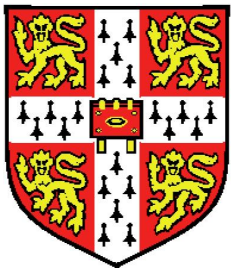
Sitewise error and constraint in mammalian comparative genomics



Gregory Jordan

European Bioinformatics Institute

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

September 29, 2011

Contents

Todo list

Chapter 1

Introduction

1.1 Introduction

Write an i
tory page
Keep it br
some old s

1.2 Biological background

[All of evolutionary biology is about understanding the history of evolution, and is thus tied to the circumstances under which such evolution occurred. Thus, a brief overview of that history is relevant and useful in this thesis.]

1.2.1 The evolutionary history of vertebrates and mammals

[Write a quick summary of the genome evolution of vertebrates and mammals. Mention 2R duplication, genome size growth, transposable elements.]

1.2.2 Mammalian population structure, adaptation and natural selection

[Introduce the concept of adaptation (molecular vs. morphological / ecological), the varied behavioral characteristics of modern day mammals (focusing on mammalian superorders and great apes, as expanded in mammals and gorilla chapters), and the impact of population structure / population size on the efficacy of natural selection.]

1.3 Mathematical methods for genomic analysis

[Introduce the importance of imperfect replication / copying as the substrate of evolution and a very convenient phenomenon for mathematical analysis. The balance between randomness and structure in evolutionary models.]

1.3.1 Codon models of evolution

[Introduce the idea of modeling protein evolution as a markov process acting on codon sequences: the incorporation of mechanistic parameters for Ts:Tv bias (kappa), dN/dS ratio (omega), or empirical models a la . Talk about heterogeneity the idea that real data may strongly violate certain models.]

1.3.2 Hypothesis testing with likelihood ratio tests

[Briefly introduce the idea of nested models and likelihood ratio tests (used for PAML in Slrsim and Gorilla chapters, and for SLR in Mammals chapters)]

1.3.3 Detecting purifying and positive selection in protein-coding sequence

[Briefly run through the history of detecting purifying / positive selection in genes and sites. Mention history of PAML models, alternative approaches, and fully describe SLR's approach.]

1.3.3.1 The Sitewise Likelihood Ratio test

SLR implements a method specifically designed for sitewise estimates which has been shown in sim- ulations to perform as well as or better than PAMLs sitewise random sites models (Massingham and Goldman, 2005). SLR models codon evolution as a continuous-time Markov process where substitutions at one site are independent of substitutions at all other sites. No assumptions are made regarding the distribution of ratios within the alignment. The value of is considered to be an independent parame- ter at each site: after first optimizing shared parameters using the whole alignment, SLR uses the shared parameters and the data at each alignment site to calculate a sitewise statistic for non-neutral evolution. This statistic is based on a likelihood-ratio test where the null model is

neutral evolution (= 1) and the alternative model is either purifying or positive selection (¡ 1 or ¿ 1, respectively). The raw statistic measures the strength of evidence for non-neutral evolution at each site; following Massingham and Goldman (2005) we use a signed version of the SLR statistic (created by negating the statistic for sites with ¡ 1) as the test statistic for positive selection.

1.3.4 Identifying biological trends in sets of genes and proteins

[Introduce the Gene Ontology and Pfam databases, which annotate genes or components of genes with structured ontologies of functions or domains, respectively. Introduce the methods for detecting enriched GO terms. Note problems and biases involved in the basic methodology and describe algorithms / corrections introduced to correct for certain biases: topGO for hierarchical GO structure and goseq for element length.]

1.3.5 Correcting for multiple testing in genome-scale datasets

[Note the issue of correcting for multiple testing in genome-scale datasets. Clarify the differences between nominal p-value, family-wise error rate, FDR. Provide examples of when / where certain methods may be more applicable than others.]

Chapter 2

The effects of alignment error and alignment filtering on the sitewise detection of positive selection

2.1 Introduction

2.1.1 Methods for detecting sitewise positive selection

2.1.2 Substitution and indel processes in simulating protein-coding sequence evolution

2.2 Models and parameters for simulating the evolution of mammalian genes

2.2.1 Distribution of selective pressures

2.2.2 Phylogenetic tree size and shape

2.2.3 Frequency and size distribution of insertions and deletions

2.3 Analysis of the alignment error simulation results

2.4 Methods for filtering alignments

2.5 Analysis of the alignment filtering simulation results

Chapter 3

The effects of alignment error and alignment filtering on detecting positive selection in genes

3.1 Introduction

3.1.1 Existing methods for detecting gene-wide positive selection

3.2 The application of sitewise estimates to the genewise detection of positive selection

3.3 Analysis of the genewise detection results

3.4 Conclusions and further work

Chapter 4

Curating a set of genome-wide orthologous mammalian gene alignments

4.1 Introduction

The first step in any evolutionary sequence analysis is the collection of homologous sequence data [TOCITE, 2011].

[The evolutionary history of vertebrate and mammalian genomes is diverse, full of genome duplications, and well-characterized relative to other species groups]

[Still, availability of orthologous coding alignments is limited, and orthology itself is often uncertain]

[To cite: Jun *et al.* [2009] — They give good reasons for why orthology inference is still difficult]

[To cite: Ruan *et al.* [2008] — Provides a non-Compara example of tree-based orthology pipelines, and show that their 17k gene trees only cover 84.5% of genes]

[Genomic alignments would seem to be ideal for use as an input source, except for the important issues of genome / segmental duplications]

[As preparation for the mammalian analysis presented in chapters 3 and 4, I undertook a small project to identify the best set of orthologs to use in an analysis of mammalian protein-coding genes. Part of the goal was to understand what taxonomic constraints best identify largely-orthologous groups of mammalian genes, and part was to evaluate whether low-coverage genomes showed high enough annotation quality to be included in

the analysis.]

4.2 Low-coverage genomes in the Ensembl database

The prevalence of missing sequence data and fragmented contigs in low-coverage genomes presents a unique set of problems for the generation of transcript annotations. In recognition of these differences, the procedure used by the Ensembl database to annotate genomes assembled from low-coverage data is distinct from the usual gene-building pipeline [Hubbard *et al.*, 2007]. Briefly, a whole-genome alignment is produced between the human genome and each low-coverage target, and gene models are projected from human to the target genome. Small frame-disrupting insertions or deletions within orthologous exons are corrected, and missing exons are padded with Ns in order to obtain the correct transcript length.

The inclusion of these error-correcting features allows intact, if not complete, coding transcripts to be generated for low-coverage genomes. The Compara gene family pipeline uses the set of transcripts from each species as its input [Vilella *et al.*, 2009], so the quality of the gene models from each species has a direct impact on the overall quality and accuracy of gene trees. Although the reliance on genome-wide alignments to, and gene annotations from, a reference genome could be criticised for potentially causing a bias towards the genomic properties of the reference, this approach is a reasonable workaround in the absence of higher-coverage sequence data or a painstakingly curated assembly. Furthermore, the gene model error-correcting features of the Ensembl pipeline are especially beneficial, making more complicated methods for correcting errors from low-coverage genomes such as those described by [Hubisz *et al.*, 2011] seem largely unnecessary.

4.3 The Ensembl Compara gene tree pipeline

All genomic data and gene trees used for this analysis were sourced from version 63 of the Ensembl Compara database [Flicek *et al.*, 2011; Vilella *et al.*, 2009]. Although a complete description of the design, implementation, and validation of the pipeline behind the Ensembl database is beyond the scope of this thesis, I will briefly outline the major aspects of the approach, focusing on a few details which are relevant to the current sitewise analysis and the ensuing discussion.

The Compara pipeline begins with a set of protein-coding transcripts collected from

each individual species' annotation database. This step is not exactly straightforward, as the prevalence of alternative splicing in Eutherian mammals makes it common for a single gene to harbor many different transcript structures. In terms of biology and evolution, alternative splicing is a very interesting phenomenon. Tightly linked to the evolutionary innovation of regulatory control and tissue-specific gene expression, the existence of multiple transcripts per gene is one of the likely substrates of biological and developmental complexity within vertebrates and mammals as compared to single-celled eukaryotes, which show less developmental complexity but largely similar numbers of genes [Csuros *et al.*, 2011]. Further evidence of the unique evolutionary characteristics of alternatively-spliced exons comes from molecular evolutionary studies which have shown such exons to show, on average, higher levels of evolutionary constraint, possibly owing to the importance of exonic splice enhancers in modulating the inclusion or exclusion of their associated exons [Parmley *et al.*, 2006].

