GENOMIC PATTERNS OF GENE EVOLUTION

by

Margaret A. Bakewell

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in The University of Michigan
2011

Doctoral Committee:

Professor Jianzhi Zhang, Chair
Professor Priscilla K. Tucker
Assistant Professor Patricia Wittkopp
Professor Jeffrey C. Long, University of New Mexico

The investigation of truth is in one sense difficult, in another easy. A sign of this is the fact that neither can one attain it adequately, nor do all fail, but each says something about the nature of things; and while each of us contributes nothing or little to the truth, a considerable amount of it results from all our contributions.

Aristotle
Metaphysics Book α
993^{a} 30; 993^{b} 1-3

This is how it works
You're young until you're not
You love until you don't
You try until you can't
You laugh until you cry
You cry until you laugh
And everyone must breathe
Until their dying breath

No, this is how it works
You peer inside yourself
You take the things you like
And try to love the things you took
And then you take that love you made
And stick it into some
Someone else's heart
Pumping someone else's blood
And walking arm in arm
You hope it don't get harmed
But even if it does
You'll just do it all again

"On the Radio"
Regina Spektor

For my favorite primates:
Russ, Rachel, Benjamin, and Samantha

Acknowledgements

For funding

EEB Department

Genome Science Training Program

For making things work when they had every inclination not to

Beverly Rathke

Deborah Goldberg

Mike Boehnke

Julia Eussen

Jane Sullivan

For chances to think about & talk about & work on other interesting things

Jo Kurdziel

Trisha Wittkopp

Gyorgyi Csankovszki

For publicizing my work

Stephen Colbert

Peter Sagal

For more than just being on the committee

Priscilla Tucker

Jeff Long

Trisha Wittkopp

For all kinds of awesomeness

Zhang lab members, past and present, especially Peng Shi, Wendy Grus, Soochin Cho, Ben-Yang Liang, Ondrej Podlaha, Xionglei He, Xiaoxia Wang, Wenfeng Qian, Ying Li, and Brian Metzger.

For the opportunity of a lifetime, and for extraordinary patience and generosity

George Zhang

For inspiration

Dad

For support

Mom

For lunch on the diag and walking to school and mix CDs and always listening

Sister

For listening, and for wise words

Brother

For joy and love

Rachel, Ben, and Sam

For always reminding me to watch the sun rise and the birds fly, and for everything else

Russ

Table of Contents

Dedication ii

Acknowledgements iii

List of Figures vii

List of Tables viii

Chapter 1: Introduction 1

Chapter 2: Did brain-specific genes evolve faster in humans than in chimpanzees? 4

- Abstract 4
- Introduction 4
- Compilation of the primate gene dataset 6
- Analysis of brain-specific genes 8
- Caveats 12
- Concluding remarks 14
- Supplementary methods 15
- Acknowledgements 19

Chapter 3: More genes underwent positive selection in chimpanzee evolution than in human evolution 25

- Abstract 25
- Introduction 26
- Results and Discussion 27
- Materials and Methods 36
- Acknowledgements 43

Chapter 4: Genomic phylostratigraphy and spurious patterns of new gene evolution.................. 51

Abstract.. 51

Introduction.. 52

Results.. 55

Discussion.. 60

Methods 64

Acknowledgements... 68

Chapter 5: Conclusions 75

Old questions, new data 75

Challenges of functional and evolutionary inference 77

Skepticism, hypotheses, and the self-correcting nature of science 80

Appendix.. 82

Genes (*n*=154) showing significant signal of positive selection in human before Bonferroni correction.. 82

Genes (*n*=233) showing significant signal of postive selection in chimpanzee before Bonferroni correction.. 88

Genes showing significant signal of postive selection at FDR = 5% 97

Genes showing significant signal of postive selection and associated with human disease 100

References.. 103

List of Figures

Figure 2.1 Evolutionary tree of human, chimpanzee and macaque monkey. 19
Figure 2.2 Venn diagram of brain-specific genes identified on the basis of five different definitions. 20
Figure 3.1 Functional differences between human and chimp unshared PSGs. 44
Figure 3.2 Frequency distribution of human and chimp PSGs across 20 peak expression tissue groups.. 45
Figure 3.3 Distributions of human and chimp PSGs among chromosomes................................. 46
Figure 3.4 Distribution of human and chimp PSGs across chromosomes. 47
Figure 4.1 Phylogeny showing the relationships of sequences simulated in gene age study........ 69
Figure 4.2 Distribution of genetic distance between *D. melanogaster* and *D. grimshawi* for genes of different ages. 70
Figure 4.3 Age group within each of 10 approximately equally sized bins of increasing alpha... 71
Figure 4.4 Distribution of 6,662 *Drosophila* genes by minimum evolutionary rate class for the gene and maximum block length of the minimum rate class (c_{max}) (A) and BLAST error rate within each group (B). 72
Figure 4.5 Log-odds ratio of expression in mesoderm, ectoderm and endoderm for each age group compared to the average of all genes studied. 73

List of Tables

Table 2.1 Evolutionary rates of brain-specific genes and other genes in humans and chimpanzees 21

Table 2.2 Comparison of the evolutionary rates of brain-specific genes and other genes in human and chimpanzee lineages using microarray 2X definition of tissue-specificity and unmodified genome sequences........ 22

Table 2.3 Tissue groups considered in the analysis of the microarray data 23

Table 2.4 Comparisons between evolutionary rates of brain-specific genes and other genes in human and chimpanzee lineages using high quality (Q20) chimpanzee sequences. 24

Table 3.1 Genic positive selection in human and chimp lineages since their split 48

Table 3.2 Association of PSGs with human disease 48

Table 3.3 The ω values used in generating the DNA sequences in computer simulation........ 48

Table 3.4 False-positive rates of the branch-site likelihood method in detecting positive selection at the 5% significance level 49

Table 3.5 Basic statistics of the three datasets 49

Table 3.6 Rate of detection of positive selection by the branch-site method when there is positive selection 49

Table 3.7 Genes showing significant signal of postive selection in both human and chimp lineages 50

Table 3.8 Numbers of genes with regard to tissue-specificity of expression........ 50

Table 4.1 Parameters and results for simulation models used. 74

Chapter 1

Introduction

When I first studied biology as an undergraduate in the early 1990's, only a handful of genomes had been sequenced, and those were of viruses and organelles (Fleischmann et al. 1995). Between then and now, a genomic revolution has occurred (Zhang 2010). Thousands of genomes have been completed, including our own (Consortium 2004), and thousands more are under way (www.genomesonline.org). In the genomic age, rock stars have their genomes sequenced (www.knome.com), as do species that have never been seen or identified except by their distinctive genome sequences (Venter et al. 2004). In other words, genomic sequences now inform us of both the diversity within species (What are the genetic differences between Ozzy Osbourne and Craig Venter?) and the diversity of life on Earth (How many species are swimming in the Sargasso Sea?). Fundamental questions in molecular evolution, such as the origin of new genes (Ohno 1970), the molecular basis of human traits (King and Wilson 1975), and the relative contributions of drift and selection (Kimura 1983b) can be studied with such data on a genomic scale, with the goal of moving beyond the understanding of individual genes and towards an understanding of the patterns of gene evolution (Zhang 2010).

In this dissertation, I examine aspects of all three of the above fundamental questions in molecular evolution. My first project (Chapter 2) examines the evolution of brain genes, using the genomic sequences of human, chimpanzee and rhesus macaque, along with several genomic gene expression datasets (Su et al. 2004; Semon et al. 2006), to first identify brain-specific genes and then compare their evolution in human and chimp, and to compare their evolution with that of other genes. A previous study (Dorus et al. 2004) had concluded that nervous system genes experienced accelerated evolution in humans, perhaps due to positive selection, but one of the flaws in that study was the small size of the dataset – only 24 genes were compared between human and chimp. Leveraging genomic sequence and expression data, I examined 686 genes expressed predominantly or specifically in the brain to address the question of whether a pattern of accelerated evolution in brain-specific genes contributed to the unique brain size and related cognitive capacities of humans.

My second project (Chapter 3) also focuses on human evolution, in this case, the prevalence of positive selection. Again using almost 14,000 genes from the human, chimpanzee, and rhesus macaque genomes, I identify positively selected genes in human and chimpanzee on a genome-wide basis. Since the neutral theory predicts that drift plays a more important role in species with a small effective population size(Kimura 1983a), like human, compared to those with a larger effective population size, like chimpanzee (Chen and Li 2001), it is interesting to compare the action of drift and positive selection in the two species. Despite the anthropocentric view that humans have many more beneficial traits than chimps, in fact, I found that chimpanzees have more genes that have experienced positive selection than humans have. This finding and other aspects of the human and chimpanzee genomes are consistent with reduced efficacy of selection in humans compared to chimps.

My final project focuses on the processes that lead to new genes by examining patterns of evolution in 12 *Drosophila* genomes and using those patterns to simulate long-term evolution of genes (Chapter 4). After evolution under gradual and punctuated models (Domazet-Loso and Tautz 2003), I assess the extent to which genes retain sequence similarity with their homologs. Findings from this study indicate that it is difficult to distinguish between new and old genes, due to both classifying old genes as new, and classifying new genes as old. In addition, this work shows that the combination of gene duplication with both gradual and punctuated evolution are not sufficient to account for the observed number of young genes in *Drosophila*, indicating that other mechanisms, such as gene origination from non-coding regions, must also be at work.

Each of these studies examines the majority of genes in multiple genomes to obtain information about how genes in general evolve. Another common theme that unites these studies is the importance of careful design in genomic studies. For example, in the study about brain genes, we found that simply defining a “brain gene” is no trivial task. We used several different definitions to try to capture a core set of brain genes, taking into account level and specificity of expression as well as functional annotations where available. In the study about positive selection in human and chimp, great care was taken to ensure that the lower quality of the chimp genome compared to the human genome did not influence the results. A naïve comparison of positive selection in human and chimp, without taking this bias into account, could only yield spurious results. Finally, in my third study about evolutionary patterns of new genes in *Drosophila*, I confront the issue of bias directly by comparing published results about genes of different ages with simulated results that incorporate biased measurement, and I find that at least some published results attributed to biological processes(Domazet-Loso et al. 2007) can be more parsimoniously explained by biased measurement of gene age. Taken together, these studies shed light on the patterns of molecular evolution, and also remind us that great care must be taken to ensure that genomic patterns reflect true biological processes rather than artifacts of the data and techniques we use.

Chapter 2

Did brain-specific genes evolve faster in humans than in chimpanzees?

Abstract

One of the most distinctive characteristics of humans among primates is the size, organization and function of the brain. A recent study has proposed that there was widespread accelerated sequence evolution of genes functioning in the nervous system during human origins. Here we test this hypothesis by a genome-wide analysis of genes that are expressed predominantly or specifically in brain tissues and genes that have important roles in the brain, identified on the basis of five different definitions of brain specificity. Although there is little overlap among the five sets of brain-specific genes, none of them supports human acceleration. On the contrary, some datasets show significantly fewer nonsynonymous substitutions in humans than in chimpanzees for brain-specific genes relative to other genes in the genome. Our results suggest that the unique features of the human brain did not arise by a large number of adaptive amino acid changes in many proteins.

Introduction

The human brain differs substantially from those of other primates in size, organization and function. For instance, in comparison to that of chimpanzees, the brain weight of humans is over 300% greater but the body is only 35% heavier (Williams 2002) (Figure 2.1). The structural asymmetry between the left and right hemispheres is especially pronounced in human brains (Deacon 1994). Humans also have speech or language and other high-order cognitive functions that are absent in non-human primates. The genetic changes that have been responsible for the emergence of these human-unique brain features are a topic of enduring interest. Generally speaking, the marked evolution of the human brain could be due to modifications of either a small or a large number of genes, where the modifications might be in gene expression or protein function.

If widespread changes in many genes were the cause of human brain evolution, the signatures of such events might be identifiable from a genome-wide analysis. Recently, Dorus et al. (Dorus et al. 2004) analyzed a set of nervous system genes at the protein sequence level and found that these genes evolved significantly faster in primates than in rodents, in hominoids than in Old World monkeys, and in humans than in chimpanzees. They further suggested that the accelerated evolution was due to positive Darwinian selection for advantageous amino acid changes. Their analysis, however, suffered from four shortcomings. First, they compared only 24 nervous system genes between human and chimpanzee – the most relevant species pair for studying evolution of the human brain. Second, their list of nervous system genes was manually compiled and might thus be incomplete or biased (see later). Third, they used house-keeping genes as controls in some of the analyses, which seems inappropriate because tissue-specific genes and house-keeping genes are expected to have different evolutionary patterns (Duret and Mouchiroud 2000; Zhang and Li 2004). Fourth, a recent comparison between the dog and mouse genomes found that 18 nervous system genes that evolved faster in primates than in rodents also evolved faster in carnivores than in rodents (Lindblad-Toh et al. 2005), suggesting that the findings of Dorus et al. (Dorus et al. 2004) might partially be due to rodent deceleration rather than primate acceleration. A more recent analysis of 5,268 genes has also found more amino acid substitutions in humans than in chimpanzees for brain-specific genes; however, the statistical significance of the difference is uncertain (P = 0.03–0.08, depending on which genes are used as controls) and the results are inconclusive (Khaitovich et al. 2005).

Here we conduct a comparison of sequence evolution of brain-specific genes between the human and chimpanzee lineages, using genome sequences of human, chimpanzee and macaque monkey, and human transcriptome data.

Compilation of the primate gene dataset

From Ensembl (http://www.ensembl.org), we obtained the DNA and amino acid sequences of all of the proteins predicted from the genome sequences of human (*Homo sapiens*), chimpanzee (*Pan troglodytes*) and macaque (*Macaca mulatta*). To identify orthologous genes, we used human proteins as queries to search chimpanzee proteins with BLASTP (see Supplementary Methods). Reciprocal best hits are considered as orthologs. Similarly, we used human sequences to search the macaque proteins with BLASTP. A total of 19,422 proteins with reciprocal best hits in both the human–chimpanzee and the human– macaque searches were found, and alignments of the human–chimpanzee–macaque orthologous proteins were obtained.

We discarded alignments containing fewer than 100 amino acids because most of these were caused by gaps in draft genome sequences. DNA sequence alignments were obtained from the protein alignments. We further removed 161 alignments that showed exceptionally high divergences among the species and were probably the results of misalignment or non-orthology (see Supplementary Methods). The proportion of brain-specific genes was lower in the removed alignments than in the remaining alignments. Finally, each protein was assigned to a gene on the basis of its Ensembl annotation, resulting in 13,955 distinct genes for further analysis. After the removal of alignment gaps, these genes contain 18,287,982 nucleotide sites or 6,095,994 codons, covering >50% of all protein coding regions in a primate genome.

We consider that a nucleotide position has a human-specific substitution if the sequence is identical between the chimpanzee and macaque but different in human at this position. We similarly define chimpanzee specific substitutions. A nucleotide substitution is then classified as either synonymous or nonsynonymous depending on whether it alters the amino acid encoded.

We observed 57,545 chimpanzee-specific nucleotide substitutions and 50,254 human-specific substitutions. Thus, the nucleotide substitution rate seems to be 1.15 times (57,545/50,254) higher in chimpanzees than in humans. This rate difference is probably due to the relatively low quality of the 4-coverage chimpanzee draft genome sequence (Consortium 2005), as compared with that of the finished human genome sequence (Consortium 2004).

A recent study estimated that the error rate in the chimpanzee genome sequence has an upper limit of 0.07% (Taudien et al. 2006), ~70 times higher than the error rate in the human sequence (Consortium 2004). The observed chimpanzee to human divergence is 0.59% in our dataset of coding sequences. If we assume that the actual substitution rates in humans and chimpanzees are identical, then the chimpanzee to human substitution rate ratio ($R_{C/H}$) might appear as high as 1.27, simply because of the 0.07% sequencing errors in the chimpanzee genome (Supplementary Methods). If we also consider that the mutation rate per year is slightly (3%) lower in humans than in chimpanzees (Elango et al. 2006), $R_{C/H}$ might appear as high as 1.30 (Supplementary Methods). Our $R_{C/H}$ value of 1.15 is within these limits. Our result is also comparable to a recent estimate of 1.11–1.18 for the $R_{C/H}$ for large numbers of intergenic sequences and introns obtained from a comparison of the draft chimpanzee genome sequence and the finished human sequence (Elango et al. 2006).

Several measures of the rate of protein sequence evolution have been well established by molecular evolutionists (Nei and Kumar 2000). For example, let n be the number of nonsynonymous substitutions for a group of genes in a particular lineage and s be the corresponding number of synonymous substitutions; and let N and S be the numbers of nonsynonymous and synonymous sites, respectively, for the group of genes (Nei and Kumar 2000). For any large group of genes in our dataset, N/S = 2.45 (Supplementary Methods). Thus, the nonsynonymous- to-synonymous rate ratio (ω), which is commonly used to measure the rate of protein evolution controlled by the mutation rate, becomes (n/N)/(s/S) = (n/s)/(N/S) = (n/s)/

2.45 = 0.408n/s. Because most genes in a genome have a ω value of <1, whereas sequencing errors are expected to have a ω value of 1, the errors cause overestimation of ω. Thus, we would see a higher ω for chimpanzees than for humans owing to chimpanzee sequencing errors. Furthermore, the bias is more serious for genes with low ω than for genes with high ω (when $\omega<1$). Because brain-specific genes tend to have lower ω values than other genes in the genome (Table 2.1), the former are affected by sequencing errors to a greater extent than the latter. Thus, we expect to observe a higher ω in chimpanzees than in humans for brain-specific genes, even when benchmarked by other genes in the genome (Table 2.2).

To rectify this problem, we add the same number of random 'sequencing errors' to the human genome sequence as the number that occurred in the chimpanzee sequence. Although sequencing errors will still affect the ω of brain specific genes more than that of other genes, the human and chimpanzee lineages can now be compared. Assuming that the total numbers of substitutions in our 13,955 genes are equal between the human and chimpanzee lineages, we estimate that the error rate in the chimpanzee sequence is 0.04%, which is equal to 7,315 errors (Supplementary Methods). We thus randomly add this number of errors to the human sequence and then compare the human and chimpanzee sequences. Although the 4.6-coverage macaque genome sequence might also contain numerous sequencing errors, these errors are not expected to bias our comparison between human and chimpanzee because the macaque is used as an outgroup.

Analysis of brain-specific genes

It is not an easy task to define those genes that function specifically in the brain. We therefore use five different definitions to examine whether they provide consistent results.

Analysis based on microarray data

Our first definition is based on a human microarray gene expression dataset (Su et al. 2004), which includes the expression signals of almost all human genes in 73 normal tissues. Because many of the 73 tissues are from the same organs, we group the tissues into 40 tissue groups (Table 2.3). For example, the brain tissue group includes 17 tissues that represent different developmental stages or parts of the brain. Brain-specific genes are defined as those genes for which the highest expression is found in one of the brain tissues and this highest expression is at least twice the expression level in any non-brain tissues. As a result, 249 brain-specific genes are identified. Similarly, we identified tissue-specific genes for the other 39 tissue groups, and the total number of these other tissue-specific genes is 1,544. The remaining 12,162 genes are referred to as non-tissue-specific genes. For our second definition, we used the same human microarray gene expression dataset but with more stringent criteria, requiring that the highest expression level in a brain tissue is at least four times that in any non-brain tissue for a gene to be called brain-specific. Because the results based on these two definitions are almost identical, below we describe in detail only those from the first definition (Table 2.1, 'Microarray [2X]'; see 'Microarray [4X]' for the results from the second definition).

We find that, for brain-specific genes, the ω value in the human lineage (ω_H) is 0.205 and that in the chimpanzee lineage (ω_C) is 0.198. Their ratio (ω_H=ω_C ¼ 1:03) is not significantly different from 1 (P > 0.5, x2-test; Table 2.1). As a comparison, ω_H=ω_C equals 1.06 ($P > 0.05$) for other tissue-specific genes, 1.12 ($P < 10^{-4}$) for non-tissue-specific genes, and 1.11 ($P < 10^{-4}$) for all of the genes considered together. The observation of ω_H=ω_C >1 for all genes together is consistent with previous findings and is explainable by a smaller effective population size and thus weaker purifying selection and a higher nonsynonymous substitution rate in the human lineage than in the chimpanzee lineage (Eyre-Walker and Keightley 1999; Consortium 2005; Khaitovich et al. 2005). We find that the ω_H=ω_C ratio of brain-specific genes is slightly lower than that of other tissue-specific genes, but the difference is not statistically significant ($P > 0.5$, simulation test; Table 2.1).

Similar results are obtained when brain-specific genes are compared with non-tissue-specific genes (Table 2.1). Because the same genes are compared between human and chimpanzee, we can compute the ratio of the number of nonsynonymous substitutions in the human lineage (n_H) to that in the chimpanzee lineage (n_C) and compare this ratio (n_H/n_C) between different groups of genes. Interestingly, we find that n_H/n_C is significantly lower for brain-specific genes than for non-tissue-specific genes (P = 0.04, χ^2-test; Table 2.1), suggesting a possible human slowdown (or chimpanzee acceleration) of the evolution of brain-specific genes, when benchmarked by non-tissue-specific genes.

The n_H/n_C values are not, however, significantly different between brain-specific genes and other tissue-specific genes, or between other tissue-specific genes and nontissue- specific genes.

Analysis based on EST data

Because the microarray data might be inaccurate (Liao and Zhang 2006), we repeated the above analysis using a third definition of brain-specific genes based on expression sequence tags (ESTs). Here, tissue-specific genes are those for which ESTs are found in only one tissue. We used a recently compiled human EST dataset that includes 4.9 million ESTs from 44 tissues (Semon et al. 2006) and classified the 13,955 primate genes into 165 brain-specific genes (i.e. ESTs are found only in the brain), 819 other tissue-specific genes, and 12,971 non-tissue-specific genes. The results from the EST data (Table 2.1) are similar to those from the microarray data. Although ω_H=ω_C is significantly greater than 1 for other tissue-specific genes and non-tissue-specific genes, it is not significantly greater than 1 for brain-specific genes. Consequently, the ω_H=ω_C ratio is slightly lower for brainspecific genes than for other genes, although the difference is not statistically significant (Table 2.1). Similarly, the n_H/ n_C ratio appears lower, although not significantly, in brainspecific genes than in other genes (Table 2.1).

Analysis based on SAGE data

We also repeated the above analysis using a fourth definition of tissue specificity based on serial analysis of gene expression (SAGE). Brain-specific genes are defined as those for which SAGE tags are detected only in the brain. On the basis of a recently compiled SAGE dataset (Semon et al. 2006) the 13,955 primate genes include 209 brain-specific genes and 632 other tissue-specific genes. The remaining genes are considered to be non-tissue-specific. The results obtained from the SAGE data (Table 2.1) are similar to those from the microarray and EST data. That is, there is no significant difference between ω_H and ω_C for brain-specific genes, regardless of whether other genes are used as controls or not. There is also no significant difference between the n_H/n_C ratios of brain-specific genes and other genes.

Analysis based on a list of nervous system genes Dorus et al. (Dorus et al. 2004) compiled a list of 214 nervous system genes on the basis of (i) literature suggesting important gene functions in the nervous system, (ii) SAGE and EST data showing gene expression exclusively or predominantly in the brain, and (iii) information on genes implicated in nervous system diseases (Dorus et al. 2004). We found 146 of these 214 genes in our list of 13,955 primate genes. Because Dorus et al. (Dorus et al. 2004) did not define other tissue-specific genes, we analyzed these nervous system genes by using the remaining 13,809 genes in our dataset as a control. We find no significant difference between ω_H and ω_C for nervous system genes, with or without comparison to other genes (Table 2.1). Dorus et al. (Dorus et al. 2004) suggested that the human lineage acceleration is particularly pronounced for a subset of genes that control nervous system development, but is absent for genes with physiological roles and minimal for the remaining (i.e. unclassified) nervous system genes. Our data, however, provide no statistical evidence for these claims (Table 2.1). We also failed to detect a difference in n_H/ n_C between nervous system genes (or developmental nervous system genes) and other genes (Table 2.1).

The main reason why we cannot repeat the result of the faster evolution of humans than chimpanzees even when we use the list of nervous system genes that Dorus et al. (Dorus et al. 2004) compiled seems to be because Dorus et al. did not compare all of the 214 nervous system genes between human and chimpanzee. Instead, between humans and chimpanzees they compared only 24 genes that were known to evolve faster in the human lineage than in the macaque lineage when the squirrel monkey was used as an outgroup. In other words, they used a small and biased gene set in their human–chimpanzee comparison.

Caveats

Although our results from the five analyses are congruent in showing that there has been no accelerated evolution of human brain-specific genes, this congruence would be expected if there were large overlaps among the five groups of brain-specific genes identified under the five different definitions. Interestingly, however, except for those identified by the two microarray-based definitions, only a few genes overlap from any two of the five groups of brainspecific genes and no genes overlap among all five groups (Figure 2.2).

Although this finding suggests that the five analyses are largely independent, it also raises the issue of how to identify brain-specific genes accurately. The level of gene expression in a tissue is a continuous variable. For the EST (or SAGE) data, we identified brain-specific genes as those that lack ESTs (or SAGE tags) in non-brain tissues, which actually means genes that have a lower expression level in non-brain tissues than in the brain. This definition is qualitatively the same as that used for the microarray data, where brain-specific genes are required to show expression at least twice as high in the brain as in any other tissue. Although it might be argued that a gene that exclusively functions in the brain could have a lower expression in this organ than in other tissues, such a situation is unlikely, particularly when expression in the brain is defined by the highest expression level among all temporal and spatial brain samples.

All five definitions that we used consider gene expression patterns, although the fifth definition also includes genes with known brain functions and genes implicated in brain diseases. On the one hand, considering gene function provides additional information that might help to reduce the reliance on gene expression, which is sometimes a poor indicator of function. On the other hand, gene function information is usually incomplete and it is difficult to know whether a gene functions exclusively in the brain. Our results suggest that it is still a challenging task to define genes that function specifically in a tissue. A potential way of increasing the accuracy of identifying brain-specific genes is to use more than one criterion. We therefore analyzed a subset of 74 genes that are brain-specific by at least two of our definitions 1, 3, 4, and 5; we excluded definition 2 because it is a subset of definition 1 (Figure 2.2). The difference between ω_H and ω_C of brain-specific genes, with or without comparison to other genes in the genome, is still not significant (Table 2.1). Interestingly, however, the n_H/n_C ratio is significantly lower for brain-specific genes than for other genes in the genome (Table 2.1).

Our analysis also highlights the intricacy of genomewide comparisons between humans and chimpanzees in the presence of sequencing errors. As eloquently articulated by Taudien et al. (Taudien et al. 2006), a small leak can sink a great ship. In our analysis, the chimpanzee sequencing errors, when not appropriately controlled, generate a significantly higher ω_C than ω_H for brain-specific genes, even when compared with other genes in the genome (Supplementary Table 2.1). This difference disappears when we add the same number of ‘sequencing errors’ to the human sequence. In our addition of sequencing errors to the human sequence, we assumed that the substitution rate for the whole set of 13,955 genes is identical between the human and chimpanzee lineages. If the mutation rate is slightly lower in humans than in chimpanzees (Elango et al. 2006) and the total substitution rate is also lower in humans than in chimpanzees, we might have added more ‘sequencing errors’ than needed, which would have raised ω_H=ω_C and favored the human acceleration hypothesis. In other words, our result of no human acceleration is conservative (see also Supplementary Methods).

To verify the results obtained from our approach of error addition, we also used the approach of error removal. We removed errors from the chimpanzee sequence by using only nucleotide sites with quality scores Q20 (or accuracy > 99%) (Consortium 2005). The new dataset contained 13,888 genes. Again, none of the analyses shows a significantly higher evolutionary rate of brain-specific genes in humans than in chimpanzees (Table 2.4).

Concluding remarks

We have analyzed almost 14,000 human, chimpanzee and macaque genes to test the hypothesis that human brainspecific genes have undergone widespread accelerated protein-sequence evolution since the human lineage separated from the chimpanzee lineage. Our results, based on five different definitions of brain-specificity, show no evidence that supports this hypothesis. Because our data include over 50% of all human genes, it is appropriate to conclude that our results reject the hypothesis of widespread accelerated sequence evolution of human brainspecific genes.

In fact, in several but not all of our analyses, the nH/nC ratio is significantly lower for brain-specific genes than for other genes in the genome, suggesting that – relative to other genes – brain-specific genes evolved more slowly in humans than in chimpanzees. This phenomenon might reflect higher importance of brain-specific genes and therefore stronger purifying selection on them in human evolution than in chimpanzee evolution. Our findings imply that the unique features of the human brain did not arise by a large number of adaptive amino acid substitutions in many proteins. This conclusion, however, does not preclude the possibility that substantial accelerations occurred in the evolution of a few nervous system genes during human origins. Indeed, several such examples are known, including genes that control brain size and speech development (Enard et al. 2002b; Zhang et al. 2002; Zhang 2003; Evans et al. 2004a; Evans et al.

2004b; Kouprina et al. 2004; Wang and Su 2004). It also remains possible that the origin of the human-unique brain features was due to expression changes (rather than coding sequence changes) of many genes, as has been suggested from some microarray data (Enard et al. 2002a; Khaitovich et al. 2005) (but see also Refs (Hsieh et al. 2003; Uddin et al. 2004; Lemos et al. 2005)).

Supplementary methods

Primate gene dataset

Protein and corresponding DNA sequences of all predicted genes in the human, chimpanzee, and macaque genome sequences were downloaded from Ensembl v 35 (November 2005, http://www.ensembl.org). To identify orthologous genes, human protein sequences (n= 33,869) were used to conduct BLASTP searches (Altschul et al. 1990) against the chimpanzee (n = 39,648) and macaque (n = 31,371) protein sequences. Reciprocal searches were performed using the chimpanzee and macaque proteins to query the human proteins. 19,422 proteins with reciprocal best hits in both human/chimpanzee and human/macaque searches were retained for further analysis.

Alignment of the human-chimpanzee-macaque orthologous proteins was performed using CLUSTALW v 1.83 (Thompson et al. 1994). DNA sequence alignments were obtained following the protein sequence alignments. Alignments containing fewer than 100 amino acids (n= 1,291) were discarded. Lineage-specific nucleotide substitutions were identified as described in the main text. Review of several alignments with exceptionally high proportions of human- or chimpanzee-specific changes revealed that they resulted from incorrect alignment or non-orthology. Therefore, alignments containing greater than 10% human or chimpanzee-specific amino acid or nucleotide changes or greater than 30% macaque-specific changes were discarded from analysis (n = 161). Finally, each protein was assigned to a gene based on the Ensembl annotation, and the protein sequence with the longest amino acid alignment was retained for each gene, resulting in the alignments of human, chimpanzee, and macaque sequences of 13,955 distinct genes. Using modified Nei-Gojobori method (Zhang et al. 1998) with a transition/transversion ratio of 2 (Rosenberg et al. 2003), we estimated that the total number of nonsynonymous sites in these 13,955 genes was N = 12,986,068 and the total number of synonymous sites was S = 5,301,914, with their ratio being N/S = 2.45. We used the parsimony approach to identify human-specific and chimpanzee-specific substitutions (see main text). This approach is justified because of the low divergence of the sequences concerned in this paper. The observed human-chimpanzee divergence is 0.59% in our dataset of coding sequences. A recent study estimated that the error rate in the chimpanzee genome sequence has an upper limit of 0.07% (Taudien et al. 2006). Assuming this error rate, the actual divergence between the two species is 0.59% – 0.07% = 0.52%. If we assume that the actual substitution rates in humans and chimpanzees are identical, the proportion of sites with chimpanzee substitutions (including sequencing errors) is 0.52% ÷ 2 + 0.07% = 0.33%, whereas the proportion of sites with human substitutions is 0.59% – 0.33% = 0.26%. The chimpanzee/human substitution rate ratio ($R_{C/H}$) becomes 0.33 ÷ 0.26 = 1.27. If we consider that the substitution rate is 2 3% lower in humans than in chimpanzees (Elango et al. 2006), the proportion of sites with chimpanzee substitutions is 0.52% ÷ 1.97 + 0.07% = 0.334%, whereas the proportion of sites with human substitutions is 0.59% – 0.334% = 0.256%. Their ratio $R_{C/H}$ is 0.334% ÷ 0.256% = 1.30.

We estimated the rate of sequencing error in the chimpanzee sequence as follows. We assume that the substitution rates are equal between humans and chimpanzees. Thus,$R_{C/H} \frac{(0.59\% - x)/2 + x}{(0.59\% - x)/2}$, where *x* is the error rate. Given the observed $R_{C/H}$ of 1.15, we estimated that *x* = 0.04%. Under the assumption of a 3% reduction of substitution rate in humans than in chimpanzees, the sequencing error rate is estimated to be 0.033%. The above $R_{C/H}$ was estimated using all substitutions. If only synonymous substitutions are used, $R_{C/H}$ = 1.10. Thus, the error rate may be lower than the above estimates, which makes our conclusion more conservative (see main text).

Microarray data

The microarray gene expression data for 73 human normal tissues and the nucleotide sequences for 27,215 probesets were obtained from ref. (Su et al. 2004). The probeset sequences were used to perform BLAST searches against the human coding sequences from Ensembl. Probesets that matched to multiple genes were considered ambiguous and discarded. 26,195 probesets were unambiguously matched to 16,605 distinct genes. Of these genes, 12,145 had human-chimpanzee-macaque alignments. For genes that matched to more than one probeset, the expression levels measured in each probeset were averaged for each tissue replicate. Two replicates were available for each tissue; these were averaged to determine the expression level of a gene in a tissue. Multiple tissues representing same organs were consolidated into tissue groups (Table 2.3). For any given gene, a single representative expression level was used for a tissue group by taking the highest expression level from the group.

