurred 0.6 myr to 0.8 myr ago (MacHugh et al. 1997), the effective population size  $N_e$  prior to domestication is estimated to have ranged between 10,000 to 90,000 (Gibbs et al. 2009; MacEachern et al. 2009a), and the generation interval is approximately 3 years. Using Haldane's estimate of 1 substitution per 300 generations for horotelic evolution (Haldane 1957), this would yield approximately 600 to 900 fixed substitutions. Kimura estimated the rate of amino acid substitutions based on neutral expectations (Kimura 1968; Kimura and Ohta 1971) and, updating his starting values for values now known for genome sizes of mammals (2.6 Gb) and the size of the combined coding sequence (~1%), this would result in approximately 1 neutral or near neutral amino acid substitution per 10 years. This would suggest 60,000 to 80,000 nonsynonymous substitutions across the genome. It would require adding together all nsSNPs, including those found only in a single bull, to be consistent with the Kimura prediction. Our results on fixed amino acid substitutions are therefore consistent with the Haldane prediction for genes under standard adaptive selective constraints. Interestingly, these substitutions had four times lower rates of deleterious substitutions than the amino acid substitutions in general. As coding sequences represent around 1% of the genome, extrapolated up to all SNP, the  $10^5$  fixed differences between taurine and indicine subspecies are therefore more consistent with polymorphisms under adaptive selection, and would be consistent with large parts of the non-coding sequence showing function, as described by the ENCODE project (Dunham et al. 2012).

Our results show that signals of positive selection are found with shorter genes. If mutations in only part of a gene are under positive selection at any one time, and this is a reasonable expectation given that random amino acid substitutions are more likely to be deleterious, then a measure of overall selection on a gene will be biased to detect negative selection the longer the transcript length of the gene. For very long genes evidence for adaptive selection would be drowned out by the evolutionary constraint on the rest of the molecule.

Some of the evidence for positive selection is consistent with balancing selection although some of it is consistent with relaxed selection. A large number of nsSNP in a transcript could be evidence of relaxed selection, i.e., where genes are not seen by selection on phenotypes and so they accumulate mutations, irrespective of how damaging those might appear to be. Under balancing selection (Wallace 1970), one would expect that many nsSNPs could accumulate on coding sequences that are maintained in polymorphic condition for tens of thousands of generations, but in this case the number of lethal alleles would be constrained. In both cases a high accumulation of nsSNP would lead to an

estimate of positive selection even though the nsSNP in some transcripts would have no adaptive significance. Analysis of severity among mutations suggests that there is a 40% increase in deleterious nsSNPs in transcripts with high SNP per bp. Nevertheless, there was both a striking excess of transcripts with high numbers of nsSNPs in some genes where none were identified as deleterious, as well as a striking excess of transcripts in other genes where a large number were identified as deleterious. The former would be consistent with balancing selection and the latter with relaxed selection. One example of balancing selection is the MHC (Bonneaud et al. 2006; Evans and Neff 2009) for which we have evidence of positive selection. In cattle, this complex contains nsSNPs with deleterious effects, which confirms that not all examples of balancing selection will be for alleles with small differences of effect.

## Materials and methods

### *Samples*

Animal Care and Use Committee approval was not obtained for this study because no new animals were handled in this experiment. The experiment was performed on DNA samples that had been collected previously. DNA samples and duplicates were available for 76 Brahman sires associated with the Cooperative Research Centres for Beef Quality and for Beef Genetic Technologies (Upton et al. 2001; Barwick et al. 2009). These samples were either of industry sires or sires from the Belmont Research Station (Burrow 1998). DNA was extracted from semen or blood (Barendse et al. 2007; Bolormaa et al. 2011). Each sample was genotyped using an Illumina Bovine HD SNP Array (Illumina, Carlsbad, CA) and quality control was performed as previously described (Bolormaa et al. 2011). The genotypes were compared using a genome relationship matrix (Yang et al. 2010) and the 32 least related individuals were chosen for genome sequencing.

