Genet. Sel. Evol. 38 (2006) 321–341

DOI: 10.1051/gse:2006007

Original article

Testing the neutral theory of molecular evolution using genomic data: a comparison of the human and bovine transcriptome

Sean MacEachern[a,b*], John McEwan[c], Andrew Mather[a], Alan McCulloch[c], Paul Sunnucks[b], Mike Goddard[a,d]

[a] Primary Industries Research Victoria, Animal Genetics and Genomics, Attwood VIC 3049, Australia
[b] Latrobe University, Department of Genetics, Bundoora VIC 3086, Australia
[c] AgResearch, Department of Genetics, Private Bag 50034, Mosgiel, New Zealand
[d] Melbourne University, School of Agriculture and Food Systems, Melbourne VIC 3000, Australia

(Received 31 August 2005; accepted 8 December 2005)

Abstract – Despite growing evidence of rapid evolution in protein coding genes, the contribution of positive selection to intra- and interspecific differences in protein coding regions of the genome is unclear. We attempted to see if genes coding for secreted proteins and genes with narrow expression, specifically those preferentially expressed in the mammary gland, have diverged at a faster rate between domestic cattle (*Bos taurus*) and humans (*Homo sapiens*) than other genes and whether positive selection is responsible. Using a large data set, we identified groups of genes based on secretion and expression patterns and compared them for the rate of nonsynonymous (*dN*) and synonymous (*dS)* substitutions per site and the number of radical (*Dr*) and conservative (*Dc*) amino acid substitutions. We found evidence of rapid evolution in genes with narrow expression, especially for those expressed in the liver and mammary gland and for genes coding for secreted proteins. We compared common human polymorphism data with human-cattle divergence and found that genes with high evolutionary rates in human-cattle divergence also had a large number of common human polymorphisms. This argues against positive selection causing rapid divergence in these groups of genes. In most cases *dN*/*dS* ratios were lower in human-cattle divergence than in common human polymorphism presumably due to differences in the effectiveness of purifying selection between long-term divergence and short-term polymorphism.

adaptive evolution / *Bos taurus* / *Homo sapiens* / mammary gland / tissue specific genes

* Corresponding author: Sean.Maceachern@dpi.vic.gov.au

1. INTRODUCTION

Adaptive evolution requires heritable phenotypic differences caused by DNA sequence variation. A major challenge in genomics is to identify variation at the DNA level that generates intra- and interspecific differences in phenotype. However, because species differ at so many sites in the genome and because most of these differences have little or no effect on phenotype, it has been difficult to identify the DNA sequence variation responsible for adaptive evolution [3].

The neutral theory of evolution [14] predicts that the majority of differences observed in the DNA sequence within and between species occurs due to random mutation and genetic drift rather than positive selection. In the simplest, completely neutral version of this theory, the rate of divergence between species would be the same at sites leading to a nonsynonymous amino acid change as those that are synonymous. That is the ratio $dN/dS = 1$, where $dS\,(dN)$ is the proportion of (non-) synonymous sites that differ between two species. If dN/dS were found to be significantly >1, this would imply that positive selection had driven the divergence between, at least, some of the sites. For the same reason, the ratio of radical (Dr) to conservative (Dc) amino acid substitutions is expected to equal one. However, neutral theory also acknowledges that biologically important sites in proteins are under strong purifying selection and therefore evolve relatively slowly. Consequently, the evolutionary ratios dN/dS and Dr/Dc are expected to be <1, even if some sites are evolving under positive selection. Thus, it is rare to find genes with Dr/Dc or $dN/dS > 1$ and this is not a powerful method to detect genes whose divergence is a result of positive selection.

Comparing divergence between species to polymorphisms within species has been suggested as a more powerful way to detect positive selection [17]. If some mutations are neutral and others are inevitably eliminated by selection, then dN/dS will be the same for divergence between species and polymorphism within species, even though both are less than 1.0. By removing polymorphisms with one low frequency allele, which are typically mildly deleterious, inflated dN/dS among polymorphisms are avoided [10]. Thus, higher evolutionary ratios (dN/dS and Dr/Dc) in divergence than in common polymorphism suggests that positive selection has driven some of the divergence. A limitation of this approach is that for individual genes there may be too few known polymorphisms to estimate dN:dS or Dr:Dc ratios with sufficient accuracy. Therefore, the methods that identify functionally related groups of genes [7, 8, 23] will have more power to find evidence of adaptive evolution in patterns of divergence and polymorphism.

Recently, higher evolutionary rates have been reported in genes that are expressed in a narrow range of tissues than those that are widely expressed [8]. This finding could be explained by two different hypotheses. Firstly, a mutation in a ubiquitously expressed gene will affect a large number of tissues and therefore is more likely to be deleterious than a mutation in a tissue specific gene (the negative selection hypothesis). Alternatively, tissue specific genes might be more able to respond to changes in selection pressure (the positive selection hypothesis). Furthermore, it has been reported that genes with secreted products evolve faster than their nonsecreted counterparts [23]. Again, this could be explained by positive selection (*e.g.* secreted genes associated with the immune system evolving in response to the evolution of pathogens in a form of genetic arms race) or by negative selection (*e.g.* nonsecreted proteins being more constrained in amino acid sequence than secreted proteins). In this paper, we used the comparison of divergence between species to polymorphism within species to distinguish between these two hypotheses.