However, in terms of organizing biological data, pervasive alternative splicing—with 34% of human genes containing at least two (and up to several dozen) transcripts per gene [Mironov *et al.*, 1999], showing tissue-specific and species-specific expression patterns, different levels of overall transcription, and sometimes comprising mutually exclusive exons—is somewhat burdensome. The first problem is the fact that primary data on alternative transcript structures (e.g., resulting from expressed sequence tags, RNA-seq, or proteomics experiments) are largely absent from most organisms with sequenced genomes. Even ignoring this lack of data, the task of incorporating multiple transcripts per gene into an evolutionary analysis is non-trivial, and leaves many unresolved questions open to debate: should all transcripts be treated as independent evolutionary entities, or should some form of meta-transcript be produced, comprising all possible transcripts for a given gene? Should expression levels and tissue-specificity be taken into account (as both factors have been correlated with evolutionary rate, e.g. [Koonin & Wolf, 2006; Zhu *et al.*, 2008])? And what is the expected evolutionary impact of the loss, gain, or modulation of the prevalence or tissue-specificity of a given exon or transcript in one lineage? Even a fairly shallow consideration of the topic quickly reveals layers of complexity that would quickly hinder many large-scale evolutionary analyses such as the current one, whose main goals are to understand the levels of evolutionary constraint of some subset of genes (or protein-coding sites) within some subset of species.

As a result of these difficulties, the current design of the Compara pipeline only incorporates one 'canonical' transcript per gene into the evolutionary analysis and the resulting

inferred gene trees. This reflects a conscious decision to sacrifice some biological fidelity for reduced design complexity and computational load (as the inclusion of multiple transcripts would inevitably require some amount of additional processing and/or calculation). Unfortunately, this only somewhat alleviates the problem, shifting the burden from "how to deal with multiple transcripts in a comparative setting" to "how to choose the best representative transcript for each gene." In the case of a gene with many transcripts of varying sizes containing many non-overlapping exons, the negative consequences of choosing a non-optimal transcript are clear: too short of a transcript could exclude important sequence information from the dataset, while transcripts with spurious exons (resulting from misannotation or erroneous experimental evidence for a transcript) could introduce potentially large amounts of non-orthologous, nonfunctional, or nonconserved sequence into the evolutionary analysis.

Fortunately, the consensus coding sequence (CCDS) project was initiated in 2005 to "identify a core set of human and mouse protein coding regions that are consistently annotated and of high quality" [Pruitt *et al.*, 2009]. Although the transcripts that satisfy these two criteria will not necessarily be the same as those which meet the desired definition of "the best representative transcript for use in an evolutionary study," the confidence that one can have in the quality and consistency of CCDS transcripts helps to reduce the prevalence of potentially damaging errors in the Compara pipeline. Thus, in the current release (version 63), the "representative" transcript used for the Compara pipeline is chosen on the basis of (a) existence within the CCDS set of transcripts and (b) the total length of the transcript's coding sequence. The combination of these two factors can be expected to identify a reasonably representative transcript, at least for the human and mouse genomes. The situation will be similar for genomes whose Ensembl annotation is derived largely from synteny and orthology to human and mouse annotated genes, but two classes of genomes—those resulting from low-coverage sequencing and those from more distant species whose annotations are derived from largely independent data sources—will still suffer from some amount error in the form of poor transcript choice.

Once the set of canonical transcripts is chosen, the Compara pipeline performs an all-against-all protein BLAST search (using the Washington University variant of BLAST) and clusters genes into groups of evolutionarily-related sequences using *hcluster_sg*, an implementation of a hierarchical clustering algorithm for sparse graphs. Sequences are aligned using MCoffee, a meta-aligner algorithm which combines the results from different aligners into one alignment using a maximum-consistency criterion [Wallace *et al.*, 2006].

The aligners used for the M-Coffee alignment include MAFFT [Katoh *et al.*, 2005], MUSCLE [Edgar, 2004], KAlign [Lassmann *et al.*, 2009], and T-Coffee [Notredame *et al.*, 2000]. Finally, the aligned sequences are input to TreeBeST, which infers a gene tree (including gene duplication and loss events) given a set of aligned sequences and a known species tree [Ruan *et al.*, 2008]. The type of the homology relationship between each pair of genes (e.g., one-to-one ortholog, one-to-many ortholog, within-species paralog) is determined using a simple set of rules based on the structure of the inferred gene tree and the annotation of ancestral nodes where a duplication event has likely occurred.

The Compara pipeline has been a part of the Ensembl ecosystem since its first introduction to Ensembl in release 42 [Birney *et al.*, 2006]. Remarkably, aside from slight tweaks to the protein clustering method and some changes in the exact aligners used, the pipeline has changed little from its original published form [Vilella *et al.*, 2009]. In part, this lack of change reflects the ease with which sets of vertebrate orthologs can be identified using the existing methodology, lying in stark contrast to the equivalent task in sets of insect or fungal genomes where divergence levels between extant sequences are much larger [Siepel *et al.*, 2005] and the shape of the underlying species tree may be uncertain and/or unknown [MacKenzie *et al.*, 2008], making the development of specialized methods or extensive manual annotation necessary [Kellis *et al.*, 2004; Rasmussen & Kellis, 2007]. This is equivalent to saying that Ensembl's pipeline, while not perfect in its orthology predictions or tree inferences (as indicated in a series of back-and-forth papers between Milinkovitch et al. [2010] and Villela et al. [2011]), has proved sufficiently accurate enough that an extensive reworking of the system has not yet been deemed necessary. Additional validation of this approach comes in the form of Treefam [Ruan *et al.*, 2008], a database of animal gene trees which applies a similar set of tools to infer gene trees from a more diverse set of genomes, with largely similar results.

[Something about Ensembl being directed at inferreing gene tree topologies, and not being vetted for use in estimates of selective constraint]

[Introduce the structure of the next few subsections: ways of massaging / filtering the Ensembl data to fit with the needs of the current project]

4.4 Identifying orthologous subtrees within large mammalian gene families

The first task in preparing the Ensembl data for sitewise analysis was to identify and extract a biologically meaningful set of orthologous mammalian subtrees from the set of gene trees within the Compara database. This was necessary because many Compara gene trees contain multiple sets of Eutherian orthologs linked by ancient gene duplication events, while I wished to study the evolution of each individual set of Eutherian orthologous genes. In other words, Compara gene trees are over-clustered with respect to the core set of Eutherian orthologs.

Evidence for this over-clustering comes from Table 4.2, which shows the number of root Compara gene trees which contain zero, one, or multiple genes in human, zebrafish and drosphila, as well as Figure 4.3, which shows the distribution of gene counts in the set of root Compara gene trees. The percentage of Compara trees with 2 or more human genes is strikingly high, at XYZ%. If each Compara tree contained one single set of Eutherian orthologs, then the proportion of trees with multiple human gene copies could only be explained by an unrealistically high rate of gene duplication. A more parsimonious explanation would be that many Ensembl trees represent not one group of Eutherian orthologs, but two or more sets of Eutherian orthologous gene trees joined by one or more ancient duplications. This explanation is further supported by Figure **??**, which shows concentrations of gene counts centered roughly around whole-integer multiples of the number of vertebrate species present in the Ensembl database (shown as gray dotted lines).