### *Next generation sequencing*

DNA for genome sequencing was checked for quality control using agarose gel electrophoresis, followed by quantification, DNA fragmentation, and size selection using a Nanodrop, a Covaris sonicator, and Qubit-HS DNA kit. DNA for genome sequencing was extracted from whole blood or semen, quantification was by both Nanodrop ND-1000 (Wilmington, DE, USA) and by Qubit dsDNA br Assay, cat no Q32853 by Molecular Probes, supplied by Thermo Scientific, USA. DNA quality was determined by running a subsample of DNA on a 1% 1 X TAE agarose gel to check for DNA smears due to degradation and the 260:280 UV absorbance ratio result on the Nano-drop was required to be between 1.8 and 2.0. Each sample was normalized to 20ng/ul and 55 ul was fragmented to an average insert length of 300 to 400 bp using a Covaris S220 ultra-sonicator (Woburn, MA, USA). DNA libraries for sequencing were prepared from at least 1.0 microgram of genomic DNA of each individual using the Illumina TruSeq DNA Sample Prep Kit v2-Set A kit following the manufacturer's instructions (Illumina Australia P/L, Scoresby, Vic). A TruSeq PE Cluster Kit v3 was used to generate sequencing clusters using the Illumina cBot on Flow Cell v3. 100 bp paired end reads were obtained using the Illumina TruSeq SBS Kit v3 - HS kit following the manufacturer's instructions on the SQ module of an Illumina HiScan instrument. Individuals were multiplexed at 8 samples per lane.



## Analysis

Initial quality control consisted of removing sequences that failed chastity filtering using Illumina Casava 1.8. Sequences passing chastity filtering were quality trimmed to remove bases below phred Q20. Using the fastq trimmer module of the Fastx toolkit ([http://hannonlab.cshl.edu/fastx\\_toolkit/commandline.html](http://hannonlab.cshl.edu/fastx_toolkit/commandline.html)) sequences less than 50 bp in length following trimming were removed from further analysis. Remaining paired end sequences were aligned to the UMD 3.1 assembly of the cattle genome (Genbank ID GCA\_000003055.3), which had been assembled using methods as described (Zimin et al. 2009). The BWA package (Li and Durbin 2009) was used to align the sequence using default values. Alignment included checking that the aligned distance between members of a pair was consistent with the fragment sizes from which libraries were constructed. Sequences were excluded that matched to more than 1 region of the genome. Single nucleotide polymorphisms (SNP) were called using previously described methods and criteria (Barris et al. 2012). The genome sequence variants in the BAM format files were output in VCF format (Li et al. 2009). DNA variation in the VCF files of each individual was annotated to sequence variants using the VEP tool from ENSEMBL (McLaren et al. 2010) and SIFT scores (Kumar et al. 2009) were obtained at the same site. While presence or absence of all the SNP or indels can be ascertained accurately with relatively low sequence coverage it requires 15x coverage to accurately score all the heterozygous bases and 30x coverage to accurately score all the homozygous bases (Bentley et al. 2008; Barris et al. 2012). Therefore the allele frequency data for the majority of scored SNP could be subject to error, whereas identification of the presence of a variant in an animal was of higher certainty. The number of bulls in which a variant was found was therefore calculated noting that this would overestimate the number of fixed differences between taurine and indicine lineages. Numbers and types of each kind of variant were obtained using customized shell and perl scripts. Contingency tables were analysed using  $G_{adj}$ , the log-likelihood test with the Williams correction (Sokal and Rohlf 1981). Goodness of fit between observed numbers and expectations of the Poisson distribution was calculated (Feller 1968). Pearson correlation coefficients were calculated. Means were compared using unpaired sample t-tests.

The comparison of nonsynonymous and synonymous DNA variants has a long history (Kimura 1968), and an excess of nonsynonymous variants compared to synonymous variants has always been considered as unambiguous evidence for adaptive selection (Kimura 1977; Nielsen and Yang 1998). Most genes, however, do not show higher rates of nonsynonymous than synonymous variants, which has led to many related approaches to identify adaptive selection in genes that are under strong purifying constraint (Kimura 1980; Li et al. 1985; Goldman and Yang 1994; Nielsen and Yang 1998; Suzuki and Gojobori 1999; Yang and Bielawski 2000; Kosakovsky Pond and Frost 2005; Meyer and Wilke 2013). Specifically, these approaches correct for the rate of transition and transversion variants, some use estimated standard errors as part of significance testing while others use the likelihood ratio test. Some models incorporate information on biological differences between variants or information from the branching pattern of evolutionary relationships. It has been shown that the differences in these methods, while methodologically interesting, do not generate much difference in output, suggesting that it does not matter too much which method is used (Kosakovsky Pond and Frost 2005). Here we evaluated the ratio of sSNP to nsSNP (Kimura 1980) where  $\omega = 2k_s/k_n$  where  $k_s$  is the

