Contents

Chapter 1

Patterns of sitewise selection in mammalian genomes

1.1 Introduction

This chapter describes the use of sitewise evolutionary estimates to characterize the global distribution of selective constraint across 38 mammalian genomes and within the major mammalian superorders. I will apply the Sitewise Likelihood Ratio (SLR) test, evaluated in Chapter **??**, to the set of mammalian orthologous gene trees from Chapter **??** to generate genome-wide sets of sitewise statistics measuring selective constraint in several groups of mammalian species. Both this chapter and the following one are concerned with the analysis of these data: here I will consider the overall distribution of constraint observed in each set of genomes, while Chapter **??** will look at the application of sitewise data to identify gene- and domain-centric evolutionary trends.

I will first introduce the scientific context of this project—namely, the sequencing and analysis of several mammalian genomes for the Mammalian Genome Project (MGP)—and outline the main questions motivating the analysis I performed.

The next section describes the preparation and alignment of the mammalian gene tree from Chapter **??** and introduce a protocol for filtering genome-wide sitewise estimates. Although the simulations from Chapter **??** showed that sequences with divergence levels above that of most mammalian proteins can be aligned without introducing many false positives due to misaligning biological insertions and deletions, the analysis of empirical sequence data involves many potential non-evolutionary sources of alignment error. A sequenced and annotated genome is not a piece of observed data; rather, it is the result of a succession of inferences (based ultimately on the observation of a pool of genomic DNA by some sequencing technology), each step along the way involving potential errors and biases. Chapter ?? looked at the identification of mammalian orthologs, showing that the inference of correct gene tree structures is fraught with difficulty and that low-coverage genomes are under-represented in gene duplications. Other sources of error, including those occurring while reading DNA bases [TOCITE, 2011], assembling genomic fragments [TOCITE, 2011], and annotating gene-coding regions [TOCITE, 2011] have all been previously highlighted as being important in the large-scale analysis of genomic data. As such, care was taken to design and evaluate a variety of filters to reduce the probability of yielding misleading results.

The third part of this chapter characterizes the global distribution of mammalian selective constraint in three ways: first, using the SLR statistic to identify sites evolving under purifying and positive selection at various confidence levels; second, fitting parametric distributions to the set of sitewise estimates to infer the distribution of selective pressures; and third, evaluating the impact of genomic variation in GC content and recombination rate on the distribution of sitewise estimates.

1.2 The Mammalian Genome Project

A major goal of mammalian comparative genomics has been to quantify, identify and understand the fraction of the human genome that is under evolutionary constraint. The first non-human mammalian genomes showed at least 5% of the human genome to be under purifying selection [Lindblad-Toh *et al.*, 2005; Mouse Genome Sequencing Consortium & Mouse Genome Analysis Group, 2002; Rat Genome Sequencing Project Consortium, 2004], but the small number of genomes available limited the extent to which regions of evolutionary constraint could be identified. The Mammalian Genome Project, a coordinated set of genome sequencing projects initiated in 2005 and organised by the Broad Institute of MIT and Harvard, was designed with the primary purpose of increasing the accuracy and confidence with which regions of the human genome that have evolved under evolutionary constraint in mammals could be identified [Margulies *et al.*, 2007]. In line with this goal, 20 mammalian species were chosen for sequencing in order to maximise the amount of evolutionary divergence available for comparative analysis when combined with the 9 already available sequenced genomes [Margulies *et al.*, 2005]. Most of the 20 additional species were only sequenced to a target twofold coverage, meaning each genomic base pair would be covered on average by two sequence reads and roughly 85% of genomic sequence would be covered by at least one read. The decision to sequence many genomes at low coverage was a deliberate choice, designed to maximize the average amount of branch length available for the identification of constrained sequence [Margulies *et al.*, 2007].

As the Mammalian Genome Project proceeded from its sequencing to analysis phase in late 2008, it was clear that the additional branch length afforded by the 29-species phylogeny would enable a number of evolutionary analyses beyond the identification of constrained non-coding regions. These included the evolutionary characterisation of gene promoters, identification of exapted non-coding elements, detection of evolutionary acceleration and deceleration in non-coding regions, and detection of purifying and positive selection in protein-coding genes. Given the prior involvement of the Goldman group in analysing the ENCODE comparative sequencing data [ENCODE Project Consortium, 2007; Margulies *et al.*, 2007] and Tim Massingham's development of the SLR software for sitewise evolutionary analysis [Massingham & Goldman, 2005], the group was recruited to perform the protein-coding evolutionary analysis for the Mammalian Genome Project, and the project turned into a portion of my PhD research. This chapter describes my work on the project, which began in late 2008; all of the work described below was performed by me, though I benefitted greatly from advice and discussion with members of the Goldman group (Nick Goldman and Tim Massingham), the EnsEMBL Compara team (Albert Vilella, Javier Herrero, Ewan Birney) and the organisers and members of the Mammalian Genome Project (especially Manolis Kellis, Kerstin Lindblad-Toh, Mike Lin, and, Katie Pollard). The major results from the initial version of this analysis have recently been published [Lindblad-Toh *et al.*, 2011]; the work presented below includes some improvements to the filtering and alignment methodology and incorporates sequence data from a number of genomes which were restricted from use in the Mammalian Genome Project analysis.

1.3 Data quality concerns: sequencing, assembly and annotation error

1.3.1 The impact of sequencing errors on error rates in detecting positive selection

The possibility that erroneously-aligned sequences might cause false positives in the detection of sitewise positive selection was a major concern for this analysis, especially given the low-coverage nature of the 20 newly-sequenced genomes. Although the SLR test and other sitewise maximum likelihood methods have been shown to be conservative in their identification of positively selected sites under most conditions, even when the amount of data is low or the null model is violated [Anisimova *et al.*, 2002, 2003; Massingham & Goldman, 2005], most evolutionary analyses are based on the assumption that all sites within an alignment column are truly homologous. This assumption can be violated in a number of ways, some of which are described below.

Of course, alignment error can result from errors in reconstructing the evolutionary history of sequences evolving with indels, causing non-homologous codons to be placed in the same alignment column. In Chapter **??** I explored the tendency of a number of progressive multiple alignment programs to produce such errors, showing that $\mathrm{PRANK}_{\mathrm{C}}$ alignments introduce few falsely identified positively-selected sites resulting from alignment errors at mammalian-like divergence levels. Thus, $\mathrm{PRANK}_{\mathrm{C}}$ was used to align all coding sequences, and the number of false positives resulting from misalignment of biological insertions and deletions was expected to be low.

However, biological indels are not the only potential source of misalignment error. Errors resulting from the inclusion of incorrect genomic sequence in coding sequences were an additional concern. Twenty of the genomes under study were sequenced at low coverage and were not assembled into chromosomes or finished to completion, making the likelihood of miscalled bases, spurious insertions or deletions, or shuffled regions due to mis-assembly relatively high [Green, 2007]. The magnitude of the effect of each of the aforementioned types of sequence errors on the detection of positive or purifying selection depends on the nature of the inference method, the type of sequencing error, and the branch length of the terminal lineage leading to the species containing the sequence error.

As most codon-based inference methods assume independence between amino acid sites, the effect of misalignment on the resulting inference will be independent between neighboring codons. Thus, one may first consider the effect—in isolation—of a single spuriously-assigned homologous codon on the maximum likelihood estimation of ω. Two distinct situations can be encountered: first, the case where a single sequence error causes one spurious nucleotide substitution within a codon, and second, the case where one or multiple sequence or assembly errors cause multiple spurious substitutions within a codon. Single spurious nucleotides, such as miscalled bases, would add noise to the estimation of ω, but as a whole they would not be expected to cause false positive positively selected codons. If we assume no large difference between the natural mutational process and the process that caused the erroneous mutation (e.g., a random distribution across codon positions and no bias in the identity of the miscalled base) then the effect would be to shift the estimated ω in the branch containing the error towards 1. This is because, on average, isolated miscalled bases would appear the same as a neutral substitution process, inflating the estimated substitution rate but not affecting the relative non-synonymous and synonymous rates.

In contrast to single spurious substitions, codons with multiple erroneous bases in one species may produce strongly elevated inferred substitution rates and ω estimates. This is due to the necessity of the codon model implemented in SLR to infer a multi-step path of single substitutions between the two codons on either side of a given evolutionary branch. The exact maximum likelihood path estimated between two completely non-homologous codons depends on the estimated codon frequencies, the branch length separating the two sequences, and the nature of the process causing misalignment of nonhomologous codons, but in general it would be reasonable to expect a greater number of false positive PSSs resulting from codons with multiple erroneous bases than from codons with single errors due to the necessary inference of a multi-step path between codons with multiple nucleotide differences.

Given the potentially greater impact of codons with multiple errors, the propensity of each of the common sequencing error types identified above (miscalled bases, spurious indels, and shuffled/repeated/collapsed regions due to mis-assembly) to cause single or multiple errors within codons could strongly affect its impact on the sitewise detection of positive selection. On its own, a miscalled nucleotide base would obviously result in a single spurious substitution. However, low-quality bases tend not to be uniformly distributed among or within sequence reads [Kircher *et al.*, 2009], increasing the probability of multiple errors within a codon resulting from miscalled bases. Spurious indels within coding regions may be even more likely than miscalled bases to cause multiple errors within a codon due to the potential for creating frameshift artifacts. Assembly errors, which result in larger-scale structural errors including missing, repeated, shuffled or inverted sequence regions [Jaffe *et al.*, 2003], are especially prone to producing codons with multiple erroneous substitutions due to the large amount of contiguous sequence data being misplaced.

For detecting positive selection, the nature of the model used for inferring positive selection and the branch lengths separating the species being tested may also have an impact on the prevalence false positives resulting from sequence errors. Sequence errors should only substantially affect the estimation of non-synonymous and synonymous substitution rates along the terminal lineage leading to the erroneous sequence data; thus, the potential impact of sequencing error on the inference of a positively selected site or gene can be estimated by considering the potential impact of an inflated rate of non-synonymous substitution along the terminal branch on the inference of positive selection with a given test. Both the branch-site test for positive selection (which is not used in this analysis) and the sitewise tests for positive selection (including PAML M8 and SLR, first described in Chapter **??**) are sensitive to erroneous substitutions occurring at individual alignment columns, with the major difference between the two types of test being that the branch-site test is highly sensitive to substitutions along the foreground branch(es) being tested for positive selection, while sitewise tests only measure the signal for positive selection across the entire evolutionary tree.

For the branch-site test, the potential effect of sequencing error should depend on the location and length of the foreground branch(es): if the terminal branch leading to the spurious sequence is within the foreground, and especially if it represents a sizeable portion of the overall foreground branch length, then false positives could easily result; if, however, the terminal branch is outside of the foreground, then it would have little direct impact on the FPR of the branch-site test aside from adding noise to the estimation of parameters in the non-foreground branches of the tree.

For site-based tests such as SLR, the effect of sequencing error should be independent of the position of the terminal branch within the tree, depending more on the magnitude of non-synonymous substitution rate elevation resulting from the sequence error and the fraction of total branch length covered by the "erroneous" terminal branch within the phylogenetic tree being studied. It would be difficult to consider each of these factors (the terminal branch length and the magnitude of non-synonymous substitution rate elevation) in isolation due to their non-independence: sequence errors in a short terminal branch may yield a strongly elevated non-synonymous substitution rate, but the impact on the overall inference of positive selection may be limited as a result of the short branch length. On the other hand, the same erroneous sequence in a species with a longer terminal branch would likely cause a smaller elevation in the non-synonymous substitution rate (due to the higher expected number of substitutions along a longer branch) yet the impact of such an elevated rate on the sitewise inference would be proportionally greater due to the higher branch length. A reasonable hypothesis would be that these opposing factors would effectively cancel each other out in the maximum likelihood calculations. In either case, the expectation that a phylogeny with a greater proportion of its branch length within terminal branches (which, in contrast to internal branches, may contain spurious substitutions resulting from sequencing errors) would be more prone to false positives should still hold.

To summarize, the expected effect of alignment errors on the sitewise detection of positive selection should be minimal when using a good aligner and analysing data within vertebrate divergence levels, but the number of false positives resulting from sequence errors depends on a number of factors including the frequency, spatial clustering, and terminal branch length associated with sequencing, assembly and annotation errors. In some cases, even a relatively large amount of sequencing error may not produce a strongly elevated FPR (e.g., when the total internal branch length is large as when analyzing all mammals or vertebrates), as the addition of a few spurious substitutions would not significantly change the estimated non-synonymous substitution rate. In other cases, however (e.g., when teh branch length is small, and/or many low-quality genomes are included), it may significantly bias results towards excess false positives.

Simulation studies similar to those I performed in Chapters **??** and **??** could improve our understanding of the relative potential of different types of sequencing errors to introduce false positives in downstream analyses, but the absolute frequency and pattern of such errors would still difficult to predict without a reliable model for their generation. This is especially true for larger-scale errors from misassembly or misannotation, which are less easily modeled than base calling errors and could have potentially more significant negative effects. Instead, an empirical approach seems more appropriate for quantifying the false positives resulting from these types of sequence errors. In particular, two empirical studies in mammals have provided convincing evidence that sequence, alignment and annotation errors can drastically increase the number of false positive PSGs in the branch-site test for positive selection.

Schneider et al. [2009] performed a genome-wide scan for positive selection in the terminal branches of 7 mammalian genomes using the branch-site test and analysed the fraction of PSGs within subsets of high- or low-quality genes according to three sequence and alignment quality metrics. They found that the fraction of PSGs was significnatly higher for genes exhibiting lower quality sequence, annotation and alignment metric, with genes in the highest-quality and lowest-quality categories showing a 7.2-fold difference in the inferred fraction of PSGs [Schneider *et al.*, 2009]. This observation provided evidence of a correlation between the chosen quality metrics and the tendency of an alignment to exhibit positive selection. It did not necessarily imply causation, however, as the same result might have been observed—even in the absence of sequence error—if some biological properties of the true PSGs caused them to yield lower quality metrics than non-PSGs. Looking at the three metrics used in their study (sequencing coverage, gene annotation status, and alignment quality according to the heads-or-tails method), it is plausible that properties associated with elevated ω ratios and positive selection, such as recent gene duplication [Beisswanger & Stephan, 2008; Casola & Hahn, 2009; Studer *et al.*, 2008], high GC content [Ratnakumar *et al.*, 2010] or functional shifts [Storz *et al.*, 2008; Wang & Gu, 2001] might have had an error-independent effect resulting in a higher proportion of PSGs in low-scoring categories. The heads-or-tails method has also been shown to be inappropriate for estimating alignment uncertainty [Fletcher & Yang, 2010], so results based on this measurement should be taken with caution. Despite these criticisms, the analysis did provide good evidence that some of the obvious sources of error may be contributing to excessive estimates of branch-specific positive selection in mammals.

Mallick et al. [2009] took a different approach to the same problem by performing a careful resequencing and reassembly of the chimpanzee genome (the initial assembly of which had lower coverage and lower quality than the human genome) and re-analysing the evidence for positive selection along the chimpanzee linegae in 59 genes which had previously been identified as chimpanzee PSGs. The authors, who were motivated by a concern that previous reports of a larger proportion of PSGs in chimpanzee than in human [Bakewell *et al.*, 2007] were the result of its lower-quality genome rather than a biologically significant difference in levels of adaptation, found that the vast majority of PSGs identified in two previous studies showed no evidence for positive selection when using their reassembled and higher-coverage version of the chimpanzee genome [Mallick *et al.*, 2009]. This suggested that the original 4x coverage chimpanzee genome contained a number of sequencing and assembly errors leading to false inferences of positive selection. A detailed analysis of 302 codons with multiple spurious non-synonymous substitutions in the original assembly showed roughly comparable effects of sequence error (explaining 23% of codons), assembly error (14% of codons) and local alignment error (30% of codons).

Taken together, the results of Schneider et al. [2009] and Mallick et al. [2009] provide strong evidence in support of the hypothesis that errors in sequencing, assembly, annotation and alignment can result in strongly elevated inferred ω values when using sensitive tests for detecting positive selection. Furthermore, the detailed identification and quantification of error sources performed by Mallick et al. [2009] provided an empirical estimate of how important each potential source of error would be in the detection of positive selection.

Although both of these studies used the branch-site test for detecting positive selection, their results could be expected to generalize well enough to guide the design of filtering methods for the present sitewise analysis. With this work in mind, I implemented three filtering steps to help identify and remove sequences and alignment regions potentially subject to the errors noted above: filtering out low-quality sequence, removing gene fragments and recent paralogs, and identifying alignment regions with extremely high numbers of clustered substitutions.

1.3.2 Filtering out low-quality sequence

Due to the presence of several low-coverage genome assemblies in the set of available mammalian genomes and the elevated sequencing error rates in such assemblies [Hubbard *et al.*, 2007], I applied a conservative filter to the set of input sequences based on sequence quality scores where available.

Most automated genome assembly pipelines, such as the Arachne tool used to sequence many of the low-coverage mammalian genomes [Jaffe *et al.*, 2003], output a set of Phred quality scores alongside the identified genome sequence, with one Phred score per base ranging in value from 0 to 50. A Phred score represents the probability, calculated by the sequencing and/or assembly program, that a given base call is incorrect. This probability is usually concisely expressed as the negative logarithm of the probability of an error multiplied by ten, or $Q = -10log_{10}P$, where Q is the Phred score and P is the probability of an incorrect base call [Cock *et al.*, 2010].