To date, large-scale comparisons of DNA sequence divergence and polymorphism have been restricted to dN/dS studies for a limited number of species with sufficient sequence data. Domestic cattle provide an interesting addition to this range of species because there is an extensive body of sequence data, large phenotypic databases, known pedigrees and we have some knowledge of the selection pressures before and after domestication. For instance, calves are much more developed at birth and grow much faster than human babies and, not surprisingly, cows produce a larger amount of milk with a higher protein concentration than do humans. These differences have been exaggerated following domestication by strong selection for increased milk production in the cow [4]. Therefore, we hypothesise that genes expressed in the mammary gland have diverged faster between humans and cattle than randomly chosen genes. In this study, we examined if evolutionary rates in the divergence of cattle and humans varied between genes expressed in different tissues, genes with different secretory motifs (anchor, secreted and nonsecreted) and genes with wide *versus* narrow expression. Secondly, we applied the McDonald-Kreitman test [17] to determine whether these differences in evolutionary rate are due to positive or negative (purifying) selection.

2. MATERIALS AND METHODS

2.1. Bovine DNA coding sequence

Our data set contained over 545 000 expressed sequence tags (EST). We extracted 342 495 *Bos taurus* EST excluding mitochondrial, sequence

tagged sites (STS) and genome survey sequences (GSS) from Genbank using ENTREZ at the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/Entrez/index.html) in late 2004. These sequences were reviewed, searching for the keywords "pseudo", "vector" and "repeat", all sequences found to be pseudogenes or those completely comprised of vectors or repeats were removed leaving 342 373 public EST sequences. The remaining 203 337 single-pass EST were commercially obtained from Genesis Research and Development (NZ) and included 50 non-normalised (high redundancy) and 1 normalised (low redundancy) cDNA libraries, all of which were collected from the domestic cow (*Bos taurus*) over a range of tissues and animals during various stages of development. The combined EST data set was checked to remove low quality sequences and sequences of non-cattle origin using the standard options of RepeatMasker [22]. To reduce any problems created by EST redundancy and to ensure the analysis was based on as many full length transcripts as possible, EST sequences were assembled into contigs using the standard options of CAP3 [13] after initial clustering of related sequences. We assembled over 40 000 contigs. We removed all contigs from the analysis with fewer than 4 EST, leaving 23 180 contigs from over 432 000 EST. This was done to remove cloning errors and improve estimates of tissue specificity of the contig concerned.

2.2. Tissue specificity patterns of bovine contigs

We removed 8103 contigs that were comprised solely of EST from NCBI or normalised libraries because we did not have any reliable information on expression patterns. Omitting these left 15 077 contigs from tissue-specific libraries. To determine the tissue specificity for each contig we counted the number of EST that were isolated from a given library and hence were expressed in the corresponding tissue. Where several libraries were from the same organ, such as the mammary gland, the number of counts for those libraries were added together: this left 32 tissues in place of the 50 libraries. Let N_{ij} be the number of EST from contig i that were found in tissue j. Assuming that EST were found randomly, we calculated the expected number for each contig and tissue $\frac{N_{i\cdot} \times N_{\cdot j}}{N_{\cdot\cdot}}$ and compared it to the observed N_{ij} with chi-squared tests. We defined genes as being over-expressed in a given tissue using two criteria. The first criterion required that contigs be significantly over-expressed in a tissue ($P < 0.01$), the second required contigs to be comprised of EST of which over 50% must come from a given tissue, to remove any contigs significantly over-expressed in more than one tissue. Housekeeping genes were defined as those that were observed in at least 16 out of the 32 tissues.

2.3. Expression breadth and secretion

Genes preferentially expressed within a tissue were further divided into those with narrow expression (expressed in only one tissue) and those with wider expression (two or more tissues). Genes containing a signal sequence for secretion or a membrane anchor and nonsecreted genes were identified using SignalP 3.0 [5].

2.4. Comparison of the bovine sequence to the human DNA sequence

Homo sapiens Genbank flat files were downloaded from the NCBI FTP site (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/human.rna.gbff.gz) from the most current draft of the human genome at the time of the analysis (NCBI build 35). Coding sequences (CDS) were assembled from the flat files using the CODERET programme from the EMBOSS software package [20]. Python and Perl scripts were used to create an automated pipeline for the high throughput analysis of evolutionary rates for pair-wise DNA sequence comparisons between species. To ensure the genes compared have evolved since their divergence from a common ancestor, and not since recent gene duplications, we identified orthologs between human and cattle via a reciprocal best hit (RBH) algorithm. The method was modified from the algorithm described by Rivera *et al.* [21]. We used BLASTN [2] to compare bovine contigs against NCBI's human RefSeq CDS, because it compares sequences at the nucleotide level and can achieve a high level of sensitivity and speed over programs like BLASTX. However, default BLASTN parameters are optimised for greater than 95% sequence identity between query and subject sequences and the typical bovine-human sequence identity is approximately 85% in CDS regions and somewhat less in non-coding regions. Therefore, the default word length, match, mismatch and gap penalties were adjusted. In this work the following BLASTN parameters were used: (-W 7 -r 17 -q -21 -f 280 -G 29 -E 22 -X 240). These are optimised for ~80% sequence identity and have been previously utilised by [24]. They are based on unpublished work by Ian Korf (http://homepage.mac.com/iankorf/) who was investigating human mouse alignments of transcribed regions. It should be noted that the expectations reported from the use of these options are only approximate (S.F. Altschul, pers. comm.).