The microarray expression data was analyzed in two ways. First, we identified 2,432 genes for which the expression level in the highest tissue group was equal to or greater than two times the expression level in the second highest tissue group. These genes are said to be tissue-specific in the highest tissue. All other genes in our primate gene dataset were treated as non-tissue-specific genes. Second, we used a more stringent criterion of tissue-specificity, requiring that the expression level in the highest tissue group be at least four times the expression level in the second highest tissue group.

EST and SAGE data

The human expression sequence tag (EST) and serial analysis of gene expression (SAGE) data were compiled by Semon et al. (Semon et al. 2006) and kindly supplied by the authors. Semon et al. (Semon et al. 2006) selected from GenBank 4.9 million ESTs from human tissues. cDNA libraries from cell culture, tumors, pooled organs, or unidentified tissues were excluded, and cDNA libraries that had been sufficiently sampled (>10,000 ESTs) were retained, resulting in 44 tissues corresponding to 141 libraries. These authors further removed those tissues with fewer than 30 tissue-specific genes, resulting in the final dataset of 2,126 tissue-specific genes from 18 tissues. The SAGE data contained 1,190 tissue-specific genes from 12 tissues. In both of these datasets, a gene was considered to be tissue-specific if its transcript was detected in only one tissue. All other genes were treated as non-tissue-specific.

Dorus et al.'s (2004) data

We also acquired the list of nervous system genes from Dorus et al. (Dorus et al. 2004). Of 214 genes in this dataset identified by gene symbol, we were able to find 146 that are present in our primate gene dataset.

Simulation tests

As described in the main text, we defined ω_H and ω_C by 0.408n/s for a group of genes in the human and chimpanzee lineages, respectively. We used the χ^2 test with 1 degree of freedom to compare ω_H and ω_C. We compared ω_H/ω_C values between two groups of genes by computer simulation. For example, we compute the ratio (r) of the $\omega H/\omega C$ value from brain-specific genes to that of other tissue-specific genes and then examine the sampling variance of r by simulation. It is obvious that n and s are both Poisson random variables when the rate of substitution is given. For a given n (or s), we generate 10,000 Poisson random numbers with the mean equal to the observed n (or s). Thus, we can obtain 10,000 random r values, which represent the variation of r. A two-tail test is then conducted. That is, we consider our observed r to be significantly different from 1 at the 5% level when fewer than 250 randomly generated r values are greater than 1 or smaller than 1. Use of one-tail tests does not change any of our conclusions in the paper. The Q20 dataset We also compiled a dataset by using those nucleotide sites in the chimpanzee genome sequence with quality scores ≥20 (Q20, or accuracy >0.99). Chimpanzee codons retained for analysis must have quality scores ≥20 at all codon positions. The quality score information was downloaded from http://www.genome.ucsc.edu. The data were further cleaned using the same criteria as described in the first section. The final dataset contained 13,888 genes.

Acknowledgements

I worked with Peng Shi and Jianzhi Zhang on the research described in this chapter. The paper we co-authored was published in *Trends in Genetics* (Shi et al. 2006). We thank Laurent Duret for providing the EST and SAGE datasets, the macaque genome sequencing team for making the draft sequence available, and Soochin Cho, Wendy Grus and Ben-Yang Liao for valuable comments. This work was supported by research grants from the University of Michigan and the National Institutes of Health to J.Z.

Figure 2.1 Evolutionary tree of human, chimpanzee and macaque monkey.
Also shown are the brains of the three species drawn to scale and the encephalization quotients (EQs). The EQ measures the brain mass relative to the total body mass and is computed by E/Pa, where E is the brain mass, P is the body mass, and a is the exponent. The EQ values are taken from (Williams 2002); a = 0.75 on the basis of previous analyses of primates (Martin 1981) or catarrhine primates (i.e. humans, apes and Old World monkeys) (Pagel and Harvey 1989). The brain images are adapted from those in the Comparative Mammalian Brain Collections (http://brainmuseum.org).

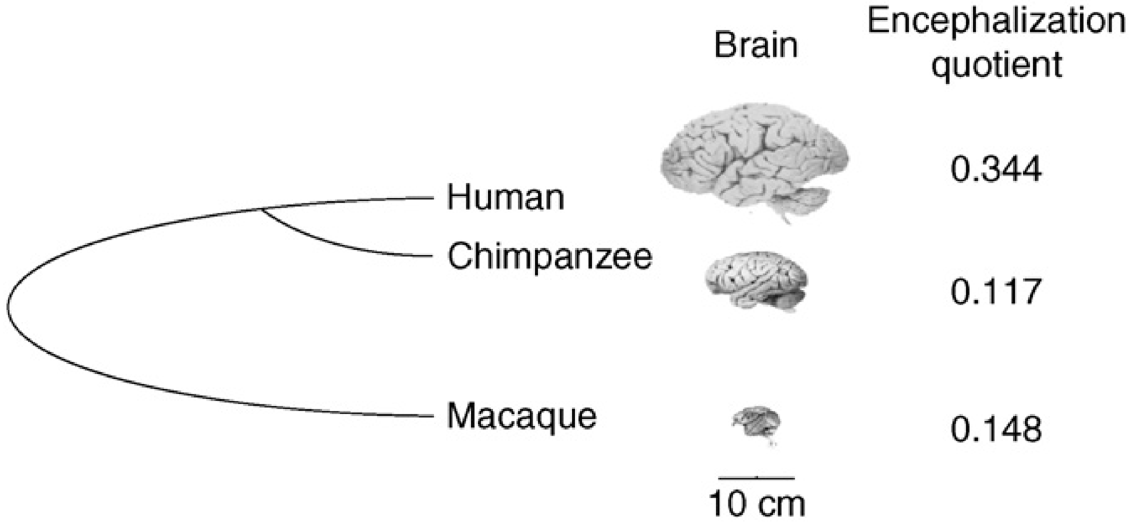


Figure 2.2 Venn diagram of brain-specific genes identified on the basis of five different definitions.

The number of genes is given in each circle. The overlapping sets of the nervous system genes (from Dorus et al. (Dorus et al. 2004); colored red) are shown in separate circles because of the difficulty in connecting all of the circles.

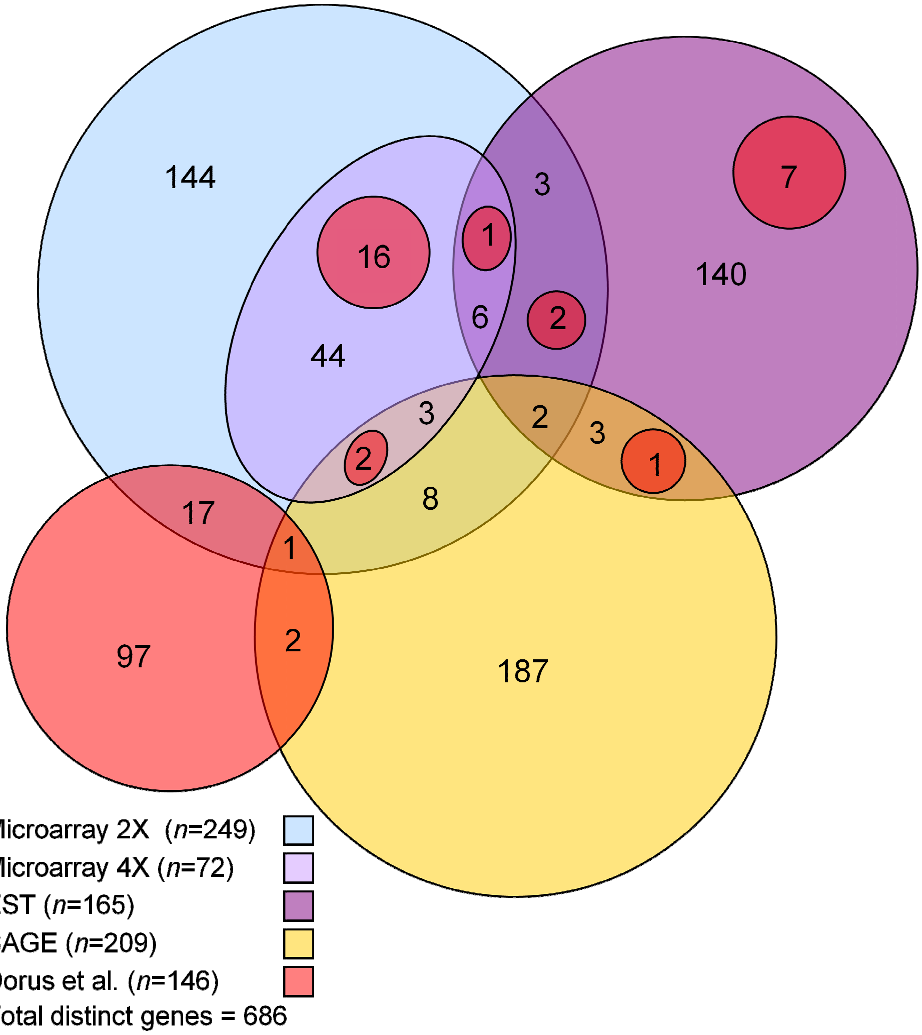
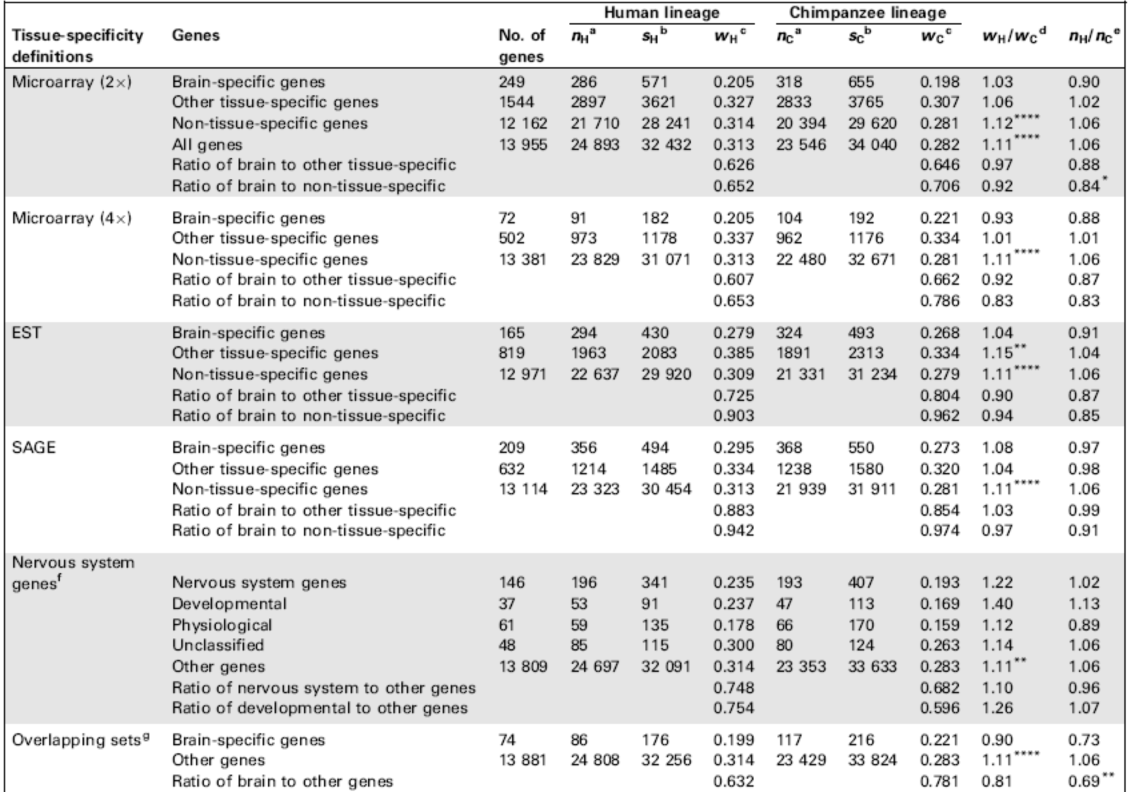


Table 2.1 Evolutionary rates of brain-specific genes and other genes in humans and chimpanzees



			Human lineage			Chimpanzee lineage				
Tissue-specificity definitions	Genes	No. of genes	n_H[a]	s_H[b]	w_H[c]	n_C[a]	s_C[b]	w_C[c]	w_H/w_C[d]	n_H/n_C[e]
Microarray (2×)	Brain-specific genes	249	286	571	0.205	318	655	0.198	1.03	0.90
	Other tissue-specific genes	1544	2897	3621	0.327	2833	3765	0.307	1.06	1.02
	Non-tissue-specific genes	12 162	21 710	28 241	0.314	20 394	29 620	0.281	1.12****	1.06
	All genes	13 955	24 893	32 432	0.313	23 546	34 040	0.282	1.11****	1.06
	Ratio of brain to other tissue-specific				0.626			0.646	0.97	0.88
	Ratio of brain to non-tissue-specific				0.652			0.706	0.92	0.84*
Microarray (4×)	Brain-specific genes	72	91	182	0.205	104	192	0.221	0.93	0.88
	Other tissue-specific genes	502	973	1178	0.337	962	1176	0.334	1.01	1.01
	Non-tissue-specific genes	13 381	23 829	31 071	0.313	22 480	32 671	0.281	1.11****	1.06
	Ratio of brain to other tissue-specific				0.607			0.662	0.92	0.87
	Ratio of brain to non-tissue-specific				0.653			0.786	0.83	0.83
EST	Brain-specific genes	165	294	430	0.279	324	493	0.268	1.04	0.91
	Other tissue-specific genes	819	1963	2083	0.385	1891	2313	0.334	1.15**	1.04
	Non-tissue-specific genes	12 971	22 637	29 920	0.309	21 331	31 234	0.279	1.11****	1.06
	Ratio of brain to other tissue-specific				0.725			0.804	0.90	0.87
	Ratio of brain to non-tissue-specific				0.903			0.962	0.94	0.85
SAGE	Brain-specific genes	209	356	494	0.295	368	550	0.273	1.08	0.97
	Other tissue-specific genes	632	1214	1485	0.334	1238	1580	0.320	1.04	0.98
	Non-tissue-specific genes	13 114	23 323	30 454	0.313	21 939	31 911	0.281	1.11****	1.06
	Ratio of brain to other tissue-specific				0.883			0.854	1.03	0.99
	Ratio of brain to non-tissue-specific				0.942			0.974	0.97	0.91
Nervous system genes[f]	Nervous system genes	146	196	341	0.235	193	407	0.193	1.22	1.02
	Developmental	37	53	91	0.237	47	113	0.169	1.40	1.13
	Physiological	61	59	135	0.178	66	170	0.159	1.12	0.89
	Unclassified	48	85	115	0.300	80	124	0.263	1.14	1.06
	Other genes	13 809	24 697	32 091	0.314	23 353	33 633	0.283	1.11**	1.06
	Ratio of nervous system to other genes				0.748			0.682	1.10	0.96
	Ratio of developmental to other genes				0.754			0.596	1.26	1.07
Overlapping sets[g]	Brain-specific genes	74	86	176	0.199	117	216	0.221	0.90	0.73
	Other genes	13 881	24 808	32 256	0.314	23 429	33 824	0.283	1.11****	1.06
	Ratio of brain to other genes				0.632			0.781	0.81	0.69**

a Number of nonsynonymous substitutions in the lineage indicated.

b Number of synonymous substitutions in the lineage indicated.

c Nonsynonymous/synonymous substitution rate ratio, computed by 0.408n/s.

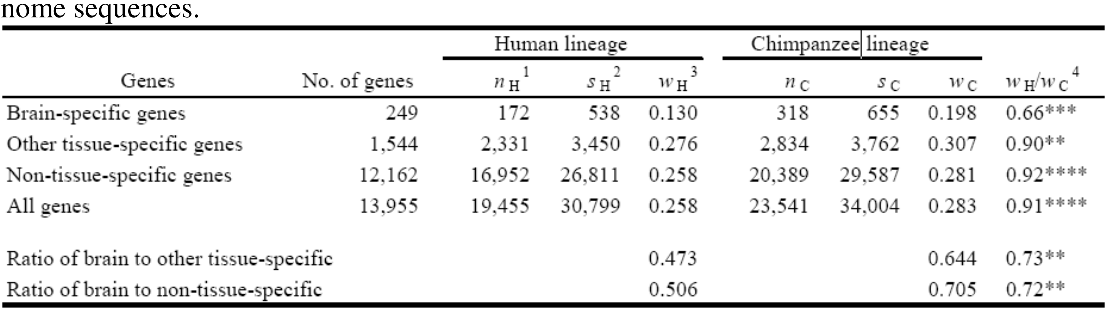
d Statistically significant deviation from 1 is indicated by asterisks: Significance level: *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001. Simulation tests are used for comparing ratios between groups of genes, whereas χ^2-tests are used within groups of genes.

e Statistically significant deviation from 1 (between groups of genes) is indicated by asterisks. χ^2-tests are used.

f From Dorus et al. (Dorus et al. 2004).

g Genes identified to be brain-specific in at least two of the four definitions ('Microarray [2X]', 'EST', 'SAGE' and 'Nervous system genes').

Table 2.2 Comparison of the evolutionary rates of brain-specific genes and other genes in human and chimpanzee lineages using microarray 2X definition of tissue-specificity and unmodified genome sequences.



		Human lineage			Chimpanzee lineage			
Genes	No. of genes	n_H[1]	s_H[2]	w_H[3]	n_C	s_C	w_C	w_H/w_C[4]
Brain-specific genes	249	172	538	0.130	318	655	0.198	0.66***
Other tissue-specific genes	1,544	2,331	3,450	0.276	2,834	3,762	0.307	0.90**
Non-tissue-specific genes	12,162	16,952	26,811	0.258	20,389	29,587	0.281	0.92****
All genes	13,955	19,455	30,799	0.258	23,541	34,004	0.283	0.91****
Ratio of brain to other tissue-specific				0.473			0.644	0.73**
Ratio of brain to non-tissue-specific				0.506			0.705	0.72**

[1] Number of nonsynonymous substitutions in the lineage indicated.

[2] Number of synonymous substitutions in the lineage indicated.

[3] Nonsynonymous/synonymous substitution rate ratio, computed by 0.408n /s .

[4] Statistically significant deviation from 1 is indicated by asterisks. Significance level: *, 5%; **, 1%; ***, 0.1%; ****, 0.01

Simulation tests are used in the bottom two rows, whereas $\chi 2$ tests are used in all other rows.

Table 2.3 Tissue groups considered in the analysis of the microarray data

Tissue Group	Tissues Combined
721B lymphoblasts	721B lymphoblasts
Adipocyte	Adipocyte
Adrenal cortex	Adrenal cortex
Adrenal gland	Adrenal gland
Appendix	Appendix
Atrioventricular node	Atrioventricular node
BM	BMCD33 myeloid, BMCD105 endothelial, BMCD34, BMCD71 early erythroid
Bone marrow	Bone marrow
Brain	temporal lobe, globus pallidus, cerebellum peduncles, cerebellum, caudate nucleus, whole brain, parietal lobe, medulla oblongata, amygdala, prefrontal cortex, occipital lobe, hypothalamus, thalamus, subthalamic nucleus, cingulated cortex, pons, fetal brain
Bronchial epithelial cells	Bronchial epithelial cells
Cardiac myocytes	Cardiac myocytes
Ciliary ganglion	Ciliary ganglion
DRG	DRG
Heart	Heart
Kidney	Kidney
Liver	liver, fetal liver
Lung	lung, fetal lung
Lymph node	Lymph node
Olfactory bulb	Olfactory bulb
Ovary	Ovary
Pancreas	pancreas, pancreatic islets
Pathogenic (excluded)	colorectal adenocarcinoma, leukemia lymphoBLASTic molt 4, lymphoma burkitts Raji, leukemia promyelocytic hl60, lymphoma burkitts Daudi, leukemia chronicmyelogenous k562
PB	PBCD14 monocytes, PBBDCA4 dentritic cells, PBCD56NK cells, PBCD4T cells, PBCD8T cells, PBCD19B cells
Pituitary	Pituitary
Placenta	Placenta
Prostate	Prostate
Salivary gland	Salivary gland
Skeletal muscle	Skeletal muscle
Skin	Skin
Smooth muscle	Smooth muscle
Spinal cord	Spinal cord
Superior cervical ganglion	Superior cervical ganglion
Testis	testis, leydig cell, germ cell, testis interstitial, seminiferous tubule
Thymus	Thymus
Thyroid	thyroid, fetal thyroid
Tongue	Tongue
Tonsil	Tonsil
Trachea	Trachea
Trigeminal Ganglion	Trigeminal Ganglion
Uterus	Uterus
Uterus Corpus	Uterus Corpus
Whole blood	Whole blood

Table 2.4 Comparisons between evolutionary rates of brain-specific genes and other genes in human and chimpanzee lineages using high quality (Q20) chimpanzee sequences.

Tissue-specificity			Human lineage			Chimpanzee lineage				
definitions	Genes	No. of genes	n_H[1]	s_H[2]	w_H[3]	n_C[1]	s_C[2]	w_C[3]	w_H/w_C[4]	n_H/n_C[5]
Microarray (2X)	Brain-specific genes	249	169	530	0.130	164	548	0.122	1.07	1.03
	Other tissue-specific genes	1,536	2,285	3,361	0.277	2,210	3,286	0.275	1.01	1.03
	Non-tissue-specific genes	12,103	16,546	26,192	0.258	15,327	25,810	0.242	1.06****	1.08
	All genes	13,888	19,000	30,083	0.258	17,701	29,644	0.244	1.06****	1.07
	Ratio of brain to other tissue-specific				0.469			0.445	1.05	1.00
	Ratio of brain to non-tissue-specific				0.505			0.504	1.00	0.95
Microarray (4X)	Brain-specific genes	72	57	170	0.137	48	166	0.118	1.16	1.19
	Other tissue-specific genes	497	783	1,097	0.291	772	1,028	0.306	0.95	1.01
	Non-tissue-specific genes	13,319	18,160	28,816	0.257	16,881	28,450	0.242	1.06****	1.08
	Ratio of brain to other tissue-specific				0.470			0.385	1.22	1.17
	Ratio of brain to non-tissue-specific				0.532			0.487	1.09	1.10
EST	Brain-specific genes	165	219	399	0.224	209	401	0.213	1.05	1.05
	Other tissue-specific genes	811	1,610	1,931	0.340	1,505	2,030	0.303	1.12*	1.07
	Non-tissue-specific genes	12,912	17,171	27,753	0.253	15,987	27,213	0.240	1.05***	1.07
	Ratio of brain to other tissue-specific				0.658			0.703	0.94	0.98
	Ratio of brain to non-tissue-specific				0.887			0.887	1.00	0.98
SAGE	Brain-specific genes	208	259	454	0.233	264	460	0.234	0.99	0.98
	Other tissue-specific genes	627	937	1,382	0.277	949	1,359	0.285	0.97	0.99
	Non-tissue-specific genes	13,053	17,804	28,247	0.257	16,488	27,825	0.242	1.06****	1.08
	Ratio of brain to other tissue-specific				0.841			0.822	1.02	0.99
	Ratio of brain to non-tissue-specific				0.905			0.969	0.93	0.91
Dorus et al. (2004)	Nervous-system genes	146	131	313	0.171	128	346	0.151	1.13	1.02
	Developmental	37	34	83	0.167	22	88	0.102	1.64	1.55
	Physiological	61	36	127	0.116	51	148	0.141	0.82	0.71
	Unclassified	48	61	103	0.242	55	110	0.204	1.18	1.11
	Other genes	13,742	18,869	29,770	0.259	17,573	29,298	0.245	1.06****	1.07
	Ratio of nervous-system to other genes				0.660			0.617	1.07	0.95
	Ratio of developmental to other genes				0.646			0.417	1.55	1.44
Overlapping sets[6]	Brain-specific genes	74	50	162	0.126	50	161	0.127	0.99	1.00
	Other genes	13,814	18,950	29,921	0.259	17,651	29,483	0.244	1.06****	1.07
	Ratio of brain to other genes				0.487			0.519	0.94	0.93

[1] Number of nonsynonymous substitutions in the lineage indicated.

[2] Number of synonymous substitutions in the lineage indicated.

[3] Nonsynonymous/synonymous substitution rate ratio, computed by 0.408*n* /*s* .

[4] Statistically significant deviation from 1 is indicated by asterisks. Significance level: *, 5%; **, 1%; ***, 0.1%; ****, 0.01%. Simulation tests are used for comparing ratios between groups of genes, whereas χ^2 tests are used within groups of genes.

[5] Statistically significant deviation from 1 (between groups of genes) is indicated by asterisks. χ^2 tests are used.

[6] Genes identified to be brain-specific in at least two of the four definitions (microarray 2X, EST, SAGE, and Dorus et al.).

Chapter 3

More genes underwent positive selection in chimpanzee evolution than in human evolution

Abstract

Observations of numerous dramatic and presumably adaptive phenotypic modifications during human evolution prompt the common belief that more genes have undergone positive Darwinian selection in the human lineage than in the chimpanzee lineage since their evolutionary divergence 6–7 million years ago. Here, we test this hypothesis by analyzing nearly 14,000 genes of humans and chimps. To ensure an accurate and unbiased comparison, we select a proper outgroup, avoid sequencing errors, and verify statistical methods. Our results show that the number of positively selected genes is substantially smaller in humans than in chimps, despite a generally higher nonsynonymous substitution rate in humans. These observations are explainable by the reduced efficacy of natural selection in humans because of their smaller long-term effective population size but refute the anthropocentric view that a grand enhancement in Darwinian selection underlies human origins. Although human and chimp positively selected genes have different molecular functions and participate in different biological processes, the differences do not ostensibly correspond to the widely assumed adaptations of these species, suggesting how little is currently known about which traits have been under positive selection. Our analysis of the identified positively selected genes lends support to the association between human Mendelian diseases and past adaptations but provides no evidence for either the chromosomal speciation hypothesis or the widespread brain-gene acceleration hypothesis of human origins.

Introduction

Although humans and their closest living relatives, chimpanzees, are highly similar at the genomic level (Chen and Li 2001; Britten 2002; Ebersberger et al. 2002; Wildman et al. 2003; Watanabe et al. 2004; Consortium 2005), they differ in many morphological, physiological, and behavioral traits (Varki and Altheide 2005). Phenotypically, modern humans appear to have changed considerably more than modern chimps from their common ancestors (King and Wilson 1975; Pilbeam 1996; Olson and Varki 2003; Varki and Altheide 2005). Many of these evolutionary modifications in humans, such as the origins of bipedalism, speech and language, and other high-order cognitive functions, are widely thought to be adaptive (Darwin 1871; Vallender and Lahn 2004; Sabeti et al. 2006). These observations led to a common belief that more genes underwent positive Darwinian selection in the human lineage than in the chimpanzee lineage. Indeed, there are more reports of positively selected genes (PSGs) in humans than in chimps (Vallender and Lahn 2004; Sabeti et al. 2006). Nonetheless, this difference may be largely due to a lack of study in chimps. To avoid such a bias, one could identify and compare all PSGs from the human and chimp genomes. Positive selection acting on a protein-coding gene may be detected by various population genetic and molecular evolutionary methods that use intraspecific polymorphism data, interspecific divergence data, or a combination of the two (Li 1997; Nei and Kumar 2000; Nielsen 2005). However, because of the paucity of polymorphism data from chimps, a fair comparison between the two species would have to be limited to the divergence data. Such data can be used to estimate the ratio of nonsynonymous to synonymous substitution rates (ω). An ω value significantly >1 indicates the action of positive selection, whereas an ω significantly <1 indicates negative (or purifying) selection. Using this approach, two earlier studies (Clark et al. 2003; Arbiza et al. 2006) pioneered the identification of human and chimp PSGs at the genomic scale, although no comparison was made between the numbers of human and chimp PSGs. In fact, the studies' results would be unsuitable for the comparison, owing to a number of deficiencies. First, both studies used the mouse as an outgroup, to distinguish between human-specific and chimp-specific nucleotide substitutions, because of the unavailability of genome sequences from any closer outgroups at that time. Because mouse is distantly related to human and chimp, this practice introduces errors. Second, one of the studies (Clark et al. 2003) was based on less reliable statistical methods and assumptions (Zhang 2004), whereas the other (Arbiza et al. 2006) used the draft chimp genome sequence (Consortium 2005) known to contain many more errors than the finished human genome sequence (Shi et al. 2006; Taudien et al. 2006). Because the majority of genes in a genome have $\omega < 1$, and sequencing errors have an expected ω of 1, the errors inflate ω and the false detection of positive selection. In this work, we first design a protocol to rectify these problems and then use the protocol to identify and compare human and chimp PSGs. Our results show substantively more PSGs in chimpanzee evolution than in human evolution.

Results and Discussion

Study Design

To compare human and chimp PSGs impartially, we made three improvements in the design of the analysis. First, to distinguish nucleotide substitutions that occurred in the human lineage from those that occurred in the chimp lineage, we used the macaque monkey as the outgroup. Because the divergence time between the macaque and human/chimp is approximately a quarter of that between the mouse and human/chimp (Goodman et al. 1998; Hedges 2002; Glazko and Nei 2003), the reliability of our analysis was expected to increase significantly. Gene orthology determination and sequence alignment among the more closely related human–chimp–macaque gene trios is also more reliable than among human–chimp–mouse trios.

Second, we applied an improved branch-site likelihood method for identifying PSGs (Zhang et al. 2005), which has been shown by computer simulation to produce good results even when some of the assumptions are violated (Zhang et al. 2005). The method requires that the branches in a phylogenetic tree be separated into foreground and background branches a priori, where foreground branches are tested for the occurrence of positive selection. The method assumes that two classes of codons, either negatively selected (class 0) or neutral (class 1), exist in the background branches. This null model is compared with an alternative model in which a proportion of class 0 codons, and the same proportion of class 1 codons, become positively selected in the foreground branches. Positive selection in foreground branches is inferred for a gene if the likelihood of the observation of the gene sequences is significantly higher under the alternative model than under the null model. To further verify the suitability of the method in the present context, we conducted additional computer simulations specifically designed to mimic the evolution of human, chimp, and macaque genes (see Materials and Methods). Our results showed that the false-positive rate is acceptable, except for extreme conditions when it slightly exceeds the nominal rate (Table 3.3 and Table 3.4).

Third, we used high-quality nucleotides from the 4X coverage chimp genome sequence to allow a fair comparison with the human sequence. Briefly, we assembled alignments of orthologous genes from human, chimp, and macaque, using publicly available genome sequences and annotations (see Materials and Methods). We then eliminated alignment gaps and those codons in which one or more chimp nucleotides did not meet our quality cutoff. Three different cutoffs, low (Q0), intermediate (Q10), and high (Q20), were used to generate three data sets. After removing alignments of <100 codons, we obtained our final data sets, containing 13,955, 13,924, and 13,888 genes for the Q0, Q10, and Q20 cutoffs, respectively (Table 3.5). Even the smallest data set (Q20) has a total alignment length of 17,995,887 nucleotides, with a mean alignment length of 432 codons (standard deviation, 339 codons). All three data sets contain >50% of genes in a primate genome and cover >50% of all protein-coding regions in the genome. Using parsimony, we inferred the numbers of nucleotide substitutions in human and chimp lineages since their split.

This inference is expected to be accurate because the three species studied here are closely related. We found that the ratio of the number of synonymous substitutions in the chimp lineage to that in the human lineage is *r* = 1.103 ± 0.009, 1.020 ± 0.008, and 0.985 ± 0.008 for the Q0, Q10, and Q20 data sets, respectively. Assuming identical mutation rates per year between human and chimp lineages, *r* is expected to be 1. If the mutation rate is 3% lower in humans than in chimps, as has been suggested (Elango et al. 2006), *r* is expected to be 1.03. Given these considerations, Q0 data, as used in an earlier study (Arbiza et al. 2006), are apparently unsuitable because the observed *r* is significantly higher than the expectation. To make our conclusion more conservative, we use Q20 rather than Q10 data. Two other independent assessments of the chimp genome sequence, one of which evaluated it against 172 kb of finished chimp sequence, also recommended the use of Q20 data for comparison with the human genome sequence (Consortium 2005; Taudien et al. 2006). Most importantly, the number of synonymous substitutions is already 1.5% lower in chimp than in human when the cutoff of Q20 is used, suggesting that the chimp sequencing errors become negligible at this quality level. The comparison between the 172 kb of draft and finished chimp sequences also showed that the use of cutoffs higher than Q20 is undesirable because many chimp-specific nucleotide changes tend to be lost (Taudien et al. 2006). This is probably because polymorphic sites in the chimp individual that was sequenced, estimated to be 0.1% of all sites (Consortium 2005), tend to have lower qualities than homozygous sites. These polymorphic sites are excluded progressively as one increases the quality cutoff, which hampers a fair comparison with human because the human genome sequence contains polymorphic sites (Consortium 2005). Note that errors in the macaque genome sequence should not affect our analysis because the probability for a macaque error to occur at a nucleotide position where human and chimp differ is small. Even when such rare events occur, they should affect human and chimp equally and hence would not bias our results. Our human–chimp comparison should not be biased by indel errors because the detection of positive selection does not use indel information.

More PSGs in Chimp Evolution than in Human Evolution

Applying the likelihood method and a P value of 5% for statistical significance (Zhang et al. 2005), we identified 154 genes that were under positive selection in the human lineage (Table 3.1 and Appendix) and 233 in the chimp lineage (Appendix). Thus, chimps have 51% more PSGs than humans have. As expected, the excess of chimp PSGs is even greater (157%) should the Q10 data be used (Table 3.5). The proportion of PSGs in the genome is 233/13,888 = 1.7% for the chimp lineage, significantly greater than that (154/13,888 = 1.1%) for the human lineage ($P < 10^{-4}$, χ^2 test). Because 13,888 statistical tests were conducted for each lineage, it is necessary to control for multiple testing. Under Bonferroni correction, two human genes and 21 chimp genes remain statistically significant (Appendix). With use of a false discovery rate of 5%, the same two human genes and 59 chimp genes are significant (Appendix). The proportion of PSGs in the chimp genome remains significantly greater than that in the human genome ($P = 10^{-4}$, χ^2 test), even after the multiple-testing corrections (Table 3.1).