Although Ensembl was used as the source for gene sequences in this analysis, it does not store quality scores from its source genome assemblies, so Phred quality scores were manually downloaded for all genomes with Phred-like quality scores made publicly available alongside the genomic sequence. Most quality scores were provided as a single file in FASTA format with one string of numerical scores per assembled contig. Since the process of filtering a single mammalian coding alignment required collecting scores from many different quality score files for many disjoint genomic locations, a custom script was written to process each quality score file to allow for faster score retrieval and better memory performance. In total, quality score files for XYZ genomes were indexed and used for quality filtering.

A suitable score threshold for filtering coding regions was chosen based on a study by Hubisz et al. [2011], who performed a detailed analysis of Phred quality scores, which are a probabilistic prediction of the error rate, and actual error rates in low-coverage mammalian genome assemblies by comparing the low-coverage assemblies to matched regions of high-quality sequence from the ENCODE comparative genomics dataset [ENCODE Project Consortium, 2007]. The authors identified a strong correlation between Phred scores and actual error rates for scores below 25, indicating that the scores were accurate predictors of the true error rate in this range. Error rates did not decrease significantly at scores above 25, however, suggesting that the use of an extremely high Phred score threshold would only minimally reduce error levels below those obtained with a moderate threshold. Furthermore, Hubisz et al. noted that 85% of bases in the low-coverage mammalian genomes contain very high Phred scores (> 45) and only 4% have low scores (< 20).

Based on these observations, a threshold Phred score of 25 was chosen as a reasonable trade-off between the potential benefit of avoiding miscalled bases and the potential cost of masking out correctly sequenced bases. For each protein-coding sequence with quality scores available, a "minimum score" approach was used to filter out whole codons: all codons containing one or more nucleotides with a score below 25 were masked out with three ambiguous nucleotides, 'NNN'.

The expected proportion of filtered nucleotides could be calculated from the fraction of bases below the Phred score threshold of 25. According to Hubisz et al. [2011], approximately 5% of bases in low-coverage mammalian genomes contain Phred scores below 25. The worst case scenario (e.g., the worst case in terms of the number of high-quality bases being masked as a result of using the minimum score approach) would be if only one base per codon had a score below the threshold. In that case, an expected 15% of nucleotides would be filtered, since 3 bases would be masked for every low-quality base. However, the distribution of low-quality bases is likely highly clustered, due to the uneven distribution of repetitiveness and GC content as well as the tendency for uncertain base calls to occur towards the end of sequence reads (all of which are known to affect read coverage and assembly performance, e.g. Teytelman *et al.* [2009]). A more clustered distribution of low-quality bases would cause fewer high-quality bases to become masked by the minimum score approach, reaching the limit of an expected 5% total filtered bases if low-quality bases always occurred in groups of three and were positioned along the boundary of codon triplets. Thus, anywhere from 5% to 15% of nucleotides from low-coverage genomes were expected to be filtered by this approach.

The above filtering scheme was applied to all coding sequences from each species for which quality scores were available, which included all of the species with low-coverage genomes as well as five with high-coverage genomes: chimpanzee, guinea pig, dog, horse,

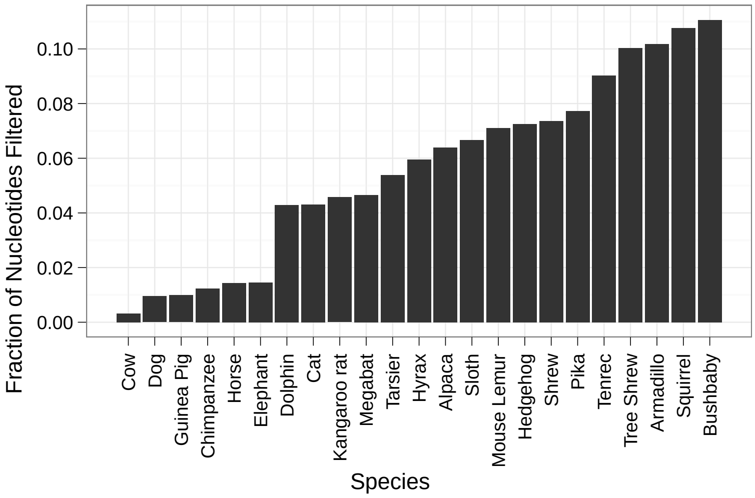

Figure 1.1

cow, and elephant. (Note that guinea pig and elephant genomes were originally sequenced at low 2x coverage for the MGP, but they have since undergone additional sequencing to produce high-coverage 7x assemblies. These assemblies were used in Ensembl version 63 and thus in the analysis described below.) The overall percentage of nucleotides filtered from each genome is shown in Figure 1.1. As expected, genomes with high-coverage sequences contained fewer bases with low Phred scores, resulting in 1-2% of nucleotides being filtered. The bulk of low-coverage genomes resulted in 4-8% of nucleotides being filtered, while five genomes (bushbaby, squirrel, tree shrew, armadillo and tenrec) showed a noticeably higher proportion of low-quality bases, with 9-11% nucleotides being filtered out. The distribution of filtered nucleotide proportions confirmed the expectation that 5-15% of nucleotides would be filtered using a Phred score threshold of 25, and the variation in filtered nucleotide proportions between different species showed that despite the uniform 2x coverage of the low-coverage mammalian genomes, different assemblies varied widely in their distributions of sequence quality scores within coding regions.

1.3.3 Removing recent paralogs

As discussed in Section **??**, the inclusion of paralogous gene relationships in a large-scale analysis of orthologous gene evolution may produce misleading signals of adaptive evolution [Lynch & Conery, 2000], artifacts resulting from gene conversion [TOCITE, 2011], and produce biases due to lineage-specific family expansion, a process which is relatively common in mammalian gene families [TOCITE, 2011]. As a result, it has traditionally been considered important to filter out recently-duplicated genes (e.g., genes duplicated after the whole-genome duplication event in the vertebrate ancestor) in large-scale evolutionary analyses. Previous genome-wide scans for positive selection involving six or fewer mammalian genomes have either required strict one-to-one orthology [Clark *et al.*, 2003; Nielsen *et al.*, 2005] or allowed very limited numbers of recent duplications in specific lineages [Kosiol *et al.*, 2008]. With larger mammalian trees, however, the requirement of strict one-to-one orthology becomes increasingly untenable: if gene duplications and deletions occur randomly in time, then the probability of observing at least one such event in a given gene family should increase linearly with the amount of branch length covered by the tree. The requirement of one-to-one orthology would result in fewer genes being available for analysis as more species are incorporated into the analysis, which is clearly an undesirable trend. As an alternative to ignoring genes which do not satisfy the requirement of strict orthology, I developed an approach, described below, for handling recently duplicated genes by removing the more-divergent paralogous copy from the the gene tree.

Before describing the method for duplications, it is worth making a point about gene deletions. Specifically, I note that gene deletions can cause problems in the branch-specific detection of positive selection, but they should not have a detrimental effect on tests for selection across the entire tree. The branch-specific effect of a gene deletion results from the merging of multiple ancestral branches into one. Take for example the inference of mutations along the evolutionary tree of human, chimpanzee and gorilla, which contains two internal nodes: HC, the human-chimpanzee ancestor, and HCG, the human-chimpanzee-gorilla ancestor. When sequences from all species are present, mutations can be separately identified as occurring along the branch from HCG to HC and along the branch from HC to the human sequence, allowing for a test to differentiate between a signal of adaptive evolution in one branch or the other. For a gene which was deleted in chimpanzee those two branches become effectively merged into one, and mutations can only be inferred to have occurred between HCG and the human sequence. The time resolution of any estimate of evolutionary rates is thus reduced, and when the identity of the branch along which synonymous and non-synonymous mutations have occurred is important to a test for positive selection, this difference can complicate the interpretation of results. Acknowledging this effect, Kosiol et al. [2008] identified a different set of orthology requirements for each branch-specific test for positive selection performed. When the test for positive selection does not depend on the identity of specific branches in the tree, however, a gene deletion would only serve to reduce the total amount of branch length available for inference. As long as the branch leading to the deleted species did not comprise a large portion of the total branch length, the effect of gene deletion on the results of tree-wide tests for selection should be minimal.

Turning back to handling gene duplications, an additional complicating factor in the current analysis was the concern that many of the apparent gene duplications were actually artifacts of the annotation of low-coverage genomes. Each low-coverage genome assembly is highly fragmented, meaning that it contains many short sequence segments that were unable to be assembled into chromosome-sized sequences due to missing sequence data. Sometimes the exons of a gene spanned the boundaries of these sequence segments, causing different parts of a gene to exist on different segments. The Ensembl annotation pipeline was not designed to merge gene annotations across different sequence segments, so each part of a gene residing on multiple sequence segments would be annotated as a separate shortened gene. These shortened genes would be treated as independent proteins by the Compara pipeline, likely being placed at very similar positions in the gene tree due to each sequence having been derived from a gene with a single correct evolutionary position. While this result might not be detrimental to sitewise analysis in itself (as each shortened gene might be correctly aligned and provide useful information to the alignment), a number of factors, including the low quality of genomic sequence and assembly within these shortened genes, problems with aligning small fractions of a gene against complete sequences, and the potential for incorrect placement of fragmented sequences within the gene tree, made it desirable to remove these shortened genes before estimating evolutionary rates. These split genes could be effectively identified by their shortened length.

Sequence divergence was the other criterion by which I selected which paralogous copy of recently-duplicated genes to retain for evolutionary analysis. A well-established theoretical model of evolution after gene duplication predicts that one of the duplicate copies retains the ancestral function (and its associated pattern of evolutionary constraint) while the other duplicate experiences relaxed constraint followed by either degradation or functional diversification [Han *et al.*, 2009]. Thus, the least-diverged copy of a recently duplicated gene should be the one most likely to have retained the pattern of evolutionary constraint shared among the mammalian species being examined in this study.

The protocol I implemented for filtering apparent paralogs used both gene length and sequence divergence to identify which gene among a set of apparent paralogous copies was most suitable to retain for sitewise analysis. Gene length was used primarily to discrim-

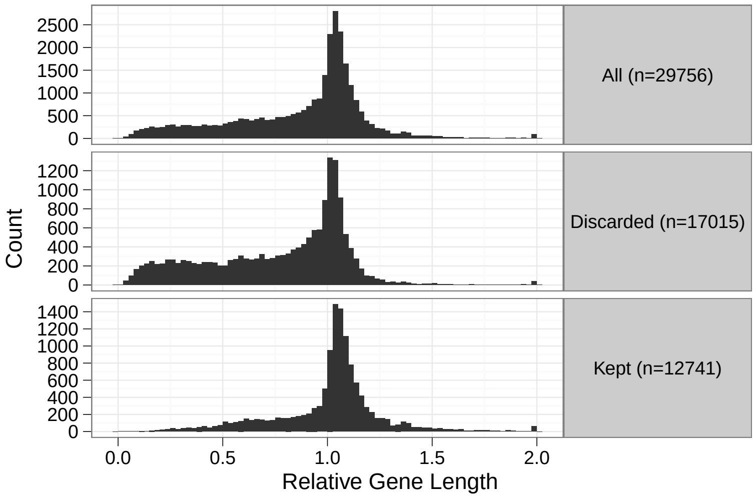

Figure 1.2: Length ratios of putative paralogs. The length ratio was calculated as the length of a putative paralogous copy divided by the mean length all sequences its corresponding gene tree. Putatively paralogous genes (top panel) were either discarded (middle panel) or kept (bottom panel) according to rules based on their length and mean sequence divergence from other aligned sequences, as described in the text.

inate spuriously shortened genes from true genes, and sequence divergence was used to distinguish between more- and less-diverged paralogs. First, the mean pairwise sequence distance was calculated for each putative paralog (hereafter referred to as the sequence distance) as the mean nucleotide distance between that sequence and every other sequence in the gene tree, using the stock Compara sequence alignments and the JC69 nucleotide model to estimate distances. Second, the ratio of the sequence length of each putative paralog to the mean sequence length across the tree (hereafter referred to as the length ratio) was also calculated.

Within each gene tree, all genes with two or more copies per species were grouped into sets of putative paralogs. Within each group of putative paralogs, a single gene was chosen to be retained for evolutionary analysis based on the following three rules, applied sequentially: (1) if only one sequence had a length ratio above 0.5 and all others had a length ratio below 0.5, the longest sequence was kept; (2) if at least one sequence yielded a sequence distance below the others, that sequence was kept; (3) if no sequence yielded a meaningful distance estimate (or if all estimated distances were identical), the longest sequence was kept.

These rules were applied to each of the 29,756 putative paralogs contained within the 16,XYZ largely orthologous gene trees from the previous chapter. Figure 1.2 shows the

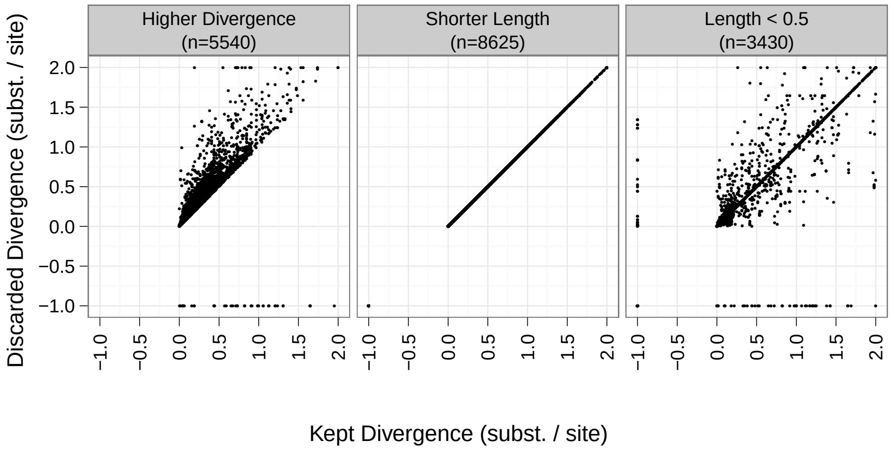

Figure 1.3: Sequence divergence of kept and discarded putative paralogs. Each point represents a gene which was discarded from the tree for one of three reasons: it had more sequence divergence than the kept gene (*Higher Divergence*; left panel), it had equal sequence divergence but shorter length than the kept gene (*Shorter Length*; middle panel), or it had a gene length (relative to the mean across all sequences) of less than 0.5 while the kept copy had a relative length greater than 0.5 (*Length* < 0.5; right panel). Divergence was measured as the mean pairwise divergence between the gene and all other sequences in the tree, and a value of -1 was assigned to genes for which no reliable divergence estimate could be attained due to a lack of sufficient data)

distributions of length ratios separately for the set of all putative paralogs, those discarded from the alignments, and those kept for subsequent analysis. The overall distribution of length ratios shows that most putative paralogs had lengths similar to the mean length across the gene tree (with a peak at or slightly above 1), but the shape of the distribution was asymmetric, with a strong bias towards shorter lengths. The filtering protocol effectively removed these shortened genes, as evidenced by the strong enrichment of lower length ratios in the distribution of discarded genes and the less skewed distribution of length ratios in the set of XYZ kept paralogs.

An alternative view of the results of the paralog filter is presented in Figure 1.3, showing the mean divergence of each discarded paralog compared that of the kept paralog. Figure 1.3 is separated into panels according to the rule used to discard the paralogous copy. The first panel corresponds to rule (1), where genes with a length ratio below 0.5 were discarded; the second panel corresponds to rule (2), where genes with higher sequence distances were removed; the third panel corresponds to rule (3), where all genes had equal mean distances and the longest gene was kept. The first panel shows that genes discarded on the basis of having a very short length contained sequence distances similar to the kept copies (i.e., the highest density is along the diagonal and there is no bias above or below the diagonal), supporting the notion that these discarded genes were smaller fragments of split genes in low-coverage genomes. The second panel shows that when paralogous copies could be differentiated by sequence distance, they tended to have low average distances (<0.5 substitutions per nucleotide site) and only a small difference between the kept and discarded copy (e.g., most of the distribution is just above the diagonal, and few points are above the dashed line corresponding to twice the distance in the discarded copy). Finally, the distribution of lengths and distances in the set of genes where length was the discriminating factor shows that most of these genes were mostly identical whether measured by sequence distance or sequence length.

These results provided evidence that many recently duplicated genes are identical or very similar to each other: for roughly 30% of putative paralogs, not enough time has elapsed since the duplication event for a significant amount of sequence change to have occurred, and the choice between retaining one copy or the other was essentially arbitrary. For the roughly 40% of putative paralogs where differences in distance to other sequences in the tree could be identified, these differences tended to be small, suggesting that massive functional divergence of recent gene duplicates has not been a common phenomenon in mammalian evolution. Still, I hypothesized that the retention of the least-divergent gene copy would minimize the potential influence of recent gene duplications on the sitewise analysis. To evaluate whether the genes containing recent duplications had an effect on the analyses performed in this chapter and the next, I identified the XYZ genes which contained the 10,821 paralogous copies discarded based on sequence distance and excluded these genes from the high-quality "stringent" dataset described in Section **??**.

