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 several groups of mammalian genomes, while Chapter **??** will look at the use of these sitewise data to identify trends in the evolution of genes and protein domains.

The first section of this chapter introduces the context in which this project was performed—namely, the sequencing and analysis of several mammalian genomes for the Mammalian Genome Project (MGP)—and outlines the main biological questions underpinning the sitewise 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], assembling genomic fragments [TOCITE], and annotating gene-coding regions [TOCITE] 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 section looks at the global distribution of mammalian selective constraint in three ways: first, using sitewise estimates from SLR 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. The consistency of these results with previous work will be explored.

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 MGP, 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 MGP 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 MGP, 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 MGP (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 MGP analysis.

In parallel with the major goal of the MGP to understand the amount of evolutionary constraint across the entire mammalian genome, the main purpose of my analysis was to better understand the distribution of evolutionary constraint within mammalian protein-coding regions—in other words, to understand what proportion of protein-coding material has been evolving under purifying, neutral, or positive selection. Proteins are well understood to evolve under strong purifying constraint due to their functional importance [Fay & Wu, 2003], but some regions of proteins, such as disordered regions between two well-folded domains, may evolve under relaxed constraints, and positive selection of beneficial substitutions can also play a role in shaping the evolutionary history of proteins [Pál *et al.*,

2006].

The first goal was a rather simplistic one: to place lower and upper boundaries on the estimated proportion of protein-coding sites that are subject to purifying and positive selection throughout mammals, and to quantify how many of those sites can be confidently identified. There has been great interest in understanding the role of adaptive evolution in shaping the genes and genomes of mammals and primates, but different studies have produced widely varying estimates of the number of genes subject to positive selection [Marques-Bonet *et al.*, 2009; **?**]. While most previous investigations of positive selection in mammals have focused on the gene as the unit of analysis, the current analysis used a primarily sitewise approach. It was hypothesized that the focus on identifying positively selected codons (PSCs) instead of the traditional identification of positively selected genes (PSGs), may allow for a more flexible quantification of levels of purifying and positive selection in mammals. One expected benefit of the sitewise approach was that fine-grained filtering on a site-by-site basis could be performed both before and after the computationally expensive step of estimating evolutionary parameters, allowing for a more extensive and flexible set of filters to be used in estimating the potential impact of alignment or annotation error on the amounts of inferred positive selection.

The second, more subtle goal of this analysis was to place the distribution of selective pressures implied by the sitewise analysis within the context of previous population genetic and comparative studies. Many population genetic studies have analyzed the distribution of selective pressures resulting from mutations in protein-coding regions, known as the distribution of fitness effects (DFE). Analyses of variation data from *Drosophila* have found that relatively few amino acid substitutions in *Drosophila* are effectively neutral, while up to 50% have apparently been due to positive selection [Eyre-Walker & Keightley, 2007; Loewe & Charlesworth, 2006]; similar studies based on variations in humans have indicated a much lower fraction of positively-selected substitutions in our recent evolutionary history [Boyko *et al.*, 2008; Eyre-Walker *et al.*, 2006]. This is in line with the expectation, based on population genetic theory, that species with higher effective population sizes experience more effective natural selection [Eyre-Walker & Keightley, 2007]. As *Drosophila* has historically had a much larger effective population size than humans and most mammals (e.g., on the order of 10^6 for *Drosophila* vs. 10^3 for humans), one would expect to see more neutral evolution, and less purifying and positive selection, in human protein-coding regions.

Although there is a strong theoretical connection between the DFE commonly from

population genetics and the dN/dS ratios more commonly estimated in comparative analyses, only one study, performed by Nielsen and Yang [2003], has explicitly estimated the DFE using data from fixed differences between species. Using data from primate mitochondrial genomes, Nielsen and Yang found that a variety of two-parameter distributions for the DFE fit the dataset equally well and that none of the best-fit distributions contained a large amount of probability mass within the range of purely neutral or beneficial selection coefficients; most of the distribution was contained within the range of moderately deleterious selection coefficients (e.g., $-3 < S < -1$, corresponding roughly to dN/dS values between 0.2 and 0.6). Unfortunately, no attempt has since been made to use comparative data to estimate the DFE; as a result, one goal of this analysis was to determine whether sitewise estimates could successfully be used to infer the DFE. Though the methods I employed for this analysis differed strongly from the approach of Nielsen and Yang, a comparison to their results could validate the use of SLR for estimating the DFE. Furthermore, it would be interesting to understand whether the differences in historical effective population sizes between mammalian subgroups, which have been shown repeatedly to affect overall dN/dS levels in primates versus rodents [Ellegren, 2009; Kosiol *et al.*, 2008], has a detectable impact on the DFE inferred from comparative data. Although the effective population size differences between mammalian subgroups are far smaller than the difference between mammals and species like *Drosophila*, a comparison of the DFE from different mammalian groups could be used to evaluate how strong of an impact the effective population size has on the proportion of protein-coding sites subject to varying levels of natural selection.

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 PRANK_{C} alignments introduce few falsely identified positively-selected sites resulting from alignment errors at mammalian-like divergence levels. Thus, PRANK_{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 es-

timated 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 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, 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 se-

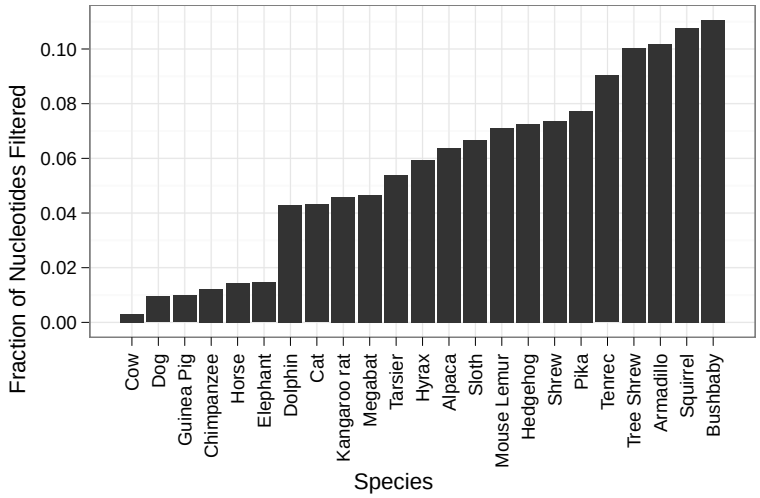

Figure 1.1

quences 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 [Casola & Hahn, 2009], and produce biases due to lineage-specific family expansion, a process which is relatively common in mammalian gene families [Gu *et al.*, 2002]. 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-specificity of estimated 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] used 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 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 discriminate 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 between each putative paralog and all other sequences in the gene tree, resulting in one mean pairwise distance estimate per putative paralog (hereafter referred to as the mean distance). For these distance calculations, the stock Compara codon

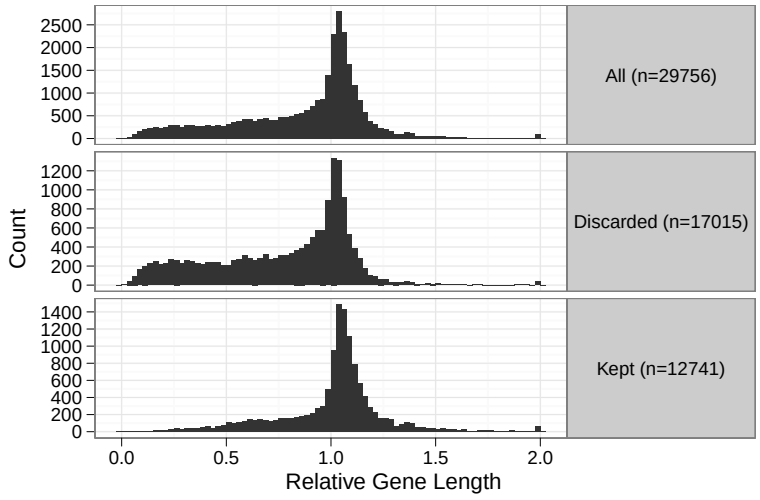

Figure 1.2: Length ratios of putative paralogs. The length ratio was calculated as the length of a putative paralogous copy divided by the mean length all sequences its corresponding gene tree. Putatively paralogous genes (top panel) were either discarded (middle panel) or kept (bottom panel) according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

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.

Genes were grouped by species within each gene tree, and any group of 2 or more genes was considered to be a set of putative paralogs .Within each set of putative paralogs, a single gene was chosen to be retained for evolutionary analysis based on three rules applied in the following order: (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 mean distance below the others, that sequence was kept; (3) if all mean distances were identical then the longest sequence was kept, or if all mean distances and length ratios were equal, an arbitrary choice was made.

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 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

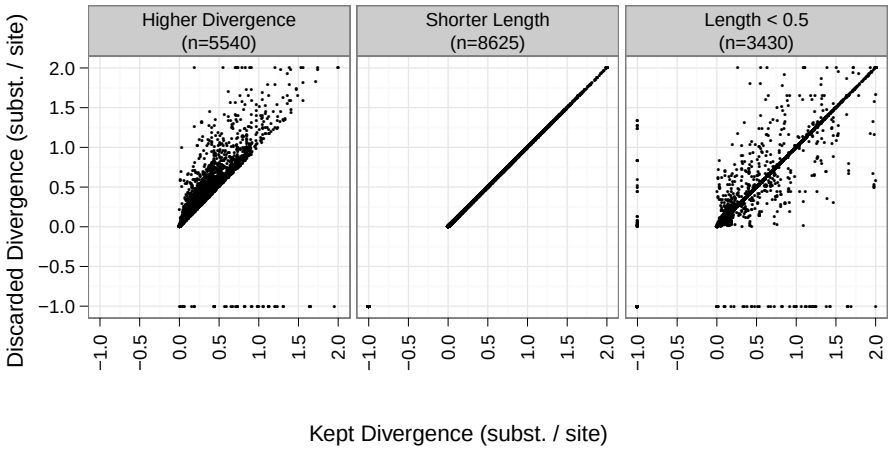

Figure 1.3: Sequence divergence of kept and discarded putative paralogs. Each point represents a gene which was discarded from the tree for one of three reasons: it had more sequence divergence than the kept gene (*Higher Divergence*; left panel), it had equal sequence divergence but shorter length than 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)

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.

To better compare the characteristics of the discarded and kept genes, a dmore detailed view of the results of the paralog filter is presented in Figure 1.3, showing a scatter plot of the mean distance and length ratio of each discarded paralog compared to that of the corresponding 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 mean distances were removed; the third panel corresponds to rule (3), where all genes had equal mean distances and the longest gene was kept (or, if all lengths were equal, an arbitrary choice was made).