2.5. Sequence evolution

Orthologous DNA sequences were aligned using the standard parameters of the NEEDLE program from the EMBOSS software package [20]. Bovine open

reading frames (ORF) were found by cross-referencing alignments of known human protein coding regions. All gaps and ambiguous characters including complete triplets left after the alignment were removed. Sequences found to contain early stop codons after the ORF was determined were considered non-orthologous comparisons and were removed from the analysis. The number of mutations per site (dN and dS) was adapted from the algorithm of Nei and Gojorbi [18]. Evolutionary ratios were calculated as $\frac{dN}{dS}$, wherein the dS average was calculated across all genes in the analysis. The average dS was used because there were no significant differences in dS, thus using gene-specific dS rates increases the error in dN/dS ratios. All mutations causing an amino acid change were classified as conservative or radical based on a 20×20 amino acid physiochemical-distance matrix [11]. A threshold value of 100 was set, above which all mutations were considered to be radical otherwise they were classed as conservative. Both Dr and Dc are not corrected for the number of sites per gene; therefore, we analysed $\frac{Dr}{(Dr+Dc)}$.

2.6. Polymorphism

Common human polymorphism data were collected for each gene from ENSEMBL (http://www.ensembl.org). The ENSMART tool was used to query a stable version of the ENSEMBL database (version 25.34e.1 from NCBI build 35, dbSNP build 123). An initial query was used to associate the NCBI nomenclature with an ENSEMBL gene name. Polymorphism data was then extracted from the ENSEMBL database for each gene. To reduce the contribution of rare deleterious alleles to evolutionary ratios, only polymorphisms where the minor allele had a frequency of 0.1 or greater were retrieved. All of the common polymorphisms were then used to estimate the number of polymorphisms per synonymous (hS) and nonsynonymous (hN) site and the number of polymorphisms causing either a radical (Hr) or conservative (Hc) substitution.

Cattle polymorphism data were collected using a PERL module we developed which was used to parse through the CAP3 output [13] and identify SNP from all of the overlapping EST. The module identifies SNP only when the variant minor allele occurs in at least two EST and at a frequency greater than 15%. This criterion is imposed to eliminate variants that are sequencing errors rather than true polymorphisms and polymorphisms that are reasonably rare in the population. All together, over 15 600 SNP were identified. However, in order to identify the cSNP from our subset of 15 600 SNP, we had to reliably identify the correct amino acid sequence and translation frame, especially if cSNP are to be broken down to their synonymous and nonsynonymous

components. We have accomplished this by cross-referencing ORF from the most recent release of annotated RefSeq mRNA from the human genome Build 35. The RBH algorithm [21] was used to identify orthologous sequences between humans and cattle and another computer module was written in PYTHON to determine if a given SNP was within the coding regions of the mRNA and if so, whether the SNP was a synonymous or nonsynonymous mutation. Orthologous sequences were aligned using the standard parameters of the NEEDLE program from the EMBOSS software package [20]. All gaps and ambiguous characters including complete triplets left after the alignment were removed. All SNP were then compared to see if they lie within the coding regions of the gene and the number of nonsynonymous (cN) and synonymous (cS) mutations per site were calculated. All mutations found to cause an amino acid change were further classified as radical (Cr) or conservative (Cc) mutations.

2.7. Statistical analysis

We compared the rate of evolution between and within species using general linear models (GLM) in GenStat For Windows 7th edition. The first model (GLM 1) was used to determine which tissues and genes were evolving rapidly within humans and between humans and cattle. The model was (1) $y = \mu + T + S + E + e$, where y: dN, dS, $dN/d\overline{S}$ and $Dr/(Dr + Dc)$ for divergence between cattle and humans or hN, hS, $hN/h\overline{S}$ and $Hr/(Hr + Hc)$ for common polymorphism within humans, μ: the mean effect and e is the residual error, T: the effect of the tissue in which the gene is expressed, S: the effect of whether the gene codes for an anchor, signal or nonsecreted protein, and E: the effect of whether the gene is expressed in multiple tissues or restricted to a single tissue. The least squares analysis finds estimates of T, S and E that minimise the error Sum of Squares (Σe^2).

An analysis of variance (ANOVA) was employed to determine if tissue, secretion or expression breadth were significant. The three factors were added sequentially and in different combinations to determine if T, S or E was associated with different evolutionary rates after correcting for the effects of one another.

A second model (GLM 2) was used to test the significance of differences in human-cattle divergence and in common human polymorphisms. The model was (2) $y = \mu + a + T + S + E + a.T + a.S + a.E + e$, where y: evolutionary ratios $dN/d\overline{S}$ or $Dr/(Dr+Dc)$ for both divergence and polymorphism data, a: the difference between divergence and polymorphism, and $v(e) = \sigma_1^2$ for divergence

data and σ_2^2 for polymorphism data, σ_1^2 and σ_2^2 were estimated from GLM 1 where polymorphism and divergence data were analysed separately. The difference in error variance between the polymorphism and divergence data was taken into account by a weighted analysis in which the weight for each data point was $1/\sigma_i^2$, where $i = 1$ for divergence data and 2 for polymorphism data. The interaction between a and T($a.T$) tests whether the difference between $dN/d\overline{S}$ and $hN/h\overline{S}$ is greater in some tissues than others. For instance, positive selection in mammary genes but not housekeeping genes should cause higher evolutionary ratios in divergence than polymorphism for mammary genes but not for housekeeping genes and hence an interaction between a and T. Similar reasoning applies to interactions between a and S or E and to $Dr/(Dr+Dc)$ instead of $dN/d\overline{S}$ data. Any significant effects found in the ANOVA were further investigated by pair-wise comparisons, which were programmed using RPAIR in GenStat, which calculates a t-test for each pair.