To further confirm our results, we analyzed the recently released 6X chimp genome assembly for the 233 chimp PSGs identified above. We found that 212 (or 91%) of them still show significant signals of positive selection (see Materials and Methods). Hence, when this new data set is used, chimps have 38% more PSGs than humans have ($P = 0.002$, χ^2 test). Note that this is a conservative estimate because we did not consider non-PSGs from the 4X sequence that may become PSGs in the 6X sequence. Such incidences are possible because potentially more nucleotides per gene can be analyzed in the 6X sequence, leading to improved statistical power in identifying PSGs. Additionally, 4X and 6X sequences may differ at polymorphic sites, which can affect the outcome of PSG identification when the number of substitutions is small. Because the analyses of the 4X and 6X sequences both indicate substantially more PSGs in chimps than in humans, and because the 6X assembly is preliminary and unpublished, our subsequent analyses use the PSGs identified from the Q20 data of the 4X assembly. An additional reason for using the

4X assembly is the finding of a number of cases in which the 4X assembly is apparently more accurate than the 6X assembly (see Materials and Methods).

We found that the mean ω of all genes is 0.259 ± 0.002 in the human lineage, significantly larger than that (0.245 ± 0.002) in the chimp lineage ($P < 10^{-4}$; Table 3.1). For the common set of 13,508 non-PSGs between humans and chimps, the mean ω is also significantly larger in human (0.252 ± 0.002) than in chimp (0.238 ± 0.002) ($P < 10^{-4}$; Table 3.1). Because the majority of non-PSGs are under negative selection, as reflected in their low ω values, the above results indicate stronger negative selection in chimps than in humans. Multiple-population genetic data indicate that the long- term effective population size of humans (in the last 1–2 million years) is several-fold smaller than that of chimps and than that of the human–chimp common ancestor ((Ferris et al. 1981; Takahata et al. 1995; Ruvolo 1997; Kaessmann et al. 1999; Chen and Li 2001; Kaessmann et al. 2001; Stone et al. 2002; Wall 2003; Fischer et al. 2004)). A recent analysis of 1 million base pairs of Neanderthal nuclear DNA also suggested that the common ancestor of modern humans and Neanderthals had a small effective population size (Green et al. 2006). It is thus probable that the effective population size is greater in the chimp lineage than in the human lineage for a large portion of the divergence time between the two lineages. Population genetic theories (Kimura 1983b) predict that both positive and negative selection are more effective in large populations than in small populations. Our observation that chimps have more PSGs but fewer nonsynonymous substitutions in non- PSGs than humans is consistent with these predictions.

Computer simulations showed that the branch-site likelihood method cannot detect all PSGs. Rather, the detection rate increases as the ω of background branches increases (Table 3.6). If the overall strength of positive selection is weaker in humans than in chimps because of smaller populations of humans than chimps, a higher average background ω is required for PSGs to be detectable in humans than in chimps. We found that in the macaque branch of the human–chimp–

macaque tree, the mean ω for all genes is 0.226±0.001. For human PSGs, the mean ω in the macaque branch is 0.294 ± 0.007, significantly greater than the mean ω in the macaque branch (0.278±0.005) for chimp PSGs (*P*<0.05). Hence, these observations are consistent with the simulation result and further support the notion that positive selection was weaker in the human lineage than in the chimp lineage. Theories also predict that recombination can increase the efficacy of selection (Hill and Robertson 1966). Indeed, PSGs tend to be located in high-recombination regions, although this effect is significant in chimps (*P* = 0.041) but not in humans (*P* = 0.32) (Figure 3.7), probably as a result of a difference in statistical power caused by the difference in the number of PSGs in the two species.

Similarities and Differences Between Human and Chimp PSGs

It has been claimed that genes of certain functional categories, such as olfaction and nuclear transport, were more frequently under positive selection in humans than in chimps, based on the ranking of all genes by their P values in the likelihood test of positive selection (Clark et al. 2003). Because genes with reduced negative selection also tend to have low *P* values (although unlikely to be as low as 0.05), such ranks potentially mix genes under positive selection with those under reduced negative selection. We took a more rigorous approach by limiting our analysis to the PSGs we detected. We found that seven genes are shared between the human and chimp PSGs (Table 3.7), significantly greater than expected by chance (2.6; *P* < 0.02, binomial test), suggesting the presence of some common targets of positive selection in the two lineages. We classified all PSGs into biological process groups and molecular function groups, as defined in the PANTHER database (Mi et al. 2005). A randomization test indicated a significant difference in distribution of human and chimp nonoverlapping PSGs among biological process groups (Figure 3.1A) and among molecular function groups (Figure 3.1B). Those groups showing the greatest differences between the two species are listed in Figure 3.1C. Interestingly, however, the majority of these groups (e.g., protein metabolism and modification, anion transport, phosphate transport, and lyase) do not correspond to the widely assumed adaptive phenotypic differences between humans and chimps (e.g., neurogenesis), suggesting the existence of yet-to-be-recognized adaptive phenotypic differences between the two species. We did not detect several previously reported PSGs that control brain size or cognitive functions (Enard et al. 2002b; Zhang et al. 2002; Zhang 2003; Evans et al. 2004b) because previous identifications of these PSGs were based on a comparison of polymorphism and divergence data, whereas only divergence data are used here. As mentioned above, due to the paucity of chimp polymorphism data, any fair genomewide comparison of human and chimp PSGs would have to be limited to divergence data at this time.

Using microarray data of human gene expression, we found that human and chimp PSGs are not significantly different in their distributions between the categories of tissue-specific genes and nonspecific genes ($P > 0.5$, χ^2 test; Table 3.8). On examining the peak-expression tissue group for each gene (Appendix), we again found no significant difference in the overall tissue distribution between human and chimp PSGs (Figure 3.2). Notably, 14 (11%) human PSGs and 13 (6.7%) chimp PSGs have peak expressions in one or more parts of the brain, but the difference is not statistically significant ($\chi^2 = 1.74$, $P = 0.19$). On the contrary, for the central nervous system outside of the brain, human (8) has fewer PSGs than chimp (14) ($\chi^2 = 0.09$, $P = 0.77$). These findings are consistent with recent comparative genomic analyses (Shi et al. 2006; Wang et al. 2007) and do not support more positive selection in humans than in chimps in regard to nervous system genes (Dorus et al. 2004).

Genome-wide identification of human and chimp PSGs helps to test several evolutionary hypotheses. First, it has been argued that PSGs are more likely than non-PSGs to underlie known Mendelian disorders in humans because the current environment of humans is considerably different from that of earlier hominins and previous adaptive changes may become deleterious today (Neel 1962; Young et al. 2005). Our data provide some support for this hypothesis. We found that 9.7% of human PSGs are disease-associated (Appendix), significantly greater than that (6.1%) among the non-PSGs examined (*P* = 0.049; Table 3.2). Consistent with the prediction of the above hypothesis, the fraction of human PSGs underlying human diseases is greater than the fraction of chimp PSGs underlying human diseases (*P* = 0.044, Fisher's exact test). Furthermore, as expected, there is no significant difference in the proportion of genes underlying human diseases between chimp PSGs and non-PSGs (*P* = 0.23; Table 3.2).

Second, a recently proposed chromosomal speciation hypothesis asserts that chromosomal rearrangements facilitated the formation of reproductive isolation between populations that eventually led to modern humans and chimps (Navarro and Barton 2003). Several predictions of this hypothesis have been examined, with mixed results (Lu et al. 2003; Navarro and Barton 2003; Marques-Bonet et al. 2004; Zhang et al. 2004; Osada and Wu 2005; Innan and Watanabe 2006; Patterson et al. 2006). One interesting prediction that has not been explicitly tested is that PSGs are preferentially located on rearranged chromosomes because such chromosomes are less likely to be introgressed after the initial separation of two lineages during speciation and thus are more likely to accumulate genes subject to local adaptations (Navarro and Barton 2003). Nine chromosomes (1, 4, 5, 9, 12, 15–18) contain pericentric inversions between humans and chimpanzees, and human chromosome 2 resulted from a fusion of two acrocentric chromosomes common to other great apes (Yunis and Prakash 1982). These chromosomes are considered as rearranged chromosomes, whereas the other chromosomes are considered as colinear chromosomes. Our data, however, do not support the chromosomal speciation hypothesis for humans and chimps because the proportion of PSGs is even slightly lower on the rearranged chromosomes than on the colinear chromosomes in both the human and chimp lineages (Figure 3.3).

Implications

In summary, our genome-wide analysis showed that substantively more genes underwent positive selection in the chimp lineage than in the human lineage since their split. Although our study could not, and did not, detect all PSGs in human and chimp evolution, particularly those beneficial alleles that are yet to be fixed (Bustamante et al. 2005; Sabeti et al. 2006; Wang et al. 2006), it provides an unbiased comparison between the two lineages. Our results have several implications. First, in sharp contrast to common belief, there were more adaptive genetic changes during chimp evolution than during human evolution. Without doubt, we tend to notice and study human-specific phenotypes more than chimp-specific phenotypes, which may have resulted in the prevailing anthropocentric view on human origins. Our finding suggests more unidentified phenotypic adaptations in chimps than in humans. Although human and chimp PSGs show different distributions among molecular functions and biological processes, the differences do not ostensibly correspond to the widely assumed adaptive phenotypes in humans. Assuming that our statistical method is equally powerful in detecting PSGs of different biological processes, the finding shows how little is currently known about which traits are adaptive. Second, although the influence of population size on negative selection has been well documented (Ohta 1995; Eyre-Walker and Keightley 1999), the present study also demonstrates the impact of population size on positive selection at the genomic scale. Interestingly, even during human evolution when so many apparently dramatic phenotypic changes took place, the laws of population genetics prevailed. This being said, it is important to recognize that other factors also influence the frequency of positive selection. For example, it is possible that as a result of the relatively recent out-of-Africa migration of modern humans, many new advantageous alleles are yet to be fixed and thus are not identified by our method. Our results thus apply largely to completed selective sweeps in human and chimp lineages. Furthermore, a higher level of polymorphism in chimps than in humans could potentially lead to more predicted PSGs in chimps than in humans. But because some chimp polymorphic sites have been removed in the Q20 data, and because the number of synonymous changes is already 1.5% lower in chimp than in human for the Q20 data, we do not think this factor has affected our result. At any rate, it will be interesting to examine in other species whether the number of PSGs is strongly dependent on population size. Third, although we only studied positive selection on protein sequence changes and did not address positive selection on gene expression evolution (Rockman et al. 2005; Khaitovich et al. 2006), a recent comparison between hominoids and murids in regard to regulatory sequence conservation showed that a reduction in population size also lowers the efficiency of natural selection on gene expression changes (Keightley et al. 2005). Most interestingly, when conserved noncoding sequences, which often regulate gene expression, are examined, chimps show more incidences of accelerated evolution than humans do (Prabhakar et al. 2006). Thus, it is likely that the total number of genes for which either the regulatory or coding regions underwent adaptive selection is also greater in chimp evolution than in human evolution.

Materials and Methods

Compilation of Human–Chimp–Macaque Gene Sequence Data

Protein and corresponding nucleotide sequences of all predicted genes in the human, chimpanzee, and macaque genome sequences were downloaded from Ensembl (version 36, December 2005; www. ensembl.org). To identify orthologous genes, human protein sequences (n = 33,869) were used to conduct BLASTP searches (Altschul et al. 1990) against the chimpanzee (n = 39,648) and macaque (n = 31,371) protein sequences. Reciprocal searches were performed using the chimpanzee and macaque proteins to query the human proteins. A total of 19,422 proteins with reciprocal best hits in both human/ chimpanzee and human/macaque searches were retained for further analysis. Alignment of the human–chimpanzee–macaque orthologous proteins was performed using CLUSTALW version 1.83 (Thompson et al. 1994). DNA sequence alignments were obtained by following the protein sequence alignments. Alignments containing <100 amino acids (n = 1,291) were discarded. Lineage-specific nucleotide substitutions were identified by parsimony as described in the next paragraph. Review of several alignments that had exceptionally high proportions of human- or chimpanzee-specific changes revealed that the apparent high level of lineage-specific changes resulted from incorrect alignment or nonorthology. Therefore, alignments containing >10% human- or chimpanzee-specific amino acid or nucleotide changes or >30% macaque-specific changes (n = 161) were discarded from analysis. Finally, each protein was assigned to a gene on the basis of the Ensembl annotation, and the protein sequence with the longest amino acid alignment was retained for each gene, resulting in the alignments of human, chimpanzee, and macaque sequences of 13,955 distinct genes (Q0 data set). Chimp genome sequence quality information was downloaded from the University of California, Santa Cruz, Bioinformatics web site (http://hgdownload.cse.ucsc.edu/goldenPath/panTro1/bigZips/chromQuals.zip). The average chimp quality score in the Q0 data set is 48.9526. The 13,955 alignments were scanned for codons in which one or more nucleotides had a chimp quality score <20 (i.e., an error rate of 1%) (Ewing et al. 1998), and these codons were removed from the alignments. After this procedure, 67 alignments contained <100 amino acids and were removed from analysis. The remaining 13,888 alignments constituted the Q20 data set. The average chimp quality score in the Q20 data set is 49.3443. We similarly obtained the Q10 data set (i.e., a maximum error rate of 10% at any nucleotide site), comprising 13,925 genes. The average chimp quality score in the Q10 data set is 49.0695.

We applied the parsimony principle to identify human-specific and chimpanzee-specific substitutions, using the macaque as the outgroup. The numbers of synonymous (s) and nonsynonymous (n) nucleotide substitutions in the human and chimp lineages were counted. Using the modified Nei–Gojobori method (Zhang et al. 1998) with a transition/transversion ratio of 2 (Rosenberg et al. 2003), we estimated that the total number of nonsynonymous sites in the

13,888 genes of the Q20 data set was N = 12,783,034 and the total number of synonymous sites was S = 5,215,415, with their ratio being N/S = 2.45. Thus, for a set of genes, the mean nonsynonymous-to-synonymous rate ratio in a lineage can be computed by (n/s)/(N/S) = (n/s)/2.45 = 0.41n/s.

Identification of PSGs

Using PAML (Yang 1997), we applied the improved branch-site test of positive selection (test 2 in ref. 25) to identify putative cases of positive selection in the human lineage among the 13,888 genes (Q20 data). When we tested positive selection in the human lineage, the human branch was designated as the foreground branch and the chimp and macaque branches were designated as background branches. We tested positive selection in the chimp lineage similarly. Bonferroni correction (Sokal and Rohlf 1995) and a false discovery rate of 5% (Storey and Tibshirani 2003) were used to correct for multiple testing. We also analyzed the Q10 data set and identified 165 human and 424 chimp PSGs.

Comparison Between Human and Chimp PSGs

Using the PANTHER database (Mi et al. 2005), we classified the 13,888 genes into different groups of biological processes and molecular functions. Note that these groups are not mutually exclusive and that a gene may belong to more than one group. To examine the distributional difference between human and chimp PSGs across PANTHER groups, we defined the statistic

$$2\sum_{i=1}^{n}(x_i - y_i)^2/(x_i + y_i)^2, \qquad [1]$$

where x_i and y_i are the number of human and chimp PSGs, respectively, in PANTHER group i, and n is the total number of PANTHER groups. Because of the nonindependence of PANTHER groups, we used a randomization test to examine whether the observed χ^2 was significantly different from the random expectation. Briefly, we randomly divided the 373 unshared human and chimp PSGs into 147 human PSGs and 226 chimp PSGs and computed χ^2 by using the above formula. We repeated this procedure 10,000 times to obtain the null distribution of χ^2, to which the observed χ^2 is compared. Similar results were obtained when the seven shared PSGs were included.

The microarray gene expression data in 79 human tissues, and the nucleotide sequences for 27,215 probe sets on the array, were obtained from (Su et al. 2004). The probe set sequences were used to perform BLAST searches against the human coding sequences annotated by Ensembl. Probe sets that matched to multiple genes were considered ambiguous and were discarded. A total of 26,195 probe sets were unambiguously matched to 16,605 distinct genes. Among these 16,605 genes, 12,099 genes, including 127 human PSGs and 195 chimp PSGs, can be found in our Q20 data set. For genes that matched to more than one probe set, the expression levels measured by different probe sets were averaged for each tissue replicate. Two replicates were available for each tissue, and these were averaged to determine the expression level of a gene in each tissue. Identification of tissue specificity can be obscured if multiple tissues with very similar expression profiles are used (Winter et al. 2004). We therefore consolidated multiple tissues representing similar areas into tissue groups and took the highest expression level from any tissue in a group as the single representative expression level score for the tissue group (Shi et al. 2006) (Table 2.3). Expression levels in pathogenic tissues were not considered. A gene was considered to be tissue-specific if the expression level in the highest tissue group was greater than or equal to twice the expression level in the second highest tissue group. The 3,299 genes meeting this criterion are said to be tissue-specific in the highest tissue. We also considered the peak expression tissue for every gene.

Online Mendelian Inheritance in Man (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) was used to identify all genes known to be involved in human Mendelian diseases. The chromosomal locations of all genes were obtained from Ensembl.

Recombination rate data for 1-megabase segments of human chromosomes were downloaded from University of California, Santa Cruz (http://genome.ucsc.edu/cgi-bin/hgTables). A recombination rate was assigned to each gene in the Q20 data set, based on the 1-megabase segment in which the midpoint of the gene lies. Of the 13,888 genes analyzed here, 13,714 are found in regions of known recombination rates. Among these 13,714 genes, 152 human and 228 chimp PSGs have available recombination rates. We then computed the mean recombination rate of the 152 human PSGs. To estimate the expected value of this mean, we randomly picked 152 genes from 13,714 genes and computed the mean. This procedure was repeated 10,000 times to estimate the probability that the observed mean is greater than the expected mean. The same procedure was applied to chimp PSGs, under the assumption that the recombination rate of a chimp gene is the same as for its human ortholog, which is probably correct for the majority of genes at the 1-megabase scale (Serre et al. 2005).

Use of the 6X Chimp Genome Assembly

The 233 chimp PSGs identified by using the Q20 data from the 4X chimp sequence were reanalyzed using sequences from the 6X chimp genome assembly (panTro2; www.genome.ucsc.edu). The 6X sequences corresponding to 4X sequences of the 233 PSGs were found by using BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start). Codons with one or more bases having a quality score less than Q20 in the 4X assembly were eliminated, as described in Materials and Methods. Of the 233 PSGs, 100 had a perfect match between the 4X and 6X assemblies. Forty-eight PSGs were aligned to the 6X assembly with no gaps, but with mismatches of 0.02-0.13%. Eighty-five PSGs were aligned to 6X with some gaps, ranging from 0.02 to 63.0%. Codons having one or more bases missing or ambiguous (i.e., N) in the 6X assembly were also eliminated, and the resulting sequence was aligned to the human and macaque sequences. This alignment was used in the branch-site test of positive selection in the chimp branch, as described in Materials and Methods.

Of the 48 PSGs with no gaps, 42 still show positive selection in chimp, using the 6X sequence, whereas 6 no longer show the signal of positive selection. Of the 85 PSGs with gaps, 65 still show positive selection in chimp, whereas 20 appeared not to be under positive selection when the 6X alignment was used. Each of these 26 (6 + 20) apparent reversals was examined manually to determine the source of discrepancies between the 4X and 6X results. In five cases (all with gaps), it was determined that the 4X assembly was more accurate because of elimination of exons or other problems in the 6X assembly. In these cases, the 4X result was retained.

Performance of the Improved Branch-Site Likelihood Method

Although there have been concerns about the performance of the likelihood method in detecting positive selection (Nei 2005), the improved branch-site likelihood method was previously shown by computer simulations to produce reasonably good results, even when some of the assumptions are violated (Zhang et al. 2005). To further verify the suitability of the method when the number of substitutions is as small, as in the present context, we conducted additional simulations specifically designed to mimic the evolution of human, chimp, and macaque genes. The simulation procedure follows ref. 2. A tree of three taxa was used. The numbers of synonymous substitutions per site for the human, chimp, and macaque branches were set as 0.006, 0.006, and 0.058, respectively, because these were the actual numbers observed from our Q20 data for the three branches. Because the 13,888 alignments have a mean length of 432 codons and a standard deviation of 339 codons, we examined three different sequence lengths (150, 400, and 1,000 codons) in the simulation. To examine the type I error (i.e., false-positives), we used model B1 (negative selection) to simulate sequence evolution in the background branches (macaque and chimp branches) and either model F1 or F2 for the foreground branch (human branch) (Table

3.3). Note that F1 and F2 do not contain sites under positive selection; rather, they represent partial and complete relaxation of negative selection, respectively. After the three sequences were generated, the likelihood method was used to detect positive selection in the human branch. Positive selection was inferred if the likelihood of the alternative model was greater than that of the null model at the 5% significance level. Four hundred simulation replications were conducted. The results showed that the type I error is lower than the nominal rate of 5% in the case of partial relaxation of negative selection (Table 3.4). In the case of complete relaxation of negative selection, the error rate is lower than, close to, and higher than the nominal rate for short, intermediate, and long sequences, respectively (Table 3.4). Because only 10% of our 13,888 genes have >800 codons, and because complete relaxation of negative selection is rare, it is expected that the slightly-higher-than-nominal type I error observed in one condition of the simulation will have only a minimal influence on our results. Although the χ^2 approximation of the likelihood ratio test depends on the large-sample assumption, our simulation showed that the approximation is justified in the present context. This may be due to two factors. First, we used χ_1^2 instead of a 50:50 mixture of point mass 0 andχ_1^2 (Self and Liang 1987; Zhang et al. 2005), thus reducing type I errors. Second, the χ^2 approximation appears insensitive to sample size, as was found previously (Zhang 1999).

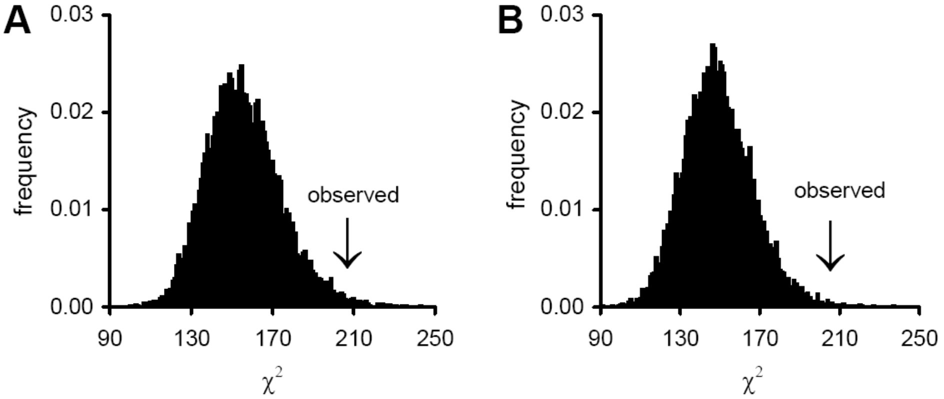
We also examined the power of the statistical test in four simulations, by changing the background and foreground models (Table 3.3). The sequence length of 400 codons was used in this set of simulations. The four background models (B3-B6) differ in the level of mean ω. The corresponding foreground models (F3-F6) also have different mean ω, but have the same level of positive selection. The results showed that a higher background ω increases the detection rate of positive selection (Table 3.6).

We noticed that when the likelihood ratio test provides statistical evidence for positive selection in a gene, the estimated ω for the positively selected sites (class 2 codons) in the foreground branch is often very large (e.g., >100). This appears biologically unreasonable. We examined the accuracy of the estimated ω by using the simulations described in the previous paragraph. We allowed 30% of codons to be under positive selection in the foreground branch, with a mean ω for these positively selected codons equal to 5 (Table 3.3). However, as shown in Table 3.6, the estimated ω for class 2 codons has a mean of several hundred and a standard deviation of several hundred among the genes in which positive selection is detected by PAML. Thus, the simulations showed that, although the likelihood ratio test of positive selection is reliable, the estimation of ω (when >1) is problematic and not trustable. For this reason, we do not present the likelihood-estimated ω values.

Acknowledgements

I worked with Peng Shi and Jianzhi Zhang on the research described in this chapter. The paper we co-authored was published in *Proceedings of the National Academy of Sciences of the United States of America* (Bakewell et al. 2007). We thank Soochin Cho, Wendy Grus, Ondrej Podlaha, Xiaoxia Wang, and especially Masatoshi Nei, for valuable suggestions; three anonymous reviewers for constructive comments; and the Washington University School of Medicine Genome Sequencing Center and Baylor College of Medicine Human Genome Sequencing Center for making available before publication the 6X chimp genome assembly and macaque genome assembly, respectively. This work was supported by the University of Michigan and by National Institutes of Health Grant GM67030 (to J.Z.).

Figure 3.1 Functional differences between human and chimp unshared PSGs.
(A and B) Human and chimp PSGs show a significantly larger difference in distribution across biological process groups (A) and molecular function groups (B) than by chance (P = 0.84% and 0.26%, respectively, one-tail randomization test). The bars show the frequency distribution of the χ^2 values in 10,000 random divisions of the 373 unshared PSGs into 147 human PSGs and 226 chimp PSGs. The arrow indicates the observed χ^2. Here, the randomization test is superior to the standard χ^2 test because the functional groups are not independent of one another, and a single gene may belong to more than one group. Similar results are obtained when the seven shared PSGs are included. (C) Biological process and molecular function groups that show the greatest differences between human and chimp unshared PSGs, as ranked by individual χ^2 values. Shown are the groups that each contribute at least 2% of the total χ^2 of all groups. Groups with a higher frequency of human PSGs than chimp PSGs are shown in red; those with a higher frequency of chimp PSGs than human PSGs are shown in blue.



C

PANTHER category	# of genes	# of PSGs Human	 Chimp	χ^2
Biological Process				
Protein metabolism & modification	2,815	9	40	9.09
mRNA transcription	1,144	5	25	6.50
Anion transport	171	6	1	6.29
Phosphate transport	80	4	0	6.15
Proteolysis	938	2	16	6.04
Ectoderm development	604	8	3	5.11
mRNA transcription regulation	891	3	17	4.99
Stress response	780	2	14	4.85
Fatty acid metabolism	169	3	0	4.61
Ion transport	578	12	7	4.49
G-protein mediated signaling	682	14	9	4.44
Molecular Function				
Lyase	153	6	0	9.22
Phosphatase	194	5	0	7.69
Nucleic acid binding	2,597	13	46	7.46
Nuclease	2,285	12	42	6.68
Transferase	1,296	5	25	6.50
Other transporter	208	6	2	4.24

Figure 3.2 Frequency distribution of human and chimp PSGs across 20 peak expression tissue groups.

The overall difference between the distributions of the two species is not statistically significant (χ^2 = 23.8, df=19, *P*=0.21). Only smooth muscle (χ^2 = 7.7, *P* = 0.0056) shows a significant difference in proportion of PSGs between the two species, but the significance disappears when multiple testing is corrected for. Pink dots show the expected distribution of PSGs when there is no enrichment of PSGs in any tissue groups.

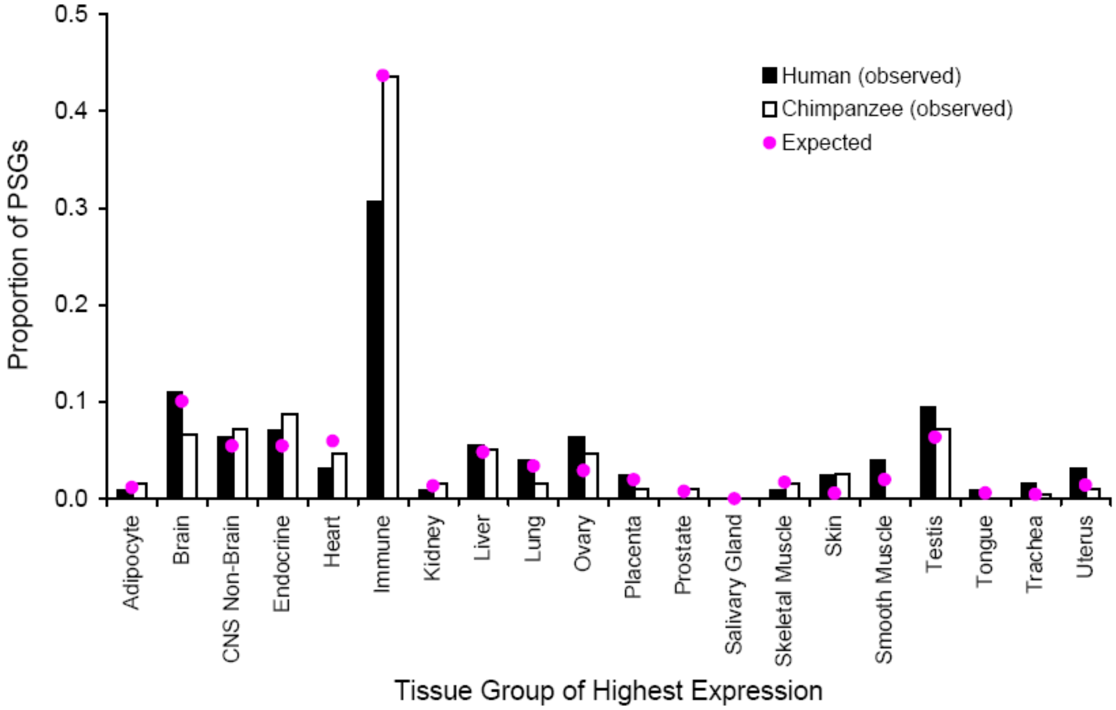


Figure 3.3 Distributions of human and chimp PSGs among chromosomes.
Contrary to the chromosomal speciation hypothesis, PSGs are slightly less abundant on rearranged chromosomes than on colinear chromosomes (P = 0.10 and 0.055 for the human and chimp lineages, respectively, χ^2 test). The human chromosome numbers are used. The expected number of PSGs on each chromosome is calculated under the assumption that the probability of a gene being targeted by positive selection is not affected by the chromosome on which it is located.

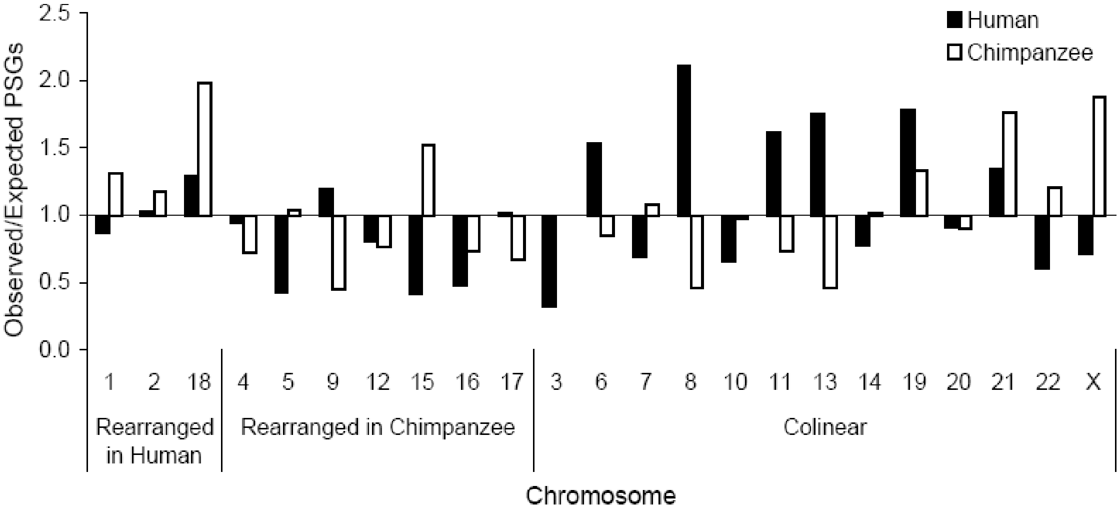


Figure 3.4 Distribution of human and chimp PSGs across chromosomes.
The human chromosome numbers are used. Shown are 13,714 genes in the Q20 dataset for which recombination rate data are available. Genes located in segments with a recombination rate in the lowest quintile of all 1-megabase segments in the genome are colored blue; genes in the second, third, fourth, and fifth quintiles of recombination rate are colored green, yellow, orange, and red, respectively. A total of 152 human PSGs (filled diamonds) and 228 chimp PSGs (open diamonds) for which recombination data are available are shown to the right of each chromosome. For all genes, position along the chromosome corresponds to the midpoint of the gene. There is a weak tendency for PSGs to have higher recombination rates than by chance (P = 0.32 and 0.04 for human and chimp, respectively; simulation test).