The prevalence of over-clustered Eutherian orthologs in the Compara database is easily explained by a combination of the *hcluster_sg* algorithm used for the hierarchical clustering step, which uses only protein distances as its source of clustering information, and the wide range of protein evolutionary rates in the vertebrate genome. As I mentioned in the previous subsection, the Compara pipeline uses all-by-all protein BLAST E-value scores and the *hcluster_sg* algorithm to produce sets of sequences containing minimal average within-group E-values. No additional biological information, such as the source species of each sequence or the overall taxonomic coverage of each cluster, is used in identifying clusters, and no attempt is made to fit clusters to an expected model of orthologous gene evolution. On the one hand, the lack of additional information and assumptions allows the algorithm to remain simple and the clustering behavior to remain consistent across

different groups of genomes; on the other hand, a number of technical (in the sense of non-biologically meaningful) parameters and thresholds must be tuned in order to result in the desired cluster sizes and contents. Importantly, even after these parameters are tuned to perform well on the dataset as a whole, the reliance on protein distances alone means that fast-evolving proteins will be more likely to be under-clustered and slow-evolving proteins will be more likely to be over-clustered. Given that the protein evolutionary rate varies widely within a genome (e.g., in a study of amino-acid substitution rates of roughly 6,000 orthologous genes in 7 eukaryotic species, Koonin et al. [Koonin *et al.* 2004] found that the middle 90% of genes showed nearly fourfold variation in evolutionary rate), the excess of over-clustered orthologs in the Compara database is understandable and even somewhat expected.

I should note that my use of the phrase "over-clustered" refers only to over-clustering with respect to the current goal of analyzing independent sets of orthologous genes within Eutherian mammals. Certainly these large "over-clustered" trees, which represent a more distant evolutionary history than a single Eutherian orthologous group, are just as accurate with respect to the true evolutionary history of the genes as more narrow groupings would be. Furthermore, the inclusion of a deeper evolutionary context may sometimes be more useful to users of the Compara database, for whom an understanding of the overall evolutionary history of a gene may be the topic of primary interest.

Take for example the gene *NBEAL2* and its human paralogs, whose gene trees, exon structures and domain classifications were extracted from Ensembl v62 and summarized in Figure 4.1. A recent medical sequencing project identified *NBEAL2*, a gene of previously unknown function, as the putative causative gene for gray platelet syndrome, a predominantly recessive platelet disorder resulting in moderate to severe bleeding [Albers *et al.*, 2011]. It was important for the purpose of this study to ensure that the *NBEAL2* gene was both well-conserved across mammals and distinct from its paralogs. The Compara pipeline clustered *NBEAL2* with three of its closest paralogs into one tree (and similarly clustered four more distant *NBEAL2* paralogs into a separate tree), yielding two views which together showed both the full taxonomic coverage of the *NBEAL2* subtree and the large amount of separation between paralogs. Had each Eutherian ortholog been displayed independently in Ensembl (using the blue "Eutherian root" nodes in Figure 4.1), it would have been more difficult for a non-expert to make such claims regarding the evolutionary history of *NBEAL2* without further analysis. Conversely, had the Compara pipeline been even more inclusive in its clustering approach and identified the hypothetical deeper

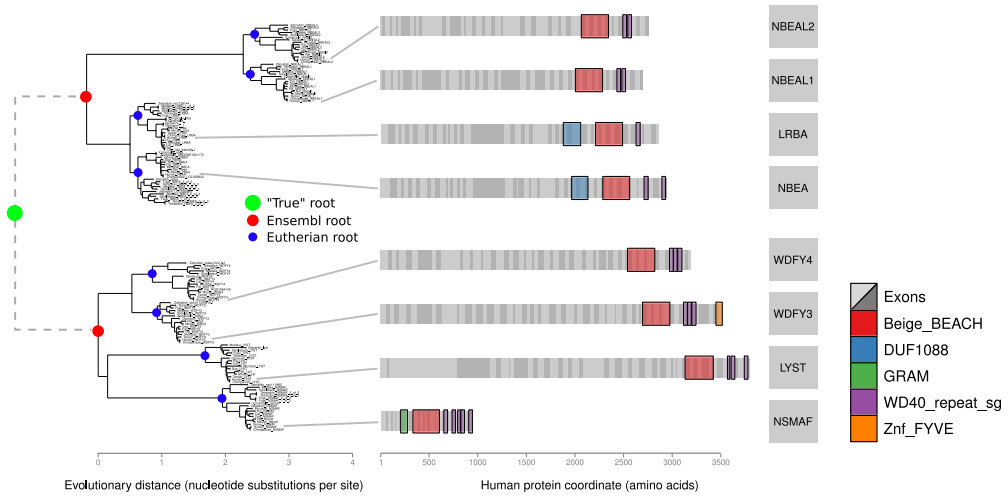


Figure 4.1: The evolutionary history of the human *neurobeachin-like 2* gene (*NBEAL2*) and its paralogs. Left, two phylogenetic trees from Ensembl Compara (release 60) are shown, summarizing the evolution of *NBEAL2* and its three paralogs (top) and *LYST*, a presumed distant paralog of *NBEAL2*, and its three paralogs (bottom) in 15 vertebrate species. The phylogeny shows that *NBEAL2* is taxonomically conserved and distinct from its paralogs. Red dots highlight the root nodes of Ensembl gene trees, blue dots highlight the root nodes of Eutherian orthologous subtrees, and a dashed line with a green dot represents the putative paralogous relationship (with a hypothetical root) between the two Ensembl gene trees. Right, the exon and domain structure of each human gene is shown: exons are displayed alternating shades of gray, and Pfam domain annotations are colored according to their Pfam identifier.

root connecting these two sets of trees (represented by the green node in Figure 4.1), the connection between these eight genes would have been even more immediately apparent.

For the purposes of the current mammalian sitewise analysis, however, it was important to isolate individual mammalian gene trees for further processing and sitewise analysis. To this end, I designed a simple scheme for splitting gene trees into non-overlapping subtrees based on flexible taxonomic coverage criteria.

I hypothesized that a relatively simple set of rules based on taxonomic coverage would be sufficient to identify most largely orthologous mammalian subtrees. This hypothesis was based on two well-established observations in mammalian genomes. First, the existence of two rounds of whole-genome duplication preceding the evolution of vertebrates [Dehal & Boore, 2005] suggested that many of the ancient duplication events contained within Ensembl gene trees occurred before the divergence of mammals, making it possible to cleanly separate out taxonomically complete mammalian subtrees in the majority of cases. This would not be possible if duplication events were common and spread evenly throughout the mammalian tree; if that were the case, many duplication events would

have occurred after the divergence of some or all of the major mammalian groups, resulting in a larger proportion of mammalian genes with "internal" duplications and, thus, fewer singly orthologous trees with high taxonomic coverage. Second, the overall low rate of gene duplication and loss in mammals [Demuth *et al.*, 2006] (excluding, of course, the aforementioned whole-genome duplication events) predicts that few mammalian gene trees will be subject to one or more gene duplication or loss events. In other words, most mammalian gene trees should contain sequences from a majority of mammalian species, so the effectiveness of using taxonomic coverage to identify mammalian subtrees should be largely unaffected by individual (i.e., post-2R) gene duplication or loss events. The potential utility of taxonomic coverage was further bolstered by the star-like shape of the mammalian tree: star-like trees contain more branch length within terminal lineages than ladder-like trees with an equivalent total branch length, making it less likely that a gene duplication or loss event (if such events occurred randomly throughout the mammalian tree) would result in a significant disruption to the taxonomic coverage of the gene tree.

The taxonomic-based tree splitting scheme works as follows. For every internal node N of each Compara gene tree, the taxonomic coverage (TC) was calculated for several vertebrate clades. The TC for node N and clade C is given by $TC(N, C) = species(N)/species(C)$, where $species(N)$ is the number of unique species represented by the sequences beneath node N and $species(C)$ is the number of species within the vertebrate clade C. The tree is traversed from root to tip, and if a given set of TC constraints (referred to as the subtree constraints) are satisfied by both subtrees below node i, then the tree is split into two subtrees at node i (with the new trees having root nodes placed at the two child nodes, i_a and i_b). The traversal continues recursively until every node is tested. If only the original root node satisfies the subtree constraints, then the entire Compara tree is included in the resulting tree set; if the entire Compara tree fails to satisfy the subtree constraints, it is excluded altogether.