The first panel of Figure 1.3 shows that genes discarded on the basis of having a very short length contained sequence distances similar to the kept copies, as the highest density is along the diagonal and there is no apparent bias for genes to lie above or below the

diagonal. This is in line with the expectationq that these discarded genes were not truly paralogous copies, but rather fragments of split genes resulting from unassembled sequence segments. The second panel shows that when paralogous copies could be differentiated by their mean distances, 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 with a slope of 2). Finally, the distribution of length ratios and mean distances in the set of genes where length was the discriminating factor (or where an arbitrary decision was made) shows that most of these genes were mostly identical whether measured by sequence distance or sequence length.

These results provided evidence that a sizeable fraction of recently duplicated mammalian 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 detectable 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 mean distance 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. Nonetheless, this protocol was designed to identify the least-diverged copy of a recently duplicated gene, and for 40% of putative paralogs the mean distance to other sequences in the gene family allowed a sensible decision to be made.

This was obviously not the most conservative approach to dealing with recent duplications—one could remove all copies from a set of putative paralogs, creating an apparent gene deletion, or one could simply ignore all gene families with any recent duplications (e.g., require one-to-one orthology allowing for gene deletions). The latter option is almost certainly too conservative, but the former option may be appropriate for a more conservative approach. As the main concern over the handling duplicated genes has been that they may introduce a bias towards elevated evolutionary rates, I marked the XYZ genes containing at least XYZ sets of putative paralogs for further evaluation. Sitewise estimates from these genes were excluded from the most conservatively-filtered sitewise dataset and examined separately for excess signal of positive selection (see Section 1.4.2), and in the next chapter I examine whether using the more conservative approach of removing all paralogous copies from genes removed the signal of positive selection from a subset of genes (see 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 many 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. 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 particuarly 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 occurring along the branch 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 could confound the effectiveness of using clustered substitutions to identify regions of non-homologous aligned sequence. First, the length of the terminal branch leading to each species determines how many lineage-specific substitutions would be expected to occur within a window of a certain size. The terminal human branch, for example, is very short (as it shares a very 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 alignment window. In contrast, a stretch of non-homologous aligned sequence should introduce, on average, a constant number of non-synonymous and synonymous substitutions into the branch ancestral to the sequence in which it exists. The end result is that it should be more difficult to distinguish homologous from non-homologous stretches in species with long terminal lineages, as species

with long terminal lineages will have higher numbers of substitutions in truly homologous regions. On the other hand, this trend should also serve to limit the negative impact of non-homologous stretches in those species on the detection of positive selection, because the resulting elevation in non-synonymous or synonymous substitutions rates would be less severe.

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 [Yang *et al.*, 1995] and to identify 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 analyzed, and the location of each substitution event 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 codon-based PRANK alignments of mammalian gene trees.

Using this set of inferred substitutions, counts of synonymous and non-synonymous substitutions within non-overlapping 15-codon alignment windows for all terminal and internal nodes were collected; the results for a selection of species and internal nodes 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*.

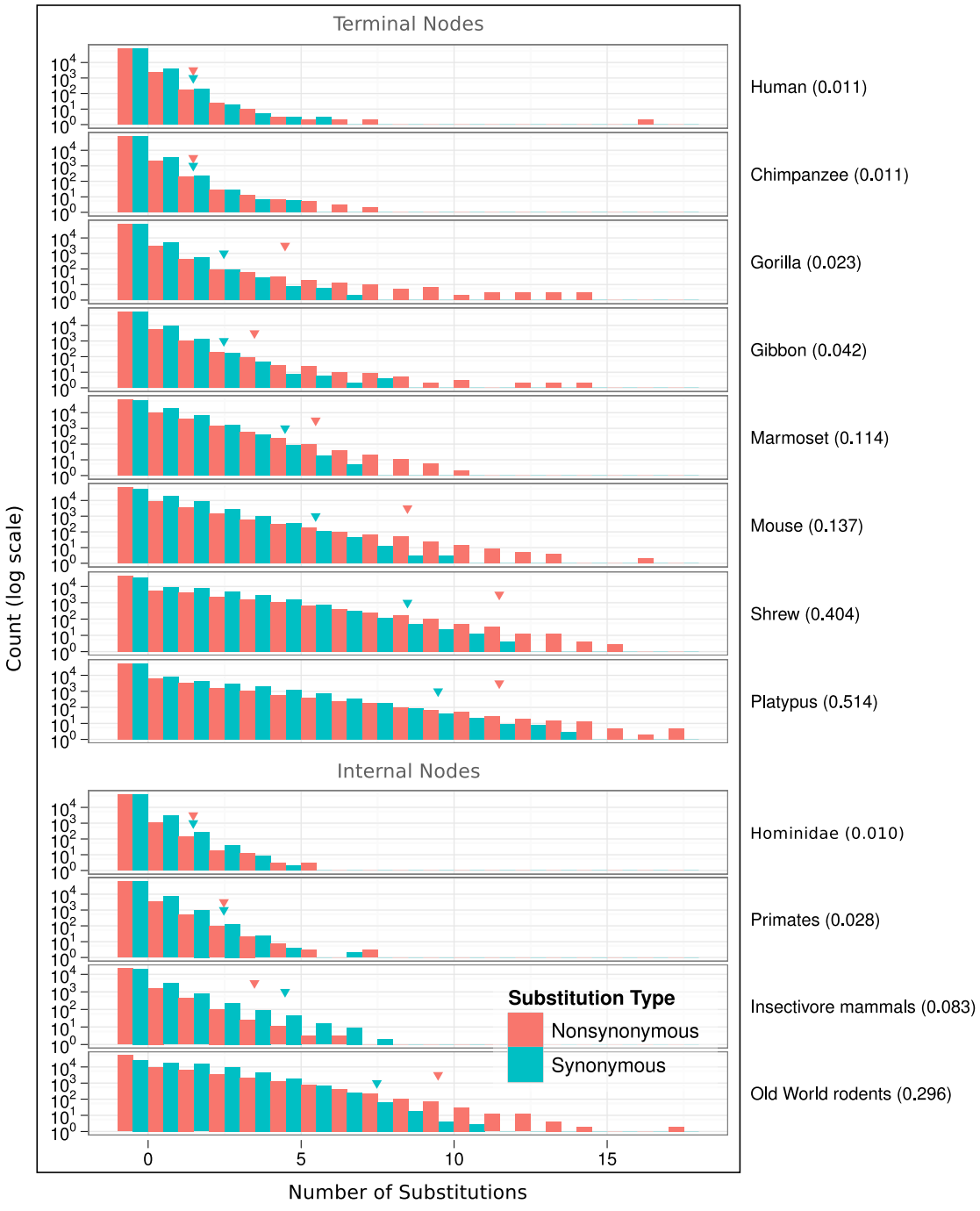

Figure 1.4: Counts of inferred non-synonymous (red bars) and synonymous (blue bars) substitutions in 15-codon windows along terminal and internal branches of the mammalian tree. The leftmost two bars correspond to windows with 0 substitutions, the next two bars correspond to windows with 1 substitution, 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.

Figure 1.4 shows that the vast majority of 15-codon windows in these alignments contained few substitutions (note that the y-axis uses a logarithmic scale), 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 than synonymous counts at 0 substitutions, lower than synonymous counts 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 compared to synonymous substitutions. The excess of windows with large numbers of non-synonymous substitutions, on the other hand, runs against the pattern of purifying selection; instead, it shows unexpectedly long clusters of non-synonymous substitutions to be a widespread feature of these 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 emphasizes the excess of highly clustered non-synonymous substitutions. Interestingly, human—which has the highest quality and best annotated genome—does not show the same level of excess seen in the other genomes analyzed.

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 possibility that the entire rodent clade contains many misaligned non-homologous stretches due to differences in gene annotations between rodent and non-rodent species.

The end result of this analysis was the identification, for each terminal node of the mammalian tree, windows with non-synonymous substitution counts above the top 0.1%

of 15-codon windows genome-wide; these windows were considered potential stretches of non-homologous aligned sequence. Despite evidence that some internal nodes might also suffer from this type of alignment artifact, most internal nodes were free from an obvious excess of clustered non-synonymous substitutions, so internal nodes were excluded from this list. And although there is no way to show that the 0.1% threshold is the most appropriate one for discriminating true from erroneous windows of clustered substitutions, manual analysis of regions containing windows at a variety of thresholds showed it to perform well.

In total, 30,XYZ [numers are approximate, will fill in later] windows containing potential stretches of non-homologous aligned sequence were identified across 3,XYZ genes, with XYZ genes containing at least 1 such window and XYZ genes containing greater than 10. The locations of these windows were stored for later use in defining the most conservatively-filtered sitewise dataset, and the impact of these potentially non-homologous windows on sitewise levels of positive selection is described in Section **??**.

1.4 Genome-wide analysis of sitewise selective pressures in mammals

1.4.1 Species groups for sitewise analysis

For each alignment of mammalin orthologs, SLR was run separately on 10 different sets of mammalian species to obtain sitewise estimates in a variety of species groups. For each species group, sequences corresponding to species within the group were extracted from the whole mammalian alignment (along with the corresponding subtree) and input to SLR, which was run with its default parameters. If fewer than two sequences were available for a given gene and species group, the sitewise analysis was skipped for that group. The species included in each group are listed in Table 1.1 alongside the MPL and total branch length of their subtrees estimated as the median value across all 16xyz gene-wise dS branch length estimates from SLR.

Three species groups (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

	Species		Median dS	
Name	Count	List	MPL	Total
Primates	10	Bushbaby, Chimpanzee, Gibbon, Gorilla, Human, Macaque, Marmoset, Mouse Lemur, Orangutan, Tarsier	0.16	0.83
Atlantogenata	5	Armadillo, Elephant, Hyrax, Sloth, Tenrec	0.26	0.97
HMRD	4	Dog, Human, Mouse, Rat	0.34	1.01
Sparse Glires	5	Guinea Pig, Kangaroo rat, Mouse, Rat, Squirrel	0.36	1.32
HQ Mammals	9	Chimpanzee, Cow, Dog, Horse, Human, Macaque, Mouse, Pig, Rat	0.31	1.61
Glires	7	Guinea Pig, Kangaroo rat, Mouse, Pika, Rabbit, Rat, Squirrel	0.40	1.90
Laurasiatheria	12	Alpaca, Cat, Cow, Dog, Dolphin, Hedgehog, Horse, Megabat, Microbat, Panda, Pig, Shrew	0.26	2.16
Sparse Mammals	7	Armadillo, Dog, Elephant, Human, Mouse, Platypus, Wallaby	0.61	2.86
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	0.35	6.43
Mammals	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	0.67	8.21

Table 1.1: Species groups used for sitewise analysis by SLR. The median MPLs and the median total branch length are shown for each species group, taken from the 15,XYZ branch lengths estimated by SLR for each gene. MPL – mean path length.

mammalian subclade, Atlantogenata, consisting of sloth, armadillo, tenrec, elephant and hyrax, was also included, but the monophyly of this group is still under debate [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. The different mammalian superorders contained a wide range of total branch lengths, with 0.83 for Primates, 0.97 for Atlantogenata, 1.90 for Glires, and 2.16 for Laurasiatheria. A slightly different ordering was found when measuring the trees by MPL, with Glires having a significantly higher MPL (0.40) than the other groups despite having fewer species and a lower total branch length than Laurasiatheria. This reflected the higher neutral evolutionary rate in the Glires group, a well-documented feature of rodent evolution likely resulting from their long-term shorter generation time, which has been strongly correlated with higher neutral evolutionary rates [Nikolaev *et al.*, 2007; Smith & Donoghue, 2008].