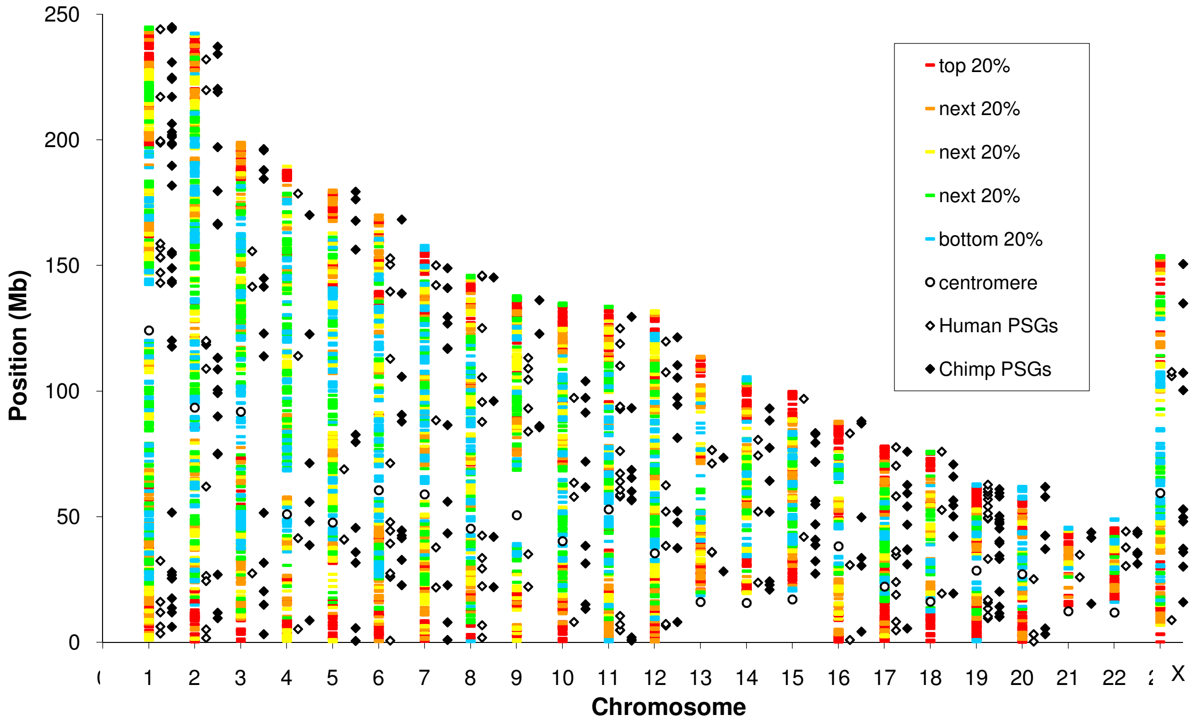
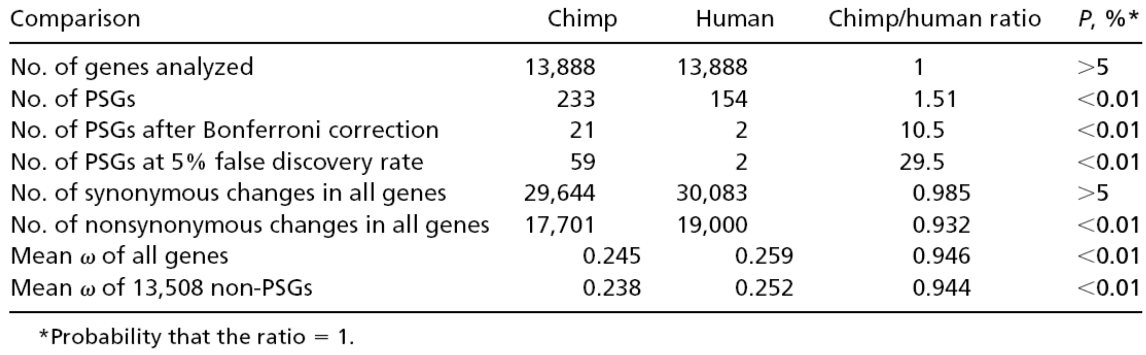


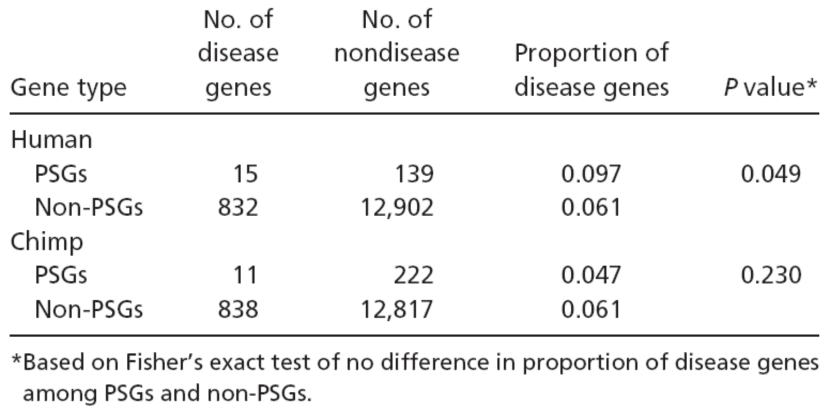
Table 3.1 Genic positive selection in human and chimp lineages since their split



Comparison	Chimp	Human	Chimp/human ratio	*P*, %*
No. of genes analyzed	13,888	13,888	1	>5
No. of PSGs	233	154	1.51	<0.01
No. of PSGs after Bonferroni correction	21	2	10.5	<0.01
No. of PSGs at 5% false discovery rate	59	2	29.5	<0.01
No. of synonymous changes in all genes	29,644	30,083	0.985	>5
No. of nonsynonymous changes in all genes	17,701	19,000	0.932	<0.01
Mean ω of all genes	0.245	0.259	0.946	<0.01
Mean ω of 13,508 non-PSGs	0.238	0.252	0.944	<0.01

*Probability that the ratio = 1.

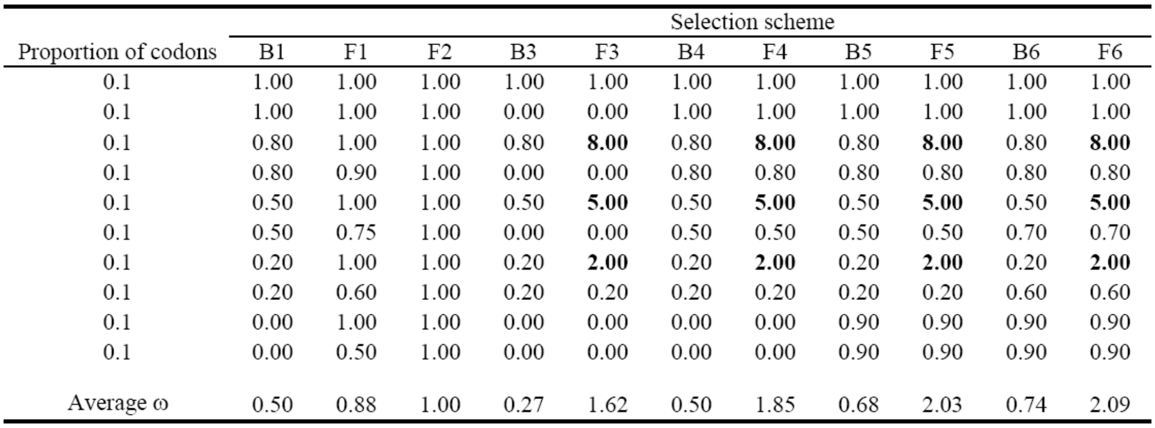
Table 3.2 Association of PSGs with human disease



Gene type	No. of disease genes	No. of nondisease genes	Proportion of disease genes	*P* value*
Human				
PSGs	15	139	0.097	0.049
Non-PSGs	832	12,902	0.061	
Chimp				
PSGs	11	222	0.047	0.230
Non-PSGs	838	12,817	0.061	

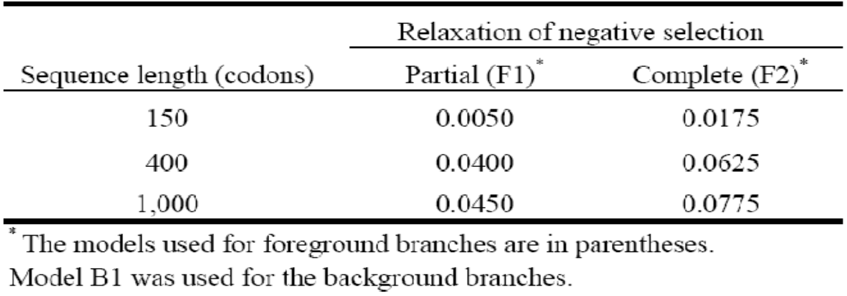
*Based on Fisher's exact test of no difference in proportion of disease genes among PSGs and non-PSGs.

Table 3.3 The ω values used in generating the DNA sequences in computer simulation



	Selection scheme										
Proportion of codons	B1	F1	F2	B3	F3	B4	F4	B5	F5	B6	F6
0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.1	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
0.1	0.80	1.00	1.00	0.80	**8.00**	0.80	**8.00**	0.80	**8.00**	0.80	**8.00**
0.1	0.80	0.90	1.00	0.00	0.00	0.80	0.80	0.80	0.80	0.80	0.80
0.1	0.50	1.00	1.00	0.50	**5.00**	0.50	**5.00**	0.50	**5.00**	0.50	**5.00**
0.1	0.50	0.75	1.00	0.00	0.00	0.50	0.50	0.50	0.50	0.70	0.70
0.1	0.20	1.00	1.00	0.20	**2.00**	0.20	**2.00**	0.20	**2.00**	0.20	**2.00**
0.1	0.20	0.60	1.00	0.20	0.20	0.20	0.20	0.20	0.20	0.60	0.60
0.1	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.90	0.90	0.90	0.90
0.1	0.00	0.50	1.00	0.00	0.00	0.00	0.00	0.90	0.90	0.90	0.90
Average ω	0.50	0.88	1.00	0.27	1.62	0.50	1.85	0.68	2.03	0.74	2.09

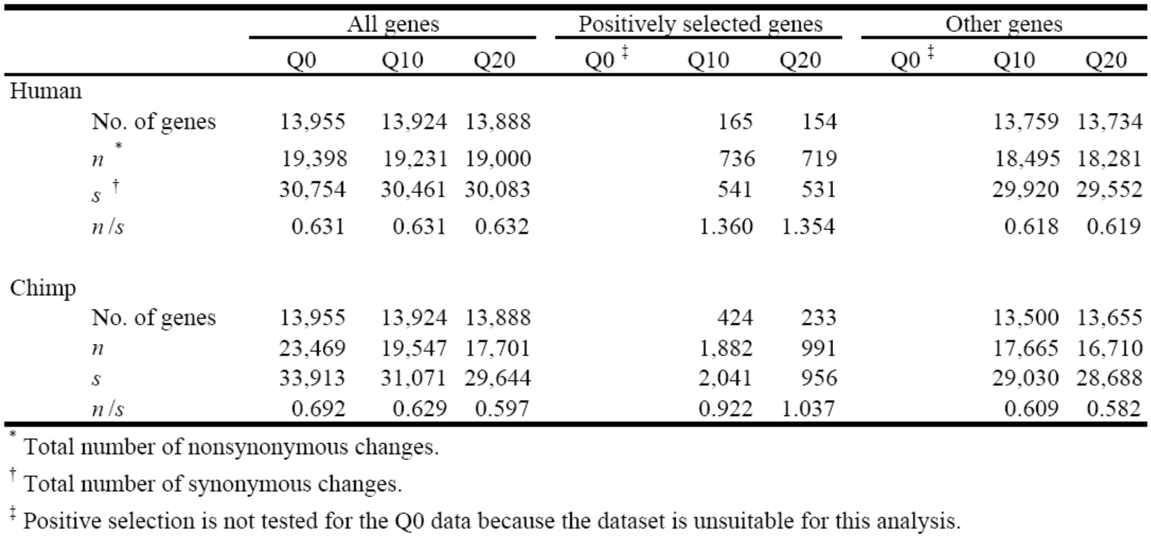
Table 3.4 False-positive rates of the branch-site likelihood method in detecting positive selection at the 5% significance level



	Relaxation of negative selection	
Sequence length (codons)	Partial (F1)[*]	Complete (F2)[*]
150	0.0050	0.0175
400	0.0400	0.0625
1,000	0.0450	0.0775

[*] The models used for foreground branches are in parentheses.
Model B1 was used for the background branches.

Table 3.5 Basic statistics of the three datasets



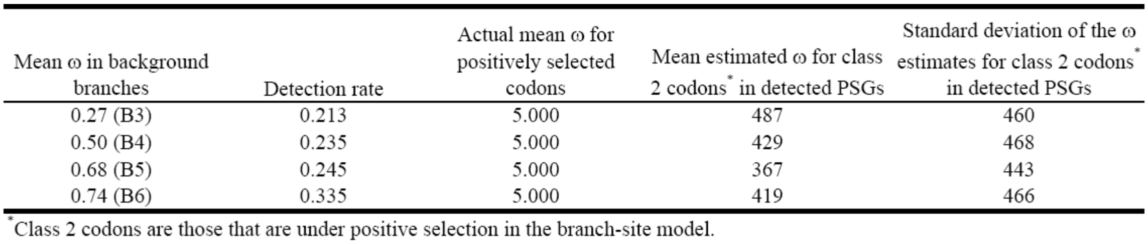
	All genes			Positively selected genes			Other genes		
	Q0	Q10	Q20	Q0 [‡]	Q10	Q20	Q0 [‡]	Q10	Q20
Human									
No. of genes	13,955	13,924	13,888		165	154		13,759	13,734
n [*]	19,398	19,231	19,000		736	719		18,495	18,281
s [†]	30,754	30,461	30,083		541	531		29,920	29,552
n/*s*	0.631	0.631	0.632		1.360	1.354		0.618	0.619
Chimp									
No. of genes	13,955	13,924	13,888		424	233		13,500	13,655
n	23,469	19,547	17,701		1,882	991		17,665	16,710
s	33,913	31,071	29,644		2,041	956		29,030	28,688
n/*s*	0.692	0.629	0.597		0.922	1.037		0.609	0.582

[*] Total number of nonsynonymous changes.

[†] Total number of synonymous changes.

[‡] Positive selection is not tested for the Q0 data because the dataset is unsuitable for this analysis.

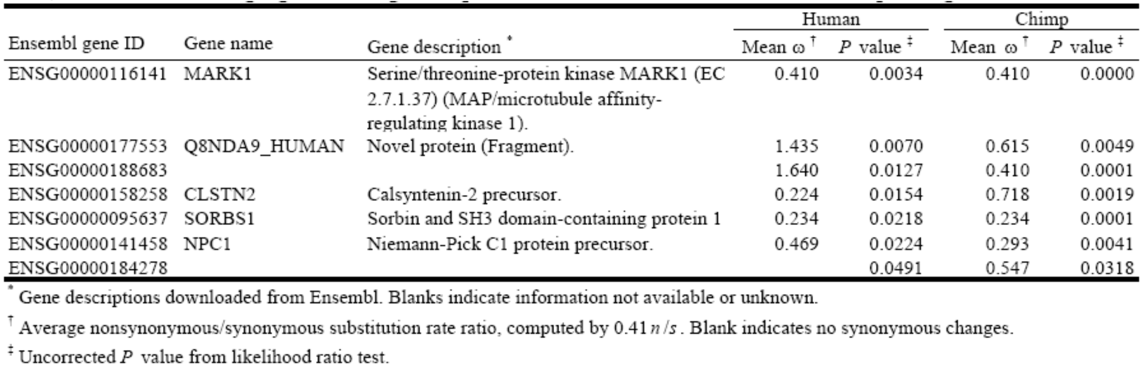
Table 3.6 Rate of detection of positive selection by the branch-site method when there is positive selection



Mean ω in background branches	Detection rate	Actual mean ω for positively selected codons	Mean estimated ω for class 2 codons[*] in detected PSGs	Standard deviation of the ω estimates for class 2 codons[*] in detected PSGs
0.27 (B3)	0.213	5.000	487	460
0.50 (B4)	0.235	5.000	429	468
0.68 (B5)	0.245	5.000	367	443
0.74 (B6)	0.335	5.000	419	466

[*] Class 2 codons are those that are under positive selection in the branch-site model.

Table 3.7 Genes showing significant signal of positive selection in both human and chimp lineages



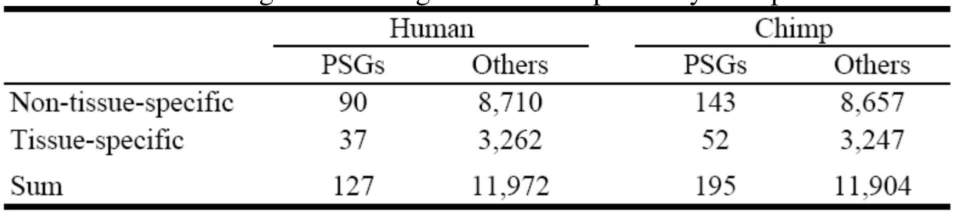
Ensembl gene ID	Gene name	Gene description *	Human		Chimp	
			Mean ω †	*P* value ‡	Mean ω †	*P* value ‡
ENSG00000116141	MARK1	Serine/threonine-protein kinase MARK1 (EC 2.7.1.37) (MAP/microtubule affinity-regulating kinase 1).	0.410	0.0034	0.410	0.0000
ENSG00000177553	Q8NDA9_HUMAN	Novel protein (Fragment).	1.435	0.0070	0.615	0.0049
ENSG00000188683			1.640	0.0127	0.410	0.0001
ENSG00000158258	CLSTN2	Calsyntenin-2 precursor.	0.224	0.0154	0.718	0.0019
ENSG00000095637	SORBS1	Sorbin and SH3 domain-containing protein 1	0.234	0.0218	0.234	0.0001
ENSG00000141458	NPC1	Niemann-Pick C1 protein precursor.	0.469	0.0224	0.293	0.0041
ENSG00000184278				0.0491	0.547	0.0318

* Gene descriptions downloaded from Ensembl. Blanks indicate information not available or unknown.

† Average nonsynonymous/synonymous substitution rate ratio, computed by $0.41\, n/s$. Blank indicates no synonymous changes.

‡ Uncorrected *P* value from likelihood ratio test.

Table 3.8 Numbers of genes with regard to tissue-specificity of expression



	Human		Chimp	
	PSGs	Others	PSGs	Others
Non-tissue-specific	90	8,710	143	8,657
Tissue-specific	37	3,262	52	3,247
Sum	127	11,972	195	11,904

Chapter 4

Genomic phylostratigraphy and spurious patterns of new gene evolution

Abstract

Phylostratigraphy is a method for estimating the phylogenetic age of a gene by the distribution of its homologs on the tree of life. Applying this method to all genes in a species constitutes genomic phylostratigraphy. Gene age estimation relies on the ability to detect distant homologs via sequence similarity, typically by BLAST and its derivatives. In spite of the expectation that homologs of fast evolving genes in distantly related species tend to be missed by BLAST, a simulation study that considered substitution rate variation among sites discovered few errors, stimulating an explosive use of genomic phylostratigraphy that has led to many reports of differential properties of genes of different ages. Unfortunately, the simulation study relied on rate heterogeneity patterns of a potentially biased set of only 14 genes. Here we reevaluate the accuracy of genomic phylostratigraphy with a genomic scale gene set and investigate the impact of its error on findings about young and old genes. We show that failure to detect distant homologs can generate spurious non-uniform distributions of various gene properties among age groups, many of which are not predicted a priori. Furthermore, phylostratigraphy also overestimates gene age substantially for new genes that have arisen from gene duplication followed by very rapid sequence evolution for 20 million years. We find that the reported large numbers of lineage-specific genes cannot possibly be explained by the model of gene duplication followed by substantive neofunctionalization. Given the high likelihood that conclusions about gene age are faulty, we advocate the use of realistic simulation to determine if observations from

genomic phylostratigraphy are explainable by a null model of biased measurement, and in all cases, critical evaluation of results.

Introduction

Compared to old genes, young genes evolve faster (Alba and Castresana 2005), are subject to less purifying selection and perhaps more positive selection (Cai and Petrov 2010), are less likely to be associated with human disease (Domazet-Loso and Tautz 2008) or to be expressed during the phylotypic stage (Domazet-Loso and Tautz 2010a), are more likely to be singletons (Prachumwat and Li 2008), and have different codon usage (Prat et al. 2009). The studies that uncovered these correlates of gene age relied on phylostratigraphy (Domazet-Loso et al. 2007) or similar methods (e.g. lineage specificity (Cai et al. 2006) or study specific variants (Prachumwat and Li 2008)) to estimate gene age. Gene age is operationally defined as the time since the divergence of the most distantly related pair of taxa in which homologs can be found. Thus, the ability to detect distant homologs is critical to the correct estimation of gene age, and to the reliability of conclusions about the properties of young and old genes. Studies of gene age may focus on one or a few closely related species and classify genes unique to those species as orphan genes, or they may take a wider view and identify genes common to a particular clade, termed lineage specific genes. Whether a study focuses on very young genes (orphans) or slightly older genes (lineage specific genes), the accuracy of gene age estimation is paramount.

Because detection of homologs depends on sequence similarity, and slowly evolving genes retain greater sequence similarity with their homologs than rapidly evolving genes of the same age, it has been suggested that the correlation of gene age with evolutionary rate is purely artifactual (Elhaik et al. 2006). Elhaik et al. simulated DNA sequences representing several taxa, all derived from a common ancestor, for a range of evolutionary rates and used BLAST to search for the homologs. In this scenario, failure of BLAST to detect a homolog is due solely to loss of sequence similarity between homologs because all simulated sequences have a homolog in each taxon. The inability to detect homologs results in gene age estimation error, that is, the assignment of an incorrect age to the gene. Error rates as high as 100% were observed, and fast evolving genes had the highest error rates, leading the authors to conclude that the statement that young genes evolve fast is a mere tautology. However, the above simulation used nucleotide sequences, whereas amino acid sequences allow more sensitive detection of distant homologs. In addition, all sites were allowed to evolve at the same rate, while in real genes some sites are more functionally constrained than others, resulting in rate heterogeneity among sites. BLAST search begins by scanning the database for high scoring words – that is, sequences of a given length (w) that achieve a minimum similarity score to a subsequence of the query. Therefore, the presence of consecutive functionally constrained sites within a sequence is expected to improve BLAST detection. The default word size (w) in protein BLAST is only 3, so even short stretches of conserved residues may improve results. A simulation study comparing gene age estimation error in sequences simulated with and without among site rate heterogeneity confirmed this prediction (Alba and Castresana 2007). Using amino acid sequences and rate heterogeneity patterns derived from alignments of vertebrate or bilaterian species, Alba and Castresana found that while gene age estimation error rates were still high in fast evolving genes, the majority of genes evolve at rates that are consistent with very low error rates.

While the latter simulation study implements a more realistic model than did the former, it has three features that make it difficult to assess to what extent the inability of BLAST to detect homologs impacts the estimation of gene age. First, the rate heterogeneity patterns used were derived from alignments of either 7 vertebrates with a most recent common ancestor approximately 450MYA, or 9 bilaterians with a most recent common ancestor about 980MYA. In both cases, these alignments may represent slowly evolving genes, as they could not have been included unless a homolog was found in each of these widely divergent taxa. A simulation using the rate heterogeneity patterns of slowly evolving genes is expected to produce BLAST results and gene age estimates similar to those of slowly evolving genes, i.e. low gene age estimation error rates and old gene ages. Second, Alba and Castresana used the rate heterogeneity patterns of only 14 genes for which alignments could be constructed at that time. Their results may thus reflect the rate heterogeneity pattern of these few genes, which may not be representative of genomic patterns of rate heterogeneity. Third, the rate heterogeneity patterns used reflect the average rate of evolution over hundreds of millions of years. One model for the evolution of orphan genes invokes a period of relaxed constraint, during which rapid sequence change reduces sequence similarity to homologs (Domazet-Loso and Tautz 2003). This model also suggests that a period of positive selection, as a gene acquires new function, could further reduce sequence similarity, resulting in the appearance of an orphan gene. After a gene has acquired its new function, it is expected to be subject to purifying selection and thus have a reduced evolutionary rate. It is not clear whether genes evolving under such a regime would differ significantly from genes evolving gradually in terms of their estimated gene age as no such model has been simulated in previous studies.

Here, we simulate the evolution of amino acid sequences using rate heterogeneity patterns derived from the alignments of 6,662 genes in 12 *Drosophila* species. These species share a most recent common ancestor about 60MYA, so they may reflect both slowly evolving genes as well as more rapidly evolving genes that would not have been included in Alba and Castresana’s study. We simulate the evolution of these genes under a gradual model using rates and rate heterogeneity patterns estimated from these alignments, and also under models of relaxation and neofunctionalization that more accurately reflect models of orphan gene evolution. Thus, we are able to examine gene age estimation error rates for genes that should be classified as old – those that have been evolving gradually throughout their history – as well as error rates expected for orphan genes – those that have undergone a period of more rapid evolution consistent with a loss of ancestral function and gain of new function. For the former, failure to detect a homolog is a false negative, while for the latter, detection of a homolog that diverged prior to the period of rapid evolution is a false positive. From these simulations, we identify factors impacting gene age estimation error rate. In addition, we compare the properties of genes identified as old and young in our simulations with those reported for old and young genes in a recent study of gene age in *Drosophila* (Domazet-Loso et al. 2007) and assess the extent to which the reported results can be explained by gene age estimation error. Finally, in order to assess how well current models of orphan gene evolution explain observed levels of lineage specific genes, we estimate the proportion of lineage specific genes expected to be observed when both gene age estimation error and true young genes are taken into account, and compare this estimate to the proportion observed in real genomes.

Results

Detection of simulated homologs by BLAST

In our study, protein sequences were simulated for each of 11 taxa represented in the phylogeny shown in Figure 4.1. Each sequence derives from an ancestral sequence; thus, a homolog is present in all 11 taxa. This process was repeated 6,662 times for each iteration of the simulation, using parameters derived from 6,662 actual *Drosophila* group alignments, resulting in a simulated proteome of 6,662 amino acid sequences for each of the 11 taxa. BLASTP searches were conducted using the simulated *D. melanogaster* sequences as query and the other 10 proteomes as database. Because homologs are present in each of the other 10 species in this simulation, failure of BLAST to detect all 10 homologs when gradual evolution is simulated is considered gene age estimation error.

Two variants of a simulation representing gradual evolution were conducted. First, randomly generated ancestral sequences created using amino acid frequencies calculated from the entire dataset and matching the lengths of actual *Drosophila* alignments were evolved using evolutionary rates and rate heterogeneity patterns estimated by Tree-Puzzle for those alignments, and indel parameters as in Alba and Castresana (2007) (main simulation). Over 10 repeats of this simulation, BLASTP failed to detect the most distant homolog for 11.8% of simulated proteins, failed to detect at least one of the 10 existing homologs for 12.2% of proteins, and failed to detect any homologs for 2.2% of proteins. (Table 4.1). A second simulation, identical to the main simulation except that actual *D. melanogaster* sequences were used as the ancestral sequences in place of the randomly generated ancestral sequences, failed to detect the most distant homolog for 10.9% of simulated proteins, failed to detect at least one of the 10 existing homologs for 11.4% of proteins, and failed to detect any homologs for 2.2% of proteins. (Table 4.1).

Since the evolutionary rate of a given site may change as selective pressures change over time, we also simulated sequences with a change in rate heterogeneity patterns. In this group of simulations, 0%, 10%, 25% or 50% of sites were randomly chosen from each protein, and the rate classes for these sites were shuffled on the branch indicated with the letter S in Figure 4.1. No indels were allowed in this group of simulations, but otherwise the methods were identical to the main simulation. In the 0%, 10%, 25%, and 50% simulations, BLAST failed to detect the most distant homolog in 9.0%, 9.5%, 10.3%, and 23.7% of proteins, respectively. The rates of failure to detect any homologs were 1.2%, 1.3%, 1.3%, and 1.3%, respectively, and the rates of failure to detect at least one of the 10 existing homologs were 9.2%, 9.8%, 10.7%, and 24.9%, respectively.

Correlates of BLAST error

It has been reported that young genes are shorter and evolve faster (Domazet-Loso and Tautz 2003; Alba and Castresana 2005; Cai et al. 2006; Cai and Petrov 2010) than old genes. In addition, the rate heterogeneity patterns of a gene influence the ability of BLAST to detect homologs of a gene, as shown above and in previous studies (Elhaik et al. 2006; Alba and

Castresana 2007). The level of rate heterogeneity is quantified by the gamma distribution shape parameter alpha; low values of alpha indicate considerable among site variability, while high values of alpha indicate less heterogeneity among sites, more similar to evolution without rate heterogeneity. The degree of rate heterogeneity among sites is expected to be negatively correlated with the average evolutionary rate of the protein (Zhang and Gu 1998). The sequence of a gene evolving with rate heterogeneity is composed of a mosaic of variable length blocks with different evolutionary rates. Long blocks with low evolutionary rates are expected to have the highest potential for BLAST detection. Thus, we investigate the correlation of gene age estimation error with gene length, evolutionary rate, alpha, and length and relative rate of the longest, slowest evolving block in each gene.

Evolutionary rate correlates well with gene age in the main simulation (r = -0.622, $P = 0$). Genes with the youngest ages, identified as having no homologs outside of *Drosophila* or insect lineages, tend to have the highest evolutionary rates (Figure 4.2). The distribution of evolutionary rates in each age class is significantly different from the genomic distribution, with those in the oldest age class evolving more slowly, and those in all other age classes evolving more rapidly (Figure 4.2, Mann-Whitney U test, $p < 0.0001$ for all tests after Bonferroni correction). Neither gene length nor alpha is well correlated with gene age in the *Drosophila* data set used in the present study (r = 0.110, $P = 0$ and -0.043, $P = 0.0005$, respectively). However, estimation of alpha is unreliable for high alpha (Gu and Zhang 1997). To examine the correlation between alpha and evolutionary rate, we ignore 44 genes with alpha > 3 and observe a positive correlation in the remaining 6,618 genes (r = 0.670, $P = 0$). Furthermore, almost all gene age estimation errors in the main simulation occur in the 10% of genes with highest alpha (Figure 4.3, last column), and alpha and gene age are correlated within this 10% of genes (r = 0.363, $P = 0$).

The 6,662 *Drosophila* genes examined show considerable variation in the minimum rate class r_{min} (calculated as evolutionary rate * minimum relative rate) and the length of the longest contiguous stretch of amino acids in that rate class (c_{max}). (Figure 4.4A). gene age estimation error rates are very low for genes with low to intermediate r_{min} and intermediate to high c_{max}, while error rates are as high as 100% for genes with the highest r_{min} and lowest c_{max} values (Figure 4.4B).

Reproduction of reported results by gene age error

We focus on three results reported in the initial paper that proposed the phylostratigraphy method (Domazet-Loso et al. 2007). First, these authors found that the proportion of genes of a particular age expressed in a given tissue may differ significantly from the genome wide proportion expressed in that tissue. Here, we also find significant differences between the genome wide proportion of endoderm, mesoderm and ectoderm expressed genes, and the proportion in genes of a specific age in our main simulation (Figure 4.5A) In the present study, significant differences after Bonferroni correction are only found when considering the sum of 10 runs of the simulation. Low sample size in individual runs of the simulation likely prevents deviations from genomic expression patterns in genes of a given age from attaining significance. Second, Domazet-Loso et al. calculated the number of novel genes arising per million years on each branch of the phylogeny. Interestingly, they found a peak in this rate of founder gene formation in the branch leading to the most recent common ancestor of Bilateria. We also calculated the rate of novel gene formation during each age interval in our simulated genes and surprisingly found a peak in the same branch leading to the bilaterian ancestor (Figure 4.5B). Finally, the authors of the previous work calculated the proportion of genes that were annotated with a biological process in each age group, and observed that this proportion was correlated with gene age. This correlation is also observed in the data from our main simulation (Figure 4.5C).

Detection of homologs after rapid evolution

Two additional classes of simulation were performed. First, we considered models of relaxation in which 44%, 90% or 100% of sites were allowed to evolve at the fastest rate observed in the actual protein. The remaining sites were somewhat constrained in the 44% and 90% relaxed models (see Methods). Second, we considered models that feature gradual evolution as in the main simulation, followed by a burst of rapid evolution of 5MY, 10MY or 20MY at the location in the phylogeny indicated by the letter B in Figure 4.1, followed by gradual evolution under a rate heterogeneity pattern different from that in the first part of the simulation (see Methods). As these simulations represent models of relaxation and/or neofunctionalization that are proposed to lead to rapid sequence divergence and orphan gene formation, here we expect BLAST not to find all homologs. Whereas in the gradual evolution models, failure of BLAST to detect the most distant homolog was considered a gene age estimation error, here, failure of BLAST to detect the most distant homolog is consistent with orphan gene formation. Strictly speaking, after the burst of rapid evolution, no homologs should be detected in taxa that diverged prior to the burst, so detection of such homologs can be considered gene age estimation error as well. In this case, the age of a young gene is overestimated. Because the location of burst of rapid evolution in our simulation is on the branch leading to *Drosophila* after its divergence from honeybee, genes should not have hits in any of the other 10 taxa. For the 44% and 90% relaxed simulations, BLAST does not find a hit in any of the other 10 taxa for 4.5% and 64.5% of proteins, respectively. 14.6% and 96.3% of proteins have no hit in the most distant (E. coli) homolog, but do have hits in one or more intermediate taxa. In this case the genes would be assigned a gene age older than the branch on which the burst of evolution took place, demonstrating that even for true orphan genes, the phylostratigraphy method may incorrectly assign their age. For the 100% relaxed simulation, a tree topologically identical to that shown in Figure 4.1 was used, but with branch lengths representing taxa that diverged from *Drosophila* 5, 10, 20, 30, 40, 50, 75, 100, 200, and 500 MYA. Rates of orphan gene formation after 200MY and 500MY are 97% and 99%, respectively (Table 4.1). After a burst of 5MY, 10MY or 20MY, the BLAST fails to detect any homolog for 2.7%, 3.3%, and 4.7% of genes, respectively.