value calculated from transitions and transversions for sSNP while  $k_n$  is the value calculated for nsSNP. Here, genes were included in the comparison if  $k$ , the overall statistic over all nucleotides, was at least twice as large as its standard error, which corresponded to genes with at least 4 cSNP. This method has advantages for genome wide analyses. It is conservative. It requires only information on the length of the coding sequence and the number of transitions and transversions for sSNP and for nsSNP. This information is readily available after processing of data into VCF and VEP format and does not require additional realignment of coding sequences. Values below 1 represent an excess of nsSNP. To graph this, however, given the visual disparity between values in the range 0 to 1 versus 1 to  $\infty$ , after examining the range of values,  $\omega$  was plotted as  $-0.5\log(\omega)$  and values of  $\omega = 0$  were truncated at 0.125 while values of  $\omega = \infty$  were truncated at 70.6, as these just exceed the values in the distribution of  $\omega$  over all genes. This generated a plot where positive selection is represented as values of  $-0.5\log(\omega) > 0$  and negative selection as values of  $-0.5\log(\omega) < 0$ . GO terms for genes with  $\omega \leq 1$  where compared to GO terms for genes with  $\omega > 1$  using GOrilla (Eden et al. 2009), via the website located at <http://cbl-gorilla.cs.technion.ac.il/>, which implements a Benjamini-Hochberg multiple testing procedure. As there is no specific database for *Bos taurus* or other artiodactyls on this website, even though GO terms are annotated for cattle in ENSEMBL, the comparison was performed for each of the Human, Mouse and Rat databases.

## Acknowledgements

The sequence acquisition was co-funded by the Cooperative Research Centre for Beef Genetic Technologies. We thank M.E. Goddard who supported and encouraged this work. B.P Dalrymple, J.W. Kijas and S.A. Lehnert read and commented on an earlier version of the manuscript.

## References

1. Elsik CG, Tellam RL, Worley KC, & Bovine Genome Sequencing Analysis Consortium (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. (Translated from English) *Science* 324(5926):522-528 (in English).
2. Gibbs RA, *et al.* (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. (Translated from English) *Science* 324(5926):528-532 (in English).
3. Barendse W, Harrison BE, Bunch RJ, Thomas MB, & Turner LB (2009) Genome wide signatures of positive selection: The comparison of independent samples and identification of regions associated to traits. *BMC Genomics* 10:178.
4. Hayes BJ, *et al.* (2009) A genome map of divergent artificial selection between *Bos taurus* dairy cattle and *Bos taurus* beef cattle. *Anim. Genet.* 40(2):176-184.
5. Flori L, *et al.* (2009) The genome response to artificial selection: a case study in dairy cattle. *PLoS ONE* 4(8):e6595.
6. MacEachern S, *et al.* (2009) Molecular evolution of the Bovini tribe (Bovidae, Bovinae): Is there evidence of rapid evolution or reduced selective constraint in Domestic cattle? *BMC Genomics* 10:179.
7. Eck SH, *et al.* (2009) Whole genome sequencing of a single *Bos taurus* animal for single nucleotide polymorphism discovery. (Translated from English) *Genome Biol.* 10(8):8 (in English).
8. Larkin DM, *et al.* (2012) Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *Proc. Natl. Acad. Sci. (USA)* 109(20):7693-7698.
9. Stothard P, *et al.* (2011) Whole genome resequencing of Black Angus and Holstein cattle for SNP and CNV discovery. *BMC Genomics* 12:559.
10. Kawahara-Miki R, *et al.* (2011) Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. *BMC Genomics* 12:103.
11. Tsuda K, *et al.* (2013) Abundant sequence divergence in the native Japanese cattle Mishima-Ushi (*Bos taurus*) detected using whole-genome sequencing. *Genomics* 102(4):372-378.
12. Jansen S, *et al.* (2013) Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. *BMC Genomics* 14:446.
13. Daetwyler HD, *et al.* (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet* 46(8):858-865.
14. Bickhart DM, *et al.* (2012) Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.* 22(4):778-790.
15. Barris W, *et al.* (2012) Next generation sequencing of African and Indicine cattle to identify single nucleotide polymorphisms. *Animal Production Science* 52(3):133-142.
16. Canavez FC, *et al.* (2012) Genome sequence and assembly of *Bos indicus*. *J. Hered.* 103(3):342-348.