I chose a variety of subtree constraints based on the structure of the vertebrate phylogeny, all of which were run against the 18,613 gene trees within the Compara database to generate several genome-wide sets of subtrees. Table 4.1 shows the details of the various subtree constraints I used; the clade names (e.g., $TC(Primates)$) are used to refer to sets of species contained within the Ensembl database, as defined by the NCBI taxonomy. The NCBI taxonomy of species contained in Ensembl is shown in Figure 4.2.

For the Ingroup and Outgroup categories of subtree constraints, a TC value of greater than 0.6 was required for a single taxonomic clade. If the required TC value for a clade

were set to 1, then all subtrees containing deletions in any species within the clade of interest would be rejected. On the other hand, requiring a TC value of less than 0.5 would allow for a truly singly-orthologous tree to be split into two subtrees, with one tree having a TC below 0.5, and the other tree (containing the other half of the species) also having a TC below 0.5. Thus, 0.6 seemed to be a reasonable TC requirement for isolating subtrees with reasonable taxonomic coverage while allowing for some amount of gene deletion.

Two additional types of constraints were designed for use in the MammalSubgroups and MammalSubgroupsPlusOutgoup methods. Inspired by the alignment filtering method from Pollard et al. [2010], which required sequence data from all three major mammalian clades (Primates, Glires, and Laurasiatheria) to be present for a column to pass through the filter, the TC_{all} constraint requires that the TC for all of the included clades is above a given threshold. To complement the TC_{all} constraint, the TC_{any} constraint requires that the TC for any of the included clades is above a given threshold. These more complicated methods were included in the analysis in case the simpler TC constraints within the Ingroup and Outgroup categories did not perform satisfactorily.

The methods within the Orthologs category of subtree constraints were implemented separately from the rest. Instead of splitting Compara trees based on taxonomic criteria, the subtrees in the Orthologs category were defined from the sets of genes annotated by Ensembl as orthologs to each gene from a given source species. Thus, for each gene from the source species, the Compara subtree containing all of the Ensembl-annoated orthologs was extracted and stored; this was guaranteed to yield exactly one subtree for every gene in the source species. I chose to include human, mouse, zebrafish, and drosophila were chosen as source species for testing. This approach differs from the tree-splitting strategy in two ways: first, it makes use of the orthology annotations resulting from Ensembl's orthology pipeline, and second, it does not guarantee that each subtree contains a completely unique set of genes. For example, a gene which was recently duplicated in humans would yield two subtrees, one for each human paralog, with identical sets of non-human genes in each tree. Although the orthology-based method might be useful when an evolutionary study is focused on a specific target or reference species, as is often done with human and mouse due to their finished genome sequence and high-quality annotation, I considered it to be less applicable to the current study due to the potential for introducing reference genome-specific biases, such as over-representation of genes with gene family expansions in the reference species or non-representation of genes which have been deleted in the reference species. Still, I expected that the sets of subtrees resulting from the Ensembl ortholog

Method		
Category	Name	Constraints
Ingroup	Primates	$TC(Primates) > 0.6$
	Glires	$TC(Glires) > 0.6$
	Laurasiatheria	$TC(Laurasiatheria) > 0.6$
	Sauria	$TC(Sauria) > 0.6$
	Fish	$TC(Clupeocephala) > 0.6$
Outgroup	Eutheria	$TC(Eutheria) > 0.6$
	Amniotes	$TC(Amniota) > 0.6$
	Vertebrates	$TC(Vertebrata) > 0.6$
	Fungi/Metazoa	$TC(Fungi/Metazoa) > 0.6$
Subgroups	MammalSubgroups	$TC_{all}(Laur., Glires, Primates) > 0.1$
	MammalSubgroupsPlusOutgroup	$TC_{all}(Laur., Glires, Primates) > 0.1$ AND $TC_{any}(Sauria, Clupeo., Ciona, Marsup.) > 0)$
Orthologs	Human Orthologs	
	Mouse Orthologs	
	Zebrafish Orthologs	
	Drosophila Orthologs	
Root Nodes	Ensembl Roots	

Table 4.1: Subtree constraints used for identifying Eutherian orthologous subtrees. Ensembl gene trees were split into subtrees based on taxonomic coverage (TC) requirements at internal nodes. Laur. - Laurasiatheria; Clupeo. - Clupeocephala; Marsup. - Marsupiala

annotations would serve as a useful reference with which to compare the other TC-based methods.

4.5 Analysis of genome-wide sets of orthologous mammalian trees

The subtree splitting schemes described in the previous subsection were applied to the 18,607 root gene trees from the Ensembl database. In this and the next section I will describe the resulting sets of trees and subtrees, discuss what they reveal about the evolutionary history of vertebrates and the feasibility of using taxonomic coverage to isolate orthologous trees for sitewise analysis, and finally, explain my reasoning for deciding to use the subtrees based on the Eutherian taxonomic coverage for the subsequent sitewise analysis.

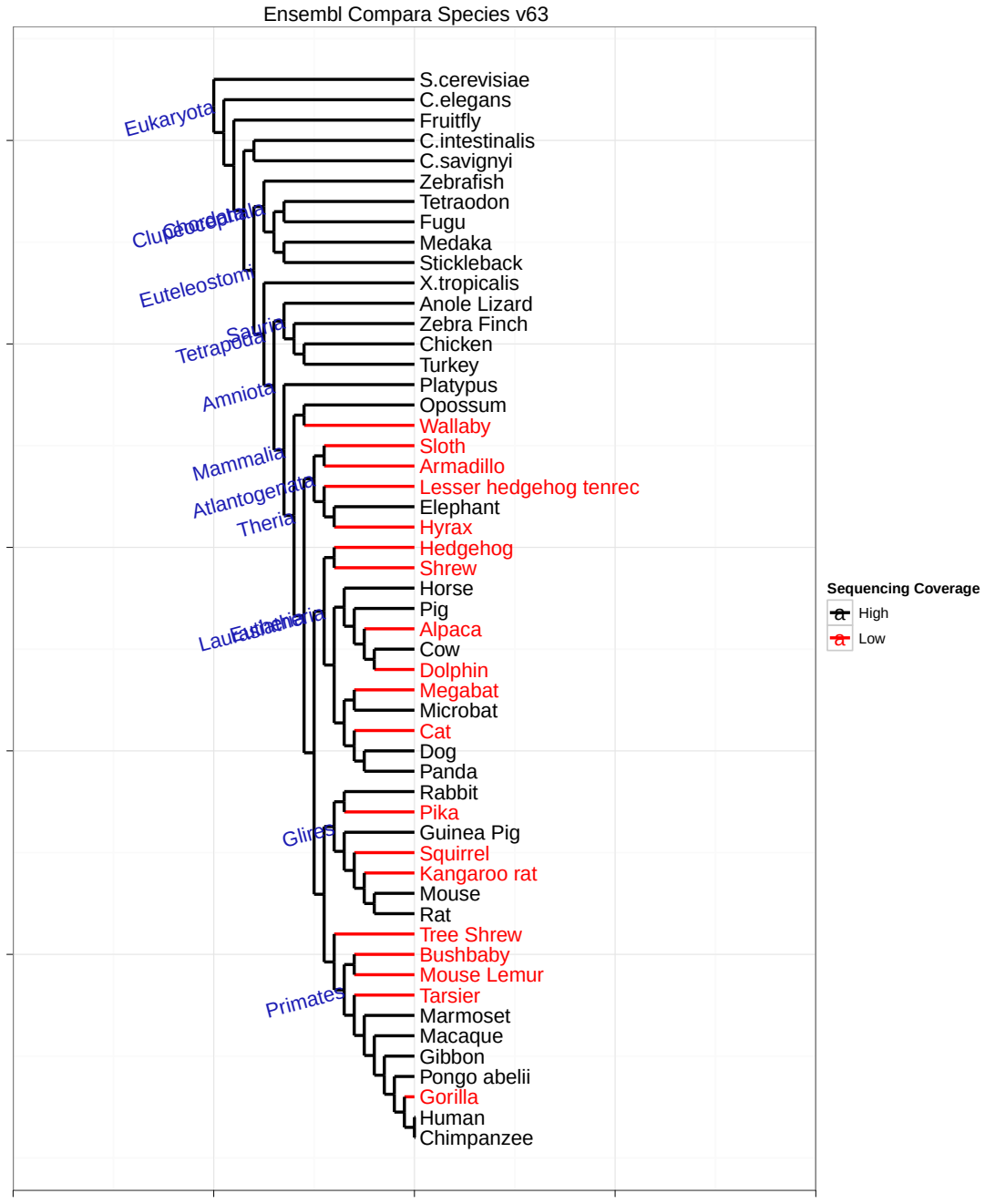


Figure 4.2: The NCBI taxonomy of species within the Ensembl Compara database. Note that branch lengths are not drawn to scale. Low-coverage genomes are labeled in red, high-coverage genomes are in black. Selected internal nodes, used are labeled in blue.