3. RESULTS

3.1. Ortholog detection

Human and cow reciprocal best hits (RBH) were obtained for each tissue group, resulting in 171 brainstem, 178 cortex and cerebellum, 52 cardiac muscle, 102 liver, 226 mammary gland, 198 skeletal muscle, 230 ovary and 265 housekeeping orthologous sequences that were used for estimating evolutionary rates. A low proportion of mammary genes (26%) were found to have RBH to human coding sequence, while the housekeeping group had a very high proportion of orthologous sequences (80%). The majority of other tissues returned RBHs between human and cow of 35–40 percent (Tab. I).

3.2. Human-cattle divergence

We applied the first model, GLM 1, to determine if there are any differences in evolutionary rates between tissues (T), modes of secretion (S) or of expression breadth (E), within and between species. Both T and S show significant effects on the rate of nonsynonymous substitutions per site (dN and hN, respectively), with divergence (dN) also showing an effect for E (Tab. II, $P < 0.05$). There were no significant effects on the number of synonymous substitutions per site. Because of the statistically homogeneous nature of the synonymous substitution rate, any differences can be attributed to random noise; therefore

Table I. Number of bovine genes for which tissue expression data is available and the number (percentage) that have orthologous reciprocal best hits in the human genome (build 35).

Gene groups	Total genes analysed	Genes with RBH (%)
Housekeeping genes	377	265 (80%)
Brainstem genes	506	171 (34%)
Cerebellum-cortex genes	494	178 (36%)
Cardiac muscle genes	153	52 (34%)
Liver genes	262	102 (39%)
Mammary gland genes	861	225 (26%)
Skeletal muscle genes	543	198 (36%)
Ovary genes	522	230 (44%)

Table II. Summary of *P*-values and error variance from an analysis of variance testing GLM1, each factor is added sequentially to determine their combined role in driving evolution.

Factors	dN	dS	$dN/d\overline{S}$	$Dr/(Dr+Dc)$	hN	hS	$hN/h\overline{S}$	$Hr/(Hr+Hc)$
T	***	0.2	***	*	**	0.4	**	0.6
S	***	0.3	***	0.3	*	0.3	*	0.4
E	*	0.3	*	*	0.8	0.1	0.8	0.9
$(V)e$	0.004	0.023	0.022	0.019	3.6e-7	7.1e-6	0.420	0.120

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

average dS should be more accurate for comparing dN/dS rates between divergence and polymorphism. Consequently, $dN/d\overline{S}$ and $hN/h\overline{S}$ are calculated using the constants $d\overline{S}$ and $h\overline{S}$ and any differences reflect the rate of dN and hN, respectively.

The radical to conservative ratio ($Dr/(Dr + Dc)$) shows significant differences between tissues and genes with different expression breadths. However, the radical to conservative ratio for common human polymorphisms ($Hr/(Hr + Hc)$) shows no significant differences in evolutionary rates for T, S or E (Tab. II).

All factors identified as generating significant differences in evolutionary rates in Table II are denoted with superscripts in the pair-wise comparisons (Tab. III). Variables without significant differences identified in the analysis of variance are not presented since they have previously been identified as statistically homogeneous (Tab. II).

Liver, mammary gland and cardiac muscle show the highest dN and $dN/d\overline{S}$ rates while housekeeping and cerebellum-cortex have the lowest (Tab. III). The association of expression breadth with significant differences in the rates of evolution (dN and $dN/d\overline{S}$) are apparent with the most widely expressed (housekeeping) genes showing significantly lower evolutionary rates than genes that are predominantly expressed within a single tissue. This phenomenon is also significant among genes expressed predominantly in one tissue; genes expressed entirely within one tissue in our EST collection (*i.e.* restricted expression breadth) evolved significantly faster than genes expressed in more than one tissue (*i.e.* wider expression breadth). For genes grouped by their secretory status, anchor and signal proteins accumulated more amino acid-changing substitutions than their nonsecreted counterparts, but, perhaps due to the small number of anchor proteins identified in our data set, no significant differences were generated between anchor and signal (secreted) genes (Tab. III).

The radical:conservative ratio shows a similar pattern to those seen in dN and $dN/d\overline{S}$, with the highest number of radical mutations found for genes expressed in the mammary gland and the lowest numbers for housekeeping genes (Tab. III). Interestingly, genes expressed in the liver, another rapidly evolving tissue, were found to have moderate $Dr/(Dr + Dc)$. For genes expressed predominantly in one tissue, the genes with wider expression had a lower $Dr/(Dr + Dc)$ than the genes with restricted expression (Tab. III).

3.3. Common human polymorphisms

The patterns seen for common human polymorphisms in Table IV show some similarity to those found for human-cattle divergence (Tab. III). No statistical differences were found for the rate of synonymous mutation per site among T, S or E. However, a statistically significant difference in the number of nonsynonymous mutations per site was detected for the tissue partition, with cerebellum-cortex and housekeeping genes showing the lowest hN and $hN/h\overline{S}$. A trend for high hN and $hN/h\overline{S}$ was detected for the mammary gland but it was nonsignificant, while the highest numbers were found in the ovaries and liver ($P < 0.05$). Significant increases in hN and $hN/h\overline{S}$ were also found for secreted proteins when compared to nonsecreted proteins, while restricted and widely expressed genes show very similar rates of evolution. The evolutionary ratio $Hr/(Hr + Hc)$ showed low values and large variances and no significant differences were found for any of the factors tested (Tab. IV).