Two larger species groups, Eutheria and Mammalia, were chosen for the purpose of measuring average sitewise selective pressures across mammals as a whole. The Eutheria group consists of the union of the mammalian superorder groups plus armadillo, and the Mammalian group adds opossum, platypus, and wallaby for a total of 38 species. The median total branch lengths for Mammalia and Eutheria were 8.21 and 6.43, respectively,

and the MPLs were 0.67 and 0.35.

Finally, to evaluate the impact of species choice and branch length on the results of the sitewise analysis, four additional "sparse" species groups were created for comparison to the main groups of interest. The species in the Sparse Glires group were chosen to create a group with species from the Glires group but having a lower overall branch length; the Sparse Mammals group was created with a similar aim, created 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 major branches within the species tree. The HQ Mammals group was similar to the Sparse Mammals group, but elephant and the deeper mammalian lineages were omitted (e.g., wallaby, platypus, armadillo) in favor of only the high-coverage Eutherian genomes (e.g., chimpanzee, cow, horse, macaque, pig, rat). Finally, the HMRD group consisted of human, mouse, rat, dog, and represented the type of phylogenetic tree that was commonly analyzed early in the last decade when only a few mammalian genome sequences were available. The HMRD group was comparable to Primates and Atlantogenata in total branch length, while HQ Mammals and Sparse Glires were more similar to Glires.

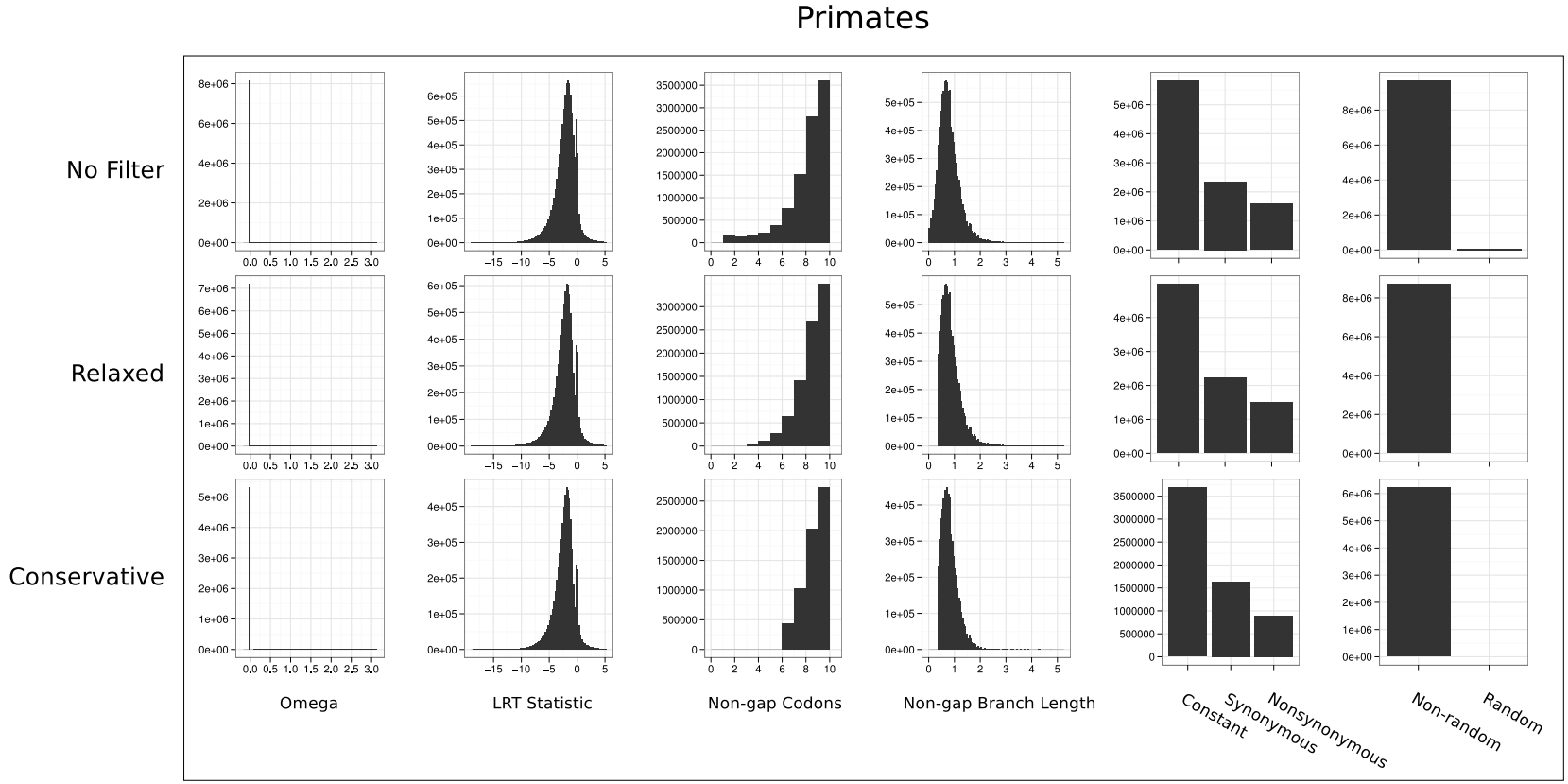

Figure 1.5: Distributions of sitewise values for the Primates species group, showing the raw data (top row) and the result of applying the relaxed (middle row) and conservative (bottom row) filters.

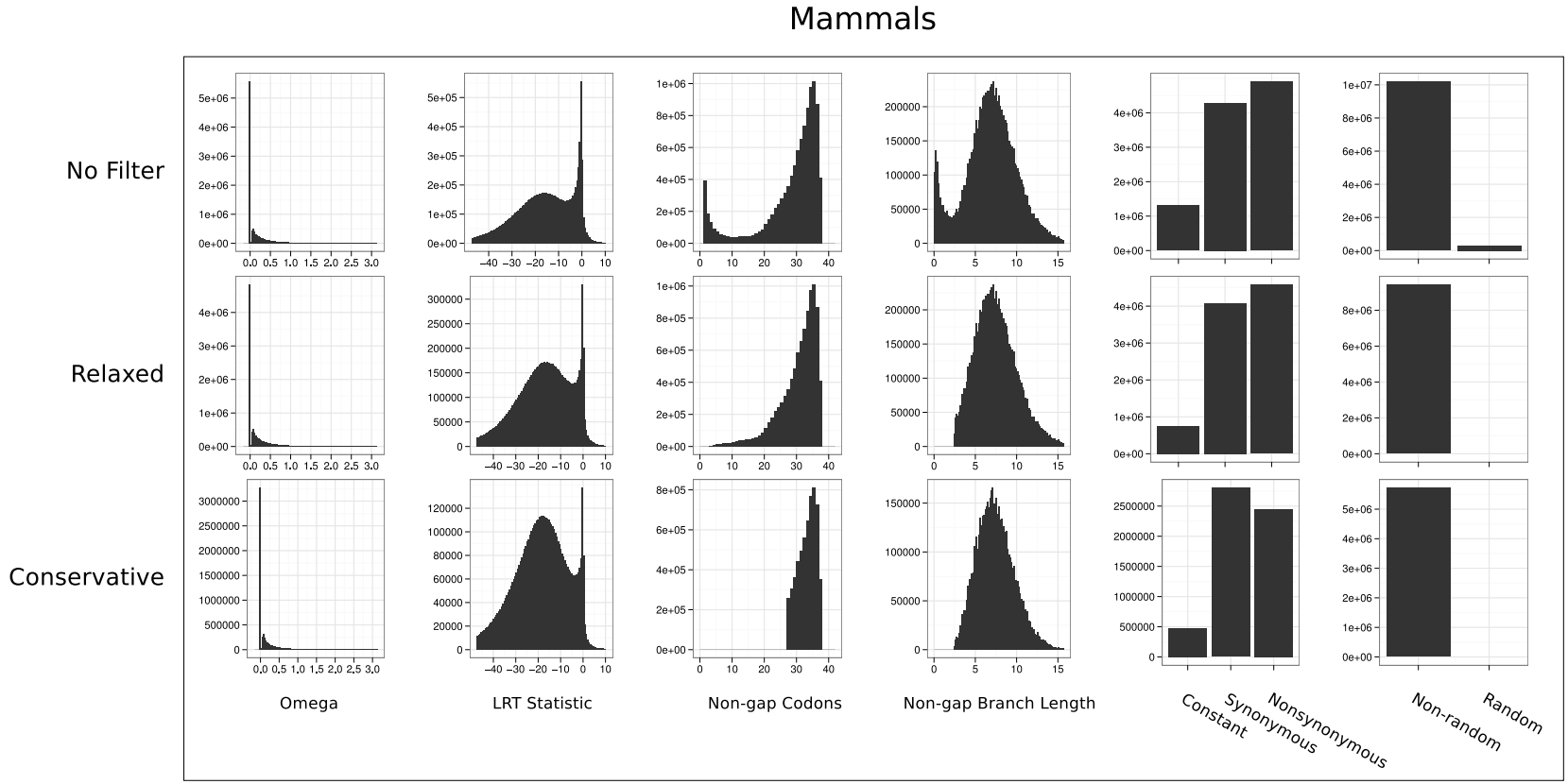

Figure 1.6: Distributions of sitewise values for the Mammals species group, showing the raw data (top row) and the result of applying the relaxed (middle row) and conservative (bottom row) filters.

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

Sitewise data were collected from SLR and stored in a database for storage and further analysis. The Mammals group, containing the most branch length of all the datasets and representing the entire set of aligned sequences, and the Primates group, containing the lowest overall branch length, were used as representative species groups to perform quality-control checks on the sitewise data and to guide the curation of filtered sitewise datasets for each species group.

Some amount of filtering is usually necessary in genomic analyses, and the situation is especially delicate in a scan for positive selection, since non-biological artifacts often appear to represent elevated evolutionary rates [Mallick *et al.*, 2009; Markova-Raina & Petrov, 2011; Schneider *et al.*, 2009]. To balance the desire to maintain as much real data as possible with the concern that a methodological bias may influence the results, two datastes were generated by processing sitewise data separately with two filters: a relaxed filter, designed to retain much of the data while filtering out the most obviously low-quality sites, and a conservative filter, designed to remove a wider set of sites and genes that showed evidence for potential errors or biases.