1.3.4 Identifying clusters of non-synonymous substitutions

After filtering for sequence quality and removing paralogous genes and shortened gene fragments, PRANK was used to align the codon sequences of each of the 16XYZ mammalian gene trees. Manual analysis of a number of these alignments revealed that many alignments contained short stretches of clearly nonhomologous sequence in one species, often flanked by stretches of perfect homology and often lying on the borders of exon junctions. An example of one such region is shown in Figure **??**. In this otherwise highly conserved region of the XYZ gene, a short 20-codon stretch of the XYZ sequence appears to contain little homology to the sequences from other species.

These obviously erroneous stretches were likely due to mis-assembly of a genomic region or mis-identification of exon boundaries within the gene of one species. These errors were particularly concerning with respect to the detection of positive selection, as the incorporation of a stretch of apparently nonhomologous material into a sequence alignment would produce many alignment columns with multiple nucleotide differences per codon. As discussed in Section 1.3.1, this type of error is particularly prone to cause false positives in the detection of positive selection.

I hypothesized that these stretches of non-homologous sequence could be identified by their impact on the pattern of substitutions within each alignment. A stretch of non-homologous aligned sequence would be expected to produce a localized cluster of apparent synonymous and nonsynonymous substitutions between the sequence containing the erroneous stretch and its ancestor. Because these substitutions would be restricted to one terminal branch in the gene tree and a region of the alignment limited to the length of the non-homologous stretch, a scan for clustered substitutions within the terminal lineages of genes might be an effective way of identifying these erroneous sequences.

Two factors may confound the ability to use **wsc!** (**wsc!**)s for identifying regions of aligned non-homologous sequence in mammalian alignments. First, the length of the terminal branch leading to each species determines how many lineage-specific substitutions would be expected to occur within a given stretch of homologous aligned sequence. The terminal human branch, for example, is very short (as it shares a recent common ancestor with chimpanzee), while the platypus branch is very long (sharing a most recent common ancestor only with the entire eutherian clade). Thus, one would expect to observe many more lineage-specific substitutions in platypus than in human for a given window of an alignment. In contrast, a stretch of aligned non-homologous sequence should introduce, on average, a constant number of non-synonymous and synonymous substitutions to the terminal lineage. This should make it more difficult to distinguish homologous from non-homologous stretches in species with long terminal lineages; on the other hand, this trend should also serve to limit the potential impact of non-homologous stretches in those species on the detection of positive selection, as the resulting elevation in non-synonymous or synonymous substitutions rates would be less.

The second confounding factor is that non-synonymous substitutions have been shown to be significantly more clustered than expected by chance in a number of genomic analyses of mammalian and insect genomes [Bazykin *et al.*, 2004; Callahan *et al.*, 2011; **?**]. Thus, a filter based on clustered non-synonymous substitutions may have a tendency to remove true clusters of non-synonymous substitutions from the dataset. The influence of this factor may be evaluated by comparing clusters of substitutions in terminal branches to those in internal branches: while both internal and terminal branches of the mammalian tree should harbor similar levels of truly clustered non-synonymous and synonymous substitutions, only the terminal lineages should contain large clusters resulting from stretches of aligned non-homologous sequence.

I investigated the distributions of non-synonymous and synonymous substitutions within windows of mammalian alignments by using *codeml* [Yang, 2007] under the M0 model (e.g., assuming one ω for all sites and all branches in the tree) to perform the marginal reconstruction of ancestral sequences at internal nodes [**?**] and storing the substitution events implied by the reconstructed ancestral sequences of each gene alignemnt. Only substitution events occurring between codons with high posterior probabilities in the marginal ancestral reconstruction (> 0.9) were included, and the location of each substitution along the alignment and within the gene tree was stored. This analysis was performed on all gene trees, yielding a large database of substitution events along internal and terminal branches of the phylogenetic tree confidently inferred from the PRANK alignments of mammalian coding regions.

Using this set of inferred substitutions, counts of synonymous and non-synonymous substitutions within non-overlapping 15-codon alignment windows for a selection of species and internal nodes were collected; the results of this analysis are shown in Figure 1.4, which plots the number of 15-codon windows containing a given number of non-synonymous and synonymous substitutions for a selection of terminal and internal nodes. The mean length of the branch ancestral to the given node, indicated in parentheses after each node name, was calculated from the set of branch lengths estimated by *codeml*.

Figure 1.4 shows that the vast majority of 15-codon windows in these alignments contained few substitutions, but a long tail of non-synonymous and synonymous substitutions were observed for some nodes. Comparing the counts of non-synonymous vs. synonymous substitutions within the terminal nodes (Figure 1.4, top panel), a pattern is seen where the non-synonymous counts (red bars) are higher at 0 substitutions, lower in the middle range of substitutions (1–5 substitutiosn), and higher again in the higher range of substitutions (>5 substitutions). The pattern in the lower range is consistent with the action of purifying selection on protein-coding regions, causing a reduced number of windows with multiple non-synonymous substitutions. The excess of windows with large numbers of non-synonymous substitutions runs against the pattern of purifying selection; instead, it shows that unexpectedly long clusters of non-synonymous substitutions were a widespread feature of the mammalian alignments. The red and blue triangles drawn in each plot mark the number of substitutions below which 99.9% of windows are contained; the shift of the non-synonymous markers to the right clearly shows the excess of highly clustered non-synonymous substitutions. Interestingly, human—which has the highest quality and best annotated genome—does not show a noticeable excess of clustered non-synonymous substitutions.

Comparing the pattern seen for terminal nodes to those from internal nodes provided further evidence for the presence of many stretches of non-homologous sequence within the mammalian alignments. For example, the terminal gorilla node is roughly equivalent in average branch length to the internal primates node (0.023 vs. 0.028), but gorilla contains windows with up to 14 non-synonymous substitutions while primates contain a maximum of 8. Looking at the non-synonymous and synonymous 99.9% quantiles, three of the four internal nodes had equal quantile positions for non-synonymous and synonymous substitutions, but the rodent ancestral node did not. This was an interesting difference, as the gene annotations for most rodent genomes were likely derived from alignments to mouse rather than human. In the case of discordant gene annotations, the entire rodent clade would share an aligned non-homologous stretch, causing clustered substitutions to be inferred along the internal rodent branch. This raises the interesting possibility that the entire rodent clade suffers from misaligned non-homologous stretches due to differences in gene annotations between rodent and non-rodent species. This appeared to be the case for at least one gene: Figure **??** shows the region surrounding a 15-codon window with 11 apparent rodent non-synonymous substitutions, likely the result of a difference in exon annotations between rodent and non-rodent genomes.

The end result of this analysis was the identification, for each terminal and internal node of the mammalian tree, windows with non-synonymous substitution counts above the top 0.1% of 15-codon windows genome-wide. The location of these windows was stored for later use in defining the most stringently-filtered sitewise dataset, described in Section **??**.

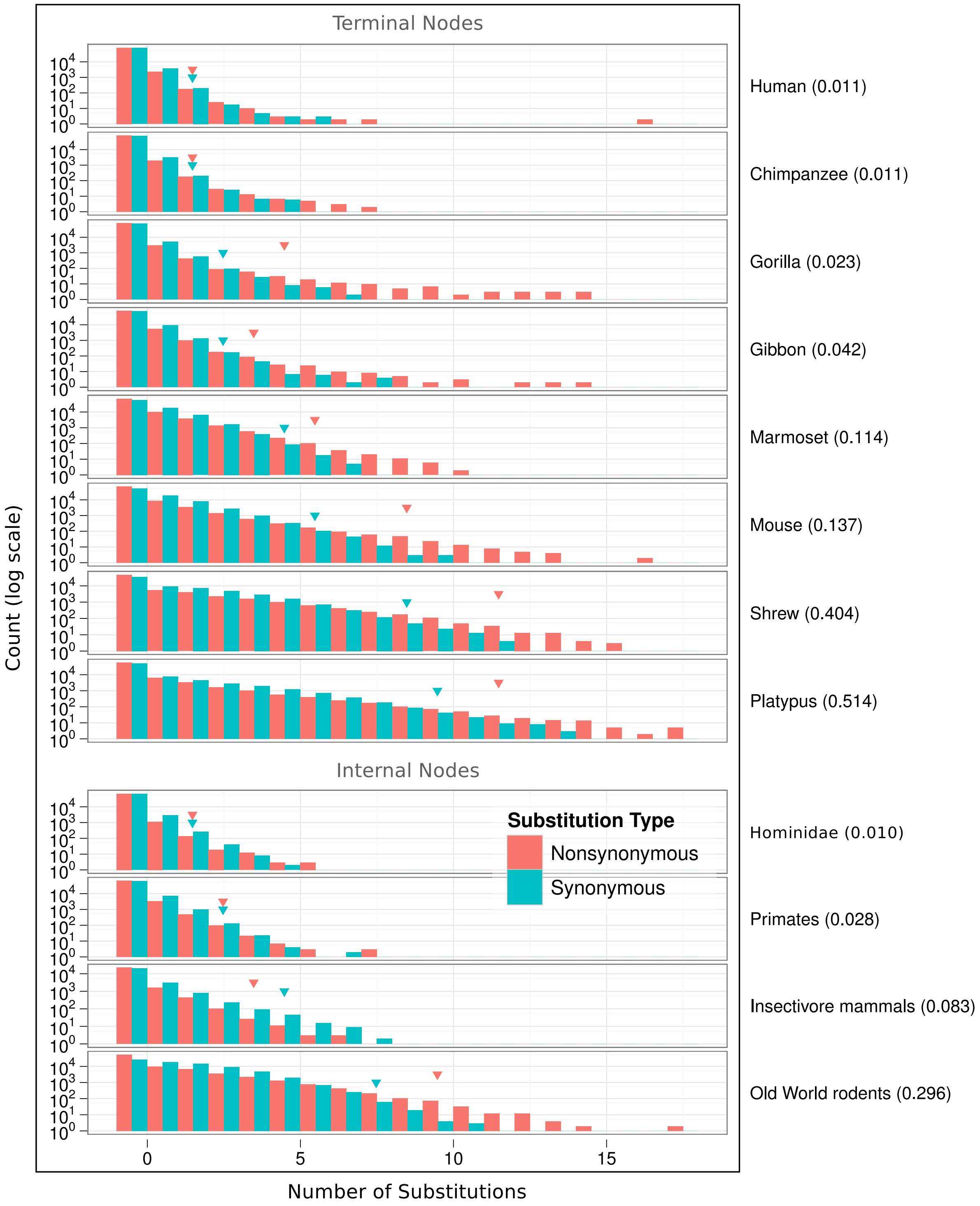

Figure 1.4: Counts of inferred non-synonymous (red bars) and synonymous (blue bars) substitutions in 15-codon windows along terminal and internal branches of the mammalian tree. The leftmost two bars correspond to windows with 0 substitutions, the next two bars correspond to windows with 1 substitution, and so on. Red and blue arrows indicate the number of non-synonymous and synonymous substitutions, respectively, corresponding to the 99.9% percentile across all windows in that node.

1.4 Genome-wide analysis of sitewise selective pressures in mammals

1.4.1 Mammalian species subsets for sitewise analysis

SLR was run sequentially on several species subsets of each alignment of mammalian orthologs. For each subset, sequences corresponding to species within the subset were extracted from the alignment along with the corresponding subtree and input to SLR. If fewer than two sequences were available for a given subset, that subset was skipped and its absence from the dataset was recorded. Eight subsets in total were selected for analysis; the species included in each subset and some phylogenetic measures of each subset are listed in Table 1.1.

Three subsets (Glires, Primates, and Laurasiatheria) were chosen because they represent the three mammalian superorders with the greatest taxonomic representation in Ensembl, providing an opportunity to compare the molecular evolutionary dynamics of three monophyletic mammalian groups containing varying levels of divergence, diverse biological characteristics, and a number of high-quality reference genomes. A fourth parallel mammalian subclade, named Atlantogenata and consisting of sloth, armadillo, tenrec, elephant and hyrax, was also included, but the monophyly of this group is still debated [Churakov *et al.*, 2009; Murphy *et al.*, 2007] and it contains only one high-coverage genome. As such, it was not considered a primary target for the mammalian superorder analysis.

Two larger species sets, Eutheria and Mammalia, were chosen for the purpose of measuring average sitewise selective pressures with high precision across a wider group of mammals. Using the Ensembl species tree as a guide, the estimated total synonymous branch lengths spanned by Ensembl species within Eutheria and Mammalia were 4.95 and 6.18, respectively. Simulations performed by Anisimova and Yang [**??**] and by myself in Chapter **??** predicted that the greater amount of branch length in the Eutherian and Mammalian trees—with two to three times the value of 1.71 for Laurasiatheria, the superorder with the largest total branch length—would result in significantly higher levels of power and accuracy for estimating sitewise ω and detecting sitewise positive selection. In this respect, Mammalia and Eutheria were more similar to each other than to any of the superorders.

However, the Mammalia and Eutheria subsets differed markedly in a different (and largely orthogonal) phylogenetic factor, the evolutionary depths of their last common ancestors. Whereas the ancestor of all Eutherian mammals lived ca. [125] mya, the Mammalian ancestor dates back to [320] mya. This suggested that a comparison between the sitewise results for the two groups might provide useful insight into the general effect of adding longer, deeper branches to a sitewise evolutionary analysis as well as some indirect evidence on the molecular evolutionary dynamics of our most distant mammalian relatives (the Eutheria and Mammalia groups only differ by the inclusion of wallaby, opossum and platypus in the Mammalia group).

Quantitatively, as measured by the MPL from the Ensembl species tree, the Eutheria subset (MPL =0.24) is far more similar to either of the three superorders (MPL from 0.13 to 0.27) than to the Mammalia subset (MPL =0.54). This is due to the striking adaptive radiation of Eutherian mammals [Archibald, 1999; Bininda-Emonds *et al.*, 2007], which caused a quick succession of speciation events around the K-T boundary and gives a largely star-like structure to the eutherian evolutionary tree. Interestingly, according to the time-resolved mammalian tree from Bininda-Edmonds et al. [2007] the Diprodontia order (containing wallaby and opossum, two outgroups to the Eutheria clade) experienced a radiation similar to, but less prounced than, the Eutherian radiation; a comparison of the evolution of the deeply-rooted Diprodontia clade to its sister Eutherian clade would be very enlightening, but the species representation of Diprodontia (currently at one high-coverage and one low-coverage genome) is too limited to allow for a powerful analysis. Nevertheless, the inclusion of the three non-Eutherian species in the Mammalian species group was expected to provide an additional data point for aiding in an understanding of the complex relationship between branch length, power and biological variability in the analysis of sitewise selective pressures.

Finally, to further investigate the combined impact of evolutionary depth, biological variability and branch length on the results of large-scale sitewise analyses, two "sparse" subsets were created to act as controls relative to two existing species subsets. The species in the Sparse Glires group were chosen to approximate the total branch length of the Primate clade with species from the Glires clade, while the Sparse Mammals subset was constructed by selecting one species (preferably with a high-coverage genome) from each major mammalian branch, greatly reducing the total branch length covered but maintaining a similar evolutionary depth and distribution of branches within the species tree. The branch lengths in Table 1.1 show that the Sparse Glires group was only somewhat successful in its goal of approximating the Primates branch length (with total branch lengths of 0.99 and 0.68, respectively) while the Sparse Mammals group achieved a threefold lower total

			Ensembl		Gene Median	
Name	(Species Count) Species List	N_E	MPL	Total	MPL	Total
Primates	(10) Bushbaby, Chimpanzee, Gibbon, Gorilla, Human, Macaque, Marmoset, Mouse Lemur, Orangutan, Tarsier	20000	0.13	0.68	0.16	0.83
Glires	(7) Guinea Pig, Kangaroo rat, Mouse, Pika, Rabbit, Rat, Squirrel	230000	0.27	1.44	0.40	1.90
Laurasiatheria	(12) Alpaca, Cat, Cow, Dog, Dolphin, Hedgehog, Horse, Megabat, Microbat, Panda, Pig, Shrew	34410	0.19	1.71	0.26	2.16
Atlantogenata	(5) Armadillo, Elephant, Hyrax, Sloth, Tenrec	30000	0.20	0.83	0.26	0.97
Eutheria	(35) Alpaca, Armadillo, Bushbaby, Cat, Chimpanzee, Cow, Dog, Dolphin, Elephant, Gibbon, Gorilla, Guinea Pig, Hedgehog, Horse, Human, Hyrax, Kangaroo rat, Macaque, Marmoset, Megabat, Microbat, Mouse, Mouse Lemur, Orangutan, Panda, Pig, Pika, Rabbit, Rat, Shrew, Sloth, Squirrel, Tarsier, Tenrec, Tree Shrew	110000	0.24	4.95	0.35	6.43
Mammalia	(38) Alpaca, Armadillo, Bushbaby, Cat, Chimpanzee, Cow, Dog, Dolphin, Elephant, Gibbon, Gorilla, Guinea Pig, Hedgehog, Horse, Human, Hyrax, Kangaroo rat, Macaque, Marmoset, Megabat, Microbat, Mouse, Mouse Lemur, Opossum, Orangutan, Panda, Pig, Pika, Platypus, Rabbit, Rat, Shrew, Sloth, Squirrel, Tarsier, Tenrec, Tree Shrew, Wallaby	120000	0.54	6.18	0.67	8.21
Sparse Glires	(5) Guinea Pig, Kangaroo rat, Mouse, Rat, Squirrel	230000	0.25	0.99	0.36	1.32
Sparse Mammalia	(7) Armadillo, Dog, Elephant, Human, Mouse, Platypus, Wallaby	120000	0.51	2.18	0.61	2.86
HQ Mammalia	(4) Dog, Human, Mouse, Rat	120000	0.28	0.81	0.31	1.61
HMRD	(9) Chimpanzee, Cow, Dog, Horse, Human, Macaque, Mouse, Pig, Rat	120000	0.22	1.23	0.34	1.01

Table 1.1: Species subsets used for sitewise analysis. Values under the "Ensembl" heading were calculated from subsets of the species tree used for evolutionary analyses by the Ensembl Compara pipeline, while values under the "Gene Median" heading were calculated as median values across the 15,XYZ gene trees analyzed (with branch lengths optimized by SLR). MPL – mean path length, Total – total branch length.

branch length compared to the full Mammalia group while maintaining a nearly identical MPL.