Discussion

Predicting BLAST error

In considering sources of error in any experiment, it is important to know whether the errors are random and merely generate noise, or if they are systematic and thus tend to bias results. In a computational experiment involving gene age, the use of BLAST to detect homologs introduces a predictable and significant source of bias. Although we know that evolutionary rate and rate heterogeneity patterns both contribute to gene age estimation error, predicting or estimating of the impact of this error in a given scenario is not trivial. For example, consider the impact of rate heterogeneity. In our simulations, BLAST often failed to find homologs for genes with high alpha. Interestingly, comparing the vertebrate and bilaterian genes used by Alba and Castresana to study the problem of BLAST error, the vertebrate genes have an average alpha (0.725) lower than average alpha in the 6,662 *Drosophila* gene used in the present study (0.997), while the bilaterian genes have a higher average alpha (1.125). We conducted additional simulations sampling from vertebrate and bilaterian rate heterogeneity patterns. Rates of gene age estimation error in the vertebrate simulation were lower than in our main simulation (8.6% vs. 11.8% unable to detect the most distant homolog), while the same measure in the bilaterian simulation was substantially higher (49.2%, Table 4.1). These findings demonstrate that gene age estimation error may play a much more important role in some contexts than in others, and thus that its specific impact must be examined on a case-by-case basis.

Proportion of false positives among orphan genes

Orphan genes may result from either gradual evolution and gene age estimation error, as in our main simulation, or by punctuated evolution and sequence divergence, as in the relaxed and burst simulations. Orphan genes resulting from these two processes may be considered false orphans and true orphans, respectively. It is interesting to know what proportion of observed orphans are true orphans. In order to estimate the proportion of true orphans expected, it is necessary to know not only the rate at which true orphans and false orphans arise, but the amount of raw material available for evolution in gradual and punctuated modes. The majority of existing genes in a genome are expected to be functionally constrained and thus evolving under the gradual model. The *D. melanogaster* genome contains ~13,600 genes (Adams et al. 2000) presumably evolving in this manner. We have observed that 2.2% of these are classified as orphans due to gene age estimation error after 377 million years (the divergence time between *Drosophila* and honeybee). Assuming constant genome size over evolutionary time, we expect gradual evolution to result in 13,600 * 0.022 = 299 *Drosophila* or Diptera specific genes (false orphans). This is a conservative estimate because the 2.2% figure is derived from 1-to-1 orthologs found in all 12 sequenced *Drosophila* genomes, and these genes are likely evolving more slowly on average than all *Drosophila* genes.

Considering the process by which true orphans arise, although some studies (Levine et al. 2006; Zhou et al. 2008; Toll-Riera et al. 2009) have identified genes formed *de novo* from intergenic sequences, the most likely source of raw material is gene duplication. It has been estimated that in *Drosophila*, 31 new duplicates arise per genome per million years, and that the half-life of these genes is 2.9MY (Lynch and Conery 2000). New duplicates may be lost to pseudogenization and deletion, or may become orphan genes, lost only in the sense that they can no longer be found in the genome due to a high level of sequence divergence. From the 100% relaxation model in our simulation, a high-end estimate of the rate at which genes may escape detection through sequence change within 5MY is 0.37%. Thus, the number of orphan genes arising in 5MY is equal to the number of recent duplicates available multiplied by this rate. The balance of genes lost during this 5MY interval are pseudogenized or deleted. Using an iterative process (see Methods), we estimate that only 63 true orphan genes arise in the *Drosophila* lineage in 377MY since its divergence from honeybee. Although this number seems very low, it is understandable given the fact that there is a very short window of time during which a gene can escape sequence detection before becoming pseudogenized or being deleted. Given limitations on the mutation rate, only a few genes are able to escape in this short period of time. It is important to note that the rate of escape by sequence change is perhaps an overestimate as it is derived from our most extreme model of 100% relaxation. Furthermore, we assume that all of these 63 escaped genes acquire new function and are retained in the genome, although this may not be the case. Thus, 63 true orphans is probably an overestimate.

After orphan gene formation, further duplication of these founder genes can result in greater numbers of observed orphan genes. Domazet-Loso et al. (2007) used all-against-all BLAST of orphan genes to identify such expansions and found that the ratio of all *Drosophila* and Diptera specific orphans to founder genes was 1.21. Multiplying this ratio by the number of founder genes estimated above, we expect to observe about 299 * 1.21 = 362 false orphans and 63 * 1.21 = 76 true orphans. Thus, the false positive rate among observed orphans is about 362/(362 + 76) = 82.6%. With such a high false positive rate, studies examining the properties of "orphan genes" are largely examining the properties of false orphans, i.e., the properties of genes that escape BLAST detection.

The overall proportion of observed *Drosophila*/Diptera specific orphan genes is 21.1% (Domazet-Loso et al. 2007), about 6.6-fold higher than the proportion (362 + 76)/13,600 = 3.2% predicted here. What can explain the difference? As mentioned, our estimate of gene age estimation error rate may be low due to the necessarily conservative nature of genes for which a 1-to-1 ortholog exists in each of 12 *Drosophila* species, which would downwardly bias the number of false orphans. Another possibility is that the contribution of non-coding genomic regions to orphan gene formation has not been sufficiently appreciated. It has been estimated that 11.9% of new genes in *Drosophila* arise from such regions (Zhou et al. 2008). In addition, ~30% of new *Drosophila* genes are chimeric in nature, including some domains from an existing gene in combination with exons from other genes and/or non-coding regions (Zhou et al. 2008). Such genes may more rapidly escape from sequence similarity with their parents because only a portion of the gene must escape. Regardless of what other sources contribute to the catalog of orphans genes, it is clear that a large proportion, conservatively estimated at 362/(13600*0.211) = 12.6%, of orphan genes are due solely to gene age estimation error, and that the contribution of gene age estimation error exceeds that of the duplication followed by rapid divergence by 4.7-fold (362/76) or more.

Finally, it is important to note that among retained duplicates, there may be many genes that have undergone significant amounts of rapid evolution, consistent with the models of radical neofunctionalization that have been proposed to result in the formation of orphan genes, but which still retain sufficient sequence similarity to be identified as homologs. Our simulations show that, for example, a burst of 20MY rapid evolution erases sequence similarity with all homologs for less than 5% of proteins (Table 4.1). If after a period of rapid evolution, a gene's functions are very different from those of its homologs, it may be more correctly considered a "young gene". Even between species as closely related as mouse and human, many homologs exhibit differences in fitness effect, which implies change in function, in the two species. Furthermore, change in fitness effect is correlated with protein sequence divergence (Liao and Zhang 2008). Although it may be debated what level of functional difference is sufficient to classify a gene as an orphan, genes with identifiable homologs that nonetheless have different functions represent a previously ignored source of error, i.e., false negatives in the search for young genes.

Gene age estimation error as a null model

Gene age estimation error alone is sufficient to produce "interesting" correlations between gene age and biologically important attributes of genes such as evolutionary rate and expression profile (Figures 4.2 and 4.5A). For example, although Domazet-Loso et al. declared that "the most parsimonious scenario" to explain their finding of significant differences in expression pattern in genes of certain ages was a fluctuation in the adaptive tendencies of different evolutionary time periods, it is clear from our simulation results (Figure 4.5A) that a much more parsimonious explanation is available: gene age estimation error. It is imperative that results based on phylostratigraphy or similar methods be rigorously compared against a null hypothesis of gene age estimation error via simulation. However, simulation design requires careful choices about model parameters and negative simulation results may reflect parameter choice. Thus, to prevent erroneous conclusions about the properties of young and old genes, a skeptical evaluation of results is essential, especially for observations that may be explained by the differential ability of BLAST to detect homologs of different evolutionary rates. For example, the observation that young genes are subject to less purifying selection (Cai and Petrov 2010) is consistent with the expectations of BLAST artifact.

Although plausible scenarios have been imagined for why young genes might evolve faster, be shorter, and be differentially expressed compared to old genes, these scenarios have arisen in the wake of observations that genes inferred to be young possess these characteristics, rather than as *a priori* hypotheses. It is critical that in the pursuit of new insights about patterns of gene evolution we evaluate our observations in the light of a nuanced understanding of the properties of the tools used to gather them. Computational experiments are not immune to the effects of biased measurement. Here, we have shown that the view of gene age has been distorted by viewing it through the lens of homology search by BLAST.

Methods

Genomic and supporting data

Protein alignments of 6,699 1-to-1 orthologs present in 12 *Drosophila* species were downloaded from FlyBase (ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/alignments/) (Clark et al. 2007). Protein alignments used in (Alba and Castresana 2007) provided by J. Castresana. Gene ontology biological process annotations were downloaded from FlyBase. Gene expression data of fly embryos (Tomancak et al. 2002) was obtained from www.fruitfly.org. Divergence dates were obtained from TimeTree (Hedges et al. 2006) (www.timetree.org).

Rate heterogeneity and main simulation

Rate heterogeneity patterns for each *Drosophila* alignment were estimated using TreePuzzle using the gamma distribution model with 16 rate categories. *D. melanogaster-D. grimishawi* genetic distance was also estimated by TreePuzzle. 37 genes with *D. melanogaster-D. grimshawi* distance = 0 were discarded from further analysis. For each site in the remaining alignments, the rate category contributing most to the likelihood was assigned. Rate heterogeneity patterns for 14 alignments from (Alba and Castresana 2007) were similarly obtained. Amino acid frequencies in the remaining 6,662 alignments were calculated.

Rose was used to simulate sequence evolution. In the main simulation (Gradual A in Table 4.1), a random sequence seed derived from calculated *Drosophila* amino acid frequencies was used as the ancestral sequence. Ancestor sequence length was set equal to the length of the 12 *Drosophila* alignment for each gene. Sequences were evolved according to the rate heterogeneity patterns estimated by TreePuzzle, and according to the tree shown in Figure 4.1. Branch lengths were scaled based on the *D. melanogaster-D. grimshawi* distance for each gene and the divergence times obtained from TimeTree. Following (Alba and Castresana 2007), gap insertion and deletion thresholds were set to 0.0001 and the PAM model of amino acid substitution was used. This process generated 6,662 sequences for each of 11 taxa included in Figure 4.1. The simulation was repeated 10 times, for a total of 6662 * 11 * 10 = 732,820 sequences in the main simulation. Sequences for taxon 1 (representing *D. melanogaster*) were queried against databases of each of the 10 other taxa individually, using BLASTP with an e-value of 0.0001. Gene age was calculated for each protein according to the divergence date of the oldest taxon in which a BLAST hit was found.

Alternate models of evolution

A simulation was conducted using actual *D. melanogaster* sequences as the ancestor sequence. In this case, the length of the ancestor sequence is shorter than in the main simulation by the length of any gaps in *D. melanogaster* relative to the other 11 *Drosophila*. Otherwise, this simulation (Gradual B) was identical to the main simulation. Gradual B was repeated twice.

Four simulations were conducted shuffling the rate categories of 0%, 10%, 25% or 50% of sites. Site to be shuffled were randomly chosen and some sites may have the same rate category after shuffling as before as the process was random and there are only 16 rate classes. In this model, Sequences were evolved as for the main simulation except that no indels or deletions were allowed, until the point indicated by the letter S in Figure 4.1, corresponding to the start of the branch ancestral to Metazoans. At this point, rate heterogeneity patterns were shuffled, and simulation was continued with the new rate assignments for the remaining branches of the phylogeny. Each of these simulations (0%, 10%, 25%, 50%) were repeated 10 times.

Three simulations were conducted to model different levels of relaxation of constraint by constructing alternative rate heterogeneity patterns. In the first model (44% relaxed), for each gene, the lowest N rate categories were selected such that at least 10% of sites were included in those rate categories, and rate category N had a relative rate > 0. The average proportion of sites selected by this procedure was 56%. The selected sites were assigned rate category N. The remaining 44% of sites were assigned rate category 16, i.e., the maximum rate category for the gene. In the second model of relaxed evolution (90% relaxed), 10% of sites were randomly selected in blocks of 5 sites. These sites retained the rate categories assigned by TreePuzzle. The remaining 90% of sites were assigned rate category 16. In the third model of relaxation (100%

relaxed), all sites were assigned rate category 16. For each of these 3 models, simulation was conducted as for the main simulation, substituting he alternate rate heterogeneity patterns described. Each relaxation simulation was repeated 10 times.

Three simulations were conducted to model a burst of rapid evolution (5MY, 10MY, and 20MY). In each of these, sequences were simulated as for the main simulation until the point indicated by the letter B in Figure 4.1, i.e. the point of divergence between *Drosophila* and honeybee. Even during a period of rapid evolution, we expect there is some level of constraint on the protein, or it would likely be lost or pseudogenized. Therefore, the sequence at this node was evolved using the same procedure as the 90% relaxed model for 5, 10 or 20 MY. Then, post-burst rate heterogeneity pattern was created by selecting a random gene from the 6,662 genes, and selecting a random starting point along the length of its alignment. Starting at this point, rate categories were assigned to the post-burst rate heterogeneity pattern for the 90% of sites that had been assigned rate category 16 during the burst. The 10% of sites that were evolving at their original rates during the burst retained those rates after the burst. Finally, the gene was evolved for an additional 372MY, 367MY, and 357MY in the 5MY, 10MY and 20MY burst models, respectively using the post-burst rate heterogeneity pattern. The burst simulation was repeated 10 times.

BLASTP search for homologs and age assignment was repeated for each iteration of the alternate model simulations as for the main simulation.

Estimation of rates of orphan gene formation

The number of new duplicate genes arising per genome per million years in *Drosophila* is 31 (Lynch and Conery 2000). Thus, mutational input in 5MY is 31 * 5 = 155. Let A = the number of recent duplicates available for founder gene formation. The total number of gene lost in a 5MY window (B) can be calculated using the half life of new duplicates (2.9MY) (Lynch and Conery

2000) with the equation for exponential decay. These lost duplicates may be lost due to sequence divergence and escape. We used the rate of escape from measured in the 100% relaxed model at 5MY ($3.75\text{x}10^{-3}$/gene/5MY) to calculate the number of genes lost by escape ($C = B * 3.75\text{x}10^{-3}$). Then, the number of genes lost to pseudogenization or deletion is $D = C - B$. The number of genes that have not been lost during this 5MY window is $E = A - B$. $A_{t+5MY} = E_t + 155$. We set $A_0 = 155$ and iterated over 380MY. Convergence is reached fairly quickly. The convergence value of C represents the equilibrium number of orphan genes formed by escape per 5MY. By multiplying this figure by 377/5 we estimate the number of orphan genes formed in this way since the divergence of *Drosophila* and honeybee.

Acknowledgements

We thank J. Castresana for providing alignments used in Alba and Castresana (2007).

Figure 4.1 Phylogeny showing the relationships of sequences simulated in gene age study. All sequences are simulated; the organism names simply illustrate the evolutionary distances involved. The branch lengths are proportional to divergence times between the indicated taxa. INT 1 and INT 2 do not represent actual taxa, but are evenly spaced between the divergence of plants and bacteria. B indicates the location of a simulated burst of rapid evolution. S represents the location where evolutionary rates were shuffled.

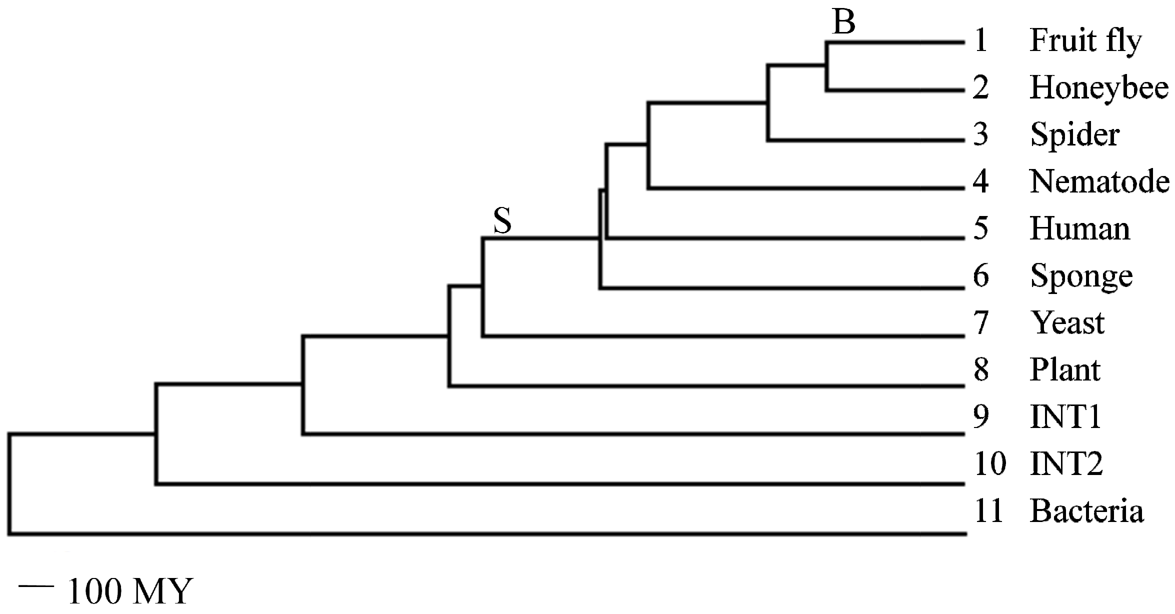


Figure 4.2 Distribution of genetic distance between *D. melanogaster* and *D. grimshawi* for genes of different ages.

First panel (ALL) shows distribution for 6,662 *Drosophila* genes used in the simulation; panels 1-11 show distribution for genes in each age group (1-11). Age groups correspond to divergence dates of branches labeled 1-11 in Figure 4.1.

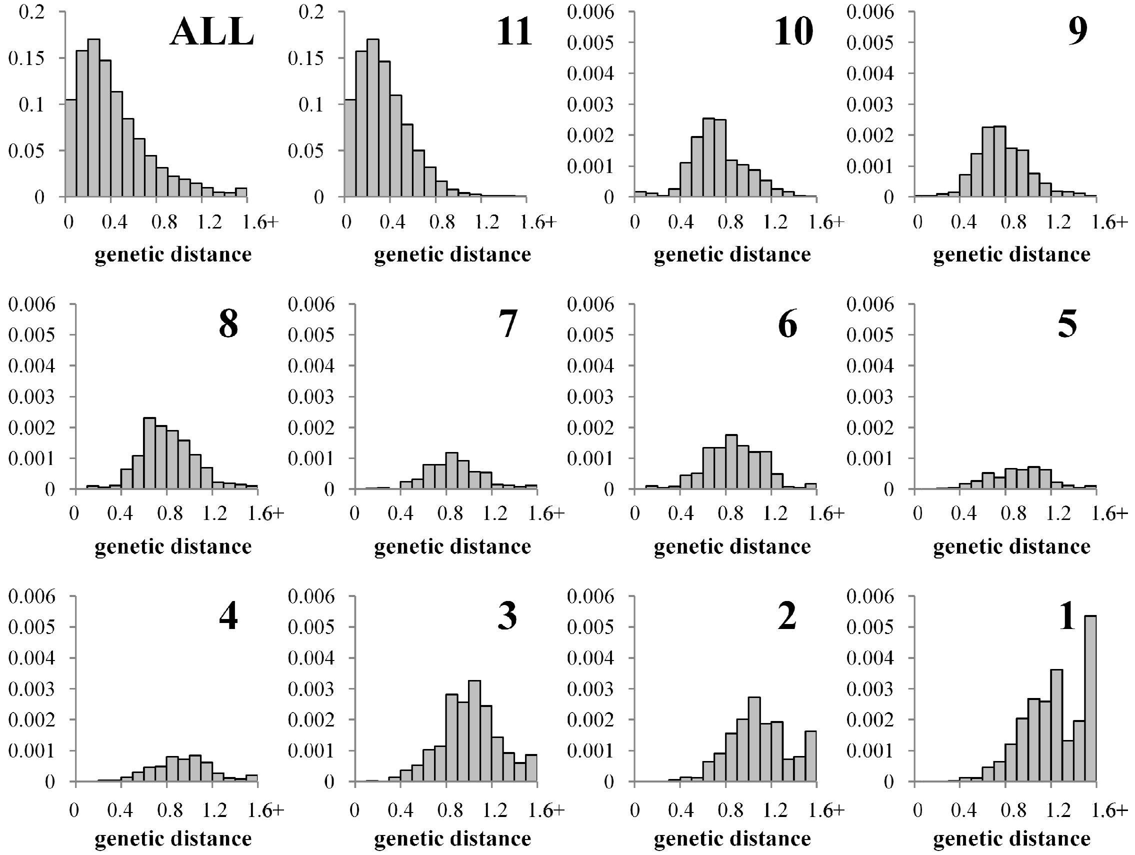


Figure 4.3 Age group within each of 10 approximately equally sized bins of increasing alpha. Age groups correspond to divergence dates of branches labeled 1-11 in Figure 4.1.

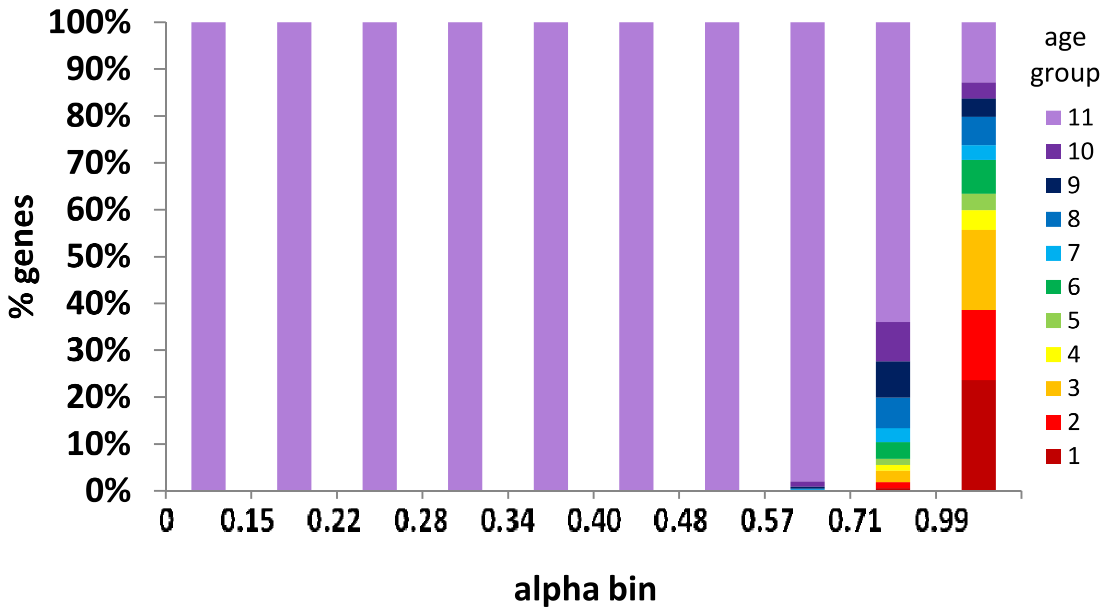
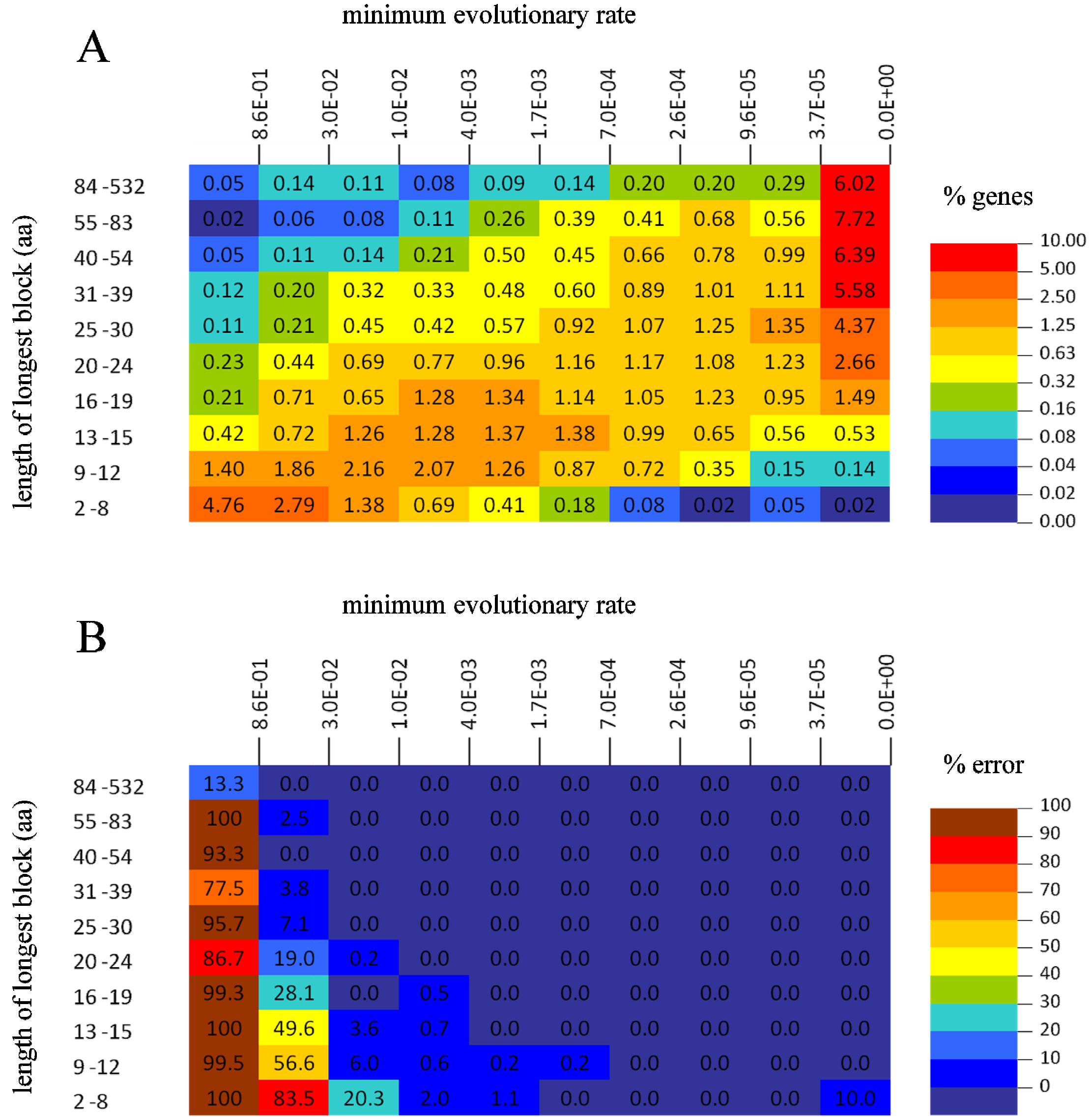


Figure 4.4 Distribution of 6,662 *Drosophila* genes by minimum evolutionary rate class for the gene and maximum block length of the minimum rate class (A) and gene age estimation error rate within each group (B).



minimum evolutionary rate

length of longest block (aa)	8.6E-01	3.0E-02	1.0E-02	4.0E-03	1.7E-03	7.0E-04	2.6E-04	9.6E-05	3.7E-05	0.0E+00
84 -532	0.05	0.14	0.11	0.08	0.09	0.14	0.20	0.20	0.29	6.02
55 -83	0.02	0.06	0.08	0.11	0.26	0.39	0.41	0.68	0.56	7.72
40 -54	0.05	0.11	0.14	0.21	0.50	0.45	0.66	0.78	0.99	6.39
31 -39	0.12	0.20	0.32	0.33	0.48	0.60	0.89	1.01	1.11	5.58
25 -30	0.11	0.21	0.45	0.42	0.57	0.92	1.07	1.25	1.35	4.37
20 -24	0.23	0.44	0.69	0.77	0.96	1.16	1.17	1.08	1.23	2.66
16 -19	0.21	0.71	0.65	1.28	1.34	1.14	1.05	1.23	0.95	1.49
13 -15	0.42	0.72	1.26	1.28	1.37	1.38	0.99	0.65	0.56	0.53
9 -12	1.40	1.86	2.16	2.07	1.26	0.87	0.72	0.35	0.15	0.14
2 -8	4.76	2.79	1.38	0.69	0.41	0.18	0.08	0.02	0.05	0.02

minimum evolutionary rate

length of longest block (aa)	8.6E-01	3.0E-02	1.0E-02	4.0E-03	1.7E-03	7.0E-04	2.6E-04	9.6E-05	3.7E-05	0.0E+00
84 -532	13.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
55 -83	100	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
40 -54	93.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
31 -39	77.5	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25 -30	95.7	7.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20 -24	86.7	19.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16 -19	99.3	28.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0
13 -15	100	49.6	3.6	0.7	0.0	0.0	0.0	0.0	0.0	0.0
9 -12	99.5	56.6	6.0	0.6	0.2	0.2	0.0	0.0	0.0	0.0
2 -8	100	83.5	20.3	2.0	1.1	0.0	0.0	0.0	0.0	10.0

Figure 4.5 Log-odds ratio of expression in mesoderm, ectoderm and endoderm for each age group compared to the average of all genes studied.
Bonferroni corrected p-value is indicated in the table above (two-tailed-hypergeometric test) (A). Rate of novel gene formation per MY by age group. (B) Percent of gene with biological process GO annotation by age group. (C)

A

p-value	< 0.01	ns	ns	ns	ns	ns	ns	ns	ns	ns	<0.001	Mesoderm
	<0.001	ns	ns	ns	ns	<0.001	<0.001	<0.001	ns	ns	<0.001	Endoderm
	ns	ns	<0.05	<0.05	ns	< 0.01	<0.05	ns	< 0.01	< 0.01	<0.001	Ectoderm

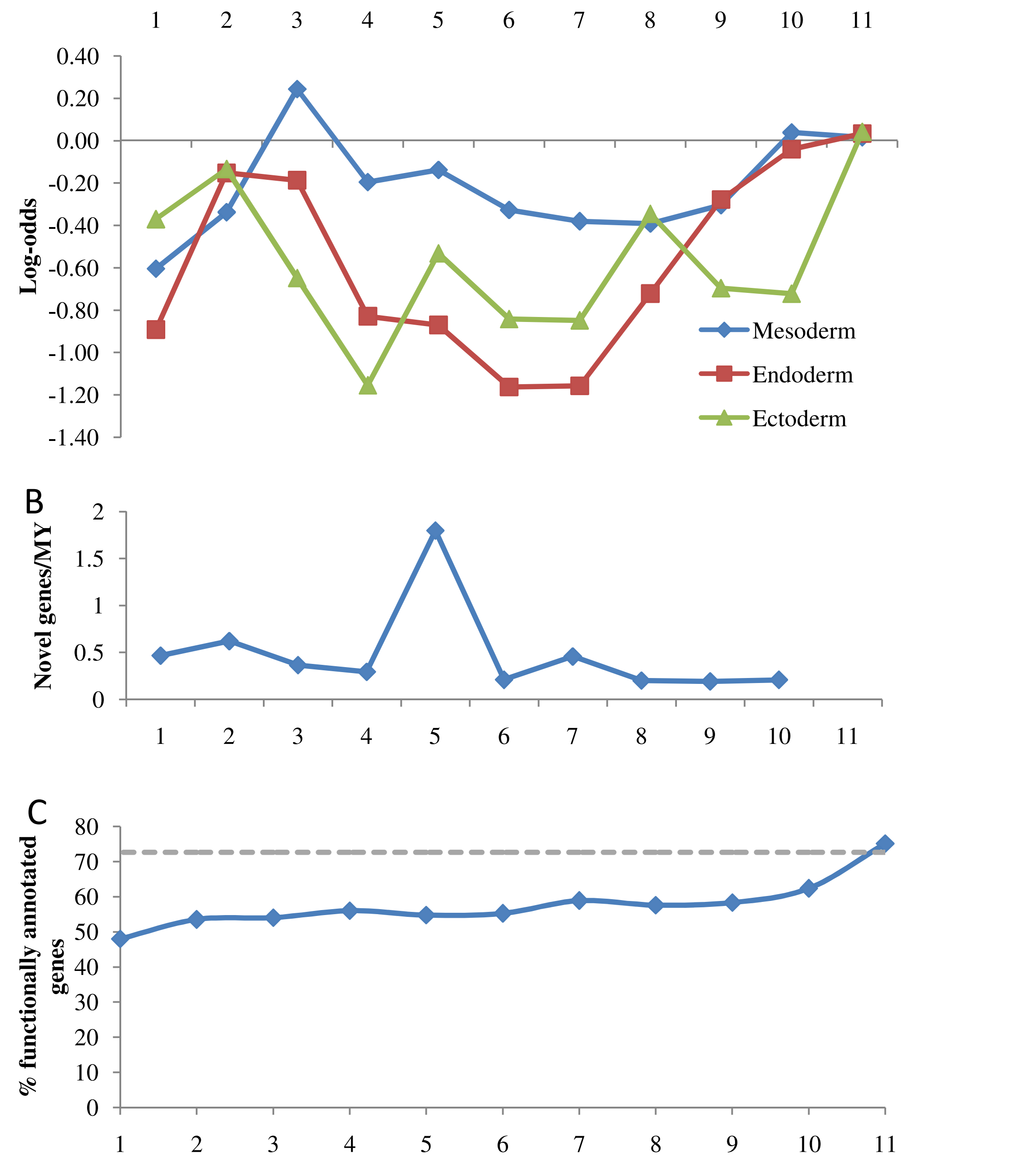


Table 4.1 Parameters and results for simulation models used.