I first examined the overall distributions of ω estimates and sitewise LRT statistics from SLR. Figures 1.5 and 1.6 show the distributions of six sitewise values for each group of species: two continuous values and two categorical values output by SLR (Omega, Signed LRT, Site Pattern and Random) and two values calculated from the codon alignment (Non-gap Codons and Non-gap Branch Length). Non-gap Codons is a count of the number of non-gap codons in each alignment columnn, and the Non-gap Branch Length value represents the total branch length connecting all non-gap sequences (using the gene-wide branch lengths optimized by SLR).

A prominent feature of the distribution of ω values for the unfiltered Mammals data, shown in the top row of Figure 1.6, was the large number of sites with a zero value for ω_{ML}, the maximum likelihood (ML) point estimate of ω. Further inspection of the data revealed that all $\omega_{ML} = 0$ sites contained either synonymous or constant site patterns. Furthermore, all sites with constant patterns (and nearly all sites with synonymous patterns) yielded a ω_{ML} estimate of zero. Intuitively, an estimate of zero for synonymous sites is appropriate, as the lack of any non-synonymous substitutions throughout the tree would 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 syn-

onymous or non-synonymous substitutions exists in the alignment column. 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 both the non-synonymous and the total substitution rate.

It is not evident from Figure 1.6, but a small proportion (ca. 0.2%) of sites containing synonymous site patterns resulted in maximum likelihood estimates greater than zero. Analysis of the alignment columns corresponding to these sites showed them all to include synonymous codons coding for serine or arginine which are separated by multiple nucleotide differences. Under the mechanistic codon model implemented by SLR, which does not allow for multiple simultaneous nucleotide changes, inferring an evolutionary path between these multiply-substituted codons required the inference of multiple non-synonymous substitutions to reach one codon from the other. This produced a non-synonymous substitution rate of greater than zero for a site with a synonymous site pattern. 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 probable 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 the Non-gap Codons and Non-gap Branch Length values in the unfiltered row of Figure 1.6 showed that most alignment columns contained sequence data from many species (with Non-gap Codons peaking at 36 and Non-gap Branch Length peaking at around 8 substitutions per site), but a noticeable portion contained only a few non-gap sequences. If the alignment columns with low non-gap codon counts represented accurate evolutionary histories, then the observed excess of highly-gapped sites might be taken as an indication that insertion events in terminal lineages or recent ancestral lineages were prominent enough throughout mammalian evolution to leave a noticeable signature of sites with very low non-gap codon counts. Given the many possible sources of error in the annotation and alignment of these sequences, however, a more likely scenario was that

	BL	Nongap BL			Nongap Codons			ω_{ML}, %		
	Quantile	25%	50%	75%	25%	50%	75%	< 1	> 1	$PSC_{5\%}$, %
Mammals	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 non-gap branch length. Sites were separated into 10 equally-sized bins of non-gap branch length and the sites within each bin were summarized by the 25^{th}, 50^{th} and 75^{th} percentiles of non-gap branch length (BL) and non-gap codons, the percentage of sites with ω estimated below or above 1, and the percentage of sites classified as positively-selected codons (PSCs) at a nominal 5% FPR. BL–branch length; PSC–positively selected codons.

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. As a result, these sites might be expected to show 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 analysis.

To test the hypothesis that sites with few non-gap sequences would be less reliable for analysis than other sites, I split the sitewise estimates from the Mammals and Primates groups into ten equally-sized bins of non-gap branch length. Sites within each bin were summarized by calculating the percentage of sites with ω_{ML} less than or greater than 1, as well as the percentage of sites showing evidence for positive selection at a nominal 5% false positive rate (FPR), hereafter referred to as PSCs. The results of this analysis are

presented in Table 1.2. The lowest bin was a clear outlier in the Mammals data, with nearly 17% of sites having $\omega_{ML} > 1$ and 2% of sites being PSCs. The other 9 bins with greater non-gap branch lengths showed fewer sites with $\omega > 1$ and less evidence for positive selection; within those 9 bins, a pattern of gradual increase in the proportion of sites with $\omega_{ML} > 1$ and PSCs was observed at progressively higher non-gap branch lengths. The increase in evidence for positive selection with increasing non-gap branch length could be explained by genes with higher overall dN/dS ratios (and presumably more PSCs) having higher branch lengths due to the increased rate of non-synonymous substitution. Overall, the pattern observed for the Mammals data was consistent with the prediction that sites with few non-gap sequences were not consistent with the general pattern of sitewise data. In terms of choosing an appropriate threshold on which to filter, Table 1.2 indicated that removing sites with the lowest 10% of non-gap branch length would remove most of the apparently anomalous sites.

Table 1.2 showed a similar trend for the Primates dataset, although the distinction between the lowest bin and the rest of the dataset was less obvious. The percentage of PSCs in the lowest decile was only slightly higher than in the next-highest decile, and the proportion of sites with $\omega_{ML} > 1$ was lower than in all other bins. Thus, despite weaker evidence in the Primates data for the anomalous nature of sites with few non-gap sequences, it still appeared that filtering sites in the bottom 10% bin would improve the overall quality and consistency of the data.

Turning back to the bulk distributions in Figure **??**, two other criteria were used to target sites for removal before analysis. First, the rightmost panels of Figures 1.6 and 1.5 depict a small set of sites designated as "random". 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. Second, all sites with fewer than four non-gap sequences were removed. This was done to avoid analyzing sites with very few sequences which were not within the bottom 10% of sites by non-gap branchlength.

At this point, all of the criteria used to define the relaxed filter have been described: non-gap branch lengths, random flags, and the number of non-gap sequences at each site. The middle rows of Figures 1.5 and 1.6 show the summary distributions resulting from applying the relaxed filter to the Mammals and Primates sitewise data.

Three additional criteria were added to create the more conservative filtered dataset. First, the threshold on non-gap sequence counts was increased: all sites with a non-gap

codon count below 75% of the maximum non-gap count for that species group were removed. Second, sites and genes containing windows of clsutered non-synonymous substitutions (as identified in Section 1.3.4) were removed: all sites overlapping the 23,116 15-codon windows with excess non-synonymous substitutions (using the 99.9% quantile based definition of excess substitutions from Section 1.3.4) were masked out, and 819 genes with greater than 10% of sites covered by windows with excess non-synonymous substitutions were removed. Finally, the 3,333 genes which contained more than 2 sets of putative paralogs were excluded.

As with the relaxed filter the result of applying the conservative filter to the Primates and Mammals datasets is shown in the bottom rows of Figures 1.5 and 1.6. Comparing between the distributions in the three rows of Figure 1.6, the most prominent effect of the two filters on the bulk distributions in was the removal of the excess of sites with low non-gap branch lengths and non-gap codon counts. The distributions of ω_{ML} estimates and LRT statistics were qualitatively unchanged unchanged, indicating that the overall characteristics of the dataset were not significantly altered by this filter.

Tables 1.3 and 1.4 provide a quantitative summary of the Mammals and Primates datasets before and after applying the two filters. Also shown is the subset of sites overlapping with Pfam domain annotations collected from Ensembl; as most Pfam domains represent well-folded protein modules [Finn *et al.*, 2010], the set of Pfam-annotated sites were expected to exhibit stronger purifying selection and be less prone to insertions or deletions and alignment error. The rows labeled in parentheses summarize the set of sites which were removed during the creation of the conservatively-filtered dataset, either due to overlap with a window of clustered substitutions (Clusters) or from being within a gene that contained more than 2 recent duplications (Paralogs).

The columns in Table 1.3 show various summary statistics of each sitewise dataset including the number of sites, the proportions of different site patterns, and the proportions of purifying and positive selection based on ω_{ML} estimates from SLR. Table 1.4 provides the number and proportion of identified PSCs (columns under the heading "Positively Selected Sites") as well as the breakdown of sites into purifying, neutral, and positively-selected at two different FPR thresholds (columns under the headings "$P_{\chi^2_1} < 0.1$" and "$P_{\chi^2_1} < 0.05$").

These views make clear the impact of extensive filtering on the genome-wide levels of positive and purifying selection observed in the data. The unfiltered data from the Primates group contained 9.07% of sites with $\omega_{ML} > 1$, and 0.59% of sites were PSCs at

a nominal 5% FPR; the evidence for positive selection was reduced in the conservatively-filtered data, showing 7.87% sites with $\omega_{ML} > 1$ and 0.41% PSCs. An even stronger effect of filtering was seen for the Mammals data, with $\omega_{ML} > 1.5$ being reduced from 5.71 to 2.73 between the unfiltered and conservatively-filtered datasets, and the percentage of PSCs reduced from 0.72% to 0.35%. The rows representing two sets of sites which were removed during the conservative filtering process showed higher signals of positive selection than the unfiltered data, suggesting that these two filtering steps were at least somewhat effective in removing potentially anomalous or untrustworthy sites from the dataset. For sites removed from being within clusters of non-synonymous substitutions, the enrichment for signals of positive selection was clear: in Primates, 18.28% of sites yielded $\omega_{ML} > 1$, and 1.47% of sites were PSCs at a 5% FPR threshold, more than three times the proportion of PSCs seen in the conservatively-filtered dataset. Sites removed as a result of being within genes containing recent duplications showed less of a signal for positive selection, but the proportions of PSCs and sites with $\omega_{ML} > 1$ were still above those seen in either the relaxed or conservatively filtered datasets for both Mammals and Primates. Thus, genes that have experienced many recent duplications in mammals contained higher levels of positive selection even after the most-divergent paralogous copies were removed.

			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
	Relaxed	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
	Conservative	6.22e+06	59.29	26.26	14.45	9	0.76	0.82	0.50	0.19	0.57	87.27	92.13	7.87	4.80
	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
	(Clusters)	9.74e+05	44.25	21.59	34.17	9	1.18	1.44	1.45	0.46	0.83	71.66	81.72	18.28	11.99
	(Paralogs)	1.64e+06	52.37	25.13	22.50	9	0.87	1.33	2.65	0.27	0.66	82.58	89.26	10.74	6.73
Mammals	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
	Relaxed	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
	Conservative	5.72e+06	8.30	48.98	42.72	34	7.28	7.48	2.42	0.15	0.30	90.86	97.27	2.73	0.91
	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
	(Clusters)	9.92e+05	4.33	22.06	73.61	29	9.50	9.68	4.86	0.40	0.52	71.95	88.84	11.15	4.69
	(Paralogs)	1.82e+06	7.18	38.51	54.31	31	7.70	8.33	6.84	0.22	0.40	85.29	94.84	5.16	2.03

Table 1.3: Summary statistics of sitewise estimates for Mammals and Primates data with various filters applied. Rows labeled (Clusters) and (Paralogs) contain sites excluded by the Conservative filter. Columns under the "ω_{ML} Below / Above" heading measure the percentage of sites with ω_{ML} below or above the indicated value. Med.—median, Const.—constant, Syn.—synonymous, Nsyn.—non-synonymous, BL—branch length.