1.4.2 Evaluation of the bulk distributions and the design of a filtering appraoch

Sitewise data were collected from SLR and stored in a database for storage and further analysis. The Mammalia subset, containing the most branch length of all the datasets and representing the entire set of aligned sequences, and the Primate subset, containing the lowest overall branch length, were used to perform quality-control checks on the sitewise data. The point of these checks were to evaluate whether any additional filtering of the sitewise results was necessary before characterizing the global distribution of constraint in this and the other species subsets. Even if the sequence and alignment filters described above were successful at reducing the number of false positives due to incorrect input alignments, the behavior of SLR when applied to large datasets of heterogeneous alignments has not been well-studied, and a number of biases might have influenced the results. A particular point of concern was that columns with different patterns of gap and non-gap sequences, especially those with few non-gap sequences, might yield different performance characteristics. Although the SLR method was sensibly designed to account for uncertainty in the estimation of ω and detection of positive selection, one might reasonably expect less-desirable statistical properties from sites with 2 non-gap sequences compared to sites with 20.

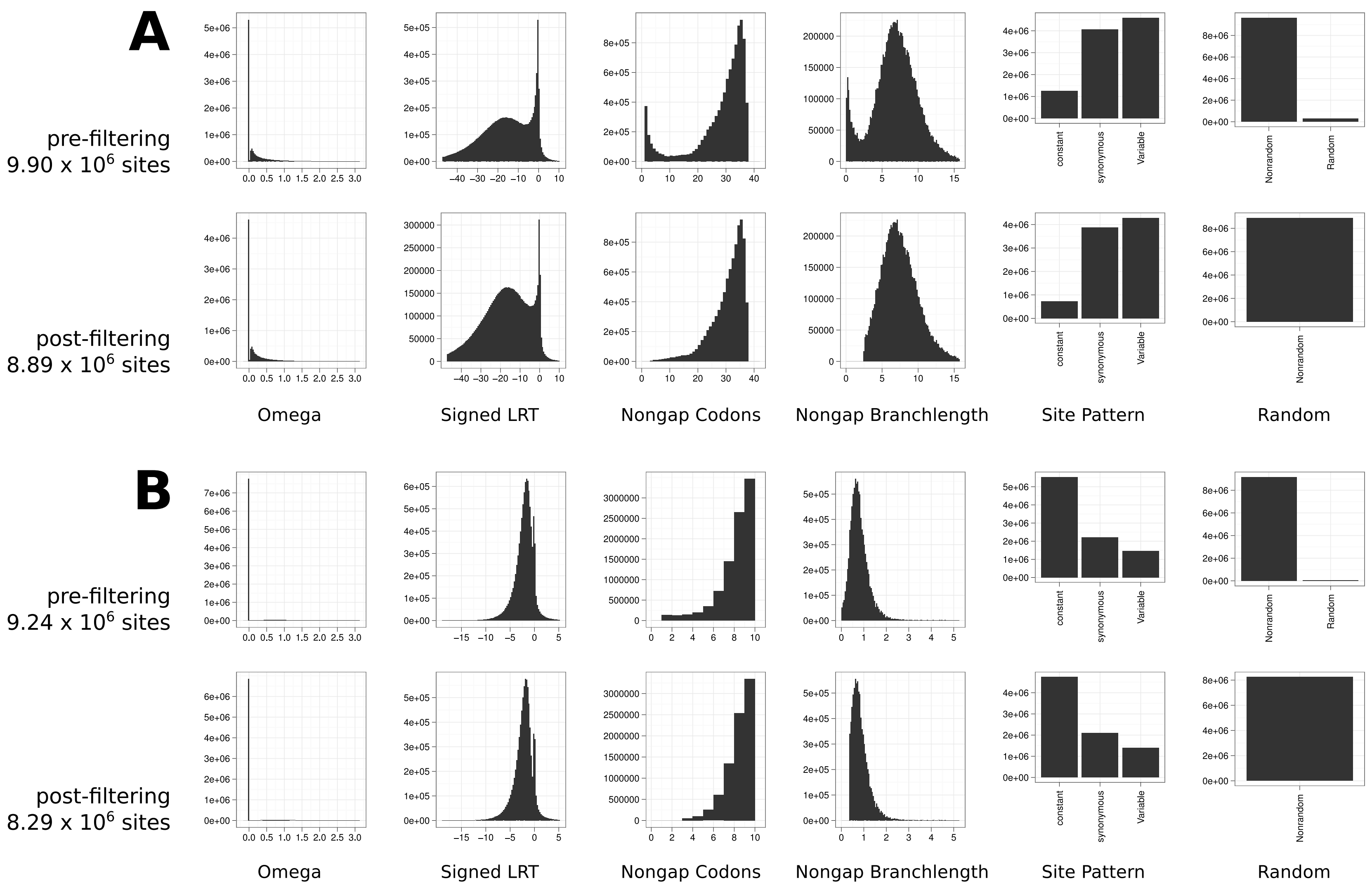

Figure 1.5: Distributions of sitewise values for Mammalia (A) and (B) Primates, before (top row) and after (bottom row) removing sites based on the filtering scheme (see text). Note: the y-axis scale varies between rows, and the x-axis scale varies between (A) and (B) for the Signed LRT, Nongap Codons and Nongap Branch Length values.

Figure 1.5 shows the distributions of six sitewise values: two continuous values output by SLR (Omega and Signed LRT), two categorical values from SLR (Site Pattern and Random), and two values calculated from the codon alignment (Nongap Codons and Nongap Branch Length). The Nongap Codons value measures the number of non-gap codons in the given alignment columnn, while the Nongap Branch Length represents the total branch length connecting all non-gap sequences (using the gene tree with branch lengths optimized by SLR).

A prominent feature of the distribution of ω values for the unfiltered Mammalian data, shown in the top panel of Figure 1.5A, was the large number of sites with a maximum-likelihood estimated ω of zero. Further inspection of the data revealed that all zero-ω sites contained either synonymous or constant site patterns, and all sites with constant patterns (and nearly all sites with synonymous patterns) yielded a maximum likelihood ω estimate of zero. An estimate of zero for synonymous sites is intuitively appropriate, as the lack of any non-synonymous substitutions throughout the tree would seem to provide no evidence for a non-synonymous substitution rate of greater than zero. For constant sites the case is less clear, because no data regarding the rate of either synonymous or non-synonymous substitutions exists. However, given SLR's assumption of a constant synonymous substitution rate throughout each gene [Massingham & Goldman, 2005], the ω value which maximizes the likelihood of observing zero substitutions is zero, since that value minimizes the non-synonymous (and total) substitution rate.

It is interesting to note that a small proportion (ca. 0.2%) of synonymous sites resulted in maximum likelihood estimates greater than zero. Manual investigation of a number of these sites [Use the sites data and 'seq' table to do a more quantitative analysis here?] showed them all to include synonymous codons with multiple nucleotide differences (such as those coding for serine and arginine), for which SLR's mechanistic codon model—which does not allow for multiple simultaneous nucleotide changes–required the inference of multiple non-synonymous substitutions and, thus, a non-synonymous substitution rate of greater than zero. The existence of multiply-substituted codons in alignments has been previously reported [Averof *et al.*, 2000; Whelan & Goldman, 2004], and empirical results have supported the notion that codon models that allow for multiple simultaneous nucleotide changes better describe evolution than those that do not [Kosiol *et al.*, 2007]. However, the very low proportion of synonymous sites requiring nonzero non-synonymous substitution rates suggested that the impact of these effects on the current dataset was minimal; this is likely due to the relatively short branch lengths separating the nodes of the mammalian tree, making it less likely that codons with multiple substitutions (whether the result of simultaneous multiple nucleotide changes or successive single changes) would be observed [Kosiol *et al.*, 2007].

The distributions of Nongap Codons and Nongap Branch Length values in the top row of Figure 1.5A showed that many alignment columns contained only a few non-gap sequences. Both distributions were bimodal with a larger peak at the upper end of the range and a smaller peak at the lower end of the range. If the sites with low non-gap codon counts represented accurate evolutionary histories, the observed excess of mostly gapped sites might be taken as an indication that insertion events in terminal lineages or recent ancestral lineages were prominent enough in mammalian evolution to leave a noticeable signature of sites with low non-gap codon counts. This would be a very interesting observation, but unfortunately it is not likely a correct one. Given the many possible sources of error in the creation of Ensembl gene trees, however, a more likely scenario was that sites with low codon counts and low branch lengths came from stretches of sequence which only exist in a few species as a result of annotation or alignment error, with a higher probability of being nonhomologous and showing spurious signals of positive selection. This would make such sites prime candidates for filtering out prior to a large-scale analysis.

To test the hypothesis that sites with few non-gap sequences are less reliable than other sites, the Mammals and Primates data were split into deciles by nongap branch length and sites within each decile were summarized by the proportion of sites showing evidence for purifying and positive selection; the results of this analysis are presented in Table 1.2. The lowest decile appeared to be a clear outlier in the Mammalia dataset, with nearly 17% of sites having an estimated ω of greater than 1 and 2% of sites showing significant evidence for positive selection at a nominal 5% error rate. Deciles with greater nongap branch lengths showed lower proportions of sites with $\omega > 1$ and less evidence for positive selection, with a gradual increase in both values at progressively higher deciles. The gradual increase in evidence for positive selection with increasing nongap branch length could be explained by genes with higher overall dN/dS ratios (and perhaps more positive selection) producing, on average, higher estimated branch lengths due to the increased non-synonymous substitution rate. Overall, these patterns were consistent with the expectation that sites with few non-gap sequences were not consistent with the bulk of the dataset, and Table 1.2 showed that removing sites with the lowest 10% of nongap branch length would remove most of the apparently anomalous sites.

The breakdown of Primates data in Table 1.2 showed a trend similar the Mammalia

	BL	Nongap BL			Nongap Codons			ω_{ML}, %		
	Quantile	25%	50%	75%	25ff%	50%	75%	< 1	> 1	PSC$_{5\%}$, %
Mammalia	0.10	0.31	0.74	1.46	2	3	6	81.87	18.13	2.09
	0.20	3.28	3.78	4.17	19	30	35	95.01	4.99	0.52
	0.30	4.77	5.02	5.24	27	33	36	96.90	3.10	0.35
	0.40	5.67	5.86	6.04	28	33	36	96.94	3.06	0.35
	0.50	6.38	6.55	6.72	29	33	36	96.75	3.24	0.39
	0.60	7.06	7.22	7.40	29	33	36	96.41	3.59	0.44
	0.70	7.77	7.95	8.15	29	33	36	96.04	3.96	0.50
	0.80	8.58	8.79	9.03	29	33	35	95.43	4.57	0.61
	0.90	9.58	9.88	10.24	29	33	35	94.57	5.43	0.79
	1.00	11.22	12.00	13.28	29	32	35	92.95	7.04	1.14
Primates	0.10	0.17	0.25	0.30	4	6	8	94.42	5.58	0.61
	0.20	0.38	0.41	0.44	8	9	10	94.39	5.61	0.32
	0.30	0.49	0.52	0.54	8	9	10	93.64	6.36	0.30
	0.40	0.59	0.61	0.63	8	9	10	93.09	6.91	0.33
	0.50	0.67	0.69	0.71	8	9	10	92.39	7.61	0.35
	0.60	0.76	0.78	0.80	8	9	10	91.29	8.71	0.46
	0.70	0.85	0.87	0.90	8	9	10	90.68	9.32	0.50
	0.80	0.97	1.00	1.04	8	9	10	89.10	10.90	0.66
	0.90	1.13	1.19	1.25	9	9	10	87.13	12.87	0.86
	1.00	1.44	1.61	1.95	8	9	10	84.64	15.36	1.24

Table 1.2: Proportions of sites with evidence for purifying and positive selection in the Mammalia and Primates datasets broken down by nongap branch length. Sites were separated into 10 equally-sized bins of nongap branch length and summarized by the 25^{th}, 50^{th} and 75^{th} percentiles of nongap branch length and nongap codons, the percentage of sites with ω estimated below or above 1, and the percentage of sites with significant evidence of positive selection at a nominal 5% FPR.

dataset, although the distinction between the lowest decile and the rest of the dataset was less clear. The F_{pos} in the lowest decile was only slightly higher than in the next-highest decile, and $F_{<1}$ was lower than in all other bins. The gradual increase of $F_{>1}$ and F_{pos} in higher branch length deciles was similar to that seen in the Mammalia dataset, however. Despite weaker evidence in the Primates data for the erroneous nature of sites with few non-gap sequences, it still appeared that filtering sites in the bottom decile would improve the overall quality and consistency of the data.

Turning back to the bulk distributions in Figure 1.5, the rightmost panel depicts the small set of sites designated as “random” by SLR. These sites were flagged by SLR as having a site pattern not significantly different from random [Massingham & Goldman, 2005], and they were also targeted for removal before analysis of the global distribution.

The final filtering protocol applied to each sitewise dataset included three steps. First, all sites within the bottom 10% of nongap branch length values were removed; second, sites flagged by SLR as "random" were removed; third, all sites with fewer than four non-gap sequences were removed.

The most prominent effect of the filter on the bulk distributions in Figures 1.5A and 1.5B was, as expected, the removal of the excess of sites with low non-gap branch lengths and non-gap codon counts. The distribution of ω estimates and Signed LRT statistics were largely unchanged, indicating that the overall characteristics of each dataset were not significantly altered by the filter. The lack of large-scale change was a somewhat reassuring result, given that the filter only removed roughly 10% of sites from each dataset.

A more detailed comparison of various summary statistics for the filtered and unfiltered datasets showed that filtering had a noticeable impact on three quantities of interest: it reduced the proportion of constant sites, lowered the mean ω, and slightly decreased the percentage of positively-selected sites. Tables **??** and 1.6 contain summary statistics and calculations performed on the filtered and unfiltered Mammalia and Primates data. Most of the data contained in these tables will be more fully described in the next subsection, but a comparison of the filtered and unfiltered rows for Primates and Mammalia provided a means by which to quantitatively assess the effect of filtering on particular aspects of the dataset. First, the percentage of constant sites was reduced in the post-filtering data, moving from 60.04% to 57.62% in Primates and from 12.62% to 8.17% in Mammalia. This was expected, as the sites removed by filtering were enriched in constant site patterns due to their lower non-gap branch lengths. Second, the mean ω value was slightly reduced in Primates (from 0.32 to 0.27) and greatly reduced in Mammalia (from 0.49 to 0.20), likely due to the removal of sites containing a small number of nonhomologous codons which might have produced abnormally high sitewise maximum likelihood ω estimates. Third, the proportion of positively-selected sites (shown for a range of significance thresholds in Table 1.6) was moderately reduced in both Primates (from 0.56% to 0.53% at $P_{\chi_1^2} < 0.05$) and Mammalia (from 0.72% to 0.56%). These three effects of filtering, each showing a shift indicating more useful data (e.g., a lower percentage of constant sites) or less evidence for positive selection (e.g., lower mean ω and proportion of positively-selected sites) in the post-filtering datasets, together provided evidence supporting the inclusion of such a filtering step prior to the analysis and comparison of sitewise estimates in different species sets. Although most quantities of interest were not noticeably changed, those that were affected by filtering shifted to more conservative values, which was taken to be a positive sign given the persistent concern regarding the presence of false positives in detecting positive selection.