17. Liao XP, *et al.* (2013) Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. *Genome* 56(10):592-598.
18. Sanders JO (1980) History and development of zebu cattle in the United States. (Translated from English) *J. Anim. Sci.* 50(6):1188-1200 (in English).
19. Kimura M & Ohta T (1971) *Theoretical Aspects of Population Genetics* (Princeton University Press, Princeton, New Jersey).
20. Eden E, Navon R, Steinfeld i, Lipson D, & Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48.
21. MacHugh DE, Shriver MD, Loftus RT, Cunningham P, & Bradley DG (1997) Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). *Genetics* 146(3):1071-1086.
22. MacEachern S, Hayes B, McEwan J, & Goddard M (2009) An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in domestic cattle. *BMC Genomics* 18:181.
23. Haldane JBS (1957) The cost of natural selection. *J. Genet.* 55(3):511-524.
24. Kimura M (1968) Evolutionary rate at molecular level. *Nature* 217(5129):624-626.
25. Dunham I, *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. (Translated from English) *Nature* 489(7414):57-74 (in English).
26. Wallace B (1970) *Genetic Load Its Biological and Conceptual Aspects* (Prentice Hall, Inc., Englewood Cliffs, New Jersey).
27. Bonneaud C, Chastel O, Federici P, Westerdahl H, & Sorci G (2006) Complex *Mhc*-based mate choice in a wild passerine. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.* 273:1111-1116.
28. Evans ML & Neff BD (2009) Major histocompatibility complex heterozygote advantage and widespread bacterial infections in populations of Chinook salmon (*Oncorhynchus tshawytscha*). *Mol. Ecol.* 18:4716-4729.
29. Upton W, Burrow HM, Dundon A, Robinson DL, & Farrell EB (2001) CRC breeding program design, measurements and database: methods that underpin CRC research results. *Aust. J. Exp. Agr.* 41:943-952.
30. Barwick SA, Wolcott ML, Johnston DJ, Burrow HM, & Sullivan MT (2009) Genetics of steer daily and residual feed intake in two tropical beef genotypes, and relationships among intake, body composition, growth and other post-weaning measures. (Translated from English) *Animal Production Science* 49(5-6):351-366 (in English).
31. Burrow HM (1998) The effects of inbreeding on productive and adaptive traits and temperament of tropical beef cattle. *Livest. Prod. Sci.* 55(3):227-243.
32. Barendse W, *et al.* (2007) A validated whole genome association study of efficient food conversion. *Genetics* 176(3):1893-1905.
33. Bolormaa S, *et al.* (2011) A genome-wide association study of meat and carcass traits in Australian cattle. *J. Anim. Sci.* 89:2297-2309.
34. Yang JA, *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. (Translated from English) *Nat. Genet.* 42(7):565-U131 (in English).

35. Zimin AV, *et al.* (2009) A whole-genome assembly of the domestic cow, *Bos taurus*. (Translated from English) *Genome Biol.* 10(4):10 (in English).
36. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
37. Li H, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. (Translated from English) *Bioinformatics* 25(16):2078-2079 (in English).
38. McLaren W, *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26(16):2069-2070.
39. Kumar P, Henikoff S, & Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* 4(8):1073-1082.
40. Bentley DR, *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. (Translated from English) *Nature* 456(7218):53-59 (in English).
41. Sokal RR & Rohlf FJ (1981) *Biometry* (W.H. Freeman & Co, San Francisco) Second Ed.
42. Feller W (1968) *An Introduction to Probability Theory and its Applications* (John Wiley and Sons, New York) Third Ed p 509.
43. Nielsen R & Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 Envelope gene. *Genetics* 148(3):929-936.
44. Kimura M (1977) Preponderance of synonymous changes as evidence for neutral theory of molecular evolution. *Nature* 267(5608):275-276.
45. Meyer AG & Wilke CO (2013) Integrating sequence variation and protein structure to identify sites under selection. *Mol. Biol. Evol.* 30(1):36-44.
46. Kosakovsky Pond SL & Frost SDW (2005) Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* 22(5):1208-1222.
47. Yang ZH & Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution* 15(12):496-503.
48. Suzuki Y & Gojobori T (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16(10):1315-1328.
49. Li WH, Wu CI, & Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2(2):150-174.
50. Goldman N & Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11(5):725-736.
51. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111-120.



## Figure Headings

**Figure 1.** The number of bulls in which a variant is found. **A.** SNP are represented by open circles, small indels by filled circles. **B.** A log-log plot of the data in 1A. **C.** Synonymous substitutions are represented by open circles, nonsynonymous substitutions by filled circles. **D.** A log-log plot of the data in 1C.