		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	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
	Relaxed	82607	(0.95)	47825	(0.55)	14619	(0.17)	104	(0.001)	33.27	65.79	0.95	15.98	83.47	0.55
	Conservative	45179	(0.73)	25710	(0.41)	7661	(0.12)	50	(0.001)	33.89	65.38	0.73	15.88	83.70	0.41
	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
	(Clusters)	23808	(2.44)	14331	(1.47)	4707	(0.48)	40	(0.004)	30.73	66.82	2.44	16.61	81.92	1.47
	(Paralogs)	19175	(1.17)	11269	(0.69)	3496	(0.21)	30	(0.002)	33.87	64.96	1.17	17.72	81.59	0.69
Mammals	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
	Relaxed	76370	(0.81)	52096	(0.55)	23323	(0.25)	1879	(0.020)	86.51	12.68	0.81	83.88	15.57	0.55
	Conservative	29075	(0.51)	19900	(0.35)	9025	(0.16)	781	(0.014)	90.61	8.88	0.51	88.54	11.11	0.35
	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
	(Clusters)	23716	(2.39)	16531	(1.67)	7648	(0.77)	659	(0.066)	70.07	27.54	2.39	65.66	32.67	1.67
	(Paralogs)	18090	(0.99)	12394	(0.68)	5551	(0.30)	432	(0.024)	83.83	15.18	0.99	80.76	18.56	0.68

Table 1.4: Proportions of sites subject to positive, purifying and neutral selection at various LRT$_{SLR}$ thresholds for Mammals and Primates data with varous filters applied. 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

To produce high-confidence sitewise estimates across the ten chosen species groups, sitewise data from each species group were processed with the conservative filter as described above. The resulting global distributions of site patterns, sitewise ω_{ML} estimates, and 95% confidence intervals are shown in Figure 1.7. 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 right panel in each row shows the distributions of ω_{ML} for sites which contained a non-synonymous site pattern.

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 1.7 showed that the branch length of each species group had a strong effect on the overall composition of the sitewise data. Groups covering little branch length, such as Primates and Atlantogenata, contained mostly constant sites, while groups covering a large amount of branch length, such as Eutheria and Mammals, contained few constant sites and roughly equal proportions of sites with synonymous and non-synonymous site patterns. Comparing the Glires and Mammals 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, for example, 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).

The distributions of ω_{ML} estimates are shown in Figure **??** as a series of histograms showing the ω_{ML} density (for nonzero values of ω_{ML} only) and a series of solid lines showing the cumulative ω_{ML} density (representing all values); the lower and upper dashed lines show the cumulative density of the lower and upper 95% confidence interval resulting from each sitewise estimate. From these distributions, it is clear that the majority of protein-coding sites have evolved under purifying selection in mammals, a fact which is most easily seen in the larger species groups. The Mammalia group, which contained only a small proportion of 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

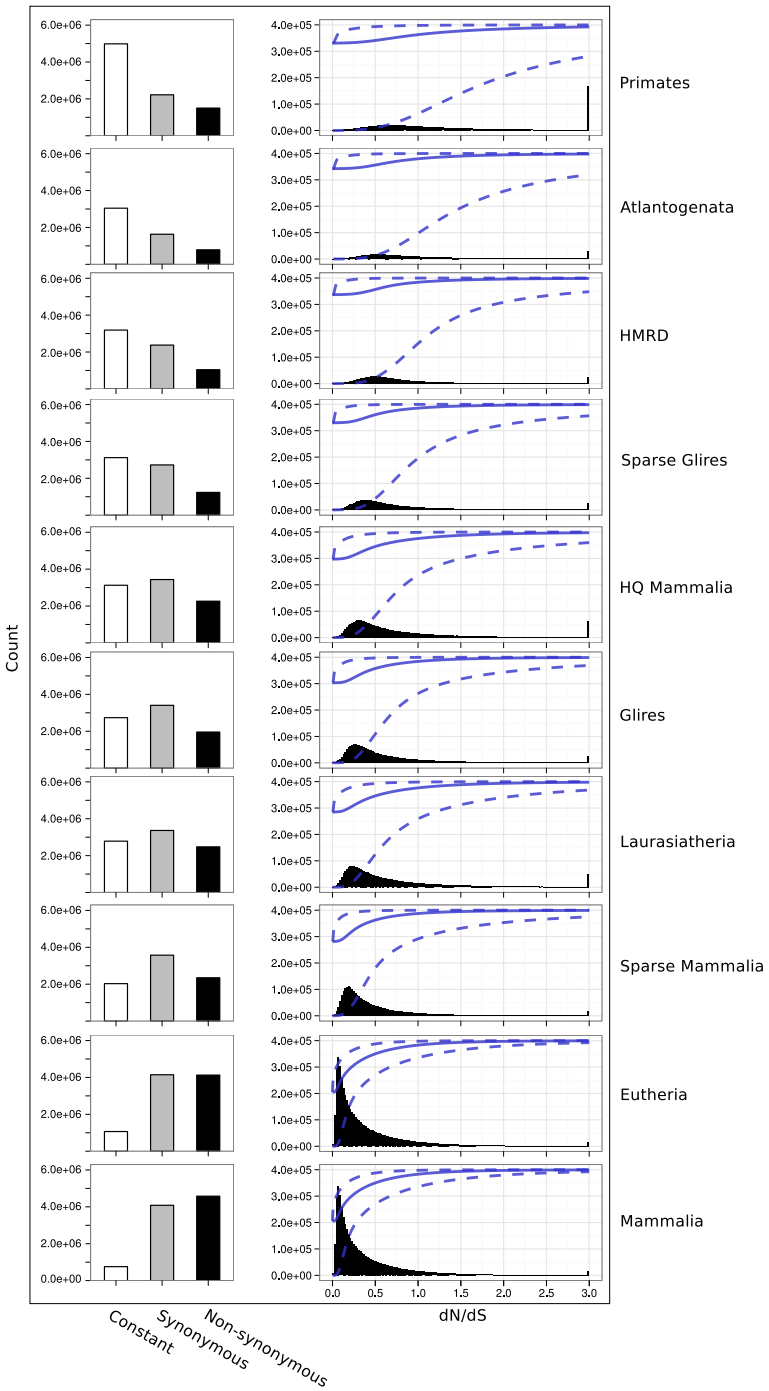

Figure 1.7: Global distributions of site patterns and ω estimates for ten species groups. Left panels: bars represent the number of sites showing constant, synonymous, and non-synonymous patterns. Note, the y-axis is held constant between rows. Right panels: bars represent histograms of ω_{ML} estimates for sites where $\omega_{ML} > 0$. Sites with $\omega_{ML} > 3$ are counted in the bin at $\omega_{ML} = 3$. A solid line is drawn showing the cumulative distribution of ω_{ML}, and dashed lines are drawn above and below the solid line showing the cumulative distributions of the lower and upper bounds, respectively, of the 95% confidence interval associated with each sitewise estimate.

selection with ω_{ML} estimates below 1. The height of the ω_{ML} cumulative distribution at $\omega = 1$ corresponds to the proportion of sites with some evidence for purifying selection. The nonzero ω_{ML} values were more evenly spread in the other species groups: Glires contained a maximum nonzero ω_{ML} density at around $\omega \approx 0.25$ and Primates at $\omega \approx 0.7$. This upwards shift in nonzero ω_{ML} estimates relative to Mammalia was likely due to the greater proportion of constant and synonymous sites in datasets with lower overall branch lengths: sites which were truly evolving with $0 < \omega < 1$, but where no non-synonymous or synonymous substitutions were observed, would have their ω_{ML} estimate "pushed" towards zero, presumably causing 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 $\omega_{ML} < 1$, as a way to sort sites according to their evidence 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 evidence for purifying selection at that site 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 an estimated ω_{ML} near zero.

To further explore this point, Figure 1.8A shows the relationship between LRT_{SLR}, ω_{ML} and the $\text{CI}_{95\%}$ width for sites from the Mammals group. 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 tightest confidence intervals 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 1.8B, reveal similar patterns, but with generally less-negative LRT_{SLR} values, higher ω_{ML}, and larger $\text{CI}_{95\%}$. These differences highlight the impact of branch length on the amount of confidence with which ω can be estimated on a per-site basis. The low branch length of the Primates clade rarely yields ω_{ML} estimates with $\text{CI}_{95\%}$ intervals smaller than 1, while the bulk of sites from the Mammalia dataset have relatively small $\text{CI}_{95\%}$s. Thus, the distribution of ω_{ML} estimates from datasets with low branch lengths (e.g., the histogram densities seen in Figure 1.7) should be interpreted with caution, as any comparison between ω_{ML} from different sites or datasets may be more sensitive to the amount of statistical confidence placed on each estimate than to any meaningful biological difference between the two sets of data.