			Site Pattern, %			Med.	Nongap BL			ω_{ML}		ω_{ML} Below / Above, %			
Name	Filter	Sites	Const.	Syn.	Nsyn.	Codons	Med.	Mean	SD	Mean	SD	< 0.5	< 1	> 1	> 1.5
Primates	None	9.76e+06	59.69	24.07	16.24	9	0.74	0.86	1.22	0.23	0.63	85.97	90.93	9.07	5.88
	Default	8.71e+06	57.24	25.51	17.25	9	0.79	0.94	1.27	0.23	0.62	85.22	90.73	9.27	5.79
	Stringent	6.10e+06	58.01	26.64	15.35	9	0.78	0.86	0.59	0.21	0.58	86.29	91.53	8.47	5.17
	Pfam	2.33e+06	58.98	27.47	13.55	9	0.77	0.94	1.72	0.17	0.52	88.91	93.52	6.48	3.87
Mammalia	None	1.05e+07	12.43	40.80	46.77	32	6.95	6.95	4.15	0.22	0.47	86.20	94.29	5.71	2.94
	Default	9.42e+06	8.03	43.36	48.61	33	7.30	7.63	3.82	0.19	0.37	87.49	95.75	4.25	1.62
	Stringent	5.54e+06	7.95	48.54	43.51	34	7.49	7.76	2.65	0.15	0.30	90.67	97.23	2.77	0.92
	Pfam	2.48e+06	8.68	50.49	40.83	33	7.28	7.66	4.81	0.13	0.30	91.94	97.49	2.51	0.92
Primates	(WCS)	3.87e+05	38.71	19.89	41.40	9	1.11	1.43	1.51	0.57	0.89	65.50	77.43	22.56	14.90
Mammalia	(WCS)	4.09e+05	2.39	14.12	83.49	30	9.17	9.47	3.92	0.47	0.55	66.13	86.22	13.78	5.81

Table 1.3: Summary statistics and maximum likelihood ω estimates for sitewise data in eight species groups. Rows labeled "Primates (raw)" and "Mammalia (raw)" were computed based on unfiltered data and are included for reference. Columns under the "ω_{ML} Above / Below, %" heading measure the cumulative percentage of sites with ω_{ML} below or above the indicated value. Med.—median, Const.—constant, Syn.—synonymous, Nsyn.—non-synonymous, ML—maximum likelihood.

		Positively Selected Sites (%)								$P_{\chi_1^2} < 0.1$, %			$P_{\chi_1^2} < 0.05$, %		
Name	Filter	$P_{\chi_1^2} < 0.1$		$P_{\chi_1^2} < 0.05$		$P_{\chi_1^2} < 0.01$		FDR< 0.05		Neg.	Neut.	Pos.	Neg.	Neut.	Pos.
Primates	None	99002	(1.01)	57919	(0.59)	18277	(0.19)	243	(0.002)	29.93	69.05	1.01	14.31	85.09	0.59
	Default	82607	(0.95)	47825	(0.55)	14619	(0.17)	104	(0.001)	33.27	65.79	0.95	15.98	83.47	0.55
	Stringent	49207	(0.81)	28203	(0.46)	8481	(0.14)	59	(0.001)	32.34	66.86	0.81	13.98	85.56	0.46
	Pfam	13988	(0.60)	8050	(0.35)	2440	(0.10)	23	(0.001)	37.73	61.67	0.60	18.80	80.85	0.35
Mammalia	None	114094	(1.09)	75509	(0.72)	30692	(0.29)	2052	(0.020)	80.21	18.71	1.09	77.03	22.25	0.72
	Default	76370	(0.81)	52096	(0.55)	23323	(0.25)	1879	(0.020)	86.51	12.68	0.81	83.88	15.57	0.55
	Stringent	28789	(0.52)	19690	(0.36)	8896	(0.16)	760	(0.014)	90.51	8.97	0.52	88.42	11.22	0.36
	Pfam	12756	(0.51)	8956	(0.36)	4353	(0.18)	428	(0.017)	91.58	7.91	0.51	89.70	9.93	0.36
Primates	(WCS)	12498	(3.23)	7605	(1.97)	2597	(0.67)	31	(0.008)	27.78	68.99	3.23	14.95	83.08	1.97
Mammalia	(WCS)	12912	(3.16)	9124	(2.23)	4389	(1.07)	425	(0.104)	64.95	31.89	3.16	60.17	37.60	2.23

Table 1.4: Proportions of sites subject to positive, purifying and neutral selection at various LRT$_{SLR}$ thresholds. The Benjamini-Hochberg method [Benjamini & Hochberg, 1995] was used to identify the LRT$_{SLR}$ threshold at which FDR<0.05. For columns under the headings "$P_{\chi_1^2} < 0.1$, %" and "$P_{\chi_1^2} < 0.05$, %", Pos. and Neg. are the percentage of sites with significant evidence for positive and negative selection, respectively, and Neut. is the percentage of "neutral" sites not showing significant evidence for non-neutral selection.

			Site Pattern, %			Med.	Nongap BL			ω_{ML}		ω_{ML} Below / Above, %			
Name	Filter	Sites	Const.	Syn.	Nsyn.	Codons	Med.	Mean	SD	Mean	SD	< 0.5	< 1	> 1	> 1.5
Primates	Default	8.71e+06	57.24	25.51	17.25	9	0.79	0.94	1.27	0.23	0.62	85.22	90.73	9.27	5.79
Atlantogenata		5.46e+06	55.74	29.87	14.39	5	0.98	1.11	0.73	0.14	0.43	89.02	94.93	5.07	2.53
HMRD		6.60e+06	48.28	35.98	15.74	4	0.99	1.13	1.44	0.13	0.39	89.15	95.97	4.02	1.94
Sparse Glires		7.08e+06	44.06	38.48	17.46	5	1.29	1.46	1.54	0.13	0.37	90.10	96.37	3.63	1.68
HQ Mammalia		8.82e+06	35.43	38.94	25.63	8	1.51	1.71	1.32	0.20	0.48	86.24	93.82	6.18	3.18
Glires		8.10e+06	33.79	42.05	24.16	7	1.82	2.03	1.61	0.15	0.38	89.33	96.02	3.98	1.75
Laurasiatheria		8.63e+06	32.29	39.01	28.70	10	2.04	2.27	1.43	0.20	0.46	86.44	94.08	5.92	2.95
Sparse Mammalia		7.95e+06	25.52	44.95	29.53	6	2.58	2.86	1.98	0.15	0.34	90.61	96.85	3.15	1.32
Eutheria		9.34e+06	11.42	44.38	44.20	30	5.87	6.23	3.31	0.20	0.39	86.93	95.25	4.75	1.92
Mammalia		9.42e+06	8.03	43.36	48.61	33	7.30	7.63	3.82	0.19	0.37	87.49	95.75	4.25	1.62

Table 1.5: Summary statistics and maximum likelihood ω estimates for sitewise data in eight species groups. Rows labeled "Primates (raw)" and "Mammalia (raw)" were computed based on unfiltered data and are included for reference. Columns under the "ω_{ML} Above / Below, %" heading measure the cumulative percentage of sites with ω_{ML} below or above the indicated value. Med.—median, Const.—constant, Syn.—synonymous, Nsyn.—non-synonymous, ML—maximum likelihood.

		Positively Selected Sites (%)								$P_{\chi^2_1} < 0.1$, %			$P_{\chi^2_1} < 0.05$, %		
Name	Filter	$P_{\chi^2_1} < 0.1$		$P_{\chi^2_1} < 0.05$		$P_{\chi^2_1} < 0.01$		FDR< 0.05		Neg.	Neut.	Pos.	Neg.	Neut.	Pos.
Primates	Default	82607	(0.95)	47825	(0.55)	14619	(0.17)	104	(0.001)	33.27	65.79	0.95	15.98	83.47	0.55
Atlantogenata		12665	(0.23)	6096	(0.11)	1242	(0.02)	0	(0.000)	47.14	52.63	0.23	24.44	75.45	0.11
HMRD		10295	(0.16)	4846	(0.07)	879	(0.01)	0	(0.000)	63.39	36.46	0.16	37.93	62.00	0.07
Sparse Glires		11095	(0.16)	5171	(0.07)	978	(0.01)	0	(0.000)	69.52	30.32	0.16	48.92	51.01	0.07
HQ Mammalia		52953	(0.60)	29773	(0.34)	8390	(0.10)	0	(0.000)	72.07	27.33	0.60	59.04	40.62	0.34
Glires		18190	(0.22)	9124	(0.11)	2025	(0.03)	0	(0.000)	76.78	23.00	0.22	65.65	34.24	0.11
Laurasiatheria		60588	(0.70)	36521	(0.42)	12175	(0.14)	69	(0.001)	74.53	24.77	0.70	64.76	34.81	0.42
Sparse Mammalia		13577	(0.17)	6775	(0.09)	1492	(0.02)	0	(0.000)	79.99	19.84	0.17	72.82	27.10	0.09
Eutheria		85819	(0.92)	58725	(0.63)	26405	(0.28)	2294	(0.025)	85.00	14.08	0.92	82.01	17.36	0.63
Mammalia		76370	(0.81)	52096	(0.55)	23323	(0.25)	1879	(0.020)	86.51	12.68	0.81	83.88	15.57	0.55

Table 1.6: Proportions of sites subject to positive, purifying and neutral selection at various LRT$_{SLR}$ thresholds. The Benjamini-Hochberg method [Benjamini & Hochberg, 1995] was used to identify the LRT$_{SLR}$ threshold at which FDR<0.05. For columns under the headings "$P_{\chi^2_1} < 0.1$, %" and "$P_{\chi^2_1} < 0.05$, %", Pos. and Neg. are the percentage of sites with significant evidence for positive and negative selection, respectively, and Neut. is the percentage of "neutral" sites not showing significant evidence for non-neutral selection.

1.4.3 The global distribution of sitewise selective pressures in mammals

Each set of sitewise data was filtered as described above. The resulting global distributions of site patterns, sitewise ω estimates, and 95% confidence intervals are shown in Figure 1.6; the left panel in each row plots the number of sites with constant, synonymous, and non-synonymous patterns. All sites with $\omega_{ML} = 0$ had constant or synonymous patterns, and all sites with $\omega_{ML} > 0$ had non-synonymous patterns; the distributions of ω_{ML} for these non-synonymous sites are shown as histograms on the right panel in each row.

1.4.3.1 Site patterns and ω_{ML} values reveal the prevalence of purifying selection in mammalian proteins

The site pattern counts in Figure 1.6 showed that the branch length of each species group had a strong effect on the overall composition of the sitewise data. Species groups covering little branch length, such as Primates and Atlantogenata, contained a majority of constant sites, while groups comprising lots of branch length, such as Eutheria and Mammalia, contained few constant sites and roughly equal proportions of synonymous and non-synonymous sites. Comparing the Glires and Mammalia data with their corresponding "sparse" datasets confirmed that this trend was largely due to branch length as opposed to biological factors: the Sparse Glires data yielded a smaller proportion of non-synonymous sites and a greater proportion of constant sites than the Glires data (17.41% versus 24.08% for non-synonymous sites, 44.21% versus 33.98% for constant sites; numbers from Table **??**), and the pattern for Sparse Mammalia and Mammalia was qualitatively the same.

Turning to the distribution of these ω_{ML} estimates, shown in Figure 1.6 as a series of histograms representing the ω_{ML} density (for nonzero values only) and a series of solid lines representing the cumulative density (for all values), it is clear that the vast majority of protein-coding sites have evolved under purifying selection in mammals. The Mammalia group, which contained a very small proportion of potentially uninformative constant sites (8.17%), showed a maximum density of nonzero ω_{ML} estimates at $\omega \approx 0.1$, and the vast majority of sites showed some evidence of purifying selection, with ω_{ML} estimates below 1. The height of the ω_{ML} cumulative distribution at $\omega = 1$ corresponds to the proportion of such sites; the exact value, included in Table **??** under the "< 1" column, is 95.74%. The nonzero ω_{ML} values were more evenly spread in the other species groups, with Glires showing a maximum nonzero ω_{ML} density at around $\omega \approx 0.25$ and Primates at $\omega \approx 0.7$.

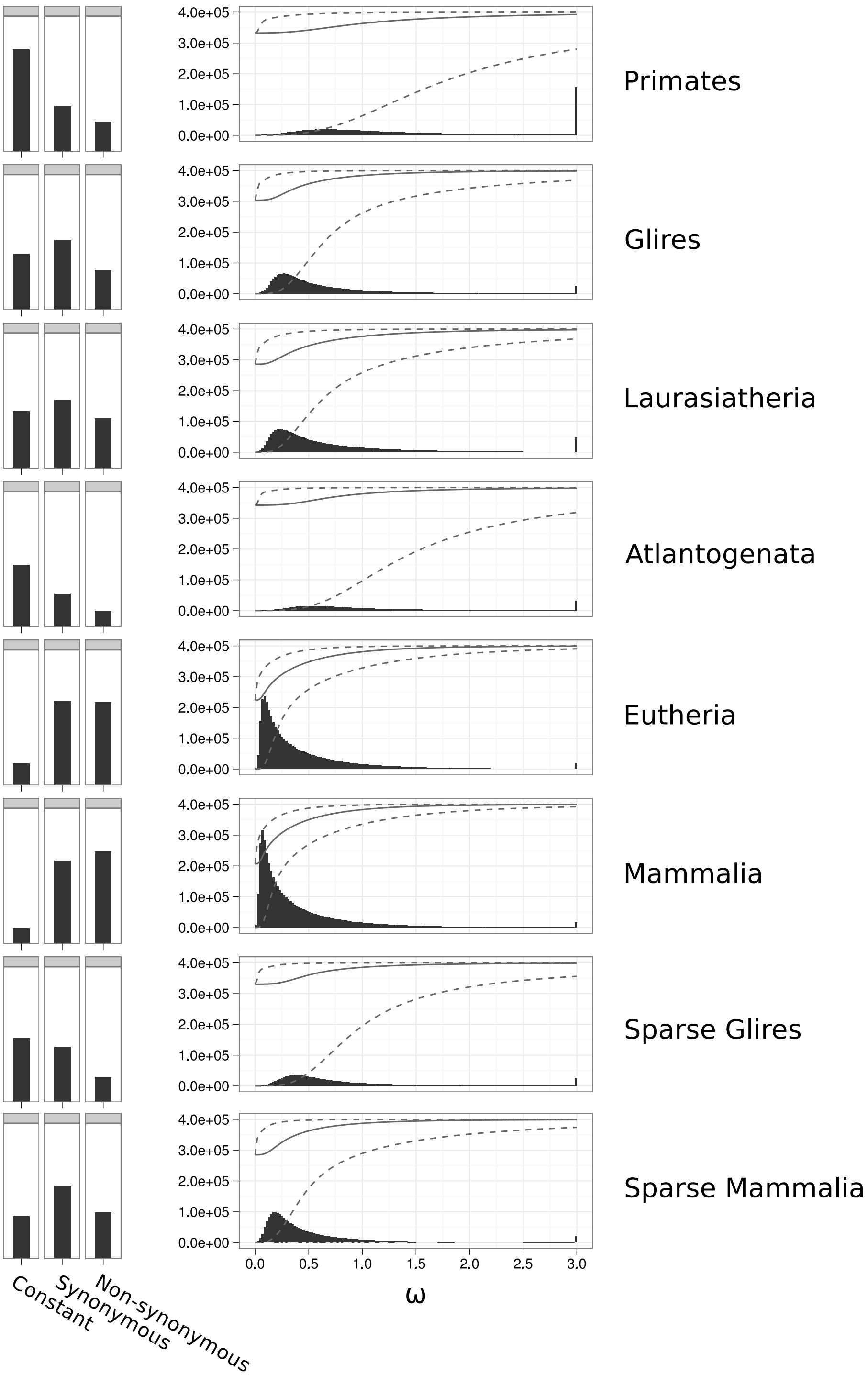

Figure 1.6: Global distributions of site patterns and ω estimates for eight species groups. Left, bars represent the number of sites showing constant, synonymous, and non-synonymous patterns. Note, the y-axis is held constant between rows. Right, bars represent a histogram of maximum likelihood ω estimates where $\omega > 0$. Sites where $\omega > 3$ are counted in the bin at $\omega = 3$. A solid line is drawn showing the cumulative proportion of sites with ω below the current value, and dashed lines are drawn above and below the solid line, showing the cumulative proportion of sites with the lower or upper range, respectively, of their 95% confidence interval below the current value.

This upwards shift in nonzero ω_{ML} estimates relative to Mammalia was likely due to the greater proportion of constant and synonymous sites in lower-branch length datasets: sites which were truly evolving with $\omega > 0$, but where no non-synonymous or synonymous substitutions were observed, would have their ω_{ML} estimate "pushed" towards zero, causing an increase in constant sites and a concomitant upwards shift in the distribution of the remaining nonzero ω_{ML} values.

1.4.3.2 Sitewise confidence intervals and LRT statistics identify sites with significant evidence for purifying and positive selection

An important component of SLR's output is the set of statistics providing information about the confidence with which purifying or positive selection was detected. These values include the lower and upper bounds of $\text{CI}_{95\%}$, the 95% confidence interval for each ω_{ML} estimate, and the LRT statistic, which corresponds to the strength of evidence for purifying or positive selection. Following Massingham [2005], I used a signed version of the LRT statistic (hereafter LRT_{SLR}), formed by negating the LRT statistic for sites where ω_{ML} < 1, as a way to sort sites according to their evidence, or lack thereof, for purifying and positive selection. Thus, sites with $\text{LRT}_{SLR} < 0$ showed at least some evidence for purifying selection and sites with $\text{LRT}_{SLR} > 0$ showed at least some evidence for positive selection. It should be noted that the LRT_{SLR} is a measure of the strength of evidence for purifying or positive selection, but not necessarily the actual strength of that selection. For example, an alignment covering a very large branch length might yield a strongly negative LRT_{SLR} for a site with ω_{ML} only moderately below 1, because the difference between dN and dS was highly statistically significant; on the other hand, a strongly-purifying site in an alignment covering less branch length might produce a much less-negative LRT_{SLR}, even with ω_{ML} near zero.