Table III. Estimated effects (and s.e.) from GLM 1 of tissue, breadth of expression and secretion status on human-cattle divergence. Levels within a factor that do not share a common superscript are significantly different ($P < 0.05$).

Estimated effect	dN	$dN/d\overline{S}$	$Dr/(Dr + Dc)$
Mean (μ)	0.071	0.184	0.169
	(0.009)	(0.024)	(0.023)
Tissue (T)			
Brainstem	0.0^{cd}	0.0^{cd}	0.0^{b}
Cerebellum &	-0.019^{ab}	-0.049^{ab}	-0.0036^{b}
Cortex	(0.006)	(0.016)	(0.015)
Cardiac muscle	$+0.002^{cde}$	$+0.039^{cde}$	-0.025^{ab}
	(0.009)	(0.024)	(0.022)
Housekeeping	-0.024^{a}	-0.063^{a}	-0.032^{a}
	(0.006)	(0.015)	(0.015)
Liver	$+0.015^{e}$	$+0.039^{e}$	-0.011^{ab}
	(0.007)	(0.019)	(0.018)
Mammary	$+0.009^{de}$	$+0.024^{de}$	$+0.009^{b}$
	(0.006)	(0.015)	(0.014)
Skeletal muscle	-0.01^{bc}	-0.026^{bc}	-0.014^{ab}
	(0.006)	(0.016)	(0.015)
Ovary	-0.002^{c}	-0.004^{c}	$+0.0002^{b}$
	(0.006)	(0.015)	(0.014)
Expression breadth (E)			
Restricted	0.0^{b}	0.0^{b}	0.0^{b}
Expression			
Widely	-0.009^{a}	-0.024^{a}	-0.026^{a}
Expressed	(0.005)	(0.012)	(0.011)
Secretory status (S)			
Anchor	0.0^{b}	0.0^{b}	0.0
Nonsecreted	-0.019^{a}	-0.05^{a}	–0.0071
	(0.008)	(0.021)	(0.019)
Signal	$+0.005^{b}$	$+0.012^{b}$	+0.0056
	(0.008)	(0.022)	(0.020)

Table IV. Estimated effects (and s.e.) from GLM 1 of tissue, breadth of expression and secretion status. Levels within a factor that do not share a common superscript are significantly different ($P < 0.05$).

Estimated effect	hN	$hN/h\overline{S}$	$Hr/(Hr + Hc)$
Mean (μ)	0.000128	0.14	–0.019
	(0.0001)	(0.102)	(0.132)
Tissue (T)			
Brainstem	0.0^{ab}	0.0^{ab}	0.0
Cerebellum &	-0.000068^{ab}	-0.073^{ab}	+0.153
Cortex	(0.0001)	(0.069)	(0.107)
Cardiac	-0.000023^{abc}	-0.025^{abc}	+0.081
muscle	(0.0001)	(0.103)	(0.162)
Housekeeping	-0.000075^{a}	-0.081^{a}	+0.069
	(0.0001)	(0.064)	(0.109)
Liver	$+0.000159^{c}$	$+0.171^{c}$	+0.013
	(0.0001)	(0.081)	(0.110)
Mammary	$+0.00003^{abc}$	$+0.034^{abc}$	+0.138
	(0.0001)	(0.066)	(0.093)
Skeletal	-0.000009^{ab}	-0.009^{ab}	+0.047
muscle	(0.0001)	(0.068)	(0.096)
Ovary	$+0.000038^{bc}$	$+0.039^{bc}$	+0.102
	(0.0001)	(0.066)	(0.094)
Expression breadth (E)			
Restricted	0.0	0.0	0.0
expression			
Widely	+0.000014	+0.015	–0.014
expressed	(0.0001)	(0.051)	(0.065)
Secretory status (S)			
Anchor	0.0^{ab}	0.0^{ab}	0.0
Nonsecreted	$+0.000028^{a}$	$+0.027^{a}$	+0.121
	(0.0001)	(0.088)	(0.136)
Signal	$+0.000132^{b}$	$+0.14^{b}$	+0.065
	(0.0001)	(0.093)	(0.139)

Table V. Summary of means and (s.e.) for evolutionary variables between human-cattle divergence and common human polymorphisms.

	dN	dS	$dN/d\overline{S}$	$Dr/(Dr + Dc)$
Divergence	0.045	0.387	0.13	0.136
	(0.002)	(0.004)	(0.01)	(0.004)
Polymorphism	0.0002	0.001	0.21	0.158
	(0.001)	(0.0001)	(0.02)	(0.02)

Table VI. Summary of P-values from an ANOVA on GLM 2 comparing divergence and polymorphism (a) for evidence of positive selection groups of genes with similar expression in tissues (T), secretion (S) and expression breadth (E).

	a	T	S	E	$a.T$	$a.S$	$a.E$	e
$dN/d\overline{S}$	***	***	***	0.06	0.55	0.28	0.46	1.00
$Dr/(Dr + Dc)$	0.37	*	0.48	*	0.66	0.30	0.60	1.00

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

3.4. Selection tests

On average, evolutionary ratios were higher in common human polymorphisms than in human-cattle divergence (Tab. V). This result means that slightly deleterious alleles can rise to a gene frequency greater than 0.1 but are unlikely to be fixed.