Simulation Name	Seed sequence	Rate heterogeneity pattehrns	Indels	No hits at all	No hit in one or more taxa	No hit to most distant homolog
Gradual evolution models				**Gene Age Estimation Error Rates**		
Gradual A	Random	6662 alignments of 12 *Drosophila*	Yes	2.20%	12.20%	11.80%
Gradual B	*D. mel* actual	Same as Gradual A	Yes	2.20%	11.40%	10.90%
Switch 0%	Random	Same as Gradual A	No	1.20%	9.20%	9.00%
Switch 10%	Random	Same as Gradual A, 10% of sites shuffled	No	1.30%	9.80%	9.50%
Switch 25%	Random	Same as Gradual A, 25% of sites shuffled	No	1.30%	10.70%	10.30%
Switch 50%	Random	Same as Gradual A, 50% of sites shuffled	No	1.30%	24.90%	23.70%
Punctuated evolution models				**Orphan Formation Rates**		
Relaxed 44%	Random	Same as Gradual A, with 56% at rate 1 and 44% at rate 16	Yes	4.50%	15.10%	14.60%
Relaxed 90%	Random	Same as Gradual A, with 10% at normal rate 1 and 90% at rate 16	Yes	64.50%	98.30%	96.30%
Relaxed 100%	Random	Same as Gradual A, with 100% at rate 16	Yes	0.10%	99.10%	99.10%
Burst 5MY	Random	Same as Gradual A, with 5MY as for Relaxed 90%, followed by 372MY at a different rate heterogeneity pattern	No	2.70%	14.50%	13.80%
Burst 10MY	Random	Same as Gradual A, with 10MY as for Relaxed 90%, followed by 367MY at a different rate heterogeneity pattern	No	3.20%	15.50%	14.70%
Burst 20MY	Random	Same as Gradual A, with 20MY as for Relaxed 90%, followed by 357MY at a different rate heterogeneity pattern	No	4.70%	17.50%	16.30%
Gradual evolution models with alternate rate heterogeneity patterns				**BLAST Error Rates**		
Vertebrate	Random	sampled from 8 alignments of 7 vertebrates	Yes	0.50%	9.40%	8.60%
Bilaterian	Random	sampled from 6 alignments of 9 bilaterians	Yes	4.20%	40.30%	39.20%

Chapter 5

Conclusions

In conclusion, I want to address three themes that unite my three projects. First, I will discuss the opportunities for revisiting old questions when new data become available. Second, I will consider the challenges of inference about function and evolutionary history from genomic sequences, high throughput expression data, and other genomic datasets. Finally, I will return to the call for careful hypothesis generation coupled with healthy skepticism that was articulated in the discussion section of Chapter 4.

Old questions, new data

Darwin attributed the success of the human species to "powers of observation, memory, curiosity, imagination, and reason" and inferred that these powers had arisen by the action of natural selection (Darwin 1871) but this conclusion was disputed by Wallace. In Chapter 2, using genome sequences that became available over 130 years after Darwin and Wallace disagreed, I examined the genomic patterns of brain gene evolution and found no evidence of widespread acceleration of the evolution of brain-specific genes, disputing the findings of Dorus and coworkers (Dorus et al. 2004). However, as noted, my results do not rule out positive selection on a subset of genes, but simply reveal the overall pattern of evolution for these genes. Nonetheless, when positive selection has been identified in specific genes known to have severe phenotypic consequences on brain morphology and function when mutated (e.g., null mutations of *ASPM* (Zhang 2003) and *microcephalin* (Evans et al. 2004a) lead to microcephaly and mental retardation), strong objections have been raised regarding the inference that positive selection in these genes was due to enhanced brain function or improved cognitive abilities (Nielsen 2009).

Given this history of conflicting conclusions and interpretations, one might wonder if it is possible to find satisfying answers about the evolutionary history of the human brain. In my opinion, the best approach takes its cue from Darwin – pile on the evidence, from as many sources as possible. For example, objections to the proposal that positive selection in *ASPM* and *microcephalin* led to increased brain size included the lack of evidence that mutations in these genes correlate with intelligence in modern humans (Nielsen 2009). As much as we might shy away from the societal implications of genetic factors that influence intelligence, the growing number of fully sequenced human genomes and associated phenomic data (Durbin et al. 2010) may make it possible to identify variants in brain genes that are relevant to brain size and/or some measure of intelligence. Another emerging resource that can help us to identify functionally important variants is the Neandertal genome sequence (Green et al. 2010). Neandertal encephalization is close to that of modern humans and even greater than that of contemporaneous archaic *Homo sapiens* (Stanyon et al. 1993). Although it is debated what cognitive abilities Neandertals possessed (e.g. (Gunz et al. 2010)), we may parsimoniously expect variants responsible for larger brain size to be present in Neandertal if they arose prior to the divergence of the two species. We may also identify additional Neandertal specific variants of genes involved in brain development and function that show signals of positive selection in that lineage. In addition, it has been suggested that because human intelligence differs in degree rather than in kind from that of other mammals, and because certain other nonprimate mammals have large brains and advanced cognitive capabilities (e.g. elephants), a phylogenomic approach examining a wide spectrum of mammals can shed light on the fundamental processes of brain evolution (Goodman and Sterner 2010). Initiatives such as the BGI "genomic zoo" (http://ldl.genomics.org.cn/page/bgi-g10k.jsp), which ambitiously proposes to sequence 10,000 vertebrate genomes by 2015, will provide the data needed to robustly implement the phylogenomic approach. By combining many types and sources of evidence, as Darwin did for evolution by natural selection in *On the Origin of Species* (Darwin 1859), the story of human brain evolution will become richer, and perhaps at some point can credibly be called a history rather than a mere story.

Although I have focused here on the ways new data will illuminate the history of human brain evolution, these and similar data will no doubt also help uncover the evolutionary history of positive selection on other traits, in human and other lineages. In addition, it will be possible to extract an accounting of the origin of new genes within the vertebrate lineage with a previously impossible level of detail by analyzing the phylogenomic and population genomic resources now being constructed. This analysis can be informative about the processes of gene duplication and orphan gene formation studied in my Chapter 4. Thus, new data can be used to address multiple old questions.

Challenges of functional and evolutionary inference

Different values of the nonsynonymous to synonymous substitution rate ratio (ω) are expected under different evolutionary regimes. For a protein with no impact on organismal fitness, fixation and loss are determined solely by genetic drift. Thus, synonymous and nonsynonymous changes are equally likely to be fixed or lost and the ratio of their substitution rates is expected to be one. However, for a protein that does impact organismal fitness, nonsynonymous changes are more likely to affect the function of a protein than synonymous changes. Nonsynoymous changes may improve function, which is understood to mean that they improve the fitness of the organism carrying such changes, and thus rise to fixation by positive selection, increasing ω compared to the neutral case. Nonsynonymous changes may also destroy or decrease the function of the protein, thereby decreasing the fitness of the organism, and be eliminated by purifying selection. These fundamental principles in molecular evolution build on two even more fundamental principles in biology: function and fitness. With the availability of large amounts of sequence data and powerful programs to detect positive selection (Yang 1997), it has been possible to catalog potential sites of positive selection without any a priori hypotheses about the function of such sites or their fitness effects, as for example in my Chapter 3. In the face of growing numbers of such reports of positive selection, it has been suggested that functional validation be required to link the molecular signature of selection to the fitness effect of the putatively selected variant (MacCallum and Hill 2006).

The challenges associated with measuring fitness and function are daunting. Even in tractable model organisms such as *Saccharomyces cerevisiae*, where fitness can be precisely assayed by competition among strains in the lab, fitness in a synthetic environment may not a good predictor of fitness in natural environments experienced in the evolutionary history of the species, the “lab-nature mismatch”. Of course, any past episodes of selection in a species depend on its fitness in the past environment. A study of thousands of simulated nutritional conditions failed to find simulated conditions that create a better correlation between evolutionary rate and condition-specific gene essentiality than in lab conditions, and thus concluded that the lab-nature mismatch hypothesis was incorrect (Wang and Zhang 2009). However, the study examined many nutritional conditions, but not other relevant aspects of environment less amenable to simulation, such as temperature, intra- and interspecific competition, toxins, pathogens, availability of water, etc. Any of these, or fluctuations in environment, may result in significant differences between lab fitness and fitness in the evolutionarily relevant environment. We are further hampered in assessing fitness by an unfortunate lack of information about even the present day environments experienced by wild yeast (Fay and Benavides 2005).

Despite these challenges, the task of assessing fitness and function in yeast seems trivial by comparison to the prospects in human. Even supposing we had perfect data about genotype-phenotype-environment relationships for some human trait such as intelligence, inferring fitness or function from that data in the context of other human traits and in the context of cultural and physical environments might still be impossible. In general, we make simplifying assumptions based on apparent trait values, as did Darwin when he gave human intelligence, and even more specifically "articulate language", primary credit for human evolutionary success (Darwin 1871).

Another challenge for making evolutionary inferences about fitness and function from molecular data derives from imperfect or nonexistent genotype-phenotype mapping. From first principles, nonsynonymous changes may be beneficial, deleterious, or even neutral, but in general we do not have a good way to tell which case applies from sequence data alone. Thus we infer the evolutionary process via observing the overall pattern of substitution in multiple lineages over multiple sites. But, because of the inherent stochasiticity of evolution, we can occasionally be misled by chance events. The levels of false positives in a test for positive selection are generally quite low (Table 3.4), but when false positives occur, it may be due to the chance occurrence of two nonsynonymous substitutions within a single codon (Mallick et al. 2009). Furthermore, when only one or a few sites are subject to positive selection, it is difficult to distinguish this from relaxation of constraint, leading to high levels of false negatives in tests for positive selection (Table 3.6).

Skepticism, hypotheses, and the self-correcting nature of science

Skepticism is a critical aspect of scientific inquiry. My analysis of the phylostratigraphic method and bias in the measurement of gene age emphasizes the need for more critical evaluation of results based on this and similar methods. Despite the publication of two simulation studies showing that measurement of gene age is biased, many studies were subsequently performed without any attempt to ascertain whether bias in measurement impacted their findings (Domazet-Loso et al. 2007; Domazet-Loso and Tautz 2008; Prachumwat and Li 2008; Cai et al. 2009; Prat et al. 2009; Cai and Petrov 2010; Domazet-Loso and Tautz 2010a; Domazet-Loso and Tautz 2010b), and thus may contain spurious results.

Bias can take other forms in addition to measurement error. For example, the existing literature on a given subject may unduly influence our expectations about future findings. The conclusion that selection is less effective in smaller populations is uncontroversial. Reports of lower effective population size in humans compared to chimpanzees or the human-chimp common ancestor have been confirmed using multiple different methods and datasets ((Ferris et al. 1981; Takahata et al.

1995; Ruvolo 1997; Kaessmann et al. 1999; Chen and Li 2001; Kaessmann et al. 2001; Stone et al. 2002; Wall 2003; Fischer et al. 2004). However, my finding, consistent with these prior findings, that both positive and purifying selection have been less efficient in the human lineage compared to the chimpanzee lineage, was considered a newsworthy surprise (Hopkin 2007). Why? Almost certainly because of an anthropocentric view of evolution, where the changes that have occurred in the human lineage are well studied and obvious even to non-scientists, and the changes that have occurred in the chimp lineage are mostly unknown. At present, 147 studies of positive selection in both human and chimpanzee are cataloged in ISI Web of Science, and 4,124 in human alone, compared to only 19 studies that focus on chimpanzee alone. Careful study of chimpanzee traits has the potential to overturn the idea that chimpanzees have changed little since their divergence from humans. For example, detailed examination of the morphological and ontological aspects of knuckle walking in chimpanzee and gorilla revealed that this is not an ancestral trait, but rather evolved independently in chimpanzees and gorillas from non-knuckle walking ancestors. In the past, knuckle walking was considered to be an ancestral trait, so the new research shows that changes in locomotion have occurred in humans, chimpanzees, and gorillas. While I do not expect or suggest that the amount of research on chimpanzee evolution will ever equal that on human evolution, it is critical that we are aware of the biases inherent in the lopsided research agenda.

In this dissertation, I present work that criticizes previous studies, as well as two published chapters (Shi et al. 2006; Bakewell et al. 2007) which have in turn been the subject of critical analysis by other authors (Mallick et al. 2009; Crespi et al. 2010). Each careful examination of data and methods may lead to further insights. Thus, I conclude that the process of scientific exchange, of open criticism and careful reading of the work of others, can continue to move us toward a clearer understanding of evolutionary processes and history.

Appendix

Genes (*n*=154) showing significant signal of positive selection in human before Bonferroni correction

Ensembl Gene ID	Gene Name	Gene Description	$w_H{}^1$	$w_H{}^2$	p-value[2]
ENSG00000165659	DACH1	Dachshund homolog 1 (Dach1).	0.205	999.000	0.0000
ENSG00000188153	COL4A5	Collagen alpha-5(IV) chain precursor.	0.205	999.000	0.0000
ENSG00000186395	KRT10	Keratin, type I cytoskeletal 10 (Cytokeratin-10) (CK-10) (Keratin-10) (K10).	0.154	268.730	0.0001
ENSG00000188712	Q8NHB3_HUMAN	Seven transmembrane helix receptor.	0.342	999.000	0.0001
ENSG00000186683			0.615	999.000	0.0001
ENSG00000147889	CDKN2A	Cyclin-dependent kinase inhibitor 2A, isoform 4 (p14ARF) (p19ARF).	1.230	999.000	0.0001
ENSG00000118298	CA14	Carbonic anhydrase 14 precursor (EC 4.2.1.1) (Carbonic anhydrase XIV) (Carbonate dehydratase XIV) (CA-XIV).	0.820	999.000	0.0002
ENSG00000180509	KCNE1	Potassium voltage-gated channel subfamily E member 1 (IKs producing slow voltage-gated potassium channel beta subunit Mink) (Minimal potassium channel) (Delayed rectifier potassium channel subunit IsK).	0.410	999.000	0.0002
ENSG00000149054	ZNF215	Zinc finger protein 215 (BWSCR2-associated zinc-finger protein 2) (BAZ 2).	1.230	999.000	0.0003
ENSG00000136003	NIFUN	NifU-like N-terminal domain-containing protein, mitochondrial precursor (NifU-like protein) (Iron-sulfur cluster assembly enzyme ISCU).	0.410	999.000	0.0003
ENSG00000164626	KCNK5	Potassium channel subfamily K member 5 (Acid-sensitive potassium channel protein TASK-2) (TWIK-related acid-sensitive K(+) channel 2).	0.051	612.830	0.0003
ENSG00000171522	PTGER4	Prostaglandin E2 receptor, EP4 subtype (Prostanoid EP4 receptor) (PGE receptor, EP4 subtype).	0.820	999.000	0.0003
ENSG00000174948	GPR149	Probable G-protein coupled receptor 149 (G-protein coupled receptor PGR10).	0.205	999.000	0.0005
ENSG00000163069	SGCB	Beta-sarcoglycan (Beta-SG) (43 kDa dystrophin-associated glycoprotein) (43DAG) (A3b).	0.137	242.742	0.0005
ENSG00000169738	DCXR	L-xylulose reductase (EC 1.1.1.10) (XR) (Dicarbonyl/L-xylulose reductase) (Kidney dicarbonyl reductase) (kiDCR) (Carbonyl reductase II) (Sperm surface protein P34H).	0.205	906.960	0.0005
ENSG00000164935	TM7SF4	Transmembrane 7 superfamily member 4 (Dendritic cell-specific transmembrane protein) (DC-STAMP) (IL-4-induced protein) (FIND).	0.820	999.000	0.0005
ENSG00000164977			0.246	999.000	0.0006
ENSG00000119596	YLPM1	YLP motif-containing protein 1 (Nuclear protein ZAP3) (ZAP113).	0.308	774.738	0.0008
ENSG00000166004	KIAA1731	CDNA FLJ37899 fis, clone CD34C3000314.	1.025	1.110	0.0008
ENSG00000119431	HDHD3	haloacid dehalogenase-like hydrolase domain containing 3	0.410	999.000	0.0008
ENSG00000185409	Q8IYB0_HUMAN		0.410	999.000	0.0009
ENSG00000165632	TAF3	TAFII140 protein (Fragment).	0.410	999.000	0.0009
ENSG00000136805	NP_055149.1		0.205	999.000	0.0010

Ensembl Gene ID	Gene Name	Gene Description	w_H[1]	w_H[2]	p-value[2]
ENSG00000165325	CCDC67		0.410	999.000	0.0012
ENSG00000177138	FAM9B	Protein FAM9B.	0.410	999.000	0.0013
ENSG00000150347	ARID5B	AT-rich interactive domain-containing protein 5B (ARID domain- containing protein 5B) (Mrf1-like) (Modulator recognition factor 2) (MRF-2).	0.137	999.000	0.0015
ENSG00000092470	WDR76	WD repeat domain 76	0.410	914.069	0.0015
ENSG00000160051	IQCC	IQ motif containing C	1.640	999.000	0.0016
ENSG00000100889	PCK2	Phosphoenolpyruvate carboxykinase [GTP], mitochondrial precursor (EC 4.1.1.32) (Phosphoenolpyruvate carboxylase) (PEPCK-M).	0.308	999.000	0.0017
ENSG00000171227	TMEM37	Voltage-dependent calcium channel gamma-like subunit (Neuronal voltage-gated calcium channel gamma-like subunit) (Transmembrane protein 37).	0.205	999.000	0.0017
ENSG00000198935			0.410	999.000	0.0018
ENSG00000090339	ICAM1	Intercellular adhesion molecule 1 precursor (ICAM-1) (Major group rhinovirus receptor) (CD54 antigen).	0.342	156.095	0.0019
ENSG00000152939	MARVELD2	MARVEL domain containing 2 isoform 1	0.615	526.677	0.0021
ENSG00000186074	CD300LF	NK inhibitory receptor precursor	0.547	264.084	0.0022
ENSG00000164393	GPR111	Probable G-protein coupled receptor 111 (G-protein coupled receptor PGR20).	1.230	999.000	0.0023
ENSG00000138029	HADHB	Trifunctional enzyme beta subunit, mitochondrial precursor (TP-beta) [Includes: 3-ketoacyl-CoA thiolase (EC 2.3.1.16) (Acetyl-CoA acyltransferase) (Beta-ketothiolase)].	0.410	999.000	0.0024
ENSG00000181908	XR_000554.1		0.410	999.000	0.0027
ENSG00000133101	CCNA1	Cyclin-A1.	0.410	663.547	0.0027
ENSG00000187475	HIST1H1T	Histone H1t (Testicular H1 histone).	2.460	999.000	0.0031
ENSG00000164944	KIAA1429		0.308	482.855	0.0032
ENSG00000131018	SYNE1	Nesprin-1 (Nuclear envelope spectrin repeat protein 1) (Synaptic nuclear envelope protein 1) (Syne-1) (Myocyte nuclear envelope protein 1) (Myne-1) (Enaptin).	0.376	1.278	0.0033
ENSG00000116141	MARK1	Serine/threonine-protein kinase MARK1 (EC 2.7.1.37) (MAP/microtubule affinity-regulating kinase 1).	0.410	404.497	0.0034
ENSG00000143851	PTPN7	Tyrosine-protein phosphatase non-receptor type 7 (EC 3.1.3.48) (Protein-tyrosine phosphatase LC-PTP) (Hematopoietic protein-tyrosine phosphatase) (HEPTP).	0.273	999.000	0.0036
ENSG00000138658	NP_060862.2		0.513	560.568	0.0039
ENSG00000163500			0.228	311.489	0.0040
ENSG00000132677	RHBG	Rhesus blood group, B glycoprotein	0.410	409.973	0.0041
ENSG00000123901	GPR83	Probable G-protein coupled receptor 83 precursor (G-protein coupled receptor 72).	0.205	144.724	0.0042
ENSG00000184574	GPR92	Probable G-protein coupled receptor 92.	0.410	256.320	0.0044
ENSG00000104804	TULP2	Tubby-related protein 2 (Tubby-like protein 2).	0.410	999.000	0.0044
ENSG00000196893	Q8N2W8_HUMAN		1.640	999.000	0.0047
ENSG00000170891	CYTL1	Cytokine-like protein 1 precursor (Protein C17).	0.205	999.000	0.0047

Ensembl Gene ID	Gene Name	Gene Description	w_H[1]	w_H[2]	p-value[2]
ENSG00000158292	GPR153	Probable G-protein coupled receptor 153 (G-protein coupled receptor PGR1).	0.308	999.000	0.0049
ENSG00000158553	POM121L2		0.923	799.779	0.0050
ENSG00000130508	PXDN	PXDN protein (Fragment).	0.082	137.188	0.0050
ENSG00000087301	KIAA1344		0.410	745.691	0.0052
ENSG00000129696	NP_079391.1		0.513	392.348	0.0053
ENSG00000130818	ZNF426	Zinc finger protein 426.	0.615	237.650	0.0054
ENSG00000182551	ADI1	1,2-dihydroxy-3-keto-5-methylthiopentene dioxygenase (EC 1.13.-.-) (Aci-reductone dioxygenase) (ARD) (Membrane-type 1 matrix metalloproteinase cytoplasmic tail-binding protein 1) (MTCBP-1) (Submergence-induced protein 2 homolog) (SIPL).	0.410	999.000	0.0055
ENSG00000135931	ARMC9		0.820	263.934	0.0056
ENSG00000033867	SLC4A7	Sodium bicarbonate cotransporter 3 (Sodium bicarbonate cotransporter 2) (Sodium bicarbonate cotransporter 2b) (Bicarbonate transporter) (Solute carrier family 4 member 7).	0.164	428.510	0.0057
ENSG00000011028	MRC2	Macrophage mannose receptor 2 precursor (Urokinase receptor-associated protein) (Endocytic receptor 180) (CD280 antigen).	0.176	128.163	0.0059
ENSG00000178977	Q8NAT9_HUMAN		2.050	999.000	0.0059
ENSG00000183005			0.957	114.349	0.0061
ENSG00000147223	Q5JRB8_HUMAN	Novel protein (Fragment).	0.410	999.000	0.0061
ENSG00000139182	CLSTN3	Calsyntenin-3 precursor.	0.103	213.710	0.0064
ENSG00000162723	SLAMF9	SLAM family member 9 precursor (CD2 family member 10) (CD2F-10) (CD84 homolog 1) (CD84-H1).	0.820	128.061	0.0065
ENSG00000185087	NP_872368.1		1.230	999.000	0.0068
ENSG00000133805	AMPD3	AMP deaminase 3 (EC 3.5.4.6) (AMP deaminase isoform E) (Erythrocyte AMP deaminase).	0.410	181.036	0.0068
ENSG00000164440	TXLNB	Beta-taxilin (Muscle-derived protein 77) (hMDP77).	0.328	179.365	0.0070
ENSG00000177553	Q8NDA9_HUMAN	Novel protein (Fragment).	1.435	999.000	0.0070
ENSG00000172769	OR5B3	Olfactory receptor 5B3 (Olfactory receptor OR11-239).	3.690	999.000	0.0072
ENSG00000167987	VPS37C	vacuolar protein sorting 37C	0.547	532.924	0.0072
ENSG00000165409	TSHR	Thyrotropin receptor precursor (TSH-R) (Thyroid-stimulating hormone receptor).	0.547	211.287	0.0074
ENSG00000118600	TMEM5	Transmembrane protein 5.		999.000	0.0074
ENSG00000160781	PAQR6	Progestin and adipoQ receptor family member VI.	0.410	184.802	0.0076
ENSG00000036672	USP2	Ubiquitin carboxyl-terminal hydrolase 2 (EC 3.1.2.15) (Ubiquitin thiolesterase 2) (Ubiquitin-specific-processing protease 2) (Deubiquitinating enzyme 2) (41 kDa ubiquitin-specific protease).	0.615	274.810	0.0076
ENSG00000196476	CT096_HUMAN		1.435	108.387	0.0079
ENSG00000197128	NP_001019767.1		0.615	204.378	0.0082
ENSG00000100141	PISD	Phosphatidylserine decarboxylase proenzyme (EC 4.1.1.65) [Contains:	0.205	999.000	0.0085

Ensembl Gene ID	Gene Name	Gene Description	w_H[1]	w_H[2]	p-value[2]
		Phosphatidylserine decarboxylase alpha chain; Phosphatidylserine decarboxylase beta chain].			
ENSG00000196242	OR2C3	Olfactory receptor 2C3.		999.000	0.0094
ENSG00000137727	ARHGAP20	Rho GTPase activating protein 20	1.025	228.974	0.0103
ENSG00000170264	NP_115556.1		0.547	163.768	0.0103
ENSG00000197586	ENTPD6	Ectonucleoside triphosphate diphosphohydrolase 6 (EC 3.6.1.6) (NTPDase6) (CD39 antigen-like 2).	0.410	999.000	0.0104
ENSG00000188996	HUS1B	Checkpoint protein HUS1B (hHUS1B).	0.683	115.765	0.0115
ENSG00000188039	NP_001007526.1		0.718	282.228	0.0117
ENSG00000018607	ZNF221	Zinc finger protein 221.	1.025	96.486	0.0119
ENSG00000198169	ZNF251	Zinc finger protein 251 (Fragment).	1.230	999.000	0.0119
ENSG00000186530	XKR5	XK-related protein 5.	3.280	999.000	0.0121
ENSG00000188683			1.640	207.826	0.0127
ENSG00000137507	LRRC32	Leucine-rich repeat-containing protein 32 precursor (GARP protein) (Garpin) (Glycoprotein A repetitions predominant).	0.820	999.000	0.0136
ENSG00000141837	CACNA1A	Voltage-dependent P/Q-type calcium channel alpha-1A subunit (Voltage- gated calcium channel alpha subunit Cav2.1) (Calcium channel, L type, alpha-1 polypeptide isoform 4) (Brain calcium channel I) (BI).	0.205	170.058	0.0140
ENSG00000163071	SPATA18	spermatogenesis associated 18 homolog	1.093	87.723	0.0143
ENSG00000178966	NP_079221.1		1.640	236.525	0.0149
ENSG00000132357	CARD6	Caspase recruitment domain-containing protein 6.	0.752	98.346	0.0150
ENSG00000198483	ANKRD35	ankyrin repeat domain 35	0.683	147.562	0.0151
ENSG00000146232	NFKBIE	NF-kappa-B inhibitor epsilon (NF-kappa-BIE) (I-kappa-B-epsilon) (IkappaBepsilon) (IKB-epsilon) (IKBE).	0.410	131.912	0.0153
ENSG00000158258	CLSTN2	Calsyntenin-2 precursor.	0.224	94.972	0.0154
ENSG00000102805	CLN5	Ceroid-lipofuscinosis neuronal protein 5 (Protein CLN5).	0.308	219.881	0.0158
ENSG00000176900	OR51T1	Olfactory receptor 51T1.	0.273	111.343	0.0159
ENSG00000129219	PLD2	Phospholipase D2 (EC 3.1.4.4) (PLD 2) (Choline phosphatase 2) (Phosphatidylcholine-hydrolyzing phospholipase D2) (PLD1C) (hPLD2).	0.205	183.490	0.0167
ENSG00000091157	WDR7	WD-repeat protein 7 (TGF-beta resistance-associated protein TRAG) (Rabconnectin-3 beta).	0.246	47.927	0.0168
ENSG00000101405	OXT	Oxytocin-neurophysin 1 precursor (OT-NPI) [Contains: Oxytocin (Ocytocin); Neurophysin 1].	0.205	52.642	0.0170
ENSG00000174038	NP_976044.1		3.690	4.740	0.0171
ENSG00000182372	CLN8	Protein CLN8.	0.137	115.068	0.0175
ENSG00000168575	SLC20A2	solute carrier family 20, member 2	0.164	143.491	0.0183
ENSG00000122971	ACADS	Acyl-CoA dehydrogenase, short-chain specific, mitochondrial precursor (EC 1.3.99.2)	0.137	80.204	0.0189

Ensembl Gene ID	Gene Name	Gene Description	w_H[1]	w_H[2]	p-value[2]
		(SCAD) (Butyryl-CoA dehydrogenase).			
ENSG00000019169	MARCO	Macrophage receptor MARCO (Macrophage receptor with collagenous structure) (Scavenger receptor class A member 2).	0.615	70.123	0.0196
ENSG00000165238	WNK2	Serine/threonine-protein kinase WNK2 (EC 2.7.1.37) (Protein kinase with no lysine 2) (Protein kinase, lysine-deficient 2).	0.351	999.000	0.0198
ENSG00000182805	NP_001013756.1		1.025	999.000	0.0202
ENSG00000197674	Q8NH68_HUMAN	Seven transmembrane helix receptor.		999.000	0.0206
ENSG00000002726	NM_001091.2	amiloride binding protein 1 (amine oxidase (copper-containing)) (ABP1), mRNA	0.293	999.000	0.0207
ENSG00000139637	MYG1_HUMAN	MYG1 protein.	0.410	166.481	0.0208
ENSG00000105501	SIGLEC5	Sialic acid-binding Ig-like lectin 5 precursor (Siglec-5) (Obesity- binding protein 2) (OB-binding protein 2) (OB-BP2) (CD33 antigen-like 2) (CD170 antigen).		999.000	0.0214
ENSG00000095637	SORBS1	Sorbin and SH3 domain-containing protein 1 (Ponsin) (c-Cbl-associated protein) (CAP) (SH3 domain protein 5) (SH3P12).	0.234	120.546	0.0218
ENSG00000196118	NP_001014979.1		0.205	999.000	0.0222
ENSG00000141458	NPC1	Niemann-Pick C1 protein precursor.	0.469	47.724	0.0224
ENSG00000189348	XP_934528.1	PREDICTED: hypothetical protein XP_934528	0.547	256.471	0.0225
ENSG00000120875	DUSP4	Dual specificity protein phosphatase 4 (EC 3.1.3.48) (EC 3.1.3.16) (Mitogen-activated protein kinase phosphatase 2) (MAP kinase phosphatase 2) (MKP-2) (Dual specificity protein phosphatase hVH2).	0.103	685.022	0.0227
ENSG00000104635	SLC39A14	solute carrier family 39 (zinc transporter), member 14	0.205	116.647	0.0227
ENSG00000008438	PGLYRP1	Peptidoglycan recognition protein precursor (PGRP-S).	2.050	999.000	0.0245
ENSG00000086288	TXNDC3	Thioredoxin domain-containing protein 3 (Spermatid-specific thioredoxin-2) (Sptrx-2) (NM23-H8).	0.615	132.598	0.0245
ENSG00000164520	RAET1E	NKG2D ligand 4 precursor (NKG2D ligand 4) (NKG2DL4) (N2DL-4) (Retinoic acid early transcript 1E) (Lymphocyte effector toxicity activation ligand) (RAE-1-like transcript 4) (RL-4).	0.410	162.719	0.0249
ENSG00000064042	NP_055803.1		0.205	226.199	0.0257
ENSG00000165125	TRPV6	Transient receptor potential cation channel subfamily V member 6 (TrpV6) (Epithelial calcium channel 2) (ECaC2) (Calcium transport protein 1) (CaT1) (CaT-like) (CaT-L).	0.513	38.372	0.0269
ENSG00000125631			3.280	4.037	0.0275
ENSG00000140950	NP_065998.2		0.410	58.150	0.0283
ENSG00000109674	NEIL3	Endonuclease VIII-like 3 (Nei-like 3) (DNA glycosylase FPG2).		999.000	0.0293
ENSG00000174655			0.615	157.318	0.0294
ENSG00000154719	MRPL39	Mitochondrial 39S ribosomal protein L39 (L39mt) (MRP-L39) (MRP-L5).	0.513	46.491	0.0301
ENSG00000100376	CV008_HUMAN		0.820	36.569	0.0306
ENSG00000187705	Q6ZUC2_HUMAN	CDNA FLJ43826 fis, clone TESTI4002703.		238.376	0.0312

Ensembl Gene ID	Gene Name	Gene Description	w_H[1]	w_H[2]	p-value[2]
ENSG00000167945	NP_001013660.1		1.640	999.000	0.0314
ENSG00000153201	RANBP2	Ran-binding protein 2 (RanBP2) (Nuclear pore complex protein Nup358) (Nucleoporin Nup358) (358 kDa nucleoporin) (P270).	0.351	133.102	0.0331
ENSG00000062370	ZNF228	Zinc finger protein 228.	0.478	58.786	0.0338
ENSG00000171936	OR10H3	Olfactory receptor 10H3.		999.000	0.0339
ENSG00000088836	SLC4A11	Sodium bicarbonate transporter-like protein 11 (Bicarbonate transporter-related protein 1) (Solute carrier family 4 member 11).	0.342	97.814	0.0342
ENSG00000162745	OLFML2B	olfactomedin-like 2B	0.319	70.723	0.0347
ENSG00000180116	NP_001026918.2			178.068	0.0371
ENSG00000182348	NP_857597.1		1.312	50.379	0.0373
ENSG00000147804	SLC39A4	Zinc transporter ZIP4 precursor (Solute carrier family 39 member 4).	2.050	14.143	0.0392
ENSG00000076382	SPAG5	Sperm-associated antigen 5 (Astrin) (Mitotic spindle-associated protein p126) (MAP126) (Deepest).		999.000	0.0397
ENSG00000124568	SLC17A1	Renal sodium-dependent phosphate transport protein 1 (Sodium/phosphate cotransporter 1) (Na(+)/PI cotransporter 1) (Renal sodium-phosphate transport protein 1) (Renal Na(+)-dependent phosphate cotransporter 1) (Solute carrier family 18 member 1) (NA/PI-4)		999.000	0.0402
ENSG00000082269	KIAA1411		0.273	74.664	0.0412
ENSG00000186453	Q86W67_HUMAN		2.050	999.000	0.0414
ENSG00000102794				999.000	0.0423
ENSG00000189424				999.000	0.0429
ENSG00000166801	FAM111A		2.870	999.000	0.0432
ENSG00000083622	NP_060296.1		3.280	50.896	0.0454
ENSG00000122490	PQLC1	PQ loop repeat containing 1	0.273	145.404	0.0463
ENSG00000164649	CDCA7L	transcription factor RAM2	0.273	151.192	0.0466
ENSG00000104974	LILRA1	Leukocyte immunoglobulin-like receptor subfamily A member 1 precursor (Leucocyte immunoglobulin-like receptor 6) (LIR-6) (CD85i antigen).	0.525	11.464	0.0486
ENSG00000184278				999.000	0.0491
ENSG00000162591	EGFL3	EGF-like domain-containing protein 3 precursor (Multiple EGF-like domain protein 3) (Multiple epidermal growth factor-like domains 6).	3.690		0.0498
ENSG00000179750	APOBEC3B	Probable DNA dC->dU-editing enzyme APOBEC-3B (EC 3.5.4.-) (Phorbolin- 1-related protein) (Phorbolin-2/3).	1.333		0.0498

[1] Gene descriptions downloaded from Ensembl. Blanks indicate information not available or unknown.