Figure 1.7A shows the empirical relationship between LRT_{SLR}, ω_{ML} and the $\text{CI}_{95\%}$ width for sites from the Mammali dataset. The left panel, comparing the LRT_{SLR} to nonzero ω_{ML} estimates, shows that the two values are highly correlated, with the greatest number of low ω_{ML} estimates occurring at sites with strongly negative LRT_{SLR}s. Correspondingly, the middle panel shows an even stronger relationship between the LRT_{SLR} magnitude and the $\text{CI}_{95\%}$ width, with the tighest windows at sites with very strong evidence for purifying selection. The rightmost panel compares the ω_{ML} of each site with the width of its $\text{CI}_{95\%}$, revealing a more linear and diffuse positive relationship between ω_{ML} and the size of the $\text{CI}_{95\%}$. The equivalent plots for Primates, shown in Figure 1.7B, reveal

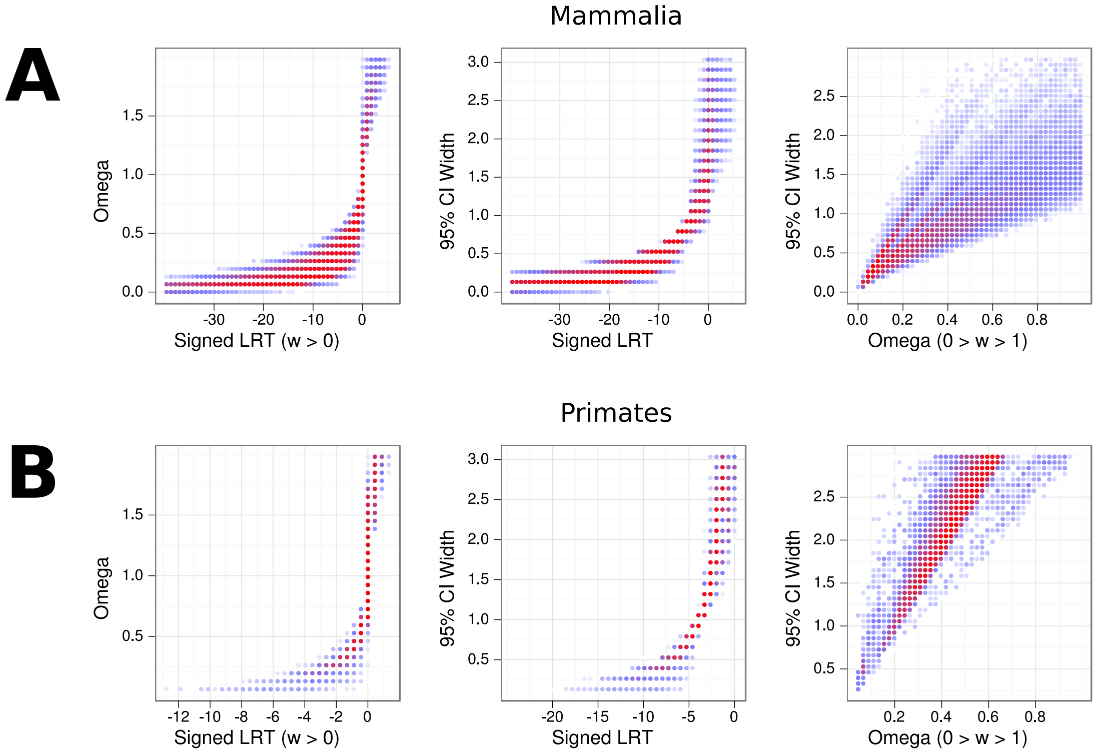

Figure 1.7: The relationship between LRT_{SLR}, ω_{ML}, and $\mathrm{CI}_{95\%}$ width in (a) Mammalia and (b) Primates datasets. Each point represents the binned density of sites; no points are drawn where no density exists, while blue and red points are drawn at areas of low and high density, respectively. The left panel shows sites where $\omega_{ML} < 0$, the middle panel shows all sites, and the right panel shows sites where $0 < \omega_{ML} < 1$. Note the change in x-axis scale between plots in (a) and (b), reflecting the paucity of sites in Primates with strong evidence (LRT_{SLR} <-12) for purifying selection.

similar patterns, but with more weight towards less-negative LRT_{SLR} values, higher ω_{ML}, and larger $\mathrm{CI}_{95\%}$. These differences highlight the impact of branch length on the amount of confidence with which ω can be estimated, showing that the low branch length of the Primates clade rarely yields ω_{ML} estimates with $\mathrm{CI}_{95\%}$ intervals smaller than 1, while the bulk of sites from the Mammalia dataset have a relatively small $\mathrm{CI}_{95\%}$. Thus, the ω_{ML} distribution from datasets with low branch lengths (e.g., Figure 1.6) should be interpreted with caution, and any comparison between different sites or datasets must be sensitive to the amount of statistical confidence placed on each estimate.

The statistical information at each site could be used to identify sites evolving under purifying or positive selection with statistical confidence. Sites with a CI_{upper}, the upper bound of the $\mathrm{CI}_{95\%}$ interval, below $\omega = 1$ showed evidence of purifying selection with an expected 5% FPR, and sites with a CI_{lower} above $\omega = 1$ showed evidence of positive selection with an expected 5% FPR. In both cases, the 5% FPR was expected under SLR's null model of neutral evolution. There was a strong relationship between CI_{upper} and the χ^2_1 approximation to the LRT_{SLR} distribution, whereby the set of sites with $\mathrm{CI}_{upper} < 1$ was exactly equivalent to the set of sites with LRT_{SLR} below the negative χ^2_1 95% critical value. Similarly, the sites with $\mathrm{CI}_{lower} > 1$ were those with LRT_{SLR} above the χ^2_1 95% critical value. Because of this equality, I will refer to LRT_{SLR} values instead of $\mathrm{CI}_{95\%}$ intervals when discussing sites with significant evidence for purifying or positive selection. In some cases, however, use of the full $\mathrm{CI}_{95\%}$ will be preferable, as the LRT_{SLR} critical values only correspond to one end of the $\mathrm{CI}_{95\%}$, depending on whether the site shows greater evidence for purifying or positive selection.

Table 1.6 summarizes the results from using the χ^2_1 approximation to the LRT_{SLR} distribution to identify sites subject to purifying selection and positive selection at various FPR thresholds. The left group of columns show the number and proportion of sites with evidence for positive selection at nominal 10%, 5%, and 1% FPR thresholds, respectively, as well as an expected 5% FDR calculated using the Benjamini Hochberg method for FDR control [Benjamini & Hochberg, 1995]. The two groups of columns on the right show the result of breaking sites into three groups (positive, negative, and neutral) based on the result of a χ^2_1 test at a given FPR threshold.

The positive selection results demonstrated that between 0.2% to 1% of sites could be identified as under positive selection in mammals at nominal FPR thresholds, but different species groups yielded strikingly different estimates of the proportion of positively-selected sites. At a 5% FPR threshold, Primates, Laurasiatheria, Eutheria, and Mammalia produced roughly comparable proportions of positively-selected sites, ranging from 0.43% to 0.72%. The proportions of positively-selected sites in these groups were higher using a 10% FPR threshold (ranging from 0.70% to 0.97% of sites) and lower using a 1% FPR threshold (ranging from 0.14% to 0.28%). When the FDR was controlled using the Benjamini-Hochberg method, however, only the Eutheria and Mammalia groups yielded a substantial number of positively-selected sites; the Primates and Laurasiatheria data were likely limited in their power to yield positively-selected sites after FDR control due to their lower total branch lengths. Interestingly, the Glires, Atlantogenata, Sparse Glires and Sparse Mammalia datasets produced much lower proportions of positively-selected sites across all FPR thresholds. At FDR<0.05, all four groups yielded zero significant PSCs.

In Mammalia, the breakdown of sites into positive, negative and neutral categories at both FPR thresholds produced a pattern similar to the ω_{ML} distribution, with overwhelming purifying constraint (83.87% of sites at 5% FPR), a small proportion of neutrally-evolving sites (15.57%), and a small fraction of positively-selected sites (0.55%). As expected given the use of a fixed LRT_{SLR} threshold to identify purifying sites, the fraction of sites confidently identified as under purifying selection showed a strong dependency on the branch length of the species set, with a much higher power in Mammalia than in Primates (83.87% vs. 15.97%).

Even for the Mammalia dataset, which encapsulated roughly 7.5 synonymous substitutions per site on average, one might reasonably expect that the power to confidently identify sites under purifying selection, though higher than in Primates, was still less than 100%. If this is the case, then the proportion of confidently identified purifying sites must be an underestimate of the true proportion of negatively-selected sites (and by symmetry, absent any methodological bias, the same should be true for positively-selected sites). As a result, the fractions of positively- and negatively-selected sites in Table 1.6 should not be taken as best estimates of the actual proportions of such sites, but more appropriately as lower bounds on those proportions. In fact, In the next two sub-sections, I will separately consider the issue of using the sets of sitewise estimates in different species groups to estimate the proportion of sites subject to purifying and positive selection in mammals.

1.4.3.3 Estimating the proportion of negatively-selected sites

For sites under purifying selection,

In , the value of ≈95% based on $\omega_{ML} < 1$ (Table **??**) is likely closer to the true number; despite the caveats involved in ignoring the uncertainty involved in ω_{ML} estimates, the 95% number was surprisingly consistent across different species groups and branch lengths, ranging from ≈91% in Primates to ≈97% in Sparse Mammalia.

1.4.3.4 Estimating the proportion of positively-selected sites

The pattern of the prevalence of positive selection across species groups was surprising. First,

, showing no sign of the expected correlation with branch length. Theory predicts, and many studies have confirmed [Anisimova *et al.*, 2001, 2002; Massingham & Goldman, 2005], that the power of LRT-based tests for non-neutral selection should increase with branch length, as the discrimination between non-synonymous and synonymous substitution rates becomes more confident with more fixed substitutions. This was certainly the case for identifying purifying selection, but there was no obvious correlation between higher branch lengths and higher numbers of confidently identified PSCs.

1.4.3.5 Correlations between branch length, effective population size and sitewise summary statistics

To quantify this observation, Spearman's rank correlation coefficients between the median branch length of each species set and each of several summary statistics were calculated (Table 1.7). Although the number of samples was small with only eight species groups, these correlations should be able to provide some indication as to which aspects of the sitewise data might be easily attributed to the effects of branch length and which aspects suggested an alternative (e.g., biological or artefactual) cause for the differences between species groups. The results were quite striking: the site classifications (constant, synonymous and non-synonymous) and the proportion of negatively-selected sites at $P_{\chi^2_1} < 0.05$ were strongly correlated with branch length, while the other factors, including mean ω_{ML}, the proportion of sites with $\omega_{ML} < 1$ and $\omega_{ML} < 0.5$, and the proportion of positively-selected sites, were weakly correlated or largely uncorrelated with branch length.

The results in Table 1.7 can be interpreted in a number of ways. First, they emphasized the unambiguous correlation between the branch length in a tree and the power to detect sitewise purifying selection. However, they also showed that various measures based on sitewise estimates contained variation between species groups that was not well-explained by branch length.

1.4.4 Modeling the global distribution of sitewise selective pressures

	Branch Length		N$_E$	
Value	Rho	P-value	Rho	P-value
Median BL	-	-	0.37	0.36
Population Size	0.37	0.36	-	-
Mean ω_{ML}	0.14	0.75	-0.54	0.17
Constant	-1.00	0.00	-0.37	0.36
Synonymous	0.88	0.01	0.48	0.23
Nsynonoymous	0.98	0.00	0.35	0.40
ω_{ML} <1	0.38	0.36	0.87	0.01
ω_{ML} <0.5	0.12	0.79	0.73	0.04
PSC$_{5\%}$	0.07	0.88	-0.70	0.05
NSC$_{5\%}$	0.98	0.00	0.48	0.23

Table 1.7: Correlations between median non-gap branch length, N$_E$, and various summary statistics of the sitewise data across eight species sets. The magnitude (Rho) and significance (P-value) of Spearman's rank correlations between variables were calculated using N$_E$ values from Table 1.1 and branch lengths and summary statistics from Tables **??** and 1.6. All summary statistics except for mean ω_{ML} were measured as a fraction of total sites. More highly significant correlations are shaded in darker blue. N$_E$ – estimated effective population size, BL – non-gap branch length, PSC$_{5\%}$– positively-selected codons at $P_{\chi^2_1} < 0.95$.

		Log-normal		Gamma		Exponential		Beta		Weibull	
Species Set	Data Type	$\bar{\omega}$	$\% > 1$	$\bar{\omega}$	$\% > 1$	$\bar{\omega}$	$\% > 1$	$\bar{\omega}$	$\% > 1$	$\bar{\omega}$	$\% > 1$
Primates	ω_{ML}	0.11	1.53	0.25	7.31	0.25	1.82	0.25	0.00	0.14	2.96
Glires		0.15	2.07	0.16	3.35	0.16	0.16	0.20	0.00	0.12	2.56
Laurasiatheria		0.33	3.52	0.20	5.34	0.20	0.74	0.23	0.00	0.19	4.08
Atlantogenata		0.04	0.59	0.15	3.33	0.15	0.12	0.20	0.00	0.07	1.30
Eutheria		0.72	6.10	0.20	4.62	0.20	0.64	0.24	0.00	0.22	5.18
Mammalia		0.77	6.57	0.19	4.26	0.19	0.57	0.23	0.00	0.22	5.16
Sparse Glires		0.06	0.87	0.13	2.64	0.13	0.06	0.18	0.00	0.08	1.46
Sparse Mammalia		0.20	2.64	0.15	2.85	0.15	0.11	0.19	0.00	0.13	2.74
Primates	$\mathrm{CI}_{95\%}$	0.40	4.09	0.42	4.56	0.37	6.83	0.42	0.00	0.41	5.55
Glires		0.21	0.14	0.21	0.17	0.18	0.44	0.21	0.00	0.20	0.27
Laurasiatheria		0.22	0.95	0.23	0.45	0.22	1.05	0.23	0.00	0.23	0.64
Atlantogenata		0.30	0.58	0.32	0.55	0.26	2.14	0.32	0.00	0.30	1.26
Eutheria		0.19	2.28	0.18	0.79	0.18	0.42	0.18	0.00	0.18	1.32
Mammalia		0.18	2.35	0.18	0.74	0.17	0.31	0.17	0.00	0.17	1.27
Sparse Glires		0.23	0.14	0.24	0.24	0.20	0.67	0.25	0.00	0.23	0.43
Sparse Mammalia		0.16	0.10	0.17	0.09	0.15	0.13	0.17	0.00	0.16	0.11

Table 1.8: Mean ω and the percentage of sites with $\omega > 1$ based on maximum-likelihood fits of parametric distributions to sitewise estimates. For each species set and each one of five distribution types (log-normal, gamma, exponential, beta, and weibull) 100 replicate datsets of 1 million sites were sampled with replacement from the genome-wide dataset and the maximum likelihood distribution parameters were numerically optimized (see text for details). Columns show, for each distribution, the median value across 100 replicates of mean ω ($\bar{\omega}$) and the probability mass with $\omega > 1$, expressed as a percentage ($\%> 1$). The top eight rows show the results based on fitting parameters to sitewise ω_{ML} estimates, and the bottom eight rows show the results based on fitting parameters to sitewise $\mathrm{CI}_{95\%}$ estimates. Note the greater consistency of $\bar{\omega}$ and $\%> 1$ across distribution types for the $\mathrm{CI}_{95\%}$-based fits.

We used the fitdistr function of the MASS package for R to fit five distributions (gamma, lognormal, beta, Weibull, and exponential) to the vertebrate dN/dS values and subsequently calculated Akaikes Information Criterion(AIC) for each fit. For all optimizations, a constant value of 0.001 was added to sites where dN/dS = 0 in order to satisfy the optimizers requirement that the probability functions have a defined value for all input data. Similarly, sites with dN/dS 1 were excluded from the analysis for the beta optimization. All distributions were also separately fit to the subset of sites with dN/dS 1; the AIC values from these optimizations were used to compare the fit of the beta distribution to the others.

The fitdistr produced the following optimized parameters for each function: gamma (shape=0.271, rate=1.203), lognormal (meanlog=-4.079, sdlog=2.863), beta (shape1=0.257, shape2=1.431), Weibull (shape=0.3882, scale=0.07151) exponential (rate=4.441). The beta distribution yielded the lowest AIC when compared to the fit of other distributions to the subset of sites where dN/dS 1 (-2.33e7 versus the next best equivalent AIC of -2.11e7 for the lognormal). Of the distributions which were fit to the whole dataset, the lognormal distribution yielded the lowest AIC (-2.11e7), followed by gamma (-2.03e7) and exponential (-6.45e6).