Tree Set	Count	Med. Size (Min / Max)	N50	Human Content 0	1	2+	Human Total	Med. MPL	Med. Species
All	18607	15 (2 / 400)	139	0.50	0.30	0.20	19995	0.55	8
($\leq$ 15)	9378	3 (2 / 15)	5	0.92	0.08	0.00	809	0.04	2
($>$ 15)	9229	54 (16 / 400)	146	0.07	0.53	0.40	19186	1.04	47

Table 4.2: Summary of the set of Ensembl Compara root trees. The 'Human Content' columns represent the fraction of trees which contain the indicated number of human genes, and 'Human Total' is the total number of human genes contained within the tree set. 'Med. Species' is the median species count across all trees. Med. - median, MPL - mean path length

4.5.1 The set of root Compara gene trees

Table 4.2 presents a summary of the set of root Compara gene trees and the subsets of trees with more or fewer than 15 sequences.

It is somewhat surprising that nearly half of all Compara gene trees contain few sequences: 9,378 out of 18,607 root trees constitute fewer than 15 sequences. Given the protein-based clustering performed by the Compara pipeline, one might expect many of these small trees to represent portions of larger fast-evolving gene trees whose high sequence divergences made the BLAST search step inaccurate or caused clustering via the *hcluster_sg* algorithm to be ineffective. Alternatively, these small clusters might have resulted from exceptional lineage-specific gene duplications or pseudogenes mis-annotated as genes, creating tight clusters of very closely-related transcripts that were identified by *hcluster_sg* as independent gene trees. Some evidence for the latter scenario comes from the median species counts and mean path lengths of the smaller versus larger trees. The subset of small root trees has a median species count of 2 compared to 47 for the large subset, indicating that the smaller trees encompass sequences from a very small taxonomic range. Furthermore, the median MPL for small trees is 0.04 compared to 1.04 for the large subset, revealing a much smaller level of sequence divergence within each tree. Together, these summary statistics indicate that the smallest trees in the Compara database consist of highly species-specific, closely-related proteins that are likely artifactual gene annotations.

Despite the existence of many small trees in the Compara database, they comprise only a small fraction of all protein-coding sequences. Only 4% of the human gene set—which we expect to be well-annotated and to contain few false positive genes due to the high level of manual curation and the large amount of continued scrutiny—is contained within the subset of small trees. This indicates that whatever process is causing the Compara

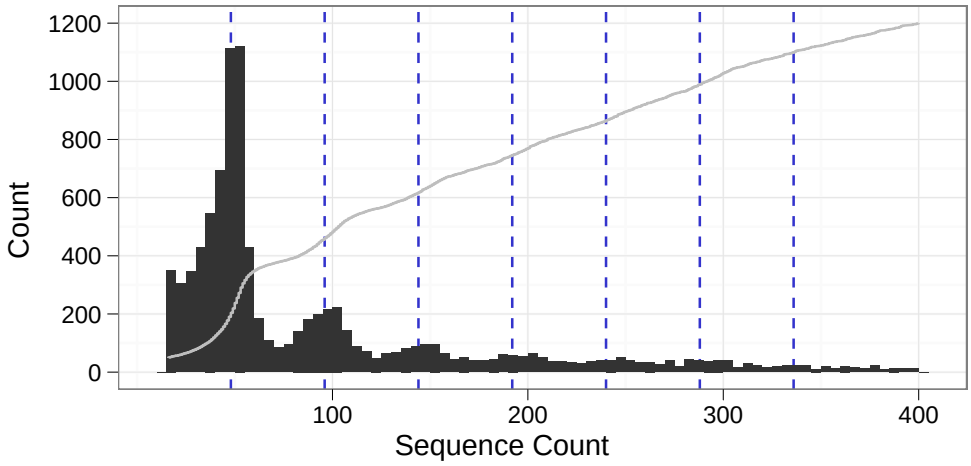


Figure 4.3: Sequence counts for the set of root Compara trees. Black bars show a histogram of sequence counts in bins of width 5, and the gray line shows the cumulative fraction of sequences contained within trees of that size or smaller. For clarity, 9,378 trees with 15 or fewer sequences are not shown. Dashed blue lines are drawn at integral multiples of 48, the number of vertebrate species within Ensembl.

pipeline to yield such a high number of small gene trees has not had too much of an impact on the placement of the most confident set of protein-coding genes within the database of root gene trees.

A closer examination of the distribution of tree sizes in the set of root Compara trees presents a clear view of the over-clustering of mammalian orthologous trees. The black bars in Figure 4.3 show the distribution of sequence counts for all trees with more than 15 sequences, with vertical dashed lines overlaid at multiples of 48 (the number of vertebrate species in Ensembl release 63). The highest peak of the histogram is at or slightly above 48 sequences, with the tree counts quickly diminishing at larger sizes. Weaker, but still discernable, peaks appear at larger tree sizes, with the location of these echo-like peaks corresponding closely to the second, third and fourth multiples of 48. The pattern of recurring peaks becomes indistinguishable at sizes above 200, but there is still a long tail of large trees extending out to a maximum size of 400 sequences. Overall, the distribution of tree sizes provides good support for the situation described above, with the Compara pipeline often clustering together two or more largely-orthologous gene trees sharing ancient homology.

It is also interesting to characterize the set of Ensembl trees by the proportion of all sequences which are covered by trees of a given size or smaller. This value is plotted in

Figure 4.3 as a gray line. First, one can see that trees with fewer than 15 sequences (which were excluded from the plot but inlcuded in the calculation of the cumulative fraction of sequences) represent a trifling fraction of the sequences within Compara; this is similar, but not identical, to the above-mentioned calculation that 4% of human genes are contained within these smaller trees. Second, the steady slope of the cumulative curve contrasts with the declining height of the histogram. This results from the increasing number of sequences encompassed by each of the larger trees: although there are relatively few trees with more than 300 sequences, together they contain around 10% of all protein-coding genes in Ensembl. Two points along this cumulative plot are of particular interest. First, one can identify the fraction of vertebrate proteins which exist as identifiable paralogs. Looking at the value along the x-axis where the largest bump in the histogram ends, at around 75 sequences, one can see that in total around 30% of proteins are covered by trees of 75 sequences or fewer. Since the pattern of bumps in the histogram correlate well with the number of Ensembl vertebrate species, it would be reasonable to state that 70% of vertebrate proteins are contained within large gene trees containing sequence-based evidence of ancient paralogy. Second, a look-up in the reverse direction can identify the tree size at which 50% of sequences are clustered. This value represents the size of tree that an "average" protein might be clustered in, and in some ways is a more accurate characterization of the set of gene trees than the median tree size. A similar calculation is often performed to characterize the size distribution of contigs (contiguous sequence blocks) within a genome assembly. This statistic, referred to as the N50 length, is the contig length for which 50% of bases are contained in contigs of that size or larger [Miller *et al.*, 2010]. For the Ensembl root trees, the N50 tree size is 139, slightly less than three times the number of vertebrate species. The N50 tree size is shown for the root trees in Table 4.3 and in the table for taxonomically-defined tree sets below.