[2] Nonsynonymous/synonymous substitution rate ratio, computed by 0.41n /s . Blank indicates no synonymous changes.

[3] Nonsynonymous/synonymous substitution rate ratio, computed by PAML for foreground site class 2b

[4] Uncorrected p-value from likelihood ratio test

Genes (*n*=233) showing significant signal of postive selection in chimpanzee before Bonferroni correction

Ensembl Gene ID	Gene Name	Gene Description[1]	w_C[2]	w_C[3]	p-value[4]
ENSG00000167522	ANKRD11	Ankyrin repeat domain-containing protein 11 (Ankyrin repeat-containing cofactor 1).	0.256	999.000	0.0000
ENSG00000126822	PLEKHG3	pleckstrin homology domain containing, family G, member 3	0.492	999.000	0.0000
ENSG00000152242	CR025_HUMAN		0.137	999.000	0.0000
ENSG00000117713	ARID1A	AT-rich interactive domain-containing protein 1A (ARID domain- containing protein 1A) (SWI/SNF-related, matrix-associated, actin- dependent regulator of chromatin subfamily F member 1) (SWI-SNF complex protein p270) (B120) (SWI-like protein) (Osa homolog	0.123	999.000	0.0000
ENSG00000185727			0.205	999.000	0.0000
ENSG00000165521	EML5	echinoderm microtubule associated protein like 5	0.252	999.000	0.0000
ENSG00000002746	HECW1	NEDD4-like ubiquitin-protein ligase 1	0.186	999.000	0.0000
ENSG00000198308	Q9NSI3_HUMAN		0.273	999.000	0.0000
ENSG00000116141	MARK1	Serine/threonine-protein kinase MARK1 (EC 2.7.1.37) (MAP/microtubule affinity-regulating kinase 1).	0.410	999.000	0.0000
ENSG00000091986	NP_955806.1	steroid-sensitive protein 1	0.351	999.000	0.0000
ENSG00000147036	LANCL3	LanC lantibiotic synthetase component C-like 3	0.205	999.000	0.0000
ENSG00000112276	BVES	Blood vessel epicardial substance (hBVES) (Popeye domain-containing protein 1) (Popeye protein 1).	0.273	999.000	0.0000
ENSG00000106415	GLCCI1	glucocorticoid induced transcript 1	0.205	999.000	0.0000
ENSG00000130227	XPO7	Exportin-7 (Ran-binding protein 16).	0.205	999.000	0.0000
ENSG00000096401	CDC5L	Cell division cycle 5-like protein (Cdc5-like protein) (Pombe cdc5- related protein).	0.164	999.000	0.0000
ENSG00000126010	GRPR	Gastrin-releasing peptide receptor (GRP-R) (GRP-preferring bombesin receptor).	0.273	999.000	0.0000
ENSG00000100345	MYH9	Myosin-9 (Myosin heavy chain, nonmuscle IIa) (Nonmuscle myosin heavy chain IIa) (NMMHC II-a) (NMMHC-IIA) (Cellular myosin heavy chain, type A) (Nonmuscle myosin heavy chain-A) (NMMHC-A).	0.112	211.772	0.0000
ENSG00000129116	PALLD	palladin	0.059	999.000	0.0000
ENSG00000197996	PADI6	Peptidylarginine deiminase type VI (EC 3.5.3.15) (Peptidylarginine deiminase type 6).	0.246	915.618	0.0000
ENSG00000198700	IPO9	Importin-9 (Imp9) (Ran-binding protein 9) (RanbP9).	0.816	999.000	0.0000
ENSG00000123607	TTC21B	tetratricopeptide repeat domain 21B	0.205	999.000	0.0000
ENSG00000178662	TAIP2_HUMAN	TGF-beta-induced apoptosis protein 2 (TAIP-2).	0.082	999.000	0.0000
ENSG00000073050	XRCC1	DNA-repair protein XRCC1 (X-ray repair cross-complementing protein 1).	0.410	999.000	0.0000
ENSG00000059588	TARBP1	TAR RNA binding protein 1	0.256	999.000	0.0000
ENSG00000144130	NT5DC4		0.342	984.278	0.0000
ENSG00000180530	NRIP1	Nuclear receptor-interacting protein 1 (Nuclear factor RIP140) (Receptor-interacting	0.273	999.000	0.0000

Ensembl Gene ID	Gene Name	Gene Description[1]	w_C[2]	w_C[3]	p-value[4]
		protein 140).			
ENSG00000102385	DRP2	Dystrophin-related protein 2.	0.273	999.000	0.0000
ENSG00000125772	K1434_HUMAN		0.137	999.000	0.0000
ENSG00000133619	NP_115923.1		0.293	379.927	0.0000
ENSG00000136014	USP44	Ubiquitin carboxyl-terminal hydrolase 44 (EC 3.1.2.15) (Ubiquitin thiolesterase 44) (Ubiquitin-specific-processing protease 44) (Deubiquitinating enzyme 44).	0.410	999.000	0.0000
ENSG00000198171	CT116_HUMAN	Protein C20orf116 precursor.	0.205	999.000	0.0000
ENSG00000198265	HELZ	Probable helicase with zinc-finger domain (EC 3.6.1.-).	0.410	999.000	0.0000
ENSG00000166012	JOSD3		0.273	999.000	0.0000
ENSG00000152422	XRCC4	DNA-repair protein XRCC4 (X-ray repair cross-complementing protein 4).	0.410	999.000	0.0000
ENSG00000185834			0.820	999.000	0.0000
ENSG00000169957	NP_078947.2		0.273	999.000	0.0000
ENSG00000105216	LSM14A	Protein FAM61A (Putative alpha synuclein-binding protein) (AlphaSNBP).	0.308	999.000	0.0000
ENSG00000135605	TEC	Tyrosine-protein kinase Tec (EC 2.7.1.112).	0.342	999.000	0.0000
ENSG00000124496	TRERF1	Transcriptional-regulating factor 1 (Transcriptional-regulating protein 132) (Zinc finger transcription factor TReP-132) (Zinc finger protein rapa).	0.308	999.000	0.0000
ENSG00000100368	CSF2RB	Cytokine receptor common beta chain precursor (GM-CSF/IL-3/IL-5 receptor common beta-chain) (CD131 antigen) (CDw131).	0.492	999.000	0.0000
ENSG00000113269	RNF130	Goliath homolog precursor (RING finger protein 130).	0.103	999.000	0.0000
ENSG00000013503	POLR3B	DNA-directed RNA polymerase III subunit 127.6 kDa polypeptide (EC 2.7.7.6) (RNA polymerase III subunit 2) (RPC2).	0.117	999.000	0.0000
ENSG00000162341	TPCN2	two pore segment channel 2	0.410	999.000	0.0000
ENSG00000189257			0.308	999.000	0.0000
ENSG00000166540	NP_060227.1	zinc finger protein 407	0.205	128.727	0.0001
ENSG00000136383	ALPK3	alpha-kinase 3	0.513	325.969	0.0001
ENSG00000182646	TMEM29		0.410	999.000	0.0001
ENSG00000116750	UCHL5	Ubiquitin carboxyl-terminal hydrolase isozyme L5 (EC 3.4.19.12) (UCH- L5) (Ubiquitin thiolesterase L5) (Ubiquitin C-terminal hydrolase UCH37).	0.410	999.000	0.0001
ENSG00000188683			0.410	941.509	0.0001
ENSG00000160299	PCNT	Pericentrin (Pericentrin B) (Kendrin).	0.308	999.000	0.0001
ENSG00000126337	KRTHA6	Keratin, type I cuticular Ha6 (Hair keratin, type I Ha6).	0.308	516.751	0.0001
ENSG00000176819			0.820	999.000	0.0001
ENSG00000152455	SUV39H2	Histone-lysine N-methyltransferase, H3 lysine-9 specific 2 (EC 2.1.1.43) (Histone H3-K9 methyltransferase 2) (H3-K9-HMTase 2) (Suppressor of variegation 3-9 homolog 2) (Su(var)3-9 homolog 2).	0.410	999.000	0.0001
ENSG00000053524	MCF2L2	Rho family guanine-nucleotide exchange factor	0.273	999.000	0.0001
ENSG00000095637	SORBS1	Sorbin and SH3 domain-containing protein 1 (Ponsin) (c-Cbl-associated protein)	0.234	999.000	0.0001

Ensembl Gene ID	Gene Name	Gene Description[1]	w_C[2]	w_C[3]	p-value[4]
		(CAP) (SH3 domain protein 5) (SH3P12).			
ENSG00000182405	PGBD4	piggyBac transposable element derived 4	0.273	999.000	0.0002
ENSG00000077063	CTTNBP2	cortactin binding protein 2	0.308	999.000	0.0002
ENSG00000139625	MAP3K12	Mitogen-activated protein kinase kinase kinase 12 (EC 2.7.1.37) (Mixed lineage kinase) (Leucine-zipper protein kinase) (ZPK) (Dual leucine zipper bearing kinase) (DLK) (MAPK-upstream kinase) (MUK).	0.410	999.000	0.0002
ENSG00000189180	ZNF11A	Zinc finger protein 33A (Zinc finger protein KOX31) (HA0946).	0.574	999.000	0.0002
ENSG00000121101	TEX14	testis expressed sequence 14 isoform a	0.492	999.000	0.0003
ENSG00000083168	MYST3	Histone acetyltransferase MYST3 (EC 2.3.1.48) (EC 2.3.1.-) (MYST protein 3) (MOZ, YBF2/SAS3, SAS2 and TIP60 protein 3) (Runt-related transcription factor-binding protein 2) (Monocytic leukemia zinc finger protein) (Zinc finger protein 220).	0.246	999.000	0.0003
ENSG00000135049	AGTPBP1	ATP/GTP binding protein 1	0.615	999.000	0.0004
ENSG00000176225	RTTN	rotatin	0.820	999.000	0.0004
ENSG00000182898	TCHHL1	trichohyalin-like 1	0.820	999.000	0.0004
ENSG00000101076	HNF4A	Hepatocyte nuclear factor 4-alpha (HNF-4-alpha) (Transcription factor HNF-4) (Transcription factor 14).	0.117	999.000	0.0004
ENSG00000132321	IQCA	IQ motif containing with AAA domain	0.234	999.000	0.0005
ENSG00000119185	ITGB1BP1	Integrin beta-1-binding protein 1 (Integrin cytoplasmic domain- associated protein 1) (ICAP-1).		999.000	0.0005
ENSG00000080511	RDH8	retinol dehydrogenase 8 (all-trans)	0.410	999.000	0.0005
ENSG00000171303	KCNK3	Potassium channel subfamily K member 3 (Acid-sensitive potassium channel protein TASK-1) (TWIK-related acid-sensitive K(+) channel 1) (Two pore potassium channel KT3.1).	1.230	999.000	0.0006
ENSG00000130595	TNNT3	Troponin T, fast skeletal muscle (TnTf) (Fast skeletal muscle troponin T) (fTnT) (Beta TnTF).	0.461	63.182	0.0008
ENSG00000117748	RPA2	Replication protein A 32 kDa subunit (RP-A) (RF-A) (Replication factor-A protein 2) (p32) (p34).	0.137	999.000	0.0008
ENSG00000180104	EXOC3	Exocyst complex component 3 (Exocyst complex component Sec6).	0.410	999.000	0.0008
ENSG00000118922	KLF12	Krueppel-like factor 12 (Transcriptional repressor AP-2rep).	0.410	999.000	0.0008
ENSG00000056586	MNAB	Membrane-associated nucleic acid-binding protein (RING finger protein 164).	0.273	492.590	0.0010
ENSG00000172732	MUS81	Crossover junction endonuclease MUS81 (EC 3.1.22.-).	1.230	999.000	0.0010
ENSG00000149115	TNKS1BP1	182 kDa tankyrase 1-binding protein.	0.410	999.000	0.0010
ENSG00000113645	WWC1	KIBRA protein	0.082	375.411	0.0011
ENSG00000159166	LAD1	Ladinin 1 (Lad-1) (120 kDa linear IgA bullous dermatosis antigen) (97 kDa linear IgA bullous dermatosis antigen) (Linear IgA disease antigen homolog) (LadA).	0.205	904.548	0.0011
ENSG00000128052	KDR	Vascular endothelial growth factor receptor 2 precursor (EC 2.7.1.112) (VEGFR-2)	0.103	999.000	0.0012

Ensembl Gene ID	Gene Name	Gene Description[1]	w_C[2]	w_C[3]	p-value[4]
		(Kinase insert domain receptor) (Protein-tyrosine kinase receptor Flk-1) (CD309 antigen).			
ENSG00000124006	OBSL1	OBSL1 protein (Fragment).	0.410	942.191	0.0012
ENSG00000174448	STARD6	StAR-related lipid transfer protein 6 (StARD6) (START domain- containing protein 6).	0.820	999.000	0.0013
ENSG00000188672	RHCE	Blood group Rh(CE) polypeptide (Rhesus C/E antigens) (Rh30A) (RhIXB) (Rh polypeptide 1) (RhPI).	2.050	999.000	0.0014
ENSG00000101844	ATG4A	Cysteine protease ATG4A (EC 3.4.22.-) (Autophagy-related protein 4 homolog A) (hAPG4A) (Autophagin-2) (Autophagy-related cysteine endopeptidase 2) (AUT-like 2 cysteine endopeptidase).	0.820	999.000	0.0014
ENSG00000135164	DMTF1	cyclin D binding myb-like transcription factor 1	0.410	638.008	0.0016
ENSG00000139508	NP_861450.1		1.230	999.000	0.0017
ENSG00000198324	NP_653272.2		0.615	999.000	0.0017
ENSG00000104059	K0574_HUMAN		3.280	999.000	0.0018
ENSG00000011260	WDR50	WD-repeat protein 50.	0.820	999.000	0.0018
ENSG00000112159	MDN1	Midasin (MIDAS-containing protein).	0.234	213.911	0.0018
ENSG00000169105	D4ST1	Carbohydrate sulfotransferase D4ST1 (EC 2.8.2.-) (Dermatan 4- sulfotransferase 1) (D4ST-1) (hD4ST).	0.205	656.829	0.0018
ENSG00000158258	CLSTN2	Calsyntenin-2 precursor.	0.718	999.000	0.0019
ENSG00000059804	SLC2A3	Solute carrier family 2, facilitated glucose transporter member 3 (Glucose transporter type 3, brain).	0.273	714.347	0.0019
ENSG00000108733	PEX12	Peroxisome assembly protein 12 (Peroxin-12) (Peroxisome assembly factor 3) (PAF-3).		999.000	0.0021
ENSG00000103995	CEP152	Centrosomal protein of 152 kDa (Cep152 protein).	0.902	237.337	0.0022
ENSG00000147081	AKAP4	A-kinase anchor protein 4 isoform 2	0.820	999.000	0.0023
ENSG00000163527	STT3B	source of immunodominant MHC-associated peptides	0.205	906.596	0.0023
ENSG00000185927			0.410	999.000	0.0027
ENSG00000001626	CFTR	Cystic fibrosis transmembrane conductance regulator (CFTR) (cAMP- dependent chloride channel) (ATP-binding cassette transporter sub- family C member 7).	0.547	421.255	0.0027
ENSG00000133069	TMCC2	Transmembrane and coiled-coil domains protein 2 (Cerebral protein 11).	0.164	308.861	0.0028
ENSG00000142178	SNF1LK	Serine/threonine-protein kinase SNF1-like kinase 1 (EC 2.7.1.37) (Serine/threonine-protein kinase SNF1LK).	0.103	999.000	0.0029
ENSG00000121486	NP_112196.2	N2,N2-dimethylguanosine tRNA methyltransferase-like	0.615	428.612	0.0031
ENSG00000185739	SRL	Sarcalumenin precursor.	1.230	999.000	0.0031
ENSG00000181804	SLC9A9	Sodium/hydrogen exchanger 9 (Na(+)/H(+) exchanger 9) (NHE-9) (Solute carrier family 9 member 9).	0.103	333.850	0.0031
ENSG00000108592	FTSJ3	Putative rRNA methyltransferase 3 (EC 2.1.1.-) (rRNA (uridine-2'-O-)-	0.410	999.000	0.0031

Ensembl Gene ID	Gene Name	Gene Description[1]	w_C[2]	w_C[3]	p-value[4]
		methyltransferase 3).			
ENSG00000187272	KRTAP9-9	Keratin-associated protein 9-4 (Keratin-associated protein 9.4) (Ultrahigh sulfur keratin-associated protein 9.4).	0.410	19.774	0.0032
ENSG00000184956	MUC6	Mucin glycoprotein (Fragment).	0.273	67.165	0.0034
ENSG00000187187	ZNF546	zinc finger protein 546	0.683	208.162	0.0034
ENSG00000138411	HECW2	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2	0.256	163.790	0.0035
ENSG00000151150	ANK3	Ankyrin-3 (ANK-3) (Ankyrin G).	0.224	721.181	0.0037
ENSG00000187820				999.000	0.0037
ENSG00000137871	SUHW4	suppressor of hairy wing homolog 4 isoform 1	0.615	492.176	0.0039
ENSG00000148840	PPRC1	PGC-1 related co-activator	0.820	192.508	0.0039
ENSG00000165164	CX022_HUMAN	Protein CXorf22.	0.410	635.085	0.0040
ENSG00000141458	NPC1	Niemann-Pick C1 protein precursor.	0.293	202.984	0.0041
ENSG00000071203	MS4A12	Membrane-spanning 4-domains subfamily A member 12.	0.615	999.000	0.0043
ENSG00000154783	FGD5	FYVE, RhoGEF and PH domain-containing protein 5 (Zinc finger FYVE domain-containing protein 23).	0.154	978.102	0.0044
ENSG00000184950			3.280	999.000	0.0044
ENSG00000174123	TLR10	Toll-like receptor 10 precursor.	0.246	444.812	0.0045
ENSG00000113905	HRG	Histidine-rich glycoprotein precursor (Histidine-proline-rich glycoprotein) (HPRG).	1.503	85.310	0.0045
ENSG00000189409	MMP23B	matrix metalloproteinase 23B precursor	0.273	148.851	0.0046
ENSG00000157950	SSX2	Protein SSX2 (Synovial sarcoma, X breakpoint 2) (SSX) (HOM-MEL-40).	0.273	460.090	0.0046
ENSG00000189136	NP_997265.1		1.640	999.000	0.0048
ENSG00000177553	Q8NDA9_HUMAN	Novel protein (Fragment).	0.615	599.959	0.0049
ENSG00000133958	NP_065869.2		0.293	195.484	0.0049
ENSG00000103569	AQP9	Aquaporin-9 (AQP-9) (Small solute channel 1).	0.410	300.882	0.0054
ENSG00000198796	ALPK2	heart alpha-kinase	0.761	164.694	0.0056
ENSG00000155275	NP_689757.1		0.615	354.045	0.0057
ENSG00000186152	LILRB3	Leukocyte immunoglobulin-like receptor subfamily B member 3 precursor (Leukocyte immunoglobulin-like receptor 3) (LIR-3) (Immunoglobulin- like transcript 5) (ILT-5) (Monocyte inhibitory receptor HL9) (CD85a antigen).	0.568	424.616	0.0060
ENSG00000196427	Q5VWK0_HUMAN	Novel protein similar to FLJ32883 containing DUF1220 domains (Fragment).	0.718	48.563	0.0060
ENSG00000179698	NP_115918.1		0.273	999.000	0.0063
ENSG00000076928	ARHGEF1	Rho guanine nucleotide exchange factor 1 (p115-RhoGEF) (p115RhoGEF) (115 kDa guanine nucleotide exchange factor) (Sub1.5).	0.137	176.505	0.0064
ENSG00000096384	HSP90AB1	Heat shock protein HSP 90-beta (HSP 84) (HSP 90).	0.154	78.444	0.0066
ENSG00000103449	SALL1	Sal-like protein 1 (Zinc finger protein SALL1) (Spalt-like transcription factor 1) (HSal1).	0.342	999.000	0.0067
ENSG00000196208	NP_149081.1	GREB1 protein isoform b	0.068	309.232	0.0067

Ensembl Gene ID	Gene Name	Gene Description[1]	w_C[2]	w_C[3]	p-value[4]
ENSG00000115526	CHST10	Carbohydrate sulfotransferase 10 (EC 2.8.2.-) (HNK-1 sulfotransferase) (HNK1ST) (HNK-1ST) (huHNK-1ST).	0.273	170.596	0.0067
ENSG00000187010	RHD	Blood group Rh(D) polypeptide (Rhesus D antigen) (RHXIII) (Rh polypeptide 2) (RhPII).	0.718	999.000	0.0071
ENSG00000181950	Q8NH17_HUMAN	Seven transmembrane helix receptor.	2.050	999.000	0.0071
ENSG00000172038	KV3H_HUMAN	Ig kappa chain V-III region CLL precursor (Rheumatoid factor).	1.025	999.000	0.0071
ENSG00000082213	NP_060826.1		0.137	228.823	0.0073
ENSG00000185507	IRF7	Interferon regulatory factor 7 (IRF-7).	0.176	302.105	0.0073
ENSG00000124731	TREM1	Triggering receptor expressed on myeloid cells 1 precursor (TREM-1) (Triggering receptor expressed on monocytes 1).		999.000	0.0073
ENSG00000135622	SEMA4F	Semaphorin-4F precursor (Semaphorin W) (Sema W) (Semaphorin M) (Sema M).	0.164	177.660	0.0074
ENSG00000173230	GOLGB1	Golgin subfamily B member 1 (Giantin) (Macrogolgin) (372 kDa Golgi complex-associated protein) (GCP372).	1.312	124.825	0.0074
ENSG00000143740	NP_444280.1		0.410	998.999	0.0074
ENSG00000130816	DNMT1	DNA (cytosine-5)-methyltransferase 1 (EC 2.1.1.37) (Dnmt1) (DNA methyltransferase HsaI) (DNA MTase HsaI) (MCMT) (M.HsaI).	0.046	164.787	0.0077
ENSG00000184459	BPIL2	Bactericidal/permeability-increasing protein-like 2 precursor.	0.308	450.330	0.0078
ENSG00000198946	SSX4	Protein SSX4.	0.683	999.000	0.0084
ENSG00000163492	NP_775919.2		0.308	229.471	0.0085
ENSG00000141690			1.230	74.633	0.0085
ENSG00000188885			1.640	999.000	0.0086
ENSG00000133056	PIK3C2B	Phosphatidylinositol-4-phosphate 3-kinase C2 domain-containing beta polypeptide (EC 2.7.1.154) (Phosphoinositide 3-Kinase-C2-beta) (PtdIns-3-kinase C2 beta) (PI3K-C2beta) (C2-PI3K).	0.082	152.423	0.0088
ENSG00000183621	ZNF438	ZNF438 transcript variant 3	0.820	160.922	0.0092
ENSG00000164151	Q6ZT40_HUMAN	CDNA FLJ44990 fis, clone BRAWH3008559.	0.293	352.398	0.0094
ENSG00000106328	FSCN3	Fascin-3 (Testis fascin).	0.410	999.000	0.0095
ENSG00000117616	NP_064713.3		0.176	433.752	0.0099
ENSG00000135951	TSGA10	testis specific, 10	0.820	999.000	0.0102
ENSG00000100829	PPP1R3E	Homeobox and leucine zipper protein Homez (Homeodomain leucine zipper-containing factor).	0.820	215.162	0.0110
ENSG00000134285	FKBP11	FK506-binding protein 11 precursor (EC 5.2.1.8) (Peptidyl-prolyl cis- trans isomerase) (PPIase) (Rotamase) (19 kDa FK506-binding protein) (FKBP-19).		999.000	0.0112
ENSG00000130779	RSN	Restin (Cytoplasmic linker protein 170 alpha-2) (CLIP-170) (Reed- Sternberg intermediate filament-associated protein) (Cytoplasmic linker protein 1).	0.176	999.000	0.0115
ENSG00000090512	FETUB	Fetuin-B precursor (IRL685) (16G2).	0.513	497.834	0.0117
ENSG00000198162	MAN1A2	Mannosyl-oligosaccharide 1,2-alpha-mannosidase IB (EC 3.2.1.113) (Processing	0.615	152.655	0.0118

Ensembl Gene ID	Gene Name	Gene Description[1]	w_C[2]	w_C[3]	p-value[4]
		alpha-1,2-mannosidase IB) (Alpha-1,2-mannosidase IB) (Mannosidase alpha class 1A member 2).			
ENSG00000164588	HCN1	Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 1 (Brain cyclic nucleotide gated channel 1) (BCNG-1).	0.082	264.456	0.0124
ENSG00000135346	CGA	Glycoprotein hormones alpha chain precursor (Anterior pituitary glycoprotein hormones common alpha subunit) (Follitropin alpha chain) (Follicle-stimulating hormone alpha chain) (FSH-alpha) (Lutropin alpha chain) (Luteinizing hormone alpha chain) (LSH-alph	0.615	289.404	0.0129
ENSG00000109205	NP_060325.2	APin protein	0.410	199.072	0.0143
ENSG00000065328	MCM10	minichromosome maintenance protein 10 isoform 2	0.574	109.016	0.0147
ENSG00000041802	LSG1		0.820	72.404	0.0149
ENSG00000173950	NP_689744.2		0.205	273.981	0.0151
ENSG00000168661	ZNF30	Zinc finger protein 30 (Zinc finger protein KOX28).	0.820	62.918	0.0152
ENSG00000183862	CNGA2	Cyclic nucleotide-gated olfactory channel (Cyclic nucleotide-gated cation channel 2) (CNG channel 2) (CNG-2) (CNG2) (Fragment).	0.615	999.000	0.0153
ENSG00000188636	LDOC1L		0.205	203.518	0.0154
ENSG00000092200	RPGRIP1	X-linked retinitis pigmentosa GTPase regulator-interacting protein 1 (RPGR-interacting protein 1).	0.574	264.144	0.0162
ENSG00000154358	OBSCN	obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF	0.888	59.721	0.0168
ENSG00000168229	PTGDR	Prostaglandin D2 receptor (Prostanoid DP receptor) (PGD receptor).	1.640	999.000	0.0171
ENSG00000198118	Q96R54_HUMAN	Olfactory receptor (Fragment).	0.820	131.097	0.0171
ENSG00000196074	SYCP2	Synaptonemal complex protein 2 (SCP-2 protein) (Synaptonemal complex lateral element protein) (hsSCP2).	4.100	173.663	0.0173
ENSG00000104853	CLPTM1	cleft lip and palate associated transmembrane protein 1	0.205	67.035	0.0181
ENSG00000085982	USP40	Ubiquitin carboxyl-terminal hydrolase 40 (EC 3.1.2.15) (Ubiquitin thiolesterase 40) (Ubiquitin-specific-processing protease 40) (Deubiquitinating enzyme 40).	0.351	161.113	0.0183
ENSG00000160856	FCRL3	Fc receptor-like 3 precursor	0.246	246.953	0.0187
ENSG00000181045	SLC26A11	solute carrier family 26, member 11	0.308	58.984	0.0191
ENSG00000135540	NHSL1	NHSL1 protein (Fragment).	2.460	734.595	0.0193
ENSG00000009790	TRAF3IP3	TRAF3-interacting JNK-activating modulator (TRAF3-interacting protein 3).	0.410	101.515	0.0195
ENSG00000087206	RXINP_HUMAN	Retinoid X receptor-interacting protein 110 (Receptor-associated protein 80) (Nuclear zinc finger protein RAP80).	1.025	107.870	0.0196
ENSG00000099399	MAGEB2	Melanoma-associated antigen B2 (MAGE-B2 antigen) (DSS-AHC critical interval MAGE superfamily 6) (DAM6) (MAGE XP-2).	2.870	999.000	0.0201
ENSG00000133773	CCDC59		0.410	184.628	0.0202
ENSG00000197734	NP_777603.1		0.273	168.421	0.0206
ENSG00000125337	KIF25	Kinesin-like protein KIF25 (Kinesin-like protein 3).	1.230	999.000	0.0206

Ensembl Gene ID	Gene Name	Gene Description[1]	w_C[2]	w_C[3]	p-value[4]
ENSG00000196735	HA25_HUMAN	HLA class II histocompatibility antigen, DQ(5) alpha chain precursor (DC-1 alpha chain).	0.478	39.209	0.0207
ENSG00000079385	CEACAM1	Carcinoembryonic antigen-related cell adhesion molecule 1 precursor (Biliary glycoprotein 1) (BGP-1) (CD66 antigen) (CD66a antigen).	0.410	217.564	0.0209
ENSG00000135968	GCC2	Ran-binding protein 2-like 4 (RanBP2L4).	0.769	54.241	0.0232
ENSG00000169021	UQCRFS1	Ubiquinol-cytochrome c reductase iron-sulfur subunit, mitochondrial precursor (EC 1.10.2.2) (Rieske iron-sulfur protein) (RISP).	0.820	167.192	0.0233
ENSG00000073067	CYP2W1	Cytochrome P450 2W1 (EC 1.14.14.-) (CYPIIW1).	0.410	57.026	0.0234
ENSG00000148411	BTBD14A	BTB (POZ) domain containing 14A	0.410	999.000	0.0236
ENSG00000100312	ACR	Acrosin precursor (EC 3.4.21.10) [Contains: Acrosin light chain; Acrosin heavy chain].	0.273	109.920	0.0239
ENSG00000175305	CCNE2	G1/S-specific cyclin-E2.	0.410	84.369	0.0243
ENSG00000100450	GZMH	Granzyme H precursor (EC 3.4.21.-) (Cytotoxic T-lymphocyte proteinase) (Cathepsin G-like 2) (CTSGL2) (CCP-X) (Cytotoxic serine protease C) (CSP-C).		999.000	0.0251
ENSG00000198703	OR10R3P	Seven transmembrane helix receptor.		999.000	0.0255
ENSG00000160229	ZNF486	Zinc finger protein 486.	1.025	105.389	0.0260
ENSG00000127838	PNKD	myofibrillogenesis regulator 1 isoform 2	0.820	999.000	0.0265
ENSG00000077935	SMC1L2	Structural maintenance of chromosome 1-like 2 protein (SMC1beta protein).		173.415	0.0270
ENSG00000175885	ZNF611	zinc finger protein 611	0.410	181.404	0.0271
ENSG00000112273	HDGFL1	hepatoma derived growth factor-like 1	1.435	51.016	0.0283
ENSG00000179588	ZFPM1	Zinc finger protein ZFPM1 (Zinc finger protein multitype 1) (Friend of GATA protein 1) (Friend of GATA-1) (FOG-1).	0.273	152.552	0.0288
ENSG00000116254	CHD5	Chromodomain helicase-DNA-binding protein 5 (EC 3.6.1.-) (ATP- dependent helicase CHD5) (CHD-5).	0.041	999.000	0.0292
ENSG00000145850	TIMD4	T cell immunoglobulin and mucin domain-containing protein 4 precursor (TIMD-4) (T cell membrane protein 4) (TIM-4).		999.000	0.0294
ENSG00000152782	PANK1	Pantothenate kinase 1 (EC 2.7.1.33) (Pantothenic acid kinase 1) (hPanK1) (hPanK).	0.205	138.333	0.0296
ENSG00000115350	POLE4	DNA polymerase epsilon subunit 4 (EC 2.7.7.7) (DNA polymerase II subunit 4) (DNA polymerase epsilon subunit p12).	0.273	36.136	0.0305
ENSG00000125520	SLC2A4RG	SLC2A4 regulator (GLUT4 enhancer factor) (GEF) (Huntington disease gene regulatory region-binding protein 1) (HDBP-1).	0.820	111.824	0.0308
ENSG00000101447	CT129_HUMAN		0.492	34.327	0.0309
ENSG00000103855	CD276	CD276 antigen precursor (Costimulatory molecule) (B7 homolog 3) (B7- H3) (4Ig-B7-H3).	0.176	37.702	0.0309
ENSG00000184278			0.547	175.020	0.0318
ENSG00000162888	NP_001020763.1		1.025	999.000	0.0324
ENSG00000182227				999.000	0.0327

Ensembl Gene ID	Gene Name	Gene Description[1]	w_C[2]	w_C[3]	p-value[4]
ENSG00000186867	GPR103	Orexigenic neuropeptide QRFP receptor (G-protein coupled receptor 103) (SP9155) (AQ27).	-2.870	999.000	0.0327
ENSG00000085832	EPS15	Epidermal growth factor receptor substrate 15 (Protein Eps15) (AF-1p protein).	0.205	70.582	0.0339
ENSG00000156574	NODAL	Nodal homolog precursor.	0.308	40.588	0.0345
ENSG00000091181	IL5RA	Interleukin-5 receptor alpha chain precursor (IL-5R-alpha) (CD125 antigen) (CDw125).	0.410	99.357	0.0354
ENSG00000129103	SUMF2	Sulfatase-modifying factor 2 precursor (C-alpha-formyglycine- generating enzyme 2).		999.000	0.0359
ENSG00000197912	SPG7	Paraplegin (EC 3.4.24.-) (Spastic paraplegia protein 7).	1.025	71.642	0.0375
ENSG00000129810	SGOL1	Shugoshin-like 1 (hSgo1) (Serologically defined breast cancer antigen NY-BR-85).		197.140	0.0376
ENSG00000161914	ZNF653	zinc finger protein 653	0.103	999.000	0.0379
ENSG00000139117	CPNE8	Copine-8 (Copine VIII).	0.082	999.000	0.0389
ENSG00000196240	OR2T2	Olfactory receptor 2T2 (Olfactory receptor OR1-43).	0.273	42.879	0.0401
ENSG00000152582	NP_653323.1	KPL2 protein isoform 2	4.920	7.647	0.0405
ENSG00000169607	NP_689728.2		0.957	110.869	0.0406
ENSG00000159433	STARD9	StAR-related lipid transfer protein 9 (StARD9) (START domain- containing protein 9) (Fragment).	0.559	525.103	0.0408
ENSG00000187753	NP_001010907.1			999.000	0.0425
ENSG00000145041	Q8TBD9_HUMAN	VPRBP protein.	0.410	999.000	0.0428
ENSG00000127366	TAS2R5	Taste receptor type 2 member 5 (T2R5).	2.050	999.000	0.0433
ENSG00000165120	NP_660311.1			999.000	0.0456
ENSG00000149418	ST14	Suppressor of tumorigenicity 14 (EC 3.4.21.-) (Matriptase) (Membrane- type serine protease 1) (MT-SP1) (Prostamin) (Serine protease TADG-15) (Tumor-associated differentially-expressed gene 15 protein).	0.256	30.260	0.0463
ENSG00000091592	NALP1	NACHT-, LRR- and PYD-containing protein 2 (Death effector filament- forming ced-4-like apoptosis protein) (Nucleotide-binding domain and caspase recruitment domain) (Caspase recruitment domain protein 7).	2.870	999.000	0.0469
ENSG00000180290	GNRHR2	Gonadotropin-releasing hormone II receptor (Type II GnRH receptor) (GnRH-II-R).		999.000	0.0477
ENSG00000188869	TMC3	TMC3 protein.	0.461	999.000	0.0482

[1] Gene descriptions downloaded from Ensembl. Blanks indicate information not available or unknown.