1.4.5 Simulations to evaluate the power to detect positive selection and estimate selective pressures

1.4.6 Evaluation of the effect of GC content, recombination rate, and codon usage on sitewise dN/dS estimates and the detection of positive selection

Species Set	Distribution	AIC	dAIC	Parameter A	Parameter B
Primates	Beta	161168.45	0.00	1.78	2.50
	Lognormal	217743.70	56575.25	-1.14	0.65
	Gamma	231082.15	13338.45	2.12	5.06
	Weibull	242149.95	11067.80	1.34	0.45
	Exponential	261904.40	19754.45	2.68	
Glires	Lognormal	312788.30	0.00	-1.76	0.59
	Beta	332776.25	19987.95	1.57	5.75
	Gamma	341872.95	9096.70	1.70	8.02
	Weibull	364256.80	22383.85	1.15	0.21
	Exponential	385083.15	20826.35	5.42	
Laurasiatheria	Lognormal	514030.30	0.00	-1.80	0.77
	Beta	528667.45	14637.15	1.25	4.27
	Gamma	558537.00	29869.55	1.49	6.52
	Weibull	570535.20	11998.20	1.12	0.23
	Exponential	575734.00	5198.80	4.56	
Atlantogenata	Beta	109348.25	0.00	2.37	4.93
	Lognormal	119888.05	10539.80	-1.34	0.53
	Gamma	126416.00	6527.95	2.75	8.72
	Weibull	143678.60	17262.60	1.33	0.33
	Exponential	175474.60	31796.00	3.84	
Eutheria	Lognormal	1180885.50	0.00	-2.40	1.20
	Weibull	1270446.50	89561.00	0.80	0.16
	Gamma	1295309.50	24863.00	0.80	4.35
	Beta	1307444.50	12135.00	0.69	3.16
	Exponential	1308964.50	1520.00	5.46	
Mammalia	Lognormal	1310778.00	0.00	-2.51	1.26
	Weibull	1403090.50	92312.50	0.78	0.15
	Gamma	1433878.50	30788.00	0.75	4.30
	Beta	1456833.00	22954.50	0.65	3.11
	Exponential	1459667.50	2834.50	5.77	
Sparse Glires	Beta	161165.45	0.00	2.04	6.12
	Lognormal	161264.00	98.55	-1.60	0.53
	Gamma	176568.30	15304.30	1.99	8.23
	Weibull	198910.90	22342.60	1.20	0.24
	Exponential	225609.40	26698.50	5.00	
Sparse Mammalia	Lognormal	403243.50	0.00	-2.09	0.68
	Gamma	447431.20	44187.70	1.31	7.74
	Beta	450936.70	3505.50	1.18	5.70
	Weibull	462813.80	11877.10	1.06	0.16
	Exponential	471324.15	8510.35	6.67	

Table 1.9: AIC values and parameters for maximum-likelihood fits of parametric distributions to sitewise $\text{CI}_{95\%}$ estimates. Distributions were fit to 100 replicate datasets for each species set as in Table 1.8. For each species set and distribution type, the median Akaike information criterion (AIC) and parameter estimates are shown. Distributions are sorted according to their median AIC value (where a lower AIC corresponds to a better fit to the data), and the difference in AIC to the next-best fitting distribution is displayed (dAIC). The lognormal and beta distributions are ranked first or second in the mammalian superorder subgroups, while lognormal, weibull and gamma distributions fit the Eutheria and Mammalia datasets best. The named parameters corresponding to parameters A and B for each distribution are as follows: lognormal (A=meanlog, B=sdlog), weibull (A=shape, B=scale), gamma (A=shape, B=rate [where rate=1/scale]), beta (A=shape1 [α], B=shape2 [β]), exponential (A=rate [λ]).

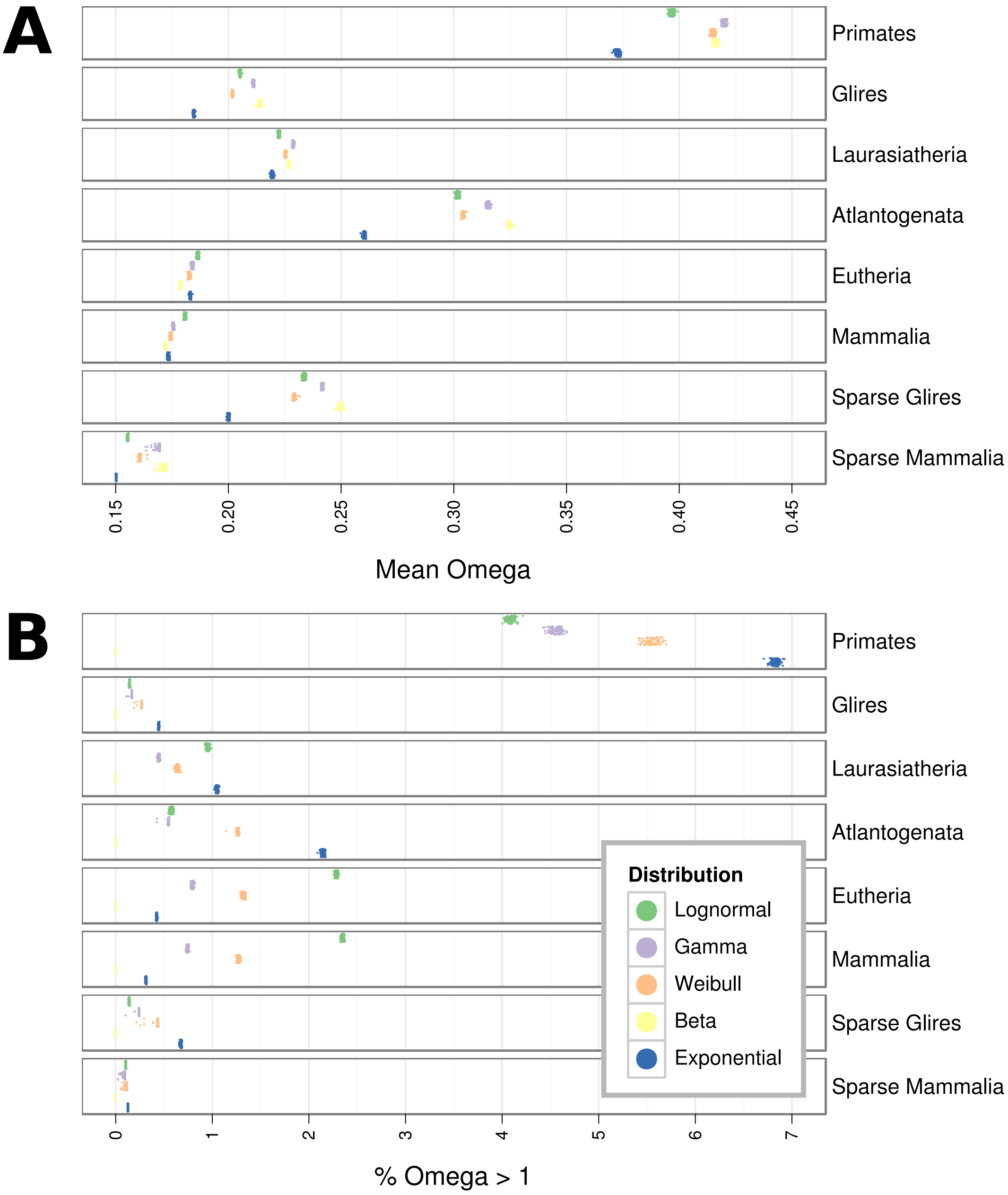

Figure 1.8: The mean value (A, Mean Omega) and the percentage of probability density with $\omega > 1$ (B, Omega> 1) from maximum likelihood fits of parametric distributions to sitewise $\mathrm{CI}_{95\%}$ estimates. Each point represents the maximum likelihood fit of one distribution to 1 million points sampled from the genome-wide dataset with replacement; 100 such fits were generated for each distribution and species set. Note that the beta distribution is only defined on the interval (0,1), so the percentage of sites with $\omega > 1$ was always zero.

	Species			Med.		ω_{ML}		ω_{ML}	Recomb., cM			Substitutions %				
Variable	Group	Quantile	Sites	BL	$NSC_{10\%}$	< 0.5	$PSC_{10\%}$	> 1.5	Male	Female	GC	Total	CpG	W-S	S-W	Nsyn
gc	Mammalia	(0,0.05]	315984	6.176	88.131	89.000	0.715	1.343	0.645	1.242	0.336	0.387	0.010	0.198	0.180	0.347
		(0.05,0.275]	1408253	6.583	86.932	87.675	0.823	1.508	0.762	1.477	0.371	0.415	0.013	0.206	0.197	0.360
		(0.275,0.5]	1389575	7.127	86.139	86.870	0.873	1.628	0.898	1.709	0.413	0.450	0.019	0.206	0.224	0.355
		(0.5,0.725]	1379927	7.530	87.514	88.229	0.747	1.466	1.012	1.815	0.469	0.463	0.029	0.181	0.253	0.340
		(0.725,0.95]	1397262	7.599	89.893	90.715	0.489	1.077	1.144	1.822	0.543	0.431	0.037	0.137	0.257	0.320
		(0.95,1]	312449	8.205	91.843	92.698	0.253	0.699	1.720	1.223	0.622	0.430	0.043	0.126	0.260	0.312
decodeM	Mammalia	(0.05,0.275]	1701616	6.923	86.696	87.790	0.800	1.578	0.076	1.143	0.450	0.424	0.023	0.170	0.231	0.351
		(0.275,0.5]	1397371	7.061	87.674	88.451	0.743	1.405	0.448	1.467	0.441	0.425	0.022	0.178	0.225	0.344
		(0.5,0.725]	1393633	7.038	88.395	89.098	0.672	1.300	0.866	1.767	0.441	0.425	0.023	0.180	0.221	0.334
		(0.725,0.95]	1372925	7.667	88.553	89.025	0.643	1.265	1.874	2.271	0.462	0.462	0.028	0.193	0.242	0.337
		(0.95,1]	304151	8.359	89.671	90.113	0.505	1.049	4.886	2.130	0.512	0.495	0.036	0.194	0.265	0.342
decodeF	Mammalia	(0,0.05]	304114	7.693	88.210	88.987	0.515	1.152	0.853	0.001	0.500	0.463	0.032	0.174	0.257	0.350
		(0.05,0.275]	1394890	6.946	87.068	88.057	0.762	1.495	0.799	0.466	0.444	0.428	0.023	0.175	0.230	0.349
		(0.275,0.5]	1393061	6.965	87.831	88.632	0.742	1.409	0.742	1.116	0.435	0.424	0.022	0.181	0.221	0.339
		(0.5,0.725]	1386168	7.306	87.545	88.278	0.771	1.472	0.893	1.833	0.456	0.443	0.025	0.182	0.236	0.347
		(0.725,0.95]	1382191	7.397	88.800	89.395	0.611	1.210	1.324	2.892	0.459	0.442	0.026	0.184	0.231	0.330
		(0.95,1]	309272	7.619	88.451	88.896	0.671	1.302	1.764	4.780	0.465	0.455	0.028	0.189	0.239	0.343

Table 1.10

	Species		Sites	Med.	$P_{\chi^2_1} < 0.1$	ω_{ML}	$P_{\chi^2_1} < 0.1$	ω_{ML}	Recombination							
Variable	Group	Quantile	Sites	BL	Neg.	< 0.5	Pos.	> 1.5	Male	Female	GC	WS	CpG	Nonsyn.	Syn.	
gc	Primates	(0,0.05]	292632	0.656	25.449	86.235	0.849	5.405	0.645	1.234	0.336	0.420	0.010	0.216	0.194	0.357
		(0.05,0.275]	1305896	0.688	24.330	85.190	0.923	5.824	0.771	1.488	0.372	0.439	0.013	0.217	0.208	0.366
		(0.275,0.5]	1289390	0.740	26.741	84.804	0.984	5.994	0.912	1.728	0.414	0.469	0.020	0.215	0.234	0.358
		(0.5,0.725]	1283434	0.797	34.863	85.763	0.882	5.444	1.033	1.815	0.470	0.483	0.030	0.189	0.264	0.342
		(0.725,0.95]	1296483	0.823	43.732	88.127	0.655	4.398	1.156	1.825	0.544	0.453	0.038	0.145	0.270	0.323
		(0.95,1]	288995	0.937	54.446	89.725	0.452	3.418	1.738	1.203	0.622	0.452	0.045	0.134	0.273	0.313
decodeM	Primates	(0.05,0.275]	1575779	0.731	29.090	85.704	0.923	5.801	0.077	1.146	0.450	0.446	0.024	0.179	0.243	0.356
		(0.275,0.5]	1299507	0.740	30.437	86.009	0.859	5.480	0.452	1.481	0.442	0.449	0.023	0.188	0.237	0.350
		(0.5,0.725]	1289984	0.737	32.063	86.581	0.810	5.155	0.876	1.769	0.441	0.447	0.024	0.191	0.232	0.338
		(0.725,0.95]	1278924	0.823	38.578	86.317	0.796	4.948	1.900	2.276	0.464	0.482	0.029	0.201	0.252	0.339
		(0.95,1]	282104	0.981	48.091	87.212	0.629	4.176	4.951	2.133	0.512	0.512	0.037	0.200	0.275	0.342
decodeF	Primates	(0,0.05]	282128	0.859	40.030	86.187	0.719	4.931	0.872	0.001	0.500	0.484	0.033	0.183	0.268	0.354
		(0.05,0.275]	1294422	0.746	29.687	85.912	0.880	5.630	0.818	0.466	0.446	0.450	0.024	0.184	0.243	0.353
		(0.275,0.5]	1293324	0.730	30.631	86.244	0.849	5.386	0.772	1.123	0.436	0.447	0.023	0.191	0.233	0.345
		(0.5,0.725]	1284474	0.774	33.634	85.823	0.895	5.438	0.886	1.840	0.456	0.464	0.026	0.191	0.247	0.351
		(0.725,0.95]	1285566	0.789	36.130	86.716	0.769	4.957	1.337	2.903	0.460	0.462	0.027	0.193	0.242	0.334
		(0.95,1]	286384	0.800	37.262	86.325	0.802	4.950	1.769	4.799	0.463	0.475	0.029	0.198	0.247	0.346

Table 1.11

Multiple lines of evidence have lent support to the hypothesis that GC-biased gene conversion (BGC) has been a major force in the evolution of mammalian genomes [Galtier et al 2001 PMID:11693127, Galtier 2003 PMID:XYZ, Dreszer et al 2007 PMID:17785536]. Both empirical and theoretical results have shown that BGC can significantly affect patterns of observed substitutions in both selectively neutral and functionally constrained sites [Galtier et al 2009 PMID:19027980, Berglund et al 2009 PMID:19175294]. Recently, Ratnakumar et al. [2010, PMID:20643747] re-analyzed the dataset of positively-selected genes from Kosiol et al. [2008, PMID:18670650] for signatures of BGC and found that up to 20% of cases of identified elevated dN/dS ratios could be due to BGC rather than adaptive evolution. However, the strongest signals of BGC were found only in genes showing signals of positive selection along short branches in the phylogenetic tree using so-called branch-site models of evolution; when the authors looked for similar BGC signatures in genes with evidence for positive selection at specific sites throughout the mammalian tree (e.g., genes with significant LRTs for PAMLs sites model) they found no evidence for a strong BGC influence [Ratnakumar et al. 2010 PMID:20643747].

The above evidence suggests that although BGC has the potential to produce misleading signals of branch-specific positive selection near recombination hotspots, the positively-selected sites we detected should not be strongly influenced by the non-adaptive effects of BGC since the dN/dS level detected by SLR is estimated from across the entire input phylogeny [Massingham 2005 PMID:15654091]. This is consistent with the observation that recombination hotspots (where most recombination in humans and other mammals occurs [Myers et al. 2005 PMID:16224025]) tend not to be maintained over long evolutionary periods, although larger-scale recombination rates are likely more conserved [Winckler et al 2005 PMID:15705809]. Still, due to the potential confounding implications of BGC on the interpretation of signals of positive selection, we found it worthwhile to to empirically test for any BGC effect on our data.

The BGC model predicts a recombination-associated drive towards the fixation of GC alleles at heterozygous sites, resulting in an expected correlation between AT to GC (or weak-to-strong, W-S) mutational bias and recombination rate [Galtier and Duret 2007, PMID:17418442]. This bias can lead to elevated dN/dS estimates in coding regions, particularly in GC-rich regions where W-S mutations are more likely to result in nonsynonymous changes [Berglund et al 2009]. Ratnakumar and colleagues identified three ways of distinguishing potential BGC effects from true signals of positive selection in protein-coding regions: (a) positive selection is not expected on its own to result in a strong W-S bias, (b) a BGC-associated W-S biased mutation pattern should extend to noncoding sites flanking the affected coding region, and (c) BGC is associated with recombination hotspots and regions of high recombination rates (and most strongly with male-specific rates) while there is no empirical evidence linking positive selection with higher recombination rates in mammals, although natural selection should theoretically be more efficient in regions of high recombination [Ratnakumar et al. 2010, PMID:20643747]. We could not use (a) or (b) to detect possible BGC influence since we did not calculate inferred ancestral mutations for either the coding or flanking noncoding regions of the mammalian gene families studied here. Instead, we turned to point (c) and tested for a correlation between signals of positive selection and an increase in recombination rates, especially the male-specific rate and in regions of high GC content. The predictions of the BGC hypothesis suggest that if our sitewise data do contain a strong BGC influence, then the positively-selected sites we detected would be expected to be associated with regions of high male-specific recombination.