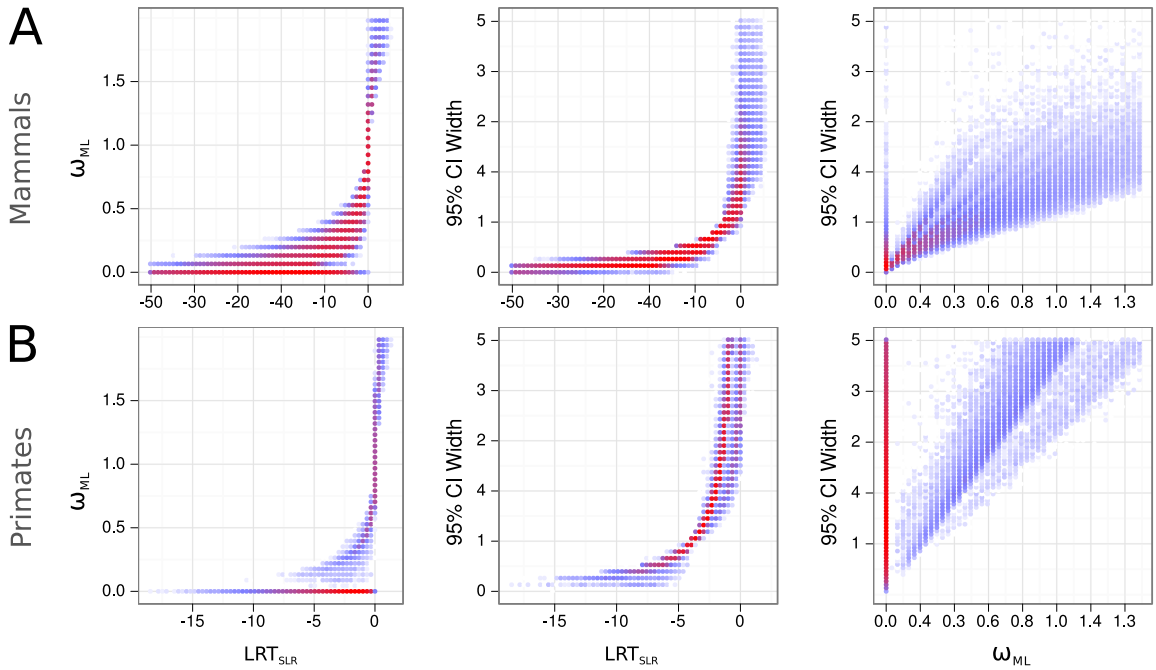

Figure 1.8: The relationship between LRT_{SLR}, ω_{ML}, and $\text{CI}_{95\%}$ width in (A) Mammalia and (B) Primates datasets. Each point represents the binned density of sites; no points are drawn where no density exists, while blue and red points are drawn at areas of low and high density, respectively. 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 scales between plots in (A) and (B), reflecting the paucity of sites in Primates with strong evidence (LRT_{SLR} <-12) for purifying 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	Conservative	6.22e+06	59.29	26.26	14.45	9	0.76	0.82	0.50	0.19	0.57	87.27	92.13	7.87	4.80
Atlantogenata		4.07e+06	57.23	30.07	12.70	5	0.94	1.01	0.37	0.13	0.41	90.01	95.38	4.62	2.29
HMRD		5.02e+06	49.75	36.41	13.84	4	0.96	1.01	0.36	0.12	0.37	90.06	96.37	3.63	1.73
Sparse Glires		5.35e+06	45.36	39.11	15.53	5	1.24	1.32	0.68	0.12	0.36	90.91	96.71	3.29	1.51
HQ Mammals		6.38e+06	37.09	40.68	22.23	8	1.46	1.55	0.64	0.17	0.43	88.31	94.97	5.03	2.50
Glires		5.79e+06	34.70	43.68	21.62	7	1.77	1.87	0.84	0.13	0.36	90.54	96.53	3.47	1.50
Laurasiatheria		5.35e+06	33.36	41.99	24.66	11	2.03	2.15	0.87	0.16	0.41	88.88	95.36	4.64	2.22
Sparse Mammals		5.65e+06	25.81	46.75	27.44	6	2.55	2.75	1.45	0.13	0.32	91.65	97.28	2.72	1.10
Eutheria		5.72e+06	11.96	49.78	38.26	32	5.80	6.01	1.96	0.15	0.33	90.17	96.76	3.24	1.18
Mammals		5.72e+06	8.30	48.98	42.72	34	7.28	7.48	2.42	0.15	0.30	90.86	97.27	2.73	0.91

Table 1.5: Summary statistics of sitewise estimates for all species groups with the conservative filter applied. Columns under the "ω_{ML} Below / Above" heading measure the percentage of sites with ω_{ML} below or above the indicated value. Med.—median, Const.—constant, Syn.—synonymous, Nsyn.—non-synonymous, BL—branch length.

		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	Conservative	45179	(0.73)	25710	(0.41)	7661	(0.12)	50	(0.001)	33.89	65.38	0.73	15.88	83.70	0.41
Atlantogenata		8143	(0.20)	3852	(0.09)	757	(0.02)	0	(0.000)	46.96	52.84	0.20	23.75	76.15	0.09
HMRD		6538	(0.13)	3040	(0.06)	534	(0.01)	0	(0.000)	63.74	36.13	0.13	37.42	62.52	0.06
Sparse Glires		7233	(0.14)	3316	(0.06)	644	(0.01)	0	(0.000)	70.34	29.52	0.14	49.07	50.87	0.06
HQ Mammals		29344	(0.46)	16374	(0.26)	4548	(0.07)	0	(0.000)	74.87	24.67	0.46	61.55	38.19	0.26
Glires		11155	(0.19)	5581	(0.10)	1221	(0.02)	0	(0.000)	78.93	20.88	0.19	67.92	31.98	0.10
Laurasiatheria		29058	(0.54)	17617	(0.33)	5944	(0.11)	41	(0.001)	78.31	21.15	0.54	68.74	30.93	0.33
Sparse Mammals		7953	(0.14)	3913	(0.07)	857	(0.02)	0	(0.000)	81.99	17.87	0.14	75.28	24.65	0.07
Eutheria		35270	(0.62)	24234	(0.42)	11006	(0.19)	999	(0.017)	89.00	10.38	0.62	86.54	13.04	0.42
Mammals		29075	(0.51)	19900	(0.35)	9025	(0.16)	781	(0.014)	90.61	8.88	0.51	88.54	11.11	0.35

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.

Instead, the confidence intervals and likelihood ratio test (LRT) statistics calculated by SLR for each site could be used to identify sites evolving under purifying or positive selection with confidence. Sites with CI_{upper}, the upper bound of the $\text{CI}_{95\%}$ interval, below $\omega = 1$ could be interpreted as having evidence of purifying selection with an expected 5% FPR; likewise, sites with CI_{lower} above $\omega = 1$ contained evidence of positive selection with an expected 5% FPR. In both cases, SLR was controlling for an expected 5% FPR under the null model of neutral evolution. As expected, there was a direct relationship between CI_{upper} and the χ^2_1 approximation to the LRT_{SLR} distribution, whereby the set of sites with $\text{CI}_{upper} < 1$ was exactly equivalent to the set of sites with LRT_{SLR} below the negative χ^2_1 95% critical value. Similarly, the sites with $\text{CI}_{lower} > 1$ were those with LRT_{SLR} above the χ^2_1 95% critical value. Because of this equality, I will refer to LRT_{SLR} values at various χ^2_1 threshold values instead of the 95% $\text{CI}_{95\%}$ intervals when discussing sites with significant evidence for purifying or positive selection.

Tables 1.5 and 1.6 provide summaries of the sitewise estimates obtained for each of the 10 mammalian species groups, showing the same values provided earlier in Tables 1.3 and 1.4 for the different filters.

Table 1.6 presents the proportions of PSCs identified at a variety of LRT_{SLR} thresholds, demonstrating that anywhere between 0.01% to 0.73% of sites could be confidently identified as under positive selection in mammals at nominal FPR thresholds between 1% and 10%. Interestingly, however, different species groups yielded strikingly different estimates of the proportion of PSCs. At a 5% FPR threshold, the Primates, HQ Mammals, Laurasiatheria, Eutheria, and Mammals groups produced broadly comparable proportions of positively-selected sites, ranging from 0.33% to 0.42%. The proportions of PSCs in these groups were higher using a 10% FPR threshold (ranging from 0.46% to 0.73%) and lower using a 1% FPR threshold (ranging from 0.07% to 0.19%). When the FDR was controlled using the Benjamini-Hochberg method, however, far fewer PSCs were identified. Only the Eutheria and Mammalia groups yielded a substantial number of positively-selected sites at this level of control; the Primates and Laurasiatheria data yielded non-zero numbers of PSCs as well, but these species groups were likely limited in their power to yield positively-selected sites after FDR control due to their lower total branch lengths.

The Atlantogenata, HMRD, Sparse Glires, Glires and Sparse Mammalia groups all produced very low proportions of positively-selected sites identified across all FPR thresholds. At FDR<0.05, all four groups yielded zero significant PSCs, and at a 1% FPR they all contained lower than 0.01% PSCs. These PSC-depleted species groups were widely dis-

tributed in the amount of total branch length they covered (ranging in median non-gap branch length from 0.94 for Atlantogenata to 2.55 for Sparse Mammals), suggesting that the lower number of PSCs was not strongly influenced by branch length; a similar point could be made of the species groups with higher proportions of PSCs, which comprised the groups with the lowest (Primates) and highest (Mammals) total branch length.

In Mammalia, the breakdown of sites into positive, negative and neutral categories at 10% and 5% significance thresholds produced a pattern similar to that seen in the ω_{ML} distributions from Figure 1.7, with a large amount of purifying constraint (83.87% of sites at 5% FPR), a small proportion of neutrally-evolving sites (15.57%), and a diminishing number 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 to confidently detect purifying selection (83.87% vs. 15.97%).

Overall, the conservatively-filtered sitewise data showed that, when using ω_{ML} estimates, between 1% to 5% of protein-coding sites are evolving under positive selection. This number varied strongly between different species groups, however. Comparing between the four phylogenetically independent mammalian superorders (Primates, Glires, Laurasiatheria, and Atlantogenata), I found that Primates showed by far the most PSCs and sites with $\omega_{ML} > 1$. Laurasiatheria showed similar proportions of sites with $\omega_{ML} > 1$, but Atlantogenata showed fewer PSCs than Laurasiatheria; this difference may reflect their different branch lengths, as the Laurasiatheria group covers twice as much branch length as Atlantogenata and thus would be expected to have more power to confidently detect positive selection. Finally, the Glires group showed strikingly lower levels of positive selection compared to the other mammalian superorders. Despite the relatively high branch length contained within Glires (median total length of 1.77 versus 2.03 for Laurasiatheria), only 0.10% of sites were identified as PSCs in Glires at a 5% FPR, compared to 0.33% in Laurasiatheria and 0.41% in Primates.

These results may be evaluated in terms of the impact of effective population size on the efficacy of natural selection in mammals [Ellegren, 2009; Nikolaev *et al.*, 2007; Popadin *et al.*, 2007]. Rodents are known to have an effective population size well above that of primates [Kosiol *et al.*, 2008], and given the strong correlation between body size, generation time and effective population size [Nikolaev *et al.*, 2007], one can infer that species within the Laurasiatheria group, with generally longer generation times and larger

body sizes than rodents [Hou *et al.*, 2009], have effective population sizes more similar to those seen in primates. The Afrotheria group, containing species ranging from small moles to elephants and manatees, is more diverse, making it difficult to estimate an expected historical effective population size. Nevertheless, Ohta's nearly neutral theory [Ohta, 1992] predicts that species with lower effective population sizes will evolve with less efficient natural selection. A comparison of the Primates and Glires data clearly revealed this effect: the proportion of sites with $\omega_{ML} < 0.5$ was 87.27% for Primates and 90.54% for Glires. Thus, the difference in the proportion of sites likely to be under purifying selection was well-explained by the difference in effective population size between primates and rodents. If the existence of PSCs in the sitewise data is interpreted as true evidence for positive selection, then the nearly neutral theory predicts that Glires should show *more* PSCs than Primates, as the efficacy of positive selection would be greater in the species with a larger effective population size. The opposite effect is seen, however, with Primates showing much greater levels of apparent positive selection as measured by both the proportion of sites with $\omega_{ML} > 1$ and by PSCs identified at various FPR thresholds.