[2] Nonsynonymous/synonymous substitution rate ratio, computed by 0.41n /s . Blank indicates no synonymous changes.

[3] Nonsynonymous/synonymous substitution rate ratio, computed by PAML for foreground site class 2b

[4] Uncorrected p-value from likelihood ratio test

Genes showing significant signal of postive selection at FDR = 5%

Lineage	Ensembl Gene ID	Gene Name	Gene Description[1]	*w*[2]	*w*[3]	p-value	
Human	ENSG00000165659	DACH1	Dachshund homolog 1 (Dach1).	0.205	999.000	0.0000	*
	ENSG00000188153	COL4A5	Collagen alpha-5(IV) chain precursor.	0.205	999.000	0.0000	*
Chimpanzee	ENSG00000167522	ANKRD11	Ankyrin repeat domain-containing protein 11 (Ankyrin repeat-containing cofactor 1).	0.256	999.000	0.0000	*
	ENSG00000126822	PLEKHG3	pleckstrin homology domain containing, family G, member 3	0.492	999.000	0.0000	*
	ENSG00000152242	CR025_HUMAN		0.137	999.000	0.0000	*
	ENSG00000117713	ARID1A	AT-rich interactive domain-containing protein 1A (ARID domain- containing protein 1A) (SWI/SNF-related, matrix-associated, actin- dependent regulator of chromatin subfamily F member 1) (SWI-SNF complex protein p270) (B120) (SWI-like protein) (Osa homolog	0.123	999.000	0.0000	*
	ENSG00000185727			0.205	999.000	0.0000	*
	ENSG00000165521	EML5	echinoderm microtubule associated protein like 5	0.252	999.000	0.0000	*
	ENSG00000002746	HECW1	NEDD4-like ubiquitin-protein ligase 1	0.186	999.000	0.0000	*
	ENSG00000198308	Q9NSI3_HUMAN		0.273	999.000	0.0000	*
	ENSG00000116141	MARK1	Serine/threonine-protein kinase MARK1 (EC 2.7.1.37) (MAP/microtubule affinity-regulating kinase 1).	0.410	999.000	0.0000	*
	ENSG00000091986	NP_955806.1	steroid-sensitive protein 1	0.351	999.000	0.0000	*
	ENSG00000147036	LANCL3	LanC lantibiotic synthetase component C-like 3	0.205	999.000	0.0000	*
	ENSG00000112276	BVES	Blood vessel epicardial substance (hBVES) (Popeye domain-containing protein 1) (Popeye protein 1).	0.273	999.000	0.0000	*
	ENSG00000106415	GLCCI1	glucocorticoid induced transcript 1	0.205	999.000	0.0000	*
	ENSG00000130227	XPO7	Exportin-7 (Ran-binding protein 16).	0.205	999.000	0.0000	*
	ENSG00000096401	CDC5L	Cell division cycle 5-like protein (Cdc5-like protein) (Pombe cdc5- related protein).	0.164	999.000	0.0000	*
	ENSG00000126010	GRPR	Gastrin-releasing peptide receptor (GRP-R) (GRP-preferring bombesin receptor).	0.273	999.000	0.0000	*
	ENSG00000100345	MYH9	Myosin-9 (Myosin heavy chain, nonmuscle IIa) (Nonmuscle myosin heavy chain IIa) (NMMHC II-a) (NMMHC-IIA) (Cellular myosin heavy chain, type A) (Nonmuscle myosin heavy chain-A) (NMMHC-A).	0.112	211.772	0.0000	*
	ENSG00000129116	PALLD	palladin	0.059	999.000	0.0000	*
	ENSG00000197996	PADI6	Peptidylarginine deiminase type VI (EC 3.5.3.15) (Peptidylarginine deiminase type 6).	0.246	915.618	0.0000	*
	ENSG00000198700	IPO9	Importin-9 (Imp9) (Ran-binding protein 9) (RanbP9).	0.824	999.000	0.0000	*
	ENSG00000123607	TTC21B	tetratricopeptide repeat domain 21B	0.205	999.000	0.0000	*
	ENSG00000178662	TAIP2_HUMAN	TGF-beta-induced apoptosis protein 2 (TAIP-2).	0.082	999.000	0.0000	
	ENSG00000073050	XRCC1	DNA-repair protein XRCC1 (X-ray repair cross-complementing protein 1).	0.410	999.000	0.0000	

Lineage	Ensembl Gene ID	Gene Name	Gene Description[1]	*w*[2]	*w*[3]	p-value
	ENSG00000059588	TARBP1	TAR RNA binding protein 1	0.256	999.000	0.0000
	ENSG00000144130	NT5DC4		0.342	984.278	0.0000
	ENSG00000180530	NRIP1	Nuclear receptor-interacting protein 1 (Nuclear factor RIP140) (Receptor-interacting protein 140).	0.273	999.000	0.0000
	ENSG00000102385	DRP2	Dystrophin-related protein 2.	0.273	999.000	0.0000
	ENSG00000125772	K1434_HUMAN		0.137	999.000	0.0000
	ENSG00000133619	NP_115923.1		0.293	379.927	0.0000
	ENSG00000136014	USP44	Ubiquitin carboxyl-terminal hydrolase 44 (EC 3.1.2.15) (Ubiquitin thiolesterase 44) (Ubiquitin-specific-processing protease 44) (Deubiquitinating enzyme 44).	0.410	999.000	0.0000
	ENSG00000198171	CT116_HUMAN	Protein C20orf116 precursor.	0.205	999.000	0.0000
	ENSG00000198265	HELZ	Probable helicase with zinc-finger domain (EC 3.6.1.-).	0.410	999.000	0.0000
	ENSG00000166012	JOSD3		0.273	999.000	0.0000
	ENSG00000152422	XRCC4	DNA-repair protein XRCC4 (X-ray repair cross-complementing protein 4).	0.410	999.000	0.0000
	ENSG00000185834			0.820	999.000	0.0000
	ENSG00000169957	NP_078947.2		0.273	999.000	0.0000
	ENSG00000105216	LSM14A	Protein FAM61A (Putative alpha synuclein-binding protein) (AlphaSNBP).	0.308	999.000	0.0000
	ENSG00000135605	TEC	Tyrosine-protein kinase Tec (EC 2.7.1.112).	0.342	999.000	0.0000
	ENSG00000124496	TRERF1	Transcriptional-regulating factor 1 (Transcriptional-regulating protein 132) (Zinc finger transcription factor TReP-132) (Zinc finger protein rapa).	0.308	999.000	0.0000
	ENSG00000100368	CSF2RB	Cytokine receptor common beta chain precursor (GM-CSF/IL-3/IL-5 receptor common beta-chain) (CD131 antigen) (CDw131).	0.492	999.000	0.0000
	ENSG00000113269	RNF130	Goliath homolog precursor (RING finger protein 130).	0.103	999.000	0.0000
	ENSG00000013503	POLR3B	DNA-directed RNA polymerase III subunit 127.6 kDa polypeptide (EC 2.7.7.6) (RNA polymerase III subunit 2) (RPC2).	0.117	999.000	0.0000
	ENSG00000162341	TPCN2	two pore segment channel 2	0.410	999.000	0.0000
	ENSG00000189257			0.308	999.000	0.0000
	ENSG00000166540	NP_060227.1	zinc finger protein 407	0.205	128.727	0.0001
	ENSG00000136383	ALPK3	alpha-kinase 3	0.513	325.969	0.0001
	ENSG00000182646	TMEM29		0.410	999.000	0.0001
	ENSG00000116750	UCHL5	Ubiquitin carboxyl-terminal hydrolase isozyme L5 (EC 3.4.19.12) (UCH- L5) (Ubiquitin thiolesterase L5) (Ubiquitin C-terminal hydrolase UCH37).	0.410	999.000	0.0001
	ENSG00000188683			0.410	941.509	0.0001
	ENSG00000160299	PCNT	Pericentrin (Pericentrin B) (Kendrin).	0.308	999.000	0.0001
	ENSG00000126337	KRTHA6	Keratin, type I cuticular Ha6 (Hair keratin, type I Ha6).	0.308	516.751	0.0001
	ENSG00000176819			0.820	999.000	0.0001
	ENSG00000152455	SUV39H2	Histone-lysine N-methyltransferase, H3 lysine-9 specific 2 (EC 2.1.1.43) (Histone H3-K9 methyltransferase 2) (H3-K9-HMTase 2) (Suppressor of variegation 3-9	0.410	999.000	0.0001

Lineage	Ensembl Gene ID	Gene Name	Gene Description[1]	w^2	w^3	p-value
			homolog 2) (Su(var)3-9 homolog 2).			
	ENSG00000053524	MCF2L2	Rho family guanine-nucleotide exchange factor	0.273	999.000	0.0001
	ENSG00000095637	SORBS1	Sorbin and SH3 domain-containing protein 1 (Ponsin) (c-Cbl-associated protein) (CAP) (SH3 domain protein 5) (SH3P12).	0.234	999.000	0.0001
	ENSG00000182405	PGBD4	piggyBac transposable element derived 4	0.273	999.000	0.0002
	ENSG00000077063	CTTNBP2	cortactin binding protein 2	0.308	999.000	0.0002
	ENSG00000139625	MAP3K12	Mitogen-activated protein kinase kinase kinase 12 (EC 2.7.1.37) (Mixed lineage kinase) (Leucine-zipper protein kinase) (ZPK) (Dual leucine zipper bearing kinase) (DLK) (MAPK-upstream kinase) (MUK).	0.410	999.000	0.0002
	ENSG00000189180	ZNF11A	Zinc finger protein 33A (Zinc finger protein KOX31) (HA0946).	0.574	999.000	0.0002

[1] Gene descriptions downloaded from Ensembl. Blanks indicate information not available or unknown.

[2] Nonsynonymous/synonymous substitution rate ratio, computed by 0.41n /s .

[3] Nonsynonymous/synonymous substitution rate ratio, computed by PAML for foreground site class 2b

* Significant after Bonferonni correction

Genes showing significant signal of postive selection and associated with human disease

Lineage	Ensembl Gene ID	Gene Name	Gene Description[1]	OMIM ID	Disease Description[2]
Both	ENSG00000141458	NPC1	Niemann-Pick C1 protein precursor.	257220	Niemann-Pick disease, type C (3)
				257220	Niemann-Pick disease, type D, 257250 (2)
Chimp	ENSG00000001626	CFTR	Cystic fibrosis transmembrane conductance regulator (CFTR) (cAMP- dependent chloride channel) (ATP-binding cassette transporter sub- family C member 7).	602421	Congenital bilateral absence of vas deferens, 277180 (3)
				602421	Cystic fibrosis, 219700 (3)
				602421	Sweat chloride elevation without CF (3)
	ENSG00000100312	ACR	Acrosin precursor (EC 3.4.21.10) [Contains: Acrosin light chain; Acrosin heavy chain].	102480	Male infertility due to acrosin deficiency (2) (?)
	ENSG00000100368	CSF2RB	Cytokine receptor common beta chain precursor (GM-CSF/IL-3/IL-5 receptor common beta-chain) (CD131 antigen) (CDw131).	138981	Pulmonary alveolar proteinosis, 265120 (3)
	ENSG00000101076	HNF4A	Hepatocyte nuclear factor 4-alpha (HNF-4-alpha) (Transcription factor HNF-4) (Transcription factor 14).	600281	MODY, type 1, 125850 (3)
				600281	Non-insulin-dependent diabetes mellitus, 125853 (3)
	ENSG00000103449	SALL1	Sal-like protein 1 (Zinc finger protein SALL1) (Spalt-like transcription factor 1) (HSal1).	602218	Townes-Brocks syndrome, 107480 (3)
	ENSG00000113905	HRG	Histidine-rich glycoprotein precursor (Histidine-proline-rich glycoprotein) (HPRG).	142640	Thrombophilia due to elevated HRG (1) (?)
	ENSG00000135346	CGA	Glycoprotein hormones alpha chain precursor (Anterior pituitary glycoprotein hormones common alpha subunit) (Follitropin alpha chain) (Follicle-stimulating hormone alpha chain) (FSH-alpha) (Lutropin alpha chain) (Luteinizing hormone alpha chain) (LSH-alph	307150	Hypertrichosis, congenital generalized (2)
	ENSG00000135605	TEC	Tyrosine-protein kinase Tec (EC 2.7.1.112).	148500	Tylosis with esophageal cancer (2)
	ENSG00000173230	GOLGB1	Golgin subfamily B member 1 (Giantin) (Macrogolgin) (372 kDa Golgi complex-associated protein) (GCP372).	303800	Colorblindness, deutan (3)
	ENSG00000197912	SPG7	Paraplegin (EC 3.4.24.-) (Spastic paraplegia protein 7).	602783	Spastic paraplegia-7 (3)

Lineage	Ensembl Gene ID	Gene Name	Gene Description[1]	OMIM ID	Disease Description[2]
Human	ENSG00000102805	CLN5	Ceroid-lipofuscinosis neuronal protein 5 (Protein CLN5).	256731	Ceroid-lipofuscinosis, neuronal-5, variant late infantile (3)
	ENSG00000122971	ACADS	Acyl-CoA dehydrogenase, short-chain specific, mitochondrial precursor (EC 1.3.99.2) (SCAD) (Butyryl-CoA dehydrogenase).	201470	Acyl-CoA dehydrogenase, short-chain, deficiency of (3)
	ENSG00000133805	AMPD3	AMP deaminase 3 (EC 3.5.4.6) (AMP deaminase isoform E) (Erythrocyte AMP deaminase).	102772	[AMP deaminase deficiency, erythrocytic] (3)
	ENSG00000138029	HADHB	Trifunctional enzyme beta subunit, mitochondrial precursor (TP-beta) [Includes: 3-ketoacyl-CoA thiolase (EC 2.3.1.16) (Acetyl-CoA acyltransferase) (Beta-ketothiolase)].	143450	Trifunctional protein deficiency, type II (3)
	ENSG00000141837	CACNA1A	Voltage-dependent P/Q-type calcium channel alpha-1A subunit (Voltage- gated calcium channel alpha subunit Cav2.1) (Calcium channel, L type, alpha-1 polypeptide isoform 4) (Brain calcium channel I) (BI).	601011	Cerebellar ataxia, pure (3)
				601011	Episodic ataxia, type 2, 108500 (3)
				601011	Hemiplegic migraine, familial, 141500 (3)
				601011	Spinocerebellar ataxia-6, 183086 (3)
	ENSG00000147889	CDKN2A	Cyclin-dependent kinase inhibitor 2A, isoform 4 (p14ARF) (p19ARF).	155600	Malignant melanoma, cutaneous (2)
				600160	Melanoma, 155601 (3)
	ENSG00000163069	SGCB	Beta-sarcoglycan (Beta-SG) (43 kDa dystrophin-associated glycoprotein) (43DAG) (A3b).	600900	Muscular dystrophy, limb-girdle, type 2E (3)
	ENSG00000165125	TRPV6	Transient receptor potential cation channel subfamily V member 6 (TrpV6) (Epithelial calcium channel 2) (ECaC2) (Calcium transport protein 1) (CaT1) (CaT-like) (CaT-L).	600184	Carnitine acetyltransferase deficiency (1) (?)
	ENSG00000165409	TSHR	Thyrotropin receptor precursor (TSH-R) (Thyroid-stimulating hormone receptor).	275200	Graves disease, 275000 (1)
				275200	Hyperthroidism, congenital (3)
				275200	Hypothyroidism, nongoitrous, due to TSH resistance (3)
				275200	Thyroid adenoma, hyperfunctioning (3)
	ENSG00000169738	DCXR	L-xylulose reductase (EC 1.1.1.10) (XR) (Dicarbonyl/L-xylulose reductase) (Kidney dicarbonyl reductase) (kiDCR) (Carbonyl reductase II) (Sperm surface protein P34H).	190685	Down syndrome (1)

Lineage	Ensembl Gene ID	Gene Name	Gene Description[1]	OMIM ID	Disease Description[2]
	ENSG00000180509	KCNE1	Potassium voltage-gated channel subfamily E member 1 (IKs producing slow voltage-gated potassium channel beta subunit Mink) (Minimal potassium channel) (Delayed rectifier potassium channel subunit IsK).	176261	Jervell and Lange-Nielsen syndrome, 220400 (3)
	ENSG00000182372	CLN8	Protein CLN8.	600143	Epilepsy, progressive, with mental retardation (2)
	ENSG00000186395	KRT10	Keratin, type I cytoskeletal 10 (Cytokeratin-10) (CK-10) (Keratin-10) (K10).	148080	Epidermolytic hyperkeratosis, 113800 (3)
	ENSG00000188153	COL4A5	Collagen alpha-5(IV) chain precursor	303630	Alport syndrome
				303630	Leiomyomatosis-nephropathy syndrome

[1] Gene descriptions downloaded from Ensembl.

[2] Disease descriptions downloaded from OMIM.

References

Adams MD Celniker SE Holt RA Evans CA Gocayne JD Amanatides PG Scherer SE Li PW Hoskins RA Galle RF et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287**(5461): 2185-2195.

Alba MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular biology and evolution* **22**(3): 598-606.

-. 2007. On homology searches by protein Blast and the characterization of the age of genes. *BMC evolutionary biology* **7**: 53.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**(3): 403-410.

Arbiza L, Dopazo J, Dopazo H. 2006. Positive Selection, Relaxation, and Acceleration in the Evolution of the Human and Chimp Genome. *PLOS Computational Biology* **2**(4): e38.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104**(18): 7489-7494.

Britten RJ. 2002. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proceedings of the National Academy of Sciences of the United States of America* **99**(21): 13633-13635.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437**(7062): 1153-1157.

Cai JJ, Borenstein E, Chen R, Petrov DA. 2009. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. *Genome Biol Evol* **1**: 131-144.

Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol Evol* **2**: 393-409.

Cai JJ, Woo PC, Lau SK, Smith DK, Yuen KY. 2006. Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. *J Mol Evol* **63**(1): 1-11.

Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American journal of human genetics* **68**(2): 444-456.

Clark AG Eisen MB Smith DR Bergman CM Oliver B Markow TA Kaufman TC Kellis M Gelbart W Iyer VN et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**(7167): 203-218.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**(5652): 1960-1963.

Consortium IHGS. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011): 931-945.

Consortium TCSaA. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**(7055): 69-87.

Crespi B, Summers K, Dorus S. 2010. Evolutionary genomics of human intellectual disability. *Evol Appl* **3**(1): 52-63.

Darwin C. 1859. *On the origin of species by means of natural selection*. J. Murray, London,.

-. 1871. *The Descent of Man and Selection in Relation to Sex*. D. Appleton, New York.

Deacon TW. 1994. The Human Brain. In *The Cambridge Encyclopedia of Human Evolution*, (ed. S Jones, RD Martin, DR Pilbeam), pp. 115-123. Cambridge University Press.

Domazet-Loso T, Brajkovic J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in genetics : TIG* **23**(11): 533-539.

Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in Drosophila. *Genome research* **13**(10): 2213-2219.

-. 2008. An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular biology and evolution* **25**(12): 2699-2707.

-. 2010a. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**(7325): 815-818.

-. 2010b. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC biology* **8**: 66.

Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, Mahowald M, Wyckoff GJ, Malcom CM, Lahn BT. 2004. Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell* **119**(7): 1027-1040.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319): 1061-1073.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular biology and evolution* **17**(1): 68-74.

Ebersberger I, Metzler D, Schwarz C, Paabo S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *American journal of human genetics* **70**(6): 1490-1497.

Elango N, Thomas JW, Yi SV. 2006. Variable molecular clocks in hominoids. *Proceedings of the National Academy of Sciences of the United States of America* **103**(5): 1370-1375.

Elhaik E, Sabath N, Graur D. 2006. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. *Molecular biology and evolution* **23**(1): 1-3.

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R et al. 2002a. Intra- and interspecific variation in primate gene expression patterns. *Science* **296**(5566): 340-343.

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Paabo S. 2002b. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**(6900): 869-872.

Evans PD, Anderson JR, Vallender EJ, Choi SS, Lahn BT. 2004a. Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size. *Human molecular genetics* **13**(11): 1139-1145.

Evans PD, Anderson JR, Vallender EJ, Gilbert SL, Malcom CM, Dorus S, Lahn BT. 2004b. Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. *Human molecular genetics* **13**(5): 489-494.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research* **8**(3): 175-185.

Eyre-Walker A, Keightley PD. 1999. High genomic deleterious mutation rates in hominids. *Nature* **397**(6717): 344-347.

Fay JC, Benavides JA. 2005. Evidence for domesticated and wild populations of Saccharomyces cerevisiae. *PLoS genetics* **1**(1): 66-71.

Ferris SD, Brown WM, Davidson WS, Wilson AC. 1981. Extensive polymorphism in the mitochondrial DNA of apes. *Proceedings of the National Academy of Sciences of the United States of America* **78**(10): 6319-6323.

Fischer A, Wiebe V, Paabo S, Przeworski M. 2004. Evidence for a complex demographic history of chimpanzees. *Molecular biology and evolution* **21**(5): 799-808.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**(5223): 496-512.

Glazko GV, Nei M. 2003. Estimation of divergence times for major lineages of primate species. *Molecular biology and evolution* **20**(3): 424-434.

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP. 1998. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Molecular phylogenetics and evolution* **9**(3): 585-598.

Goodman M, Sterner KN. 2010. Colloquium paper: phylogenomic evidence of adaptive evolution in the ancestry of humans. *Proceedings of the National Academy of Sciences of the United States of America* **107 Suppl 2**: 8918-8923.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH et al. 2010. A draft sequence of the Neandertal genome. *Science* **328**(5979): 710-722.

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M et al. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**(7117): 330-336.

Gu X, Zhang J. 1997. A simple method for estimating the parameter of substitution rate variation among sites. *Molecular biology and evolution* **14**(11): 1106-1113.

Gunz P, Neubauer S, Maureille B, Hublin JJ. 2010. Brain development after birth differs between Neanderthals and modern humans. *Curr Biol* **20**(21): R921-922.

Hedges SB. 2002. The origin and evolution of model organisms. *Nature reviews Genetics* **3**(11): 838-849.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**(23): 2971-2972.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res* **8**(3): 269-294.

Hopkin M. 2007. Chimps lead evolutionary race. *Nature* **446**(7138): 841.

Hsieh WP, Chu TM, Wolfinger RD, Gibson G. 2003. Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* **165**(2): 747-757.

Innan H, Watanabe H. 2006. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Molecular biology and evolution* **23**(5): 1040-1047.

Kaessmann H, Wiebe V, Paabo S. 1999. Extensive nuclear DNA sequence diversity among chimpanzees. *Science* **286**(5442): 1159-1162.

Kaessmann H, Wiebe V, Weiss G, Paabo S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature genetics* **27**(2): 155-156.

Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS biology* **3**(2): e42.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**(5742): 1850-1854.

Khaitovich P, Tang K, Franz H, Kelso J, Hellmann I, Enard W, Lachmann M, Paabo S. 2006. Positive selection on gene expression in the human brain. *Curr Biol* **16**(10): R356-358.

Kimura M. 1983a. *The Neutral Theroy of Molecular Evolution*. Combridge Univ Press, Cambridge, UK.

-. 1983b. Rare variant alleles in the light of the neutral theory. *Molecular biology and evolution* **1**(1): 84-93.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**(4184): 107-116.

Kouprina N, Pavlicek A, Mochida GH, Solomon G, Gersch W, Yoon YH, Collura R, Ruvolo M, Barrett JC, Woods CG et al. 2004. Accelerated evolution of the ASPM gene controlling brain size begins prior to human brain expansion. *PLoS biology* **2**(5): E126.

Lemos B, Meiklejohn CD, Caceres M, Hartl DL. 2005. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution* **59**(1): 126-137.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America* **103**(26): 9935-9939.

Li W. 1997. In *Molecular Evolution*, pp. 237-267. Sinauer, Sunderland, MA.

Liao BY, Zhang J. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular biology and evolution* **23**(3): 530-540.

-. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. *Proceedings of the National Academy of Sciences of the United States of America* **105**(19): 6987-6992.

Lindblad-Toh K Wade CM Mikkelsen TS Karlsson EK Jaffe DB Kamal M Clamp M Chang JL Kulbokas EJ, 3rd Zody MC et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**(7069): 803-819.

Lu J, Li WH, Wu CI. 2003. Comment on "Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes". *Science* **302**(5647): 988; author reply 988.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**(5494): 1151-1155.

MacCallum C, Hill E. 2006. Being Positive about Selection. *PLoS Biol* **4**(3): e87.

Mallick S, Gnerre S, Muller P, Reich D. 2009. The difficulty of avoiding false positives in genome scans for natural selection. *Genome research* **19**(5): 922-933.

Marques-Bonet T, Caceres M, Bertranpetit J, Preuss TM, Thomas JW, Navarro A. 2004. Chromosomal rearrangements and the genomic distribution of gene-expression divergence in humans and chimpanzees. *Trends in genetics : TIG* **20**(11): 524-529.

Martin RD. 1981. Relative brain size and basal metabolic rate in terrestrial vertebrates. *Nature* **293**(5827): 57-60.

Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ et al. 2005. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic acids research* **33**(Database issue): D284-288.

Navarro A, Barton NH. 2003. Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. *Science* **300**(5617): 321-324.

Neel JV. 1962. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American journal of human genetics* **14**: 353-362.

Nei M. 2005. Selectionism and neutralism in molecular evolution. *Molecular biology and evolution* **22**(12): 2318-2342.

Nei M, Kumar S. 2000. In *Molecular Evolution and Phylogenetics*, pp. 51-71. Oxford University Press, New York.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet* **39**: 197-218.

-. 2009. Adaptionism-30 years after Gould and Lewontin. *Evolution* **63**(10): 2487-2490.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, Berlin.

Ohta T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J Mol Evol* **40**(1): 56-63.

Olson MV, Varki A. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat Rev Genet* **4**(1): 20-28.

Osada N, Wu CI. 2005. Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics* **169**(1): 259-264.

Pagel MD, Harvey PH. 1989. Taxonomic differences in the scaling of brain on body weight among mammals. *Science* **244**(4912): 1589-1593.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**(7097): 1103-1108.

Pilbeam D. 1996. Genetic and morphological records of the Hominoidea and hominid origins: a synthesis. *Mol Phylogenet Evol* **5**(1): 155-168.

Prabhakar S, Noonan JP, Paabo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**(5800): 786.

Prachumwat A, Li WH. 2008. Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes. *Genome research* **18**(2): 221-232.

Prat Y, Fromer M, Linial N, Linial M. 2009. Codon usage is associated with the evolutionary age of genes in metazoan genomes. *BMC evolutionary biology* **9**: 285.

Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, Wray GA. 2005. Ancient and recent positive selection transformed opioid cis-regulation in humans. *PLoS biology* **3**(12): e387.

Rosenberg MS, Subramanian S, Kumar S. 2003. Patterns of transitional mutation biases within and among mammalian genomes. *Molecular biology and evolution* **20**(6): 988-993.

Ruvolo M. 1997. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. *Molecular biology and evolution* **14**(3): 248-265.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* **312**(5780): 1614-1620.

Self SG, Liang KY. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc* **82**: 605-610.

Semon M, Lobry JR, Duret L. 2006. No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Molecular biology and evolution* **23**(3): 523-529.

Serre D, Nadon R, Hudson TJ. 2005. Large-scale recombination rate patterns are conserved among human populations. *Genome research* **15**(11): 1547-1552.

Shi P, Bakewell MA, Zhang J. 2006. Did brain-specific genes evolve faster in humans than in chimpanzees? *Trends Genet* **22**(11): 608-613.

Sokal RR, Rohlf FJ. 1995. *Biometry: The principles and Practice of Statistics in Biological Research*. Freeman, New York.

Stanyon R, Consigliere S, Morescalchi M. 1993. Cranial capacity in hominid evolution. *Human Evolution* **8**(3): 205-216.

Stone AC, Griffiths RC, Zegura SL, Hammer MF. 2002. High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proceedings of the National Academy of Sciences of the United States of America* **99**(1): 43-48.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**(16): 9440-9445.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**(16): 6062-6067.

Takahata N, Satta Y, Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor Popul Biol* **48**(2): 198-221.

Taudien S, Ebersberger I, Glockner G, Platzer M. 2006. Should the draft chimpanzee sequence be finished? *Trends in genetics : TIG* **22**(3): 122-125.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**(22): 4673-4680.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Alba MM. 2009. Origin of primate orphan genes: a comparative genomics approach. *Molecular biology and evolution* **26**(3): 603-612.

Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE et al. 2002. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome biology* **3**(12): RESEARCH0088.

Uddin M, Wildman DE, Liu G, Xu W, Johnson RM, Hof PR, Kapatos G, Grossman LI, Goodman M. 2004. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **101**(9): 2957-2962.

Vallender EJ, Lahn BT. 2004. Positive selection on the human genome. *Hum Mol Genet* **13 Spec No 2**: R245-254.

Varki A, Altheide TK. 2005. Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res* **15**(12): 1746-1758.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**(5667): 66-74.

Wall JD. 2003. Estimating ancestral population sizes and divergence times. *Genetics* **163**(1): 395-404.

Wang HY, Chien HC, Osada N, Hashimoto K, Sugano S, Gojobori T, Chou CK, Tsai SF, Wu CI, Shen CK. 2007. Rate of evolution in brain-expressed genes in humans and other primates. *PLoS biology* **5**(2): e13.

Wang X, Grus WE, Zhang J. 2006. Gene losses during human origins. *PLoS Biol* **4**(3): e52.

Wang YQ, Su B. 2004. Molecular evolution of microcephalin, a gene determining human brain size. *Human molecular genetics* **13**(11): 1131-1137.

Wang Z, Zhang J. 2009. Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS genetics* **5**(1): e1000329.

Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, Kuroki Y, Noguchi H, BenKahla A, Lehrach H, Sudbrak R et al. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429**(6990): 382-388.

Wildman DE, Uddin M, Liu G, Grossman LI, Goodman M. 2003. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo. *Proceedings of the National Academy of Sciences of the United States of America* **100**(12): 7181-7188.

Williams MF. 2002. Primate encephalization and intelligence. *Med Hypotheses* **58**(4): 284-290.

Winter EE, Goodstadt L, Ponting CP. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res* **14**(1): 54-61.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**(5): 555-556.

Young JH, Chang YP, Kim JD, Chretien JP, Klag MJ, Levine MA, Ruff CB, Wang NY, Chakravarti A. 2005. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS genetics* **1**(6): e82.

Yunis JJ, Prakash O. 1982. The origin of man: a chromosomal pictorial legacy. *Science* **215**(4539): 1525-1530.

Zhang J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Molecular biology and evolution* **16**(6): 868-875.

-. 2003. Evolution of the human ASPM gene, a major determinant of brain size. *Genetics* **165**(4): 2063-2070.

-. 2004. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol* **21**(7): 1332-1339.

-. 2010. Positive Darwinian selection in gene evolution. In *Darwin's Heritage Today: Proceedings of the Darwin 200 Beijing International Conference*, (ed. ea M. Long), pp. 288-309. High Education Press, Beijing.

Zhang J, Gu X. 1998. Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics* **149**(3): 1615-1625.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**(12): 2472-2479.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences of the United States of America* **95**(7): 3708-3713.

Zhang J, Wang X, Podlaha O. 2004. Testing the chromosomal speciation hypothesis for humans and chimpanzees. *Genome Res* **14**(5): 845-851.

Zhang J, Webb DM, Podlaha O. 2002. Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. *Genetics* **162**(4): 1825-1835.

Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular biology and evolution* **21**(2): 236-239.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in Drosophila. *Genome research* **18**(9): 1446-1455.