We combined the sitewise codon data with male, female, and sex-averaged recombination rates derived from the deCODE map (using rates averaged over genomic bins of 1Mb downloaded from the UCSC human genome browser hg19 release) and human GC content calcualted in 10-kb windows and analyzed sites within various quantiles of GC content, mean recombination rate, and sitewise statistics. Supplementary Table S17.13 contains summaries for each subset. The LRT statistic section shows that sites with higher LRT statistics (which corresponds to weaker purifying selection when the value is below zero and stronger positive selection when the value is above zero) show decreasing recombination rates; this trend holds true even for the highest quantile (mean signed_lrt between 3.648 and 108.850), which is composed entirely of sites with evidence for positive selection. In other words, the bulk of positively-selected sites are in regions of lower than average male recombination rates – exactly opposite what would be expected in the face of strong BGC effects. The Male Recombination quantiles show a similar trend, with the mean dN/dS, mean signed LRT and the proportion of sites identified as positively-selected (pos_f) all consistently decreasing as the recombination rate increases. The GC content quantiles showed a slightly different pattern. Although the mean LRT decreased and male recombination increased monotonically with increasing GC content, the mean dN/dS and fraction of positive sites started low, increased to a maximum in the middle range of GC content, and decreased again in regions of high GC content. Thus, although the GC content quantiles were similar to the male recombination quantiles in their higher range (with similar mean dN/dS, mean LRT, and pos_f values), they differed slightly in their lower range (with lower dN/dS and pos_f for low GC quantiles). Although the exact reason for such a pattern is unclear, it is consistent with the existence of altered or constrained selective or mutational dynamics at the extreme ends of the genomic distribution of GC content. As GC content has been shown to correlate with myriad structural and evolutionary features of mammalian genomes [Xia et al. 2009 PMID:19521505], the existence of other (possibly unrelated) confounding influences such as CpG mutability or isochore structure is likely.

Theoretical and empirical evidence pointed towards an increased sensitivity of dN/dS estimates to BGC influence in regions of high GC content, so we separated out the top 10% of sites by GC content and analyzed them according to quantiles of male recombination rate (Supplementary Table S17.13, High GC, Male Recombination. The middle four recombination quantiles showed a similar pattern to that observed for all GC contents, with mean LRT decreasing with increased male recombination and mean dN/dS and pos_f decreasing or hovering around values slightly lower than those observed across all GC contents (e.g., mean dN/dS in the 1-25% bin is 0.207 for the top 10% GC sites, but 0.249 for the same recombination bin across all sites). The highest recombination bin of the top 10% GC sites showed a strikingly different pattern, however, with mean dN/dS=0.348, mean LRT=-11.262, and pos_f=0.0338. These values suggest a strong shift towards higher dN/dS values and more positively-selected sites. This jump in values in the highest recombination bin is not seen in the highest male recombination bin across all GC contents (mean dN/dS=0.207, mean LRT=-16.616, pos_f=0.00932) or for the highest female recombination bin for the top 10% of GC sites (mean dN/dS=0.164, mean LRT=-18.475, pos_f=0.00449). Although the small number of sites in the bin of interest compared to other bins suggests possible stochastic artifacts, the shift is dramatic, directly opposite to the trends observed for the female recombination rates and for male recombination rates in regions of lower GC content, and is in agreement with the BGC prediction of elevated dN/dS estimates in regions of high GC content and male-specific recombination rates. This evidence raises the interesting possibility that BGC may have a detectable, if rather minor, impact on sitewise dN/dS estimates across the mammalian phylogeny. It is highly unlikely, however, that any such effect – which in our analysis was only detectable in 0.05% of sites with the most extreme GC content and recombination rates – has contaminated our codon-specific estimates with more than a negligible amount of noise resulting from the neutral but biased process of BGC.

1.5 Conclusions

[To cite: **?** — Showed, through constraint analysis of various sequence types, that there is higher selective constraint in 4-fold sites in priamtes compared to murids. Quote: "It is well established that in several organisms, mutations at 4-fold sites are selected against (Chamary et al. 2006; Rocha 2006; Drummond and Wilke 2008) and as a consequence the dN/ds ratio, which has been frequently used to detect the strength and direction of selection (e.g., Dorus et al. 2004; Wang et al. 2006), may be underestimated. Our result of higher 4-fold constraint in hominids suggests that this bias more strongly affects hominid estimates and it may well exceed 20%."]

[To cite: **??** — The Keightley et al. 2011 paper (ABC to estimate mutation rate parameters) cited Ohta 1993, 1995 and Eory et al. 2010 for the effective population size and efficacy of selection in primates vs. murids]

[To cite: **?** — Wolf et al. GBE 2009, used pairwise dN dS counts to try to show that trends in dN/dS ratios are a result of branch length, at least when calculated in a pairwise fashion. Slightly unconvincing stuff... could be cited as somehow relating to the discussion regarding eff. pop. size, branch length, and selection]

[To cite: **?** — Berglund et al. 2009 looked at hotspots of biased substitutions in humans. Showed that exons with accelerated rates in humans have a tendency towards clusters of AT-to-GC (weak-to-strong) substitutions. Did some simulations showing that this effect is strongest in GC-poor regions, though the impact on overall dN/dS is probably minimal (e.g., genes with overall high dN/dS didn't show BGC, only the most accelerated exons did) and the effect on dN/dS is highest in high-GC regions. The most-accelerated exons tend to reside in high-male (but not female) recombination, and ¡50kb from hotspots. Upshot: these biased clusters seem to show up in isolated regions (exons), rather than spread throughout entire genes. Probably not a huge impact on overall apparent constraint.]

[To cite: **?** — Duret and Arndt 2008 use nonreversible nucleotide models to estimate NEUTRAL rates correlated with recombination, GC, and GC*. Lots of stuff here, but the important bits: overall mutation rate increases with increasing GC content (due to overall higher rates of S-W substitution); recombination should have a strong impact on W-S substitution, but weak impact on S-W substitutions; CpG deamination varies by factor of two, very low in GC-poor regions and very high in GC-rich ones.]

[To cite: **?** — Galtier et al. TRIG 2009 is similar to Berglund et al. in many ways – find accelerated exons in a primate branch, and identify significantly higher male recombination rates there. The number of accelerated exons is small – 100 in each of four branches – and not all of these accelerated exons showed strongly elevated dN/dS ratios. Only 19 exons at the 1% level. However, they do some nice modeling (mostly in the supp. material) which shows that the effect of BCG on dN/dS ratio at different GC contents – it has more effect in GC-rich genes.]

[To cite: **?** — Capra and Pollard quantified BDS (biased divergent substitutions) across metazoans, additionally using recombination rate data. Dog has the strongest, mouse has the weakest BDS scores. (This could be due to lower rec. rate in mouse, e.g. Coop and Przeworski 2006)]

[To cite: **?** — Nordborg et al. 1996 (Genet. Res.) modeled the effect of background selection on variation in neutral linked loci. They showed that weakly selected mutations, rather than strongly selected ones, are more likely to produce regional patterning of variation in response to local recombination rate. Should have a large effect in Drosophila but small effect in mammals, though in mammals "local reductions in regions of reduced recombination might be detectable."]

[To cite: **?** — Chun and Fay 2011 (PLoS Gen) looked at neutral and deleterious SNP density according to local recombination rate, showing that in 'hitchhiking' regions there are fewer neutral, but as many deleterious, polymorphisms. That stuff is boring, but they also show that the deleterious SNP density stays constant throughout the range of recombination rates, while the neutral and synonymous SNP density decreases. Thus, slightly deleterious mutations are less effectively purged in regions of low recombination.]

[To cite: **?** — Bullaughey et al. (2008, Gen. Res.) looked at gene-wide dN/dS ratios in primates and recombination rates. They found no significant correlation between broad- or fine-scale recomb. rates and rates of protein evolution, **once GC content is taken into account**.]

[To cite: **?** — Spencer et al. (2006 PLoS Gen) Quote: "In short, while there is a strong relationship between recombination and GC content, most of the relationship is explained by scales broader than recombination hotspots (16 to 256 kb; unpublished data) and may well result from interactions of both factors with additional processes such as chromatin organisation or replication timing. Similar arguments apply to the question of whether a GC bias in recombination-associated mutation can explain the relationship between GC content and recombination."]

[...]

Bibliography

Anisimova, M., Bielawski, J. & Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*, **18**, 1585–92. 39

Anisimova, M., Bielawski, J. & Yang, Z. (2002). Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol*, **19**, 950–8. 5, 39

Anisimova, M., Nielsen, R. & Yang, Z. (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, **164**, 1229–36. 5

Archibald, J.D. (1999). Divergence times of eutherian mammals. *Science*, **285**, 2031. 23

Averof, M., Rokas, A., Wolfe, K. & Sharp, P. (2000). Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science*, **287**, 1283–6. 27

Bakewell, M., Shi, P. & Zhang, J. (2007). More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci U S A*, **104**, 7489–94. 9

Bazykin, G., Kondrashov, F., Ogurtsov, A., Sunyaev, S. & Kondrashov, A. (2004). Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature*, **429**, 558–62. 18

Beisswanger, S. & Stephan, W. (2008). Evidence that strong positive selection drives neofunctionalization in the tandemly duplicated polyhomeotic genes in drosophila. *Proc Natl Acad Sci U S A*, **105**, 5447–52. 9

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300. 32, 33, 38

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L. & Purvis, A. (2007). The delayed rise of present-day mammals. *Nature*, **446**, 507–512. 23

Callahan, B., Neher, R., Bachtrog, D., Andolfatto, P. & Shraiman, B. (2011). Correlated evolution of nearby residues in drosophilid proteins. *PLoS Genet*, **7**, e1001315. 18

Casola, C. & Hahn, M. (2009). Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. *J Mol Evol*, **68**, 679–87. 9

Churakov, G., Kriegs, J., Baertsch, R., Zemann, A., Brosius, J. & Schmitz, J. (2009). Mosaic retroposon insertion patterns in placental mammals. *Genome Res*, **19**, 868–75. 22

Clark, A., Glanowski, S., Nielsen, R., Thomas, P., Kejariwal, A., Todd, M., Tanenbaum, D., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T., Sninsky, J., Adams, M. & Cargill, M. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–3. 13

Cock, P., Fields, C., Goto, N., Heuer, M. & Rice, P. (2010). The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*, **38**, 1767–71. 10

TOCITE (2011). Citation will be inserted at a later point in time. 3, 12, 13

ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816. 4, 11

Fletcher, W. & Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol*, **27**, 2257–67. 9

Green, P. (2007). 2x genomes–does depth matter? *Genome Res*, **17**, 1547–9. 5

HAN, M., DEMUTH, J., MCGRATH, C., CASOLA, C. & HAHN, M. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome Res*, **19**, 859–67. 14

HUBBARD, T., AKEN, B., BEAL, K., BALLESTER, B., CACCAMO, M., CHEN, Y., CLARKE, L., COATES, G., CUNNINGHAM, F., CUTTS, T., DOWN, T., DYER, S., FITZGERALD, S., FERNANDEZ-BANET, J., GRAF, S., HAIDER, S., HAMMOND, M., HERRERO, J., HOLLAND, R., HOWE, K., HOWE, K., JOHNSON, N., KAHARI, A., KEEFE, D., KOKOCINSKI, F., KULESHA, E., LAWSON, D., LONGDEN, I., MELSOPP, C., MEGY, K., MEIDL, P., OUVERDIN, B., PARKER, A., PRLIC, A., RICE, S., RIOS, D., SCHUSTER, M., SEALY, I., SEVERIN, J., SLATER, G., SMEDLEY, D., SPUDICH, G., TREVANION, S., VILELLA, A., VOGEL, J., WHITE, S., WOOD, M., COX, T., CURWEN, V., DURBIN, R., FERNANDEZ-SUAREZ, X., FLICEK, P., KASPRZYK, A., PROCTOR, G., SEARLE, S., SMITH, J., URETA-VIDAL, A. & BIRNEY, E. (2007). Ensembl 2007. *Nucleic Acids Res*, **35**, D610–7. 10

HUBISZ, M., LIN, M., KELLIS, M. & SIEPEL, A. (2011). Error and error mitigation in low-coverage genome assemblies. *PLoS One*, **6**, e17034. 10, 11

JAFFE, D., BUTLER, J., GNERRE, S., MAUCELI, E., LINDBLAD-TOH, K., MESIROV, J., ZODY, M. & LANDER, E. (2003). Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res*, **13**, 91–6. 6, 10

KIRCHER, M., STENZEL, U. & KELSO, J. (2009). Improved base calling for the illumina genome analyzer using machine learning strategies. *Genome Biol*, **10**, R83. 6

KOSIOL, C., HOLMES, I. & GOLDMAN, N. (2007). An empirical codon model for protein sequence evolution. *Mol Biol Evol*, **24**, 1464–79. 27, 28

KOSIOL, C., VINAR, T., DA FONSECA, R., HUBISZ, M., BUSTAMANTE, C., NIELSEN, R. & SIEPEL, A. (2008). Patterns of positive selection in six mammalian genomes. *PLoS Genet*, **4**, e1000144. 13

LINDBLAD-TOH, K., WADE, C.M., MIKKELSEN, T.S., KARLSSON, E.K., JAFFE, D.B., KAMAL, M., CLAMP, M., CHANG, J.L., KULBOKAS, E.J., ZODY, M.C., MAUCELI, E., XIE, X., BREEN, M., WAYNE, R.K., OSTRANDER, E.A., PONTING, C.P., GALIBERT, F., SMITH, D.R., DEJONG, P.J., KIRKNESS, E., ALVAREZ, P., BIAGI, T., BROCKMAN, W., BUTLER, J., CHIN, C.W., COOK, A., CUFF, J., DALY, M.J., DECAPRIO, D., GNERRE, S., GRABHERR, M., KELLIS, M., KLEBER, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.P., Parker, H.G., Pollinger, J.P., Searle, S.M.J., Sutter, N.B., Thomas, R., Webber, C., Baldwin, J., Broad Sequencing Platform Members & Lander, E.S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803–819. 3

Lindblad-Toh, K., Garber, M., Zuk, O. & ,et al. (64 co-authors) (2011). A high-resolution map of evolutionary constraint in the human genome based on 29 eutherian mammals (in press). *Nature*, **0**, 0. 4

Lynch, M. & Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–5. 12

Mallick, S., Gnerre, S., Muller, P. & Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res*, **19**, 922–33. 9

Margulies, E., Vinson, J., NISC Comparative Sequencing Program, Miller, W., Jaffe, D., Lindblad-Toh, K., Chang, J., Green, E., Lander, E., Mullikin, J. & Clamp, M. (2005). An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci U S A*, **102**, 4795–800. 3

Margulies, E., Cooper, G., Asimenos, G., Thomas, D., Dewey, C., Siepel, A., Birney, E., Keefe, D., Schwartz, A., Hou, M., Taylor, J., Nikolaev, S., Montoya-Burgos, J., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Brown, J., Bickel, P., Holmes, I., Mullikin, J., Ureta-Vidal, A., Paten, B., Stone, E., Rosenbloom, K., Kent, W., Bouffard, G., Guan, X., Hansen, N., Idol, J., Maduro, V., Maskeri, B., McDowell, J., Park, M., Thomas, P., Young, A., Blakesley, R., Muzny, D., Sodergren, E., Wheeler, D., Worley, K., Jiang, H., Weinstock, G., Gibbs, R., Graves, T., Fulton, R., Mardis, E., Wilson, R., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D., Chang, J., Lindblad-Toh, K., Lander, E., Hinrichs, A., Trumbower, H., Clawson, H., Zweig, A., Kuhn, R., Barber, G., Harte, R., Karolchik, D., Field, M., Moore, R., Matthewson, C., Schein, J., Marra, M., Antonarakis, S., Batzoglou, S., Goldman, N., Hardison, R., Haussler, D., Miller, W., Pachter, L., Green, E. & Sidow, A. (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res*, **17**, 760–74. 3, 4

Massingham, T. & Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–62. 4, 5, 27, 29, 36, 39

Mouse Genome Sequencing Consortium & Mouse Genome Analysis Group (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–62. 3

Murphy, W., Pringle, T., Crider, T., Springer, M. & Miller, W. (2007). Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res*, **17**, 413–21. 22

Nielsen, R., Bustamante, C., Clark, A., Glanowski, S., Sackton, T., Hubisz, M., Fledel-Alon, A., Tanenbaum, D., Civello, D., White, T., J Sninsky, J., Adams, M. & Cargill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*, **3**, e170. 13

Rat Genome Sequencing Project Consortium (2004). Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521. 3

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L. & Webster, M. (2010). Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci*, **365**, 2571–80. 9

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. & Graur, D. (2009). Estimates of positive darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol*, **1**, 114–8. 8, 9

Storz, J., Hoffmann, F., Opazo, J. & Moriyama, H. (2008). Adaptive functional divergence among triplicated alpha-globin genes in rodents. *Genetics*, **178**, 1623–38. 9

Studer, R., Penel, S., Duret, L. & Robinson-Rechavi, M. (2008). Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res*, **18**, 1393–402. 9

Teytelman, L., Ozaydin, B., Zill, O., Lefrançois, P., Snyder, M., Rine, J. & Eisen, M. (2009). Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One*, **4**, e6700. 11

Wang, Y. & Gu, X. (2001). Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics*, **158**, 1311–20. 9

Whelan, S. & Goldman, N. (2004). Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*, **167**, 2027–43. 27

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**, 1586–91. 19