**Figure 2.** The number of SNP in transcripts of various lengths. **A.** Transcript length in base pairs plotted against number of coding SNP of both types. The line has intercept  $a=3.29$ , slope  $b=0.003$ ,  $r=0.49$  and  $n=17,881$ . **B.** Transcript length plotted against SNP per bp, showing excessive SNP per bp is restricted to very short transcripts.

**Figure 3.** The relationship between positive and negative selection and gene characteristics. **A** Genes with positive selection (black circles) and negative selection (gray circles) plotted against number of SNP per bp in each gene. **B.** Genes with positive selection (black circles) and negative selection (gray circles) plotted against transcript length.

**Figure 4.** Distribution of  $-0.5\log(\omega)$  in genes across the bovine genome. Positive selection is shown as brown on odd numbered chromosomes and black on even numbered chromosomes. Negative selection is shown as wheat colored on odd numbered chromosomes and gray on even numbered chromosomes.

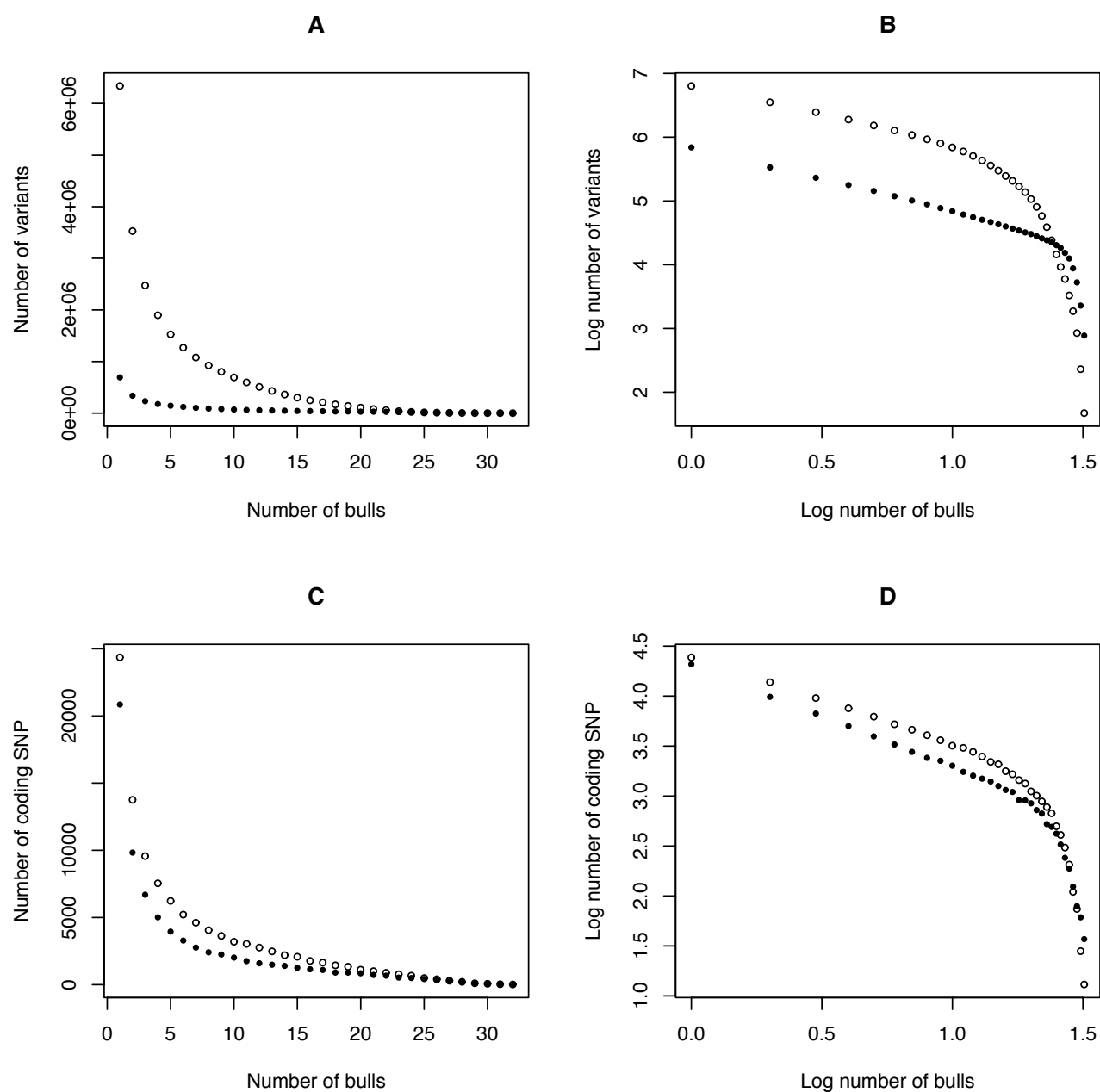


Figure 1

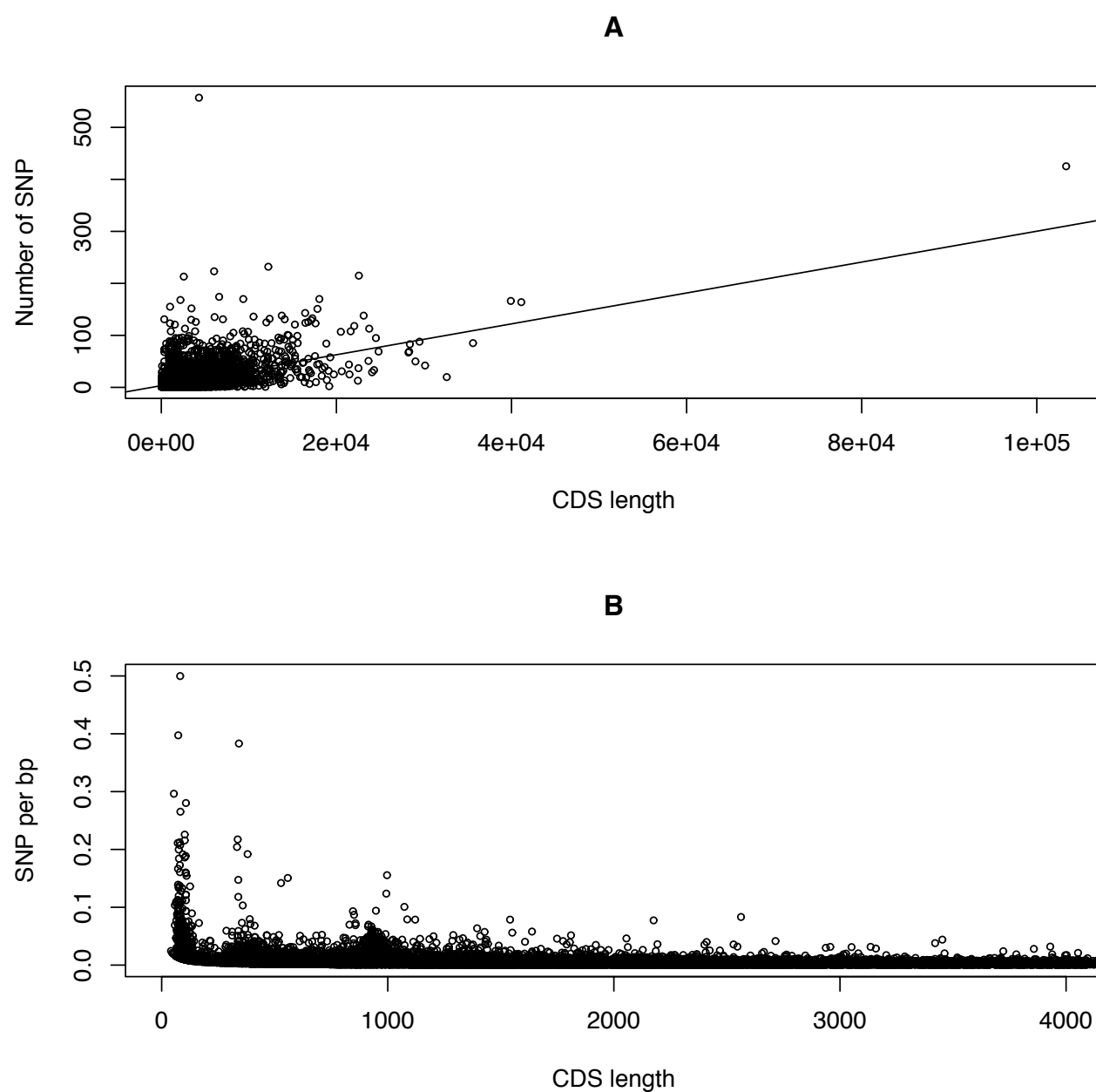


Figure 2

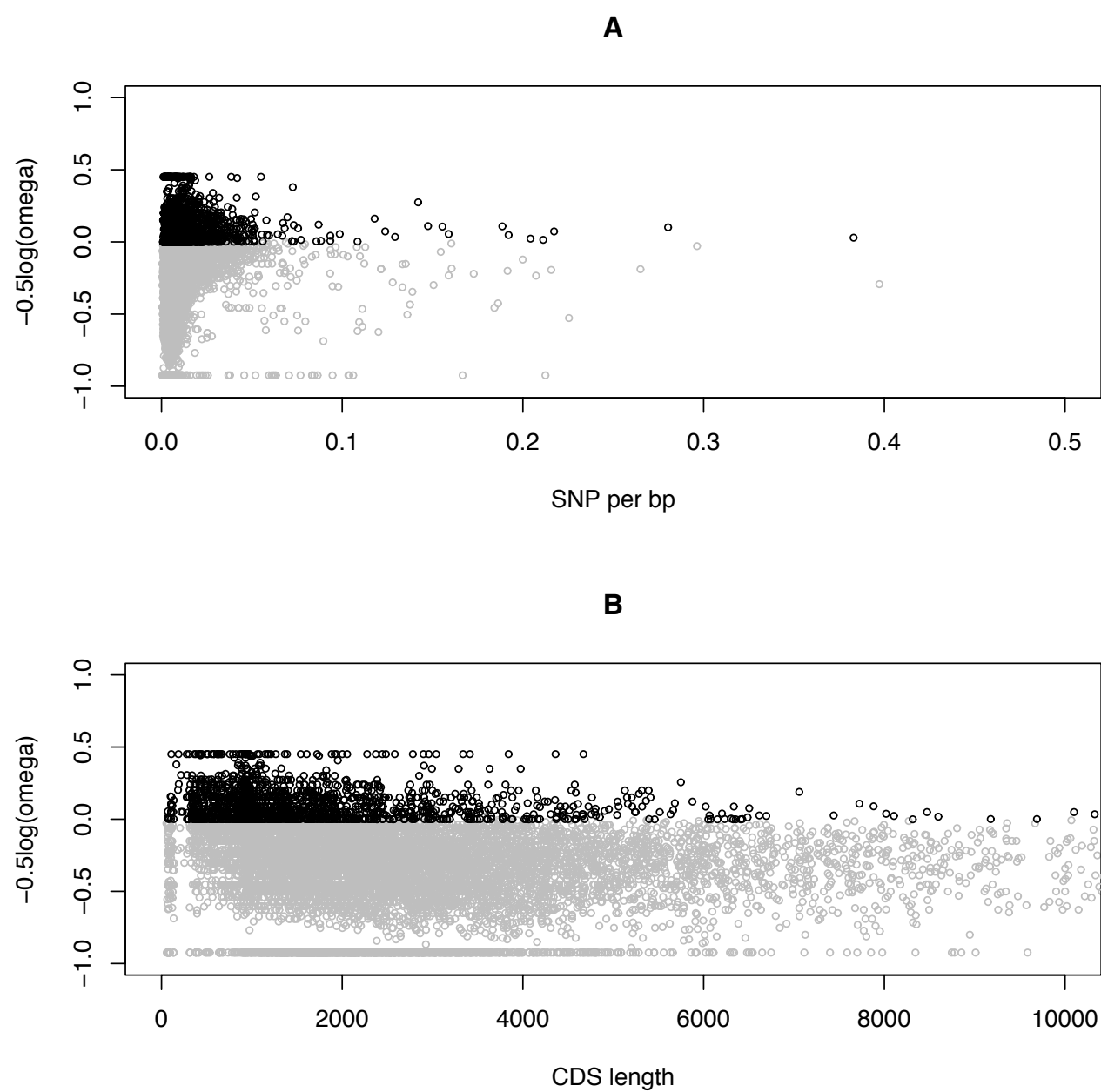


Figure 3

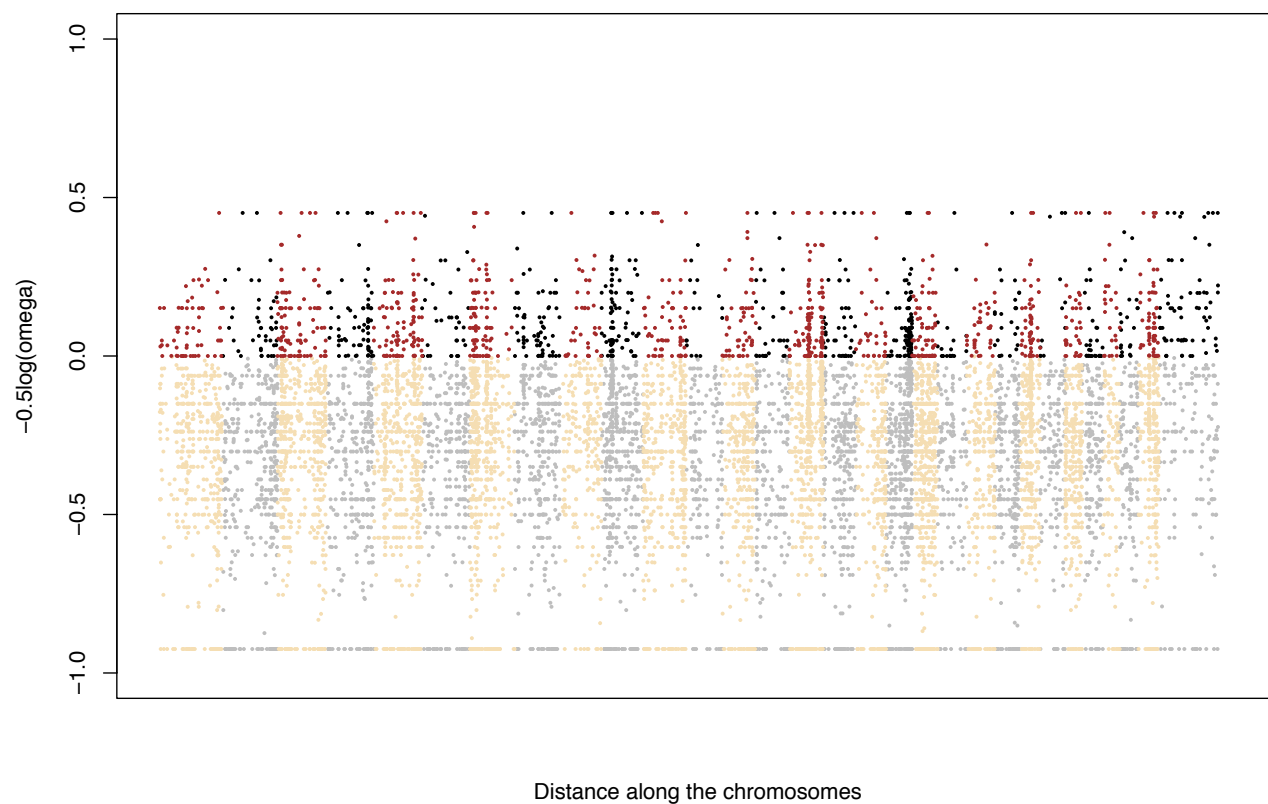


Figure 4



**Table 1. SNP and indel variability in the sample of Brahman bulls**

<b>Type</b>	<b>Mean</b>	<b>SD</b>	<b>MIN</b>	<b>MAX</b>	<b>CV</b>