Another way to characterize the distribution of gene trees is across the taxonomic space. A question of particular interest to the identification and analysis of mammalian orthologs is whether levels of gene presence and absence are consistent across different species and different levels of assembly quality. To investigate this question in the context of the root Ensembl trees, data were collected by counting the number of sequences from each species contained within each gene tree. Results were tabulated for each species and are presented in Figure 4.4, showing the number of trees containing 0, 1, 2, or more than 3 genes from each of the 53 species in Ensembl. Comparing the range of values in the panels for each copy count (labeled 0, 1, 2 and 3+), one finds that most trees (8,000-11,000

within vertebrates) contain zero copies from a given species, fewer trees contain one copy (4,000-6,000) and several thousand contain two, three or more copies (ca. 1,000-1,500 for 2 copies and 1,500-2,000 for 3+). The plethora of trees with zero copies from a given species is again a result of the existance of many small, species-specific trees within the root Ensembl set. Similarly, the high number of trees with many copies from each species reflects the clustering of multiple orthologous sub-trees together.

A comparison of values across the range of species in Figure 4.4 reveals that the zero-copy count tends to increase along with evolutionary distance from human, while the 1, 2 and 3+ copy counts tend to decrease as the distance from human increases. Both trends are most striking at the distant end of the tree where the five non-vertebrate species begin. For the increase in zero-copy trees and the decrease in single-copy trees, the strength of the trend at the highest level of divergence can be partly explained by the very long branch lengths connecting those species to each other and to the more well-represented vertebrate clade: the distance-based clustering algorithm might reasonably be expected to produce more false negatives in longer branches for a number of reasons including the behavior of the *hcluster_sg* algorithm, inaccurate BLAST E-values at larger distances, and heterogeneity in evolutionary rates across lineages [Whelan, 2008]. However, the dearth of 2 and 3+ copy counts in non-vertebrates is most likely a signal resulting from the 2R event at the basal vertebrate lineage, with the non-vertebrate species strongly depleted of multi-copy duplicates compared to their vertebrate relatives.

It is slightly concerning that human and its close primate relatives contain fewer zero-copy genes and more one-copy and two-copy genes than any other group of vertebrates in the set of Ensembl trees. There is no *ab initio* biological reason to expect this to be the case, and I suspect that the existence of such a pattern, which is fairly small in effect, is due to the widespread reliance on human annotation and protein experimental data in the annotation of non-human genomes. There is one region where this trend does not appear to be the case: in the 3+ copy count for the fish species, which is instead a result of gene duplicates retained after the third round of genome duplication which occurred in the teleost ancestor [Jaillon *et al.*, 2004]. The signal resulting from the teleost genome duplication event is clearer in the sets of taxonomically-defined subtrees, so I will defer its discussion to the next subsection where those sets of trees are described.

Finally, the differences in copy counts between species with low- and high-coverage genome sequences show the tendency of low-coverage genome sequences to yield false negatives in the gene annotation, as low-coverage species contain more zero-copy, roughly

the same number of one-copy, and noticeably fewer multi-copy genes than high-coverage species. These clear effects of low sequencing coverage show that gene absence in low-coverage genomes should not be taken as evidence for actual gene loss and that gene duplications are systematically underrepresented in low-coverage genomes. The former point was emphasized in a recent critical analysis of the effect of low-coverage genomes on gene duplication inference [Milinkovitch *et al.*, 2010], but the latter point was largely ignored. Again, this signal is also stronger in the more stringent set of mammalian orthologous subtrees and will be revisited in the next subsection.

The preceding analysis of the set of root Ensembl trees, in which I characterised the distribution of trees with respect to size (i.e. sequence count) and across the taxonomic space, showed that despite the over-representation of small, species-specific trees, most sequences are contained in trees with biologically plausible sizes given the history of vertebrate genome duplications. The tree-based equivalent of the N50 statistic was developed for summarizing the distribution of differently-sized trees, and two main views of this distribution were introduced (in Figures 4.4 and 4.3), providing evidence for the clustering of paralogous mammalian sub-trees and for species-based and genome coverage-based trends in the breakdown of gene copy counts within these trees.

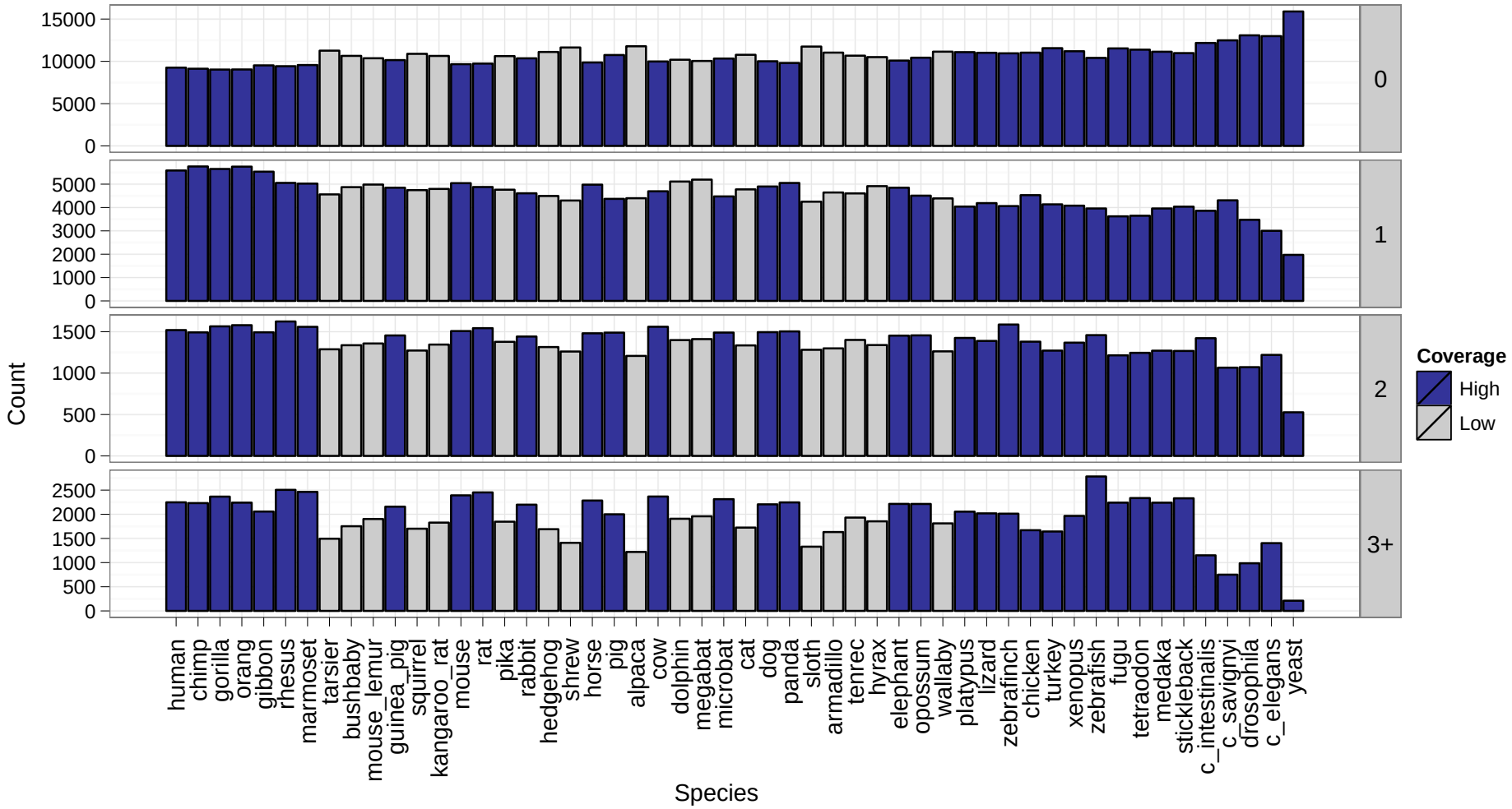


Figure 4.4: Taxonomic distribution of gene copy counts for the root Ensembl trees. The number of trees containing 0, 1, 2 or more than 3 sequences from each species is shown. Bars are colored blue and gray for species with high- and low-coverage genomes, respectively. Note that the y-axis scale is not the same for each panel.

4.5.2 Sets of subtrees defined by taxonomic coverage and orthology annotation

The sets of trees resulting from applying the various subtree extraction methods to the root Ensembl gene trees are summarised in Table 4.3, with the original Ensembl trees included at the bottom for comparison.