Using GLM 2 we compared the evolutionary ratios $dN/d\overline{S}$ and $Dr/(Dr + Dc)$ between divergence and polymorphism and checked for evidence of significant interactions (Tab. VI). A significant difference between divergence and polymorphism was seen for the evolutionary ratio $dN/d\overline{S}$ (Tab. VI, $P < 0.001$). However, as indicated in Table V, this difference is in the opposite direction for any interpretation of positive selection. Significant differences were detected for T and S ($P < 0.001$) for $dN/d\overline{S}$, and T and E for $Dr/(Dr + Dc)$ indicating that the majority of the effects in GLM 1 are still significant in GLM 2. The generation of these differences has previously been explained in Tables II, III and IV. A lack of significant interaction between polymorphism versus divergence (a) and T, S or E for $dN/d\overline{S}$ and $Dr/(Dr + Dc)$ shows that the effects of T, S or E are not significantly different in polymorphism than in divergence data. Therefore, there is no evidence of positive selection generating the majority of differences in molecular evolution between humans and cattle.

The comparison of divergence and polymorphism has been based on human polymorphism only because this data is more extensive than that on bovine polymorphism. However, a comparison of common bovine and human polymorphisms is presented in Table VII. We have found a slightly higher rate of polymorphism for nonsynonymous and synonymous mutations per site and high evolutionary ratios in cattle than in humans. Thus, the lower $dN/d\overline{S}$ rate in human-cattle divergence is not due to abnormally high polymorphisms, and the use of bovine or human polymorphisms would not have increased the evidence for positive selection.

4. DISCUSSION

4.1. Rate of evolution

We confirm and extend previous findings [8, 23] that the most widely expressed (housekeeping) genes evolve at a slower rate than genes expressed in one or a few tissues. Also, among genes predominantly expressed in one tissue, genes expressed in >1 tissue show significantly lower $dN/d\overline{S}$ and $Dr/(Dr+Dc)$ than genes expressed in only one tissue. We also find that genes that code for secreted proteins are more likely to accumulate amino acid changes than are nonsecreted ones.

We found that the tissue in which genes are predominantly expressed strongly affects the evolutionary rate even after controlling for differences in breadth of expression and secretion. Genes expressed in the mammary gland apparently evolved more quickly than genes expressed in the brain's cerebellum and cortex. Not only did genes expressed in the mammary gland have a high $dN/d\overline{S}$ and $Dr/(Dr + Dc)$, but also a high proportion of mammary expressed genes did not even meet the RBH criterion and so were excluded from further study (Tab. I). Failure to meet the RBH criterion may indicate that these genes were orthologous but had diverged greatly between humans and cattle. Initial comparisons of genes failing the RBH criterion seem to have similarities in that their human best hit has another cattle best hit which is not typically expressed in the mammary gland. These drastic changes may hint at adaptive evolution being driven by gene duplications and rearrangements, which may not be detected by typical mutational analyses. Since more information is made available through the cattle genome project, this hypothesis will have to be tested further by comparing a number of species over a range of mammary specific genes.

Winter *et al.* identified the mammary gland as a tissue evolving at a moderate pace when compared to rapidly evolving tissues like the liver and slowly

Table VII. Summary of the mean number of polymorphisms per site for humans and cows.

	hN	cN	$h\overline{S}$	$c\overline{S}$	$hN/h\overline{S}$	$cN/c\overline{S}$	$Hr/(Hr+Hc)$	$Cr/(Cr+Cc)$
Averages	0.0002	0.0004	0.00092	0.0016	0.21	0.27	0.16	0.26
Tissues								
Brainstem	0.00019	0.00045	0.00083	0.0016	0.211	0.290	0.058	0.244
Cerebellum-Cortex	0.00012	0.00029	0.00074	0.0017	0.127	0.184	0.238	0.241
Heart	0.00015	0.00021	0.00155	0.0024	0.167	0.134	0.167	0.2
Housekeeping	0.00011	0.00036	0.00084	0.0015	0.119	0.23	0.155	0.237
Liver	0.00049	0.00061	0.00103	0.0022	0.528	0.392	0.103	0.389
Mammary	0.00023	0.00059	0.00084	0.0014	0.245	0.379	0.213	0.258
Muscle	0.00018	0.00041	0.00087	0.0012	0.199	0.266	0.13	0.192
Ovary	0.00025	0.00031	0.0012	0.0015	0.266	0.201	0.176	0.313

evolving tissues like skeletal muscle in human-mouse and rat-mouse comparisons [23]. Similarly in human-cattle divergence, we found evidence of rapid evolution in the liver. However, we also detected high rates of molecular evolution in the mammary gland. Based on the less commonly used radical:conservative ratio, we identified a very high rate of radical mutation in the mammary gland when compared to other rapidly-evolving tissues like the liver. This excess of radical mutations may be of evolutionary importance and highlights the need to utilise both ratios when comparing species for evidence of rapid evolution.

4.2. Synonymous mutations and mutation rate

One possible explanation for rapid evolution in a group of genes is that they are subject to higher than normal mutation rates. However, our finding that synonymous substitution rates did not vary between groups of genes argues against this explanation. Our findings confirm that synonymous substitution rates are similar in different genes [6], which suggests that different evolutionary rates are not a result of differences in mutation rates. Synonymous substitutions are thought to be selectively neutral in mammals and should therefore be representative of the underlying rate of mutation [15, 19]. Despite some evidence of selection on silent sites in invertebrates [1], a recent study that looked at the rate of synonymous mutation in different species of mammals found the rate to be constant between different tissues for human-mouse and human-cattle comparisons [8]. A steady rate of synonymous substitution in the majority of the categories we analysed confirmed previous findings that synonymous codon usage is not constrained by selection in mammals [8]. The increased dS reported for humans and mice [23], may be a result of the increased mutation rate reported in rodents [15] and thus reflects differences in the rate of mutation between some lineages.