<b>SNP</b>					
<b>Intergenic/All</b>	0.6628	0.0074	0.6430	0.6789	0.011
<b>NS/All</b>	0.0025	0.0002	0.0022	0.0029	0.077
<b>INDELS</b>					
<b>Intergenic/All</b>	0.6498	0.0064	0.6309	0.6610	0.010
<b>NS/All</b>	0.00028	$3.31 \times 10^{-5}$	0.00020	0.00037	0.119

Intergenic/All Intergenic SNP divided by All SNP for each animal

NS/All Non-synonymous SNP divided by All SNP for each animal

Intergenic/All Intergenic indels divided by All indels for each animal

NS/All Non-synonymous indels divided by All indels for each animal

**Table 2. The number of genes under positive and negative selection with low proportions of SNP**

<b>Type</b>	<b>SPB <math>\geq 0.01</math></b>	<b>0.01 &gt; SBP &gt; 0.001</b>	<b>SBP &lt; 0.001</b>
<b>Positive selection</b>	771	817	8
<b>Negative selection</b>	1,706	7,767	73

SBP SNP per bp

**Table 3. The average length of genes under positive and negative selection**

<b>Type</b>	<b>N</b>	<b>Transcript length (bp)</b>	<b>S.D.</b>	<b>MIN</b>	<b>MAX</b>
<b>Positive selection</b>	1,596	1,663.4	1,470.1	68	18,849
<b>Positive selection, CDS &gt; 500 bp</b>	1,442	1,804.1	1,478.2	501	18,849
<b>Negative selection</b>	9,546	2,899.9	2,769.6	54	103,346
<b>Negative selection, CDS &gt; 500 bp</b>	9,302	2,969.1	2,772.1	501	103,346