The Ensembl Roots and Drosophila Orthologs sets are two clear outliers, with much higher N50 values than any other set (139 and 125 vs. the next highest value of 56) and more trees with multiple human copies (0.20 and 0.43 vs. the next highest value of 0.14). In fact, the major differences between these sets are all attributable to the excess of small species-limited trees in the Ensembl Roots group: the Drosophila Orthologs set contains fewer trees than the Ensembl Roots (9,210 vs. 18,607) and a larger average tree size (60 vs. 15), closely resembling the set of Ensembl Roots with small trees removed as summarised in the third row of Table 4.2.

Within the Ingroup category of methods, the methods based on mammalian TC values (Primates, Glires and Laurasiatheria) produced largely similar sets of trees, with the Primate set containing around 2,000 more trees and covering around 1,000 more human genes than the other two sets. A reason for the higher number and human coverage of Primate trees is not immediately apparent, although it may speculatively be due to an excess of primate-specific gene trees that are not captured by non-primate TC-based criteria. Further investigation of the trees unique to this set might reveal the root cause of this slight discrepancy.

The Sauria and Fish tree sets stood in strong contrast to the mammal-based methods from the Ingroup category. The Sauria clade is represented by only four Ensembl species and diverged from the mammalian ancestor at an early point in the evolution of amniotes. The moderately lower number of trees (13,046 vs. 15,764 for Laurasiatheria) and the increased proportion of trees containing multiple human genes (0.14 vs. 0.09 for Laurasiatheria) are presumably consequences of the lower clade size, which could affect the TC calculation, and the long branch separating Sauria from the other vertebrate clades. The fish-based subtree constraint produced a strikingly different set of trees resulting from the impact of the teleost-specific whole genome duplication on the structure of fish gene trees. Although the Fish tree set contains a N50 value of 49 which is no different from the N50 of the other Ingroup sets, Table 4.3 highlights three major differences in the Fish set: it contains many more trees, a higher proportion of trees with zero human copies, and a lower total human gene count than the other Ingroup sets.

The reason for the drastically different Fish tree set is that the tree splitting procedure identifies largest non-overlapping subtrees that satisfy the given TC criteria. Genes that were duplicated in the teleost lineage and retained in duplicate form (as opposed to one or both copies being lost in either of the descendant duplicate chromosomes) would result in a gene tree with two teleost-specific subtrees, each containing a high TC value for the Clupeocephala clade. In this case, the splitting procedure would result in two small Fish subtrees, "missing" the single subtree of mammalain orthologs because two non-overlapping trees already exceeded the TC threshold of 0.6. If, however, one of the duplicate gene copies were lost, then the tree would resemble a typical singly-orthologous vertebrate gene tree, and the splitting procedure would select a subtree encompassing the entire vertebrate clade. It follows that the presence of small, teleost-specific gene trees in the Fish set is a signal of retained duplicate copies, and the size distribution of trees from the Fish set, shown in Figure 4.5, shows that several thousand trees fit the expected model. If we assume that all trees from the Fish subset which contain zero human copies, span 5 or fewer species, and contain 40 or fewer sequences are likely retained duplicate genes, a total of 6,980 retained duplicates are identified, yielding a retention rate of 17.5%, which is very much in line with a previously published estimate of 15% based on a comparison of tetraodon, fugu and zebrafish genes [Brunet *et al.*, 2006].

The sets of subtrees resulting from the Outgroup methods were of special interest, as the clades used to define these TC constraints contained all or nearly all of the mammalian species whose orthologous genes I wished to study. The resulting sets of subtrees showed little variation, owing perhaps to the large sizes of the clades and their similar composition. Each set contained between 15 to 17 thousand trees, N50 values of around 49, and greater than 90% of trees containing exactly one human sequence. These measures provided good evidence that the tree-splitting method was effectively isolating singly orthologous mammalian trees. Some slight trends were apparent, however, with the tree count decreasing, the proportion of trees with human duplications increasing, and the overall human gene coverage decreasing as the clade size used for the TC calculation increased. These trends could understandably be the result of the minimum required tree size increasing along with the clade size, ranging from 21 for Eutheria to 32 for Fungi/Metazoa.

The Subgroups methods did not appear to produce subtrees of any higher quality or more biological interest than the Outgroup methods. The MammalsSubgroups set was more numerous than the Outgoups sets, but the N50 was slightly lower (46 vs. 49) and the proportion of zero-humany trees was higher (0.18 vs. 0.01), suggesting that the additional

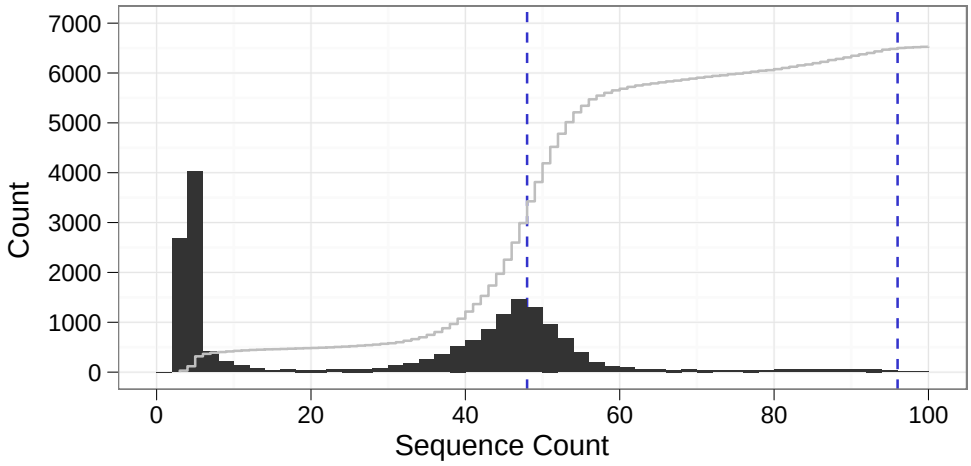


Figure 4.5: Sequence counts for the set of subtrees identified using the Fish clade taxonomic coverage constraint, showing an excess of small subtrees resulting from the teleost genome duplication. Black bars show a histogram of sequence counts in bins of width 2, a gray line shows the cumulative fraction of sequences contained within trees of that size or smaller, and dashed blue lines are drawn at integral multiples of 48, the number of vertebrate species within Ensembl. The 255 trees with more than 100 sequences are not shown.

trees were spurious results containing fragmented species coverage. The addition of an outgroup requirement to the MammalSubgroupsPlusOutgroup method produced a tree set more closely resembling the Outgroup methods, but the human gene coverage was lower than that for any Outgroup method despite the overall higher tree count.

Finally, the ortholog annotation-derived subtrees provided for an interesting comparison between three different ortholog sources and between the overlapping and non-overlapping sets of subtrees. As I mentioned at the beginning of this section, the *Drosophila* ortholog set was highly contrasted with the vertebrate sets due to the two rounds of whole genome duplication. There was minimal variation among the other ortholog sets, although it is interesting to note that Ensembl contained 21,873 mouse protein-coding genes while human contained only 19,991. Zebrafish, on the other hand, contained 24,540 genes, in line with the 17.5% rate of duplicate gene retension I estimated earlier. Overall, 76% and 81% of mouse and zebrafish genes have an apparent one-to-one ortholog in human, which is slightly lower than the 92% of Eutheria subtrees containing one human sequence.