This suggestes an alternative interpretation: that perhaps the different levels of positive selection could be due mainly to the relaxation of selective constraint in Primates and other species with low effective population sizes. A difference in effective population sizes should have its main effect on slightly deleterious mutations, with a greater proportion of slightly deleterious mutations (e.g., mutations with fitness effects corresponding to dN/dS values slightly below 1) becoming effectively neutral in a species with a low effective population size. In comparing the Primates and Glires groups, the expected result is that a subset of mutations which were under purifying selection in Glires would be effectively neutral in Primates, bringing the expected ω from < 1 to 1. If this class of sites were large enough, it might significantly interfere with the resulting FPR for detecting PSCs, as sites with $\omega = 1$ are the most prone to produce false positives.

Thus, the relaxed constraint argument tempers the interesting observation of strong differences in the numbers of PSCs between different species groups. A lower historical effective population size for the Primates and Laurasiatheria species groups may explain some of the increase in the number of PSCs detected, even in the absence of true variation in the prevalence of positive selection between the species groups investigated here. Still, the argument may be made that statistical methods for controlling error rates, such as the Benjamini-Hochberg method for FDR control used to identify PSCs at an expected FDR< 0.05 in Table 1.6 [Benjamini & Hochberg, 1995], should account for the potential

confounding effects of relaxed constraint noted above. For this reason, the observation that Primates and Laurasiatheria both yielded non-zero numbers of PSCs at FDR<0.05 may be taken as some indication of a true difference in the levels of positive selection between the species groups investigated here.

1.5 Conclusions

This chapter described the filtering, alignment, and analysis of a comprehensive set of mammalian orthologs across 38 species.

In order to ensure that false signals of positive selection were avoided as much as possible, several levels of filtering were applied before and after the estimation of sitewise selective pressures using SLR: low-quality genomic sequence was masked out, short or divergent apparent paralogous copies were removed, and alignment columns showing evidence of clustered non-synonymous substitutions or low amounts of evolutionary information were excluded from the analysis. A comparison of the levels of purifying and positive selection contained within sites filtered at various thresholds showed the importance of thorough filtering prior to genome-wide analysis, highlighting especially the ability of stretches of mis-annotated or mis-assembled sequence to introduce strong (and incorrect) signals of localized positive selection into evolutionary analyses. I showed that a novel approach, based on the identification of lineage-specific clusters of excessive non-synonymous substitutions within short alignment windows, could effectively target these erroneous regions for removal.

I applied the conservative filter to sitewise estimates obtained from several groups of mammalian species. This allowed the impact of the total branch length of a species group on the estimation of sitewise selective pressures to be clearly seen, with the Mammals group containing many more non-constant alignment sites and more realistic ω_{ML} estimates than groups with little branch length, such as Primates. Relating my results back to the MGP, which used the HMRD and HQ Mammals groups as reference points by which to estimate the increase in power to detect genome-wide constraint resulting from the additional mammals sequenced at low coverage, I found that the addition of low-coverage genomes increased the ability to detect purifying constraint in protein-coding regions by 43.85% and 136% compared to the HQ Mammals and HMRD species groups, respectively, at a 5% FPR. Although I found the levels of positive selection between species groups to be highly dependent on the species sampling (and thus, a comparison of "power" to be less

meaningful), the Mammals species group identified 21.5% and 550% more PSCs than the HQ Mammals and HMRD species groups, respectively, at a 5% FPR. Thus, the additional branch length resulting from the sequencing of low-coverage genomes greatly improved the power to detect purifying and positive selection in mammalian proteins.

Finally, I analyzed the levels of purifying and positive selection within four phylogenetically independent mammalian species groups, identifying strong differences between different groups, likely resulting from differences in effective population size. Although the impact of effective population size is well known and has been previously studied in mammalian superorders, the work described in this chapter represented a careful and quantitative analysis of levels of purifying and positive selection in these species groups. The observation that the Glires group showed less positive selection than all other groups suggested a connection between high numbers of PSCs and relaxed constraint, although Primates and Laurasiatheria both showed evidence for strong PSCs even at a very stringent FDR threshold.

Although more work needs to be done to evaluate what might be causing these differences between species and to correctly control for the possible effects of relaxed constraint, I have shown that the analysis of sitewise estimates is an intuitive and informative approach to evaluating signals of purifying and positive selection in mammalian genomes.

[...]

Chapter 2

Characterizing the evolution of genes and domains in mammals using sitewise selective pressures

2.1 Introduction

This chapter describes the use of sitewise data to identify trends in the evolution of protein-coding genes and domains. I will first develop a number of methods for quantifying signals of positive and purifying selection from sitewise estimates within genes and domains and apply these methods to the sitewise data generated in Chapter 1. To provide a higher-level interpretation of these results, in the next section I will use functional gene annotations to identify categories enriched for genes with evidence of positive selection or accelerated evolution. Lastly, I will quantitatively evaluate these results in the context of previously-published studies investigating the distribution of positive selection across mammalian genomes.

Since the first non-human mammalian genomes were sequenced, there has been great interest in using comparative data to identify genes showing signatures of positive selection in mammals. Much of this interest stems from the prospect that such genes may reflect the historical impact of natural selection acting to fix beneficial mutations within a population over time—a major driving force in the modern molecular interpretation of Darwin's theory of natural selection [Endo *et al.*, 1996; Hughes, 1999]. Previous scans for positive selection in primate genomes have revealed enrichments for PSGs related to sensory perception and olfaction [Clark *et al.*, 2003], apoptosis and spermatogenesis [Nielsen *et al.*, 2005],

and iron ion binding and keratin formation [Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007]; analyses in other mammalian genomes have revealed largely similar patterns [Kosiol *et al.*, 2008; Li *et al.*, 2009]. To explain the increased dN/dS values observed within PSGs, three distinct evolutionary dynamics have commonly been invoked: an evolutionary arms race between host and parasite interacting genes [Yang, 2005], sexual selection or genetic conflict between the sexes [Clark & Civetta, 2000; Wyckoff *et al.*, 2000], and functional adaptation following gene duplication [Zhang *et al.*, 2002].

As the power of phylogenetic analysis using codon models depends strongly on the amount of branch length encompassed by the species being compared [Anisimova *et al.*, 2001, 2002], there was some reason to believe *a priori* that the detection of PSGs using mammalian alignments incorporating low-coverage genomes would be more powerful than in previous whole-genome analyses, which typically included 12 or fewer species across mammals and lower total branch length [ELLEGREN, 2008]. However, differences in the specific models used to detect positive selection are expected to affect the sensitivities of one study compared to another [Anisimova & Kosiol, 2009], so the set of genes identified using the current methodology would necessarily be expected to be a superset of those identified in previous studies. Most large-scale studies have used the branch-site test for positive selection [Zhang *et al.*, 2005], while the results described in this chapter were generated using SLR. I showed in Chapter **??** that SLR has similar power to the site-based test implemented in PAML for detecting sitewise positive selection, but no analysis has yet compared the differences in PSGs identified by site-specific and branch-site methods on a large scale. For this reason, I hoped that a quantitative comparison between PSGs identified using the current methodology and those found in previously-published studies may improve our understanding of how similar or different the PSGs identified by different methods can be.

2.2 Methods for combining sitewise estimates to identify positive selection

In Chapter **??** I covered the curation and analysis of several highly filtered sets of genome-wide sitewise selective pressures generated from different groups of mammalian species. These sitewise estimates were used to characterize the global distribution of evolutionary constraint and to compare overall levels of purifying and positive selection between groups of mammalian species. The focus on individual codons as an evolutionary unit of investiga-

tion is relatively uncommon, but it allowed for large-scale differences in evolutionary trends between species groups to be identified and for the impact of different filtering schemes on overall signals of positive selection to be easily evaluated.

The more traditional approach in comparative genomics has been to model the protein-coding gene, as opposed the protein-coding amino acid site, as the unit of analysis. For detecting positive selection, the grouping of alignment sites into genes—which results in identification of PSGs instead of PSCs—has three main advantages. First, the combined analysis of many alignment sites improves the accuracy of estimated evolutionary parameters and boosts the power likelihood ratio (LR)-based tests for detecting positive selection. This can be easily seen in the simulations of Anisimova and Yang [2001; 2001], which showed large power differences for detecting positive selection in alignments simulated with 100, 200, and 500 codons. Second, detailed studies of sitewise selective pressures in genes with strong signals of positive selection have usually observed clusters of positively-selected sites [Kosiol *et al.*, 2008; Sawyer *et al.*, 2005], suggesting that the evolutionary dynamics creating detectable signals of positive selection tend to affect many functionally or structurally related amino acid sites within a gene as opposed to a single site. These studies represent empirical evidence that combining sitewise estimates within genes is biologically sensible. The third argument in support a gene-centric analysis of positive selection is that in the absence of complete protein structure information, much more tends to be known about entire genes (through the results of high-throughput studies and experiments in model organisms) than is known about individual protein-coding sites. Thus, a gene-centric analysis allows a dataset to be more easily analyzed in connection with abundant external functional data, benefitting the biological interpretation of results.

2.2.1 One-sided and two-sided p-values

[TODO...]

2.2.2 Combining multiple sitewise tests within genes

A major issue in combining sitewise estimates to identify PSGs is that of correcting for performing multiple sitewise tests per gene. The SLR method performs an independent statistical test at each site, producing a sitewise statistic which can be compared to a χ^2_1 distribution to yield a p-value representing the strength of evidence against strict neutral evolution [Massingham & Goldman, 2005]. When combining these p-values to decide

whether a gene contains significant evidence for positive seletcion, one must take into account the number of tests performed. For example, a 100-codon gene evolving under the null model ($\omega = 1$) would be expected to produce 5 sites with p-values at a nominal FPR of 0.05; correspondingly, there would be a 99.4% chance that at least one site within the gene would have $p < 0.05$. This comes from the complement of the probability that no sites out of n have $p < x$, which is $(1 - x)^n$. Thus, if the set of genes containing at least one site with nominal $p < 0.05$ were called PSGs, nearly all genes evolving under the true null model would be selected. In contrast, the LRTs for positive selection implemented in PAML only perform one statistical test per gene and do not suffer from the same multiple testing problem. Clearly, some procedure for correcting or combining the results from multiple tests must be applied in order to correctly identify PSGs using sitewise data.

I tested three different methods which are capable of correcting for multiple sitewise tests within genes to identify PSGs: first, adjusting significance thresholds to control the family-wise error rate (FWER), second, combining p-values from multiple tests to produce a single p-value summarizing the overall evidence against the null hypothesis, and third, estimating empirical gene-wise p-values based on the genome-wide distribution of sitewise estimates. Each approach makes different use of the sitewise data from each gene to identify a set of significant PSGs and may thus yield a unique set of PSGs. The remainder of this section provides some background on each approach and describes how it was applied to the current dataset.