4.3. Purifying selection

We found strong evidence for purifying selection as an important force in controlling the rate of molecular evolution between humans and cows. The results published for similar studies based on other mammals also show purifying selection's importance by keeping the majority of evolutionary ratios far below one [8, 23]. The reason evolutionary rates are so small is presumably that most nonsynonymous mutations to a gene are deleterious, and thus the majority of amino acids are prevented from changing due to strong purifying

selection. For instance, the average $dN/d\overline{S}$ of 0.13 implies that 87% of amino acid-changing mutations are deleterious.

Widely expressed genes, nonsecreted genes and genes expressed in tissues like the brain and skeletal muscle presumably have more amino acids that cannot be changed without deleterious effects. In contrast genes expressed in the mammary gland or liver, genes with secreted products and narrowly expressed genes are more likely to undergo amino acid changes without having much effect on fitness.

The $dN/d\overline{S}$ rates between cattle and humans are even lower than the analogous rate in common human polymorphisms. Presumably, despite their slightly deleterious effects, some of these mutations can reach frequencies >0.1 in a population but purifying selection reduces the probability of fixation. This conclusion was further reinforced when we used all polymorphisms and not just common ones (data not shown). The $dN/d\overline{S}$ rates were much higher when rare polymorphisms were not removed from the analysis as reviewed in some other studies [10]. However, our results and our interpretation of selective forces rely heavily on the accuracy of present public human SNP databases and any inaccuracies or biases towards disease-causing alleles may have the potential to influence the results. Therefore, the strength of the method will increase, as more coding SNP become publicly available.

4.4. Positive selection

The fact that $dN/d\overline{S}$ is lower in cattle-human divergence than in common human polymorphisms argues against the importance of selection driving this divergence in general, but what about in particular classes of genes? Rapid evolution of genes expressed in the mammary gland or liver, narrowly expressed genes and genes with secreted products could be due to positive selection, or reduced selective constraint. Significantly higher $dN/d\overline{S}$ and $Dr/(Dr + Dc)$ ratios among common human polymorphisms in all of the groups of genes showing rapid divergence between humans and cattle argue that the faster divergence is due to less constraint on the proteins and hence the genes coding for them. In general a higher rate of polymorphism in common cattle polymorphisms than in humans also justifies this conclusion. A lack of significant interactions in GLM 2 between common human polymorphism and human-cattle divergence for tissue, expression breadth or secretion status means that there is no evidence that positive selection was driving divergence in any particular class of genes. Therefore, we conclude that differences in evolutionary rate between groups of genes based on expression pattern and secretion status are due mainly to strong purifying selection on widely expressed genes,

genes expressed in the brain and genes for non-secreted proteins. This result was in agreement with the neutral theory of evolution [14] because it seems that there is a considerable amount of selective pressure that has acted against some mutations in the divergence of humans and cattle.

The recent findings of Ho *et al.* [12] may have a similar explanation. They found that substitution rates per year were higher for recently separated populations than for widely divergent species. This could be due to polymorphisms existing within species and contributing to short term divergence so that it is greater than expected. This phenomenon would be exaggerated if some polymorphisms were not neutral. A deleterious allele might exist in the population for some generations before it is eliminated by natural selection. Alternatively, some polymorphisms (*e.g.* sickle cell anemia) might be maintained by selection. In either case, the dN/dS ratio among polymorphisms is higher than expected from neutral polymorphisms. Consequently the McDonald-Kreitman test is a conservative test for positive selection.

Recent studies have reported that genes expressed in the brain have low evolutionary rates [8, 23]. However, we expected to observe positive selection in genes expressed in the human cortex and cerebellum, because these regions are associated with some of the unique complexities of the human brain. Surprisingly, we found that the evolutionary rate was lower in the cortex-cerebellum than the brain stem. Despite the rapid evolution of the human cortex, it appears that most of the genes expressed there are heavily constrained in their amino acid sequence. This result may stress the necessity of comparing species over an appropriate time frame; because of the short time frame and the rapid evolution of the human brain, a stronger signature of selection may be present in comparisons of humans to closely related species such as the chimpanzee. Even human polymorphism data may contain evidence of recent selection in human evolution [9].

Despite recent findings identifying positive selection in two genes related to the immune response in cattle [16], little is known about the contribution of positive selection to the uniqueness of domestic cattle. As far as we know, our bovine gene dataset combines the largest genomic screen for positive selection in this agriculturally important species. Because of the conservative nature of our test we removed any genes that did not return a RBH between human and cattle in an attempt to reduce the possibility of comparing paralogs. By doing so some of the most divergent, and possibly the most interesting genes were removed from the analysis. Thus our results and conclusions could be underestimates of the true underlying differences between humans and cattle. For instance, genes that had duplicated in either the cattle or human lineage

may be important in the evolution of the two species, but would probably be eliminated by our RBH criterion. Also we must acknowledge that our analysis did not incorporate any information regarding insertions and deletions (indels), or mutations in promoter and non-coding regions all of which potentially have important roles in the divergence of cattle and humans, which promise to be fruitful areas of research.

The test for positive selection might also be more powerful if the number of base substitutions on the branches leading to cattle and to humans were counted separately. To do this requires a suitable out species, the RBH criterion would have reduced the size of the data set greatly.

Despite these shortcomings we believe our results fairly reflect the species divergence in protein coding DNA and the most parsimonious explanation is that few of the single base substitutions are due to positive selection. Our results suggest it will be very difficult to detect positive selection in species as divergent as humans and cattle.