2.2.3 Controlling the FWER

The FWER is defined as the probability, for a given set of tests performed, of one or more tests producing a false positive result. In the example of a 100-codon gene evolving under the null model, the FWER at a nominal p-value of 0.05 was 0.994. Assuming an appropriate uniform null distribution of p-values and independence between tests, the Sidak method of p-value adjustment (to which the popular Bonferroni correction is an approximation) identifies the maximum nominal p-value x expected when the FWER is equal to or below the desired level α: if the FWER expected for a family of n tests thresholded at a nominal p-value of x is $\alpha = 1-(1-x)^n$, then the maximum p-value expected while controlling for a desired FWER can be found by rearranging the equation: $x = 1-(1-\alpha)^{1/n}$. A similar but more powerful approach is the step-up method from Hochberg; this method is implemented internally in SLR [Hochberg, 1988; Massingham & Goldman, 2005]. I used the Hochberg method, as implemented in the *p.adjust* function from the R statistical project [TOCITE],

to identify PSGs at 1%, 5%, and 10% FWER thresholds. Genes identified using this approach will be referred to as $\text{PSG}_{H_{1\%}}$, $\text{PSG}_{H_{5\%}}$, and $\text{PSG}_{H_{10\%}}$.

A drawback of using FWER control to identify PSGs is that it only identifies a per-test significance threshold at which the FWER is expected to be below a certain value. Applying the significance threshold to sitewise data results naturally in a binary classification of genes into PSGs (which contain at least one site with a statistic more extreme than the threshold) and non-PSGs (which contain no significant sites at the FWER-controlling threshold). Information is lost regarding the number and distribution of significant sites within genes, however, and the set of PSGs cannot be ranked in order of their overall evidence for positive selection. The identification of PSGs at multiple FWER thresholds can help by identifying genes significant at a variety of thresholds, but this approach is still inflexible when compared to the gene-wise p-values resulting from the site-based LRTs in PAML.

2.2.4 Combining p-values

The second approach to multiple testing directly addresses this problem by combining p-values from a series of independent tests, producing an overall p-value for the null hypothesis given the set of tests performed. The motivation behind such methods is that moderately significant results from independent tests of a common null hypothesis should be considered as good or better evidence than one strongly significant test. The two most popular methods for performing this type of combination are Fisher's combined probability test, based on the product of p-values from multiple tests, and Stouffer's method, based on a normal transformation of p-values (Fisher, 1932; Stouffer *et al.*, 1949; reviewed in Whitlock, 2005). The performance of these methods depends highly on the distribution of input p-values, however; it has been noted that a relatively small number of large p-values can limit the power of Fisher's test [Zaykin *et al.*, 2002], and the Stouffer method is equally sensitive to small and large p-values. Since the majority of mammalian protein-coding sites showed moderately strong signals of purifying selection, the distribution of one-sided p-values for positive selection would be heavily weighted towards 1 for most genes. As a result, both the Fisher and Stouffer methods were expected to lack power, failing to yield significant p-values in the face of a dominant signal of purifying selection even when several sites showed strong evidence of positive selection.

Myriad other protocols for combining p-values exist (see Cousins [2007] for an extensive review of alternative methods), but given the...

2.2.5 Assigning empirical p-values based on the global sitewise distribution

2.3 A comparison of sitewise results to previously described sets of positively selected genes

In order to evaluate the prevalence of positive selection in mammals for various domain structures, we used the Pfam domain mappings from the Ensembl database (release version 54) to annotate the site-wise dN/dS values with domain assignments. We mapped Pfam protein domain annotations from all sequences in a gene tree onto the alignment, keeping only features with a hit score greater than 20 and alignment sites with greater than 4 columns and an inferred dN/dS value of less than 50. We then removed any domains with fewer than one thousand annotated sites, to avoid errors resulting from small sample sizes.

2.3.1 Identifying protein domains subject to positive selection

2.3.2 Identifying protein domains subject to strong or weak purifying selection

2.4 Identifying genes under unusual selective pressures in mammalian superorders

Bibliography

Anisimova, M. & Kosiol, C. (2009). Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol*, **26**, 255–71. 47

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. 47, 48

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, 47

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

Averof, M., Rokas, A., Wolfe, K. & Sharp, P. (2000). Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science*, **287**, 1283–6. 29

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. 10

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. 20

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. 10

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. 34, 40, 43

Boyko, A., Williamson, S., Indap, A., Degenhardt, J., Hernandez, R., Lohmueller, K., Adams, M., Schmidt, S., Sninsky, J., Sunyaev, S., White, T., Nielsen, R., Clark, A. & Bustamante, C. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, **4**, e1000083. 4

Callahan, B., Neher, R., Bachtrog, D., Andolfatto, P. & Shraiman, B. (2011). Correlated evolution of nearby residues in drosophilid proteins. *PLoS Genet*, **7**, e1001315. 20

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. 10, 13

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. 24

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, 46

Clark, A.G. & Civetta, A. (2000). Evolutionary biology: Protamine wars. *Nature a - z index*, **403**, 261–263. 47

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. 11

TOCITE (????). Citation will be inserted at a later point in time. 2, 49

Cousins, R. (2007). Annotated bibliography of some papers on combining significances or p-values. 50

ELLEGREN, H. (2008). Comparative genomics and the study of evolution by natural selection. *Molecular Ecology*, **17**, 4586–4596. 47

Ellegren, H. (2009). A selection model of molecular evolution incorporating the effective population size. *Evolution*, **63**, 301–5. 5, 42

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. 3, 11

Endo, T., Ikeo, K. & Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. *Molecular Biology and Evolution*, **13**, 685–690. 46

Eyre-Walker, A. & Keightley, P. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet*, **8**, 610–8. 4

Eyre-Walker, A., Woolfit, M. & Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, **173**, 891–900. 4

Fay, J. & Wu, C. (2003). Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet*, **4**, 213–35. 3

Finn, R., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J., Gavin, O., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E., Eddy, S. & Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Res*, **38**, D211–22. 32

Fisher, R. (1932). *Statistical methods for research workers*. Oliver and Boyd, London. 50

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. 10

Green, P. (2007). 2x genomes–does depth matter? *Genome Res*, **17**, 1547–9. 6

Gu, X., Wang, Y. & Gu, J. (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet*, **31**, 205–9. 13

Han, M., Demuth, J., McGrath, C., Casola, C. & Hahn, M. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome Res*, **19**, 859–67. 15

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple significance testing. *Biometrika*, **75**, 800–803. 49

Hou, Z., Romero, R. & Wildman, D. (2009). Phylogeny of the ferungulata (mammalia: Laurasiatheria) as determined from phylogenomic data. *Molecular phylogenetics and evolution*, **52**, 660–664. 43

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. 11

Hubisz, M., Lin, M., Kellis, M. & Siepel, A. (2011). Error and error mitigation in low-coverage genome assemblies. *PLoS One*, **6**, e17034. 11, 12

Hughes, A. (1999). *Adaptive evolution of genes and genomes*. Oxford University Press, USA. 46

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. 7, 11

Kircher, M., Stenzel, U. & Kelso, J. (2009). Improved base calling for the illumina genome analyzer using machine learning strategies. *Genome Biol*, **10**, R83. 7

Kosiol, C., Holmes, I. & Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Mol Biol Evol*, **24**, 1464–79. 29

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. 5, 14, 42, 47, 48

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O.A., Leung, F.C.C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C.C., Lam, T.T.Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M.W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.W., Yiu, S.M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G.K.S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J. & Wang, J. (2009). The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317. 47

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. 2

Lindblad-Toh, K., Garber, M., Zuk, O. & ,et al. (64 co-authors) (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*. 3

Loewe, L. & Charlesworth, B. (2006). Inferring the distribution of mutational effects on fitness in drosophila. *Biol Lett*, **2**, 426–30. 4

Lynch, M. & Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–5. 13

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. 10, 28

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. 2

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. 2, 3

Markova-Raina, P. & Petrov, D. (2011). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 drosophila genomes. *Genome Research*, **21**, 863–874. 28

Marques-Bonet, T., Ryder, O.A. & Eichler, E.E. (2009). Sequencing primate genomes: What have we learned? *Annual Review of Genomics and Human Genetics*, **10**, 355–386. 4

Massingham, T. & Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–62. 3, 5, 29, 31, 37, 48, 49

Mouse Genome Sequencing Consortium & Mouse Genome Analysis Group (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–62. 2

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. 24

Nielsen, R. & Yang, Z. (2003). Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol*, **20**, 1231–9. 5

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. 14, 46

Nikolaev, S., Montoya-Burgos, J., Popadin, K., Parand, L., Margulies, E., National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program & Antonarakis, S. (2007). Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc Natl Acad Sci U S A*, **104**, 20443–8. 24, 42

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, **23**, 263–286. 43

Pál, C., Papp, B. & Lercher, M. (2006). An integrated view of protein evolution. *Nat Rev Genet*, **7**, 337–48. 3

Popadin, K., Polishchuk, L., Mamirova, L., Knorre, D. & Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc Natl Acad Sci U S A*, **104**, 13390–5. 42

Rat Genome Sequencing Project Consortium (2004). Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521. 2

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. 10

BIBLIOGRAPHY

Rhesus Macaque Genome Sequencing and Analysis Consortium (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–34. 47

Sawyer, S.L., Wu, L.I., Emerman, M. & Malik, H.S. (2005). Positive selection of primate trim5α identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 2832–2837. 48

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. 9, 10, 28

Smith, S. & Donoghue, M. (2008). Rates of molecular evolution are linked to life history in flowering plants. *Science*, **322**, 86–9. 24

Storz, J., Hoffmann, F., Opazo, J. & Moriyama, H. (2008). Adaptive functional divergence among triplicated alpha-globin genes in rodents. *Genetics*, **178**, 1623–38. 10

Stouffer, S., DeVinney, L. & Suchmen, E. (1949). *The American soldier: Adjustment during army life*, vol. 1. Princeton University Press, Princeton, NJ. 50

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. 10

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. 12

Wang, Y. & Gu, X. (2001). Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics*, **158**, 1311–20. 10

Whelan, S. & Goldman, N. (2004). Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*, **167**, 2027–43. 29

Whitlock, M. (2005). Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *J Evol Biol*, **18**, 1368–73. 50

Wyckoff, G.J., Wang, W. & Wu, C.I. (2000). Rapid evolution of male reproductive genes in the descent of man. *Nature*, **403**, 304–309. 47

Yang, Z. (2005). The power of phylogenetic comparison in revealing protein function. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 3179–3180. 47

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**, 1586–91. 20

Yang, Z., Kumar, S. & Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**, 1641–50. 20

Zaykin, D., Zhivotovsky, L., Westfall, P. & Weir, B. (2002). Truncated product method for combining p-values. *Genet Epidemiol*, **22**, 170–85. 50

Zhang, J., Zhang, Y. & Rosenberg, H. (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet*, **30**, 411–5. 47

Zhang, J., Nielsen, R. & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, **22**, 2472–9. 47