In conclusion, widely expressed genes, genes expressed in some tissues like brain and skeletal muscle and genes coding for nonsecreted proteins evolved more slowly than other genes because their amino acid sequence is more constrained. The high $dN/d\overline{S}$ and $Dr/(Dr + Dc)$ rates for polymorphism when compared to species divergence implies that long-term purifying selection is better at eliminating unfavourable mutations than short-term purifying selection is at removing unfavourable polymorphisms within the human population. This phenomenon probably explains the findings of Ho *et al.* and suggests that some caution be used when interpreting the results of McDonald-Kreitman tests between distantly related species. Thus, we conclude that the positive selection that occurred in the evolution of humans and cattle is not apparent, because it is overshadowed by neutral mutations that have led to substitutions within species.

ACKNOWLEDGEMENTS

We would like to thank the staff of AgResearch in Dunedin NZ for their valuable help and input for this project; in particular Ken Dodds and Nauman Maqbool who have had a great deal of input with the cluster analysis and contig generation. Finally we would also like to thank Jason Stajich, Keith Savin, Phil Bowman, Kon Konstantinov and both Martin and Judith Schweitzer for their valuable comments and help in the development of the analysis pipeline.

REFERENCES

[1] Akashi H., Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA, Genetics 139 (1995) 1067–1076.
[2] Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J., Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[3] Andersson L., Georges M., Domestic-animal genomics: deciphering the genetics of complex traits, Nat. Rev. Genet. 5 (2004) 202–212.
[4] Beja-Pereira A., Luikart G., England P.R., Bradley D.G., Jann O.C., Bertorelle G., Chamberlain A.T., Nunes T.P., Metodiev S., Ferrand N., Erhardt G., Gene-culture coevolution between cattle milk protein genes and human lactase genes, Nat. Genet. 35 (2003) 311–313.
[5] Bendtsen J.D., Nielsen H., von Heijne G., Brunak S., Improved prediction of signal peptides: SignalP 3.0, J. Mol. Biol. 340 (2004) 783–795.
[6] Bulmer M., Wolfe K.H., Sharp P.M., Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders, Proc. Natl. Acad. Sci. USA 88 (1991) 5974–5978.
[7] Clark A.G., Glanowski S., Nielsen R., Thomas P.D., Kejariwal A., Todd M.A., Tanenbaum D.M., Civello D., Lu F., Murphy B., Ferriera S., Wang G., Zheng X., White T.J., Sninsky J.J., Adams M.D., Cargill M., Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios, Science 302 (2003) 1960–1963.
[8] Duret L., Mouchiroud D., Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate, Mol. Biol. Evol. 17 (2000) 68–74.
[9] Enard W., Przeworski M., Fisher S.E., Lai C.S., Wiebe V., Kitano T., Monaco A.P., Paabo S., Molecular evolution of FOXP2, a gene involved in speech and language, Nature 418 (2002) 869–872.
[10] Fay J.C., Wu C.I., The neutral theory in the genomic era, Curr. Opin. Genet. Dev. 11 (2001) 642–646.
[11] Grantham R., Amino acid difference formula to help explain protein evolution, Science 185 (1974) 862–864.
[12] Ho S.Y., Phillips M.J., Cooper A., Drummond A.J., Time dependency of molecular rate estimates and systematic overestimation of recent divergence times, Mol. Biol. Evol. 22 (2005) 1561–1568.
[13] Huang X., Madan A., CAP3: A DNA sequence assembly program, Genome Res. 9 (1999) 868–877.
[14] Kimura M., The neutral theory of molecular evolution, Cambridge University Press, Cambridge, 1983.
[15] Kimura M., Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics, Proc. Natl. Acad. Sci. USA 88 (1991) 5969–5973.
[16] Lynn D.J., Freeman A.R., Murray C., Bradley D.G., A Genomics Approach to the Detection of Positive Selection in Cattle – Adaptive Evolution of the T Cell and NK Cell Surface Protein, CD2, Genetics 170 (2005) 1189–1196.

[17] McDonald J.H., Kreitman M., Adaptive protein evolution at the Adh locus in Drosophila, Nature 351 (1991) 652–654.
[18] Nei M., Gojorbi T., Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions, Mol. Biol. Evol. 3 (1986) 418–426.
[19] Nei M., Kumar S., Molecular Evolution and Phylogenetics, Oxford University Press, Oxford, 2000.
[20] Rice P., Longden I., Bleasby A., EMBOSS: the European Molecular Biology Open Software Suite, Trends Genet. 16 (2000) 276–277.
[21] Rivera M.C., Jain R., Moore J.E., Lake J.A., Genomic evidence for two functionally distinct gene classes, Proc. Natl. Acad. Sci. USA 95 (1998) 6239–6244.
[22] Smit A.F.A., Hubley R., Green P., RepeatMasker, http://repeatmasker.org, 2005.
[23] Winter E.E., Goodstadt L., Ponting C.P., Elevated rates of protein secretion, evolution, and disease among tissue-specific genes, Genome Res. 14 (2004) 54–61.
[24] Zhao S., Shatsman S., Ayodeji B., Geer K., Tsegaye G., Krol M., Gebregeorgis E., Shvartsbeyn A., Russell D., Overton L., Jiang L., Dimitrov G., Tran K., Shetty J., Malek J.A., Feldblyum T., Nierman W.C., Fraser C.M., Mouse BAC ends quality assessment and sequence analyses, Genome Res. 11 (2001) 1736–1745.