Method			Med. Size		Human Content			Human	Med.	Med.
Category	Name	Count	(Min / Max)	N50	0	1	2+	Total	MPL	Species
Ingroup	Primates	17673	46 (6 / 388)	48	0.02	0.93	0.05	19024	0.68	42
	Glires	15786	48 (8 / 391)	49	0.02	0.90	0.08	17904	0.73	44
	Laurasiatheria	15764	48 (8 / 391)	49	0.01	0.90	0.09	17952	0.73	44
	Sauria	13046	49 (3 / 391)	51	0.06	0.80	0.14	14988	0.78	45
	Fish	18291	40 (3 / 391)	49	0.43	0.52	0.06	12183	0.58	38
Outgroup	Eutheria	16477	47 (21 / 391)	49	0.01	0.92	0.07	18343	0.71	43
	Amniotes	15899	48 (26 / 391)	49	0.01	0.91	0.08	18094	0.73	44
	Vertebrata	15634	48 (29 / 391)	49	0.01	0.91	0.08	17938	0.74	44
	Fungi/Metazoa group	14957	48 (32 / 391)	50	0.01	0.90	0.09	17623	0.76	44
Subgroups	MammalSubgroups	21179	40 (4 / 159)	46	0.18	0.79	0.03	18595	0.54	37
	MammalSubgroupsPlusOutgroup	17155	46 (5 / 159)	48	0.05	0.90	0.05	17640	0.71	43
Orthologs	Human Orthologs	19991	49 (2 / 367)	52	0.00	1.00	0.00	19991	1.07	44
	Mouse Orthologs	21873	50 (2 / 352)	54	0.10	0.81	0.09	28256	1.01	43
	Zebrafish Orthologs	24540	51 (2 / 392)	56	0.11	0.76	0.13	30063	1.14	46
	Drosophila Orthologs	9210	60 (2 / 399)	125	0.08	0.49	0.43	17625	1.22	50
Root Nodes	Ensembl Roots	18607	15 (2 / 400)	139	0.50	0.30	0.20	19995	0.55	8

Table 4.3: Summary of Ensembl subtrees identified using taxonomic criteria or Ensembl ortholog annotations. The set of Ensembl root trees ("Ensembl Roots") from Table 4.2 is included for comparison. Cells in numeric columns are shaded according to their value relative to other rows, with low values in white and high values in blue. The 'Human Content' columns represent the fraction of trees which contain the indicated number of human genes. 'Med. Species' is the median species count across all trees. Med. – median, MPL – mean path length

Figure 4.6 shows the taxonomic distribution of gene copy counts for the trees resulting from each of the subtree methods tested. By way of reference, the values shown in the separate panels of Figure 4.4 appear in Figure 4.6 as different-colored bars in the bottom panel. Although the various characteristics of each of the subtree methods have already been discussed at length, the taxonomic view reveals some salient features of the patterns of gene deletion and duplication within the tree sets and shows the pervasive impact of genome-wide duplications on the evolution of vertebrate genes. The large fraction of species with multiple copies in Drosophila Orthologs subtrees is a result of the two rounds of vertebrate genome evolution, while the elevated fraction of multi-copy fish trees in the Outgroup subtrees shows the impact of the teleost-specific duplication event.

Furthermore, the relative prevalence of zero-copy and multi-copy trees can provide some indication of whether gene deletion or gene duplication is a more common process in vertebrate genomes. Focusing on the four Outgroup subtree methods, the observation of a greater number of multi-copy trees than zero-copy genes, valid across all four subtree methods and throughout all mammalian species except platypus, can be interpreted as tentative evidence for a greater number of gene duplications than gene deletions in the evolution of mammalian genomes. This pattern does not hold for vertebrates more distantly related to human, however: vertebrates beyond opossum show a distinct and consistent increase in zero-copy trees, and birds appear to exhibit a slight clade-specific drop in the proportion of multi-copy trees. Of course, both of these trends could be methodological artifacts related to the *hcluster_sg* algorithm or to the methods used to assemble and annotate more distantly-related genomes.

The distributions in Figure 4.6 also reveal the pig to harbor a very high number of apparent gene deletions, unmatched by other mammalian species and nearing the proportion of zero-copy trees seen in platypus and more distant vertebrates. Given the consistently low proportion of zero-copy trees for other closely-related species, I would expect this number to change once a finished-quality pig genome sequence is included in the Ensembl pipeline [Archibald *et al.*, 2010].

In the end, the set of Eutheria subtrees was chosen as the final set for use in the downstream evolutionary analysis, due to the slightly larger number of trees and better coverage of human genes in the Eutheria set compared to the other Outgroup methods. The distribution of tree sizes for the Eutheria set of subtrees is shown in Figure 4.8 and the full taxonomic distribution of copy counts is included in Figure 4.7.

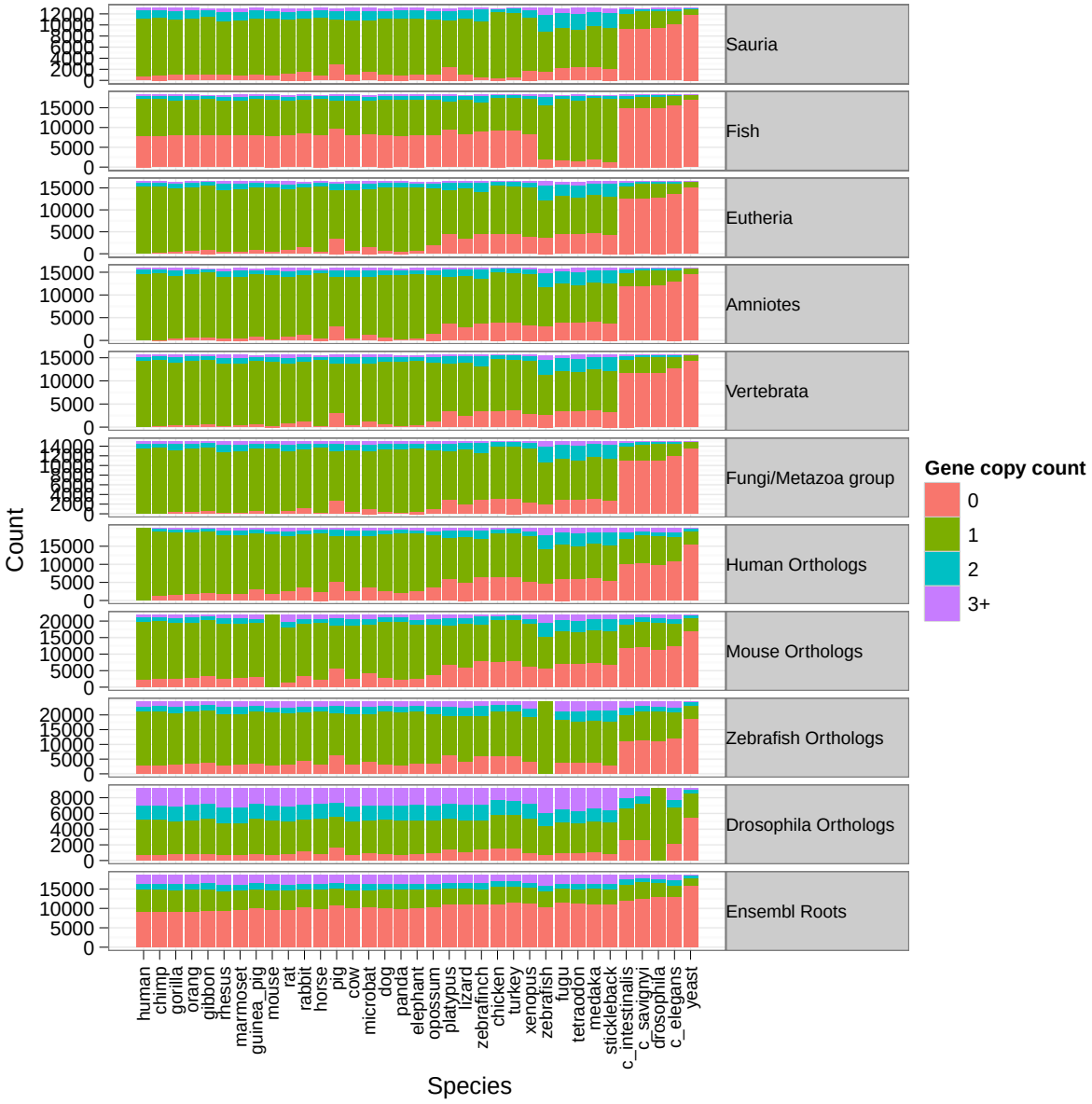


Figure 4.6: Taxonomic distribution of gene copy counts for different subtree methods. The numbers of trees containing 0 (red), 1 (green), 2 (blue) or more than 3 (purple) sequences from each species are shown as stacked colored bars. The Ingroup and Subgroups methods were omitted for clarity, as were species with low-coverage genomes. Note that the y-axis scale is not the same for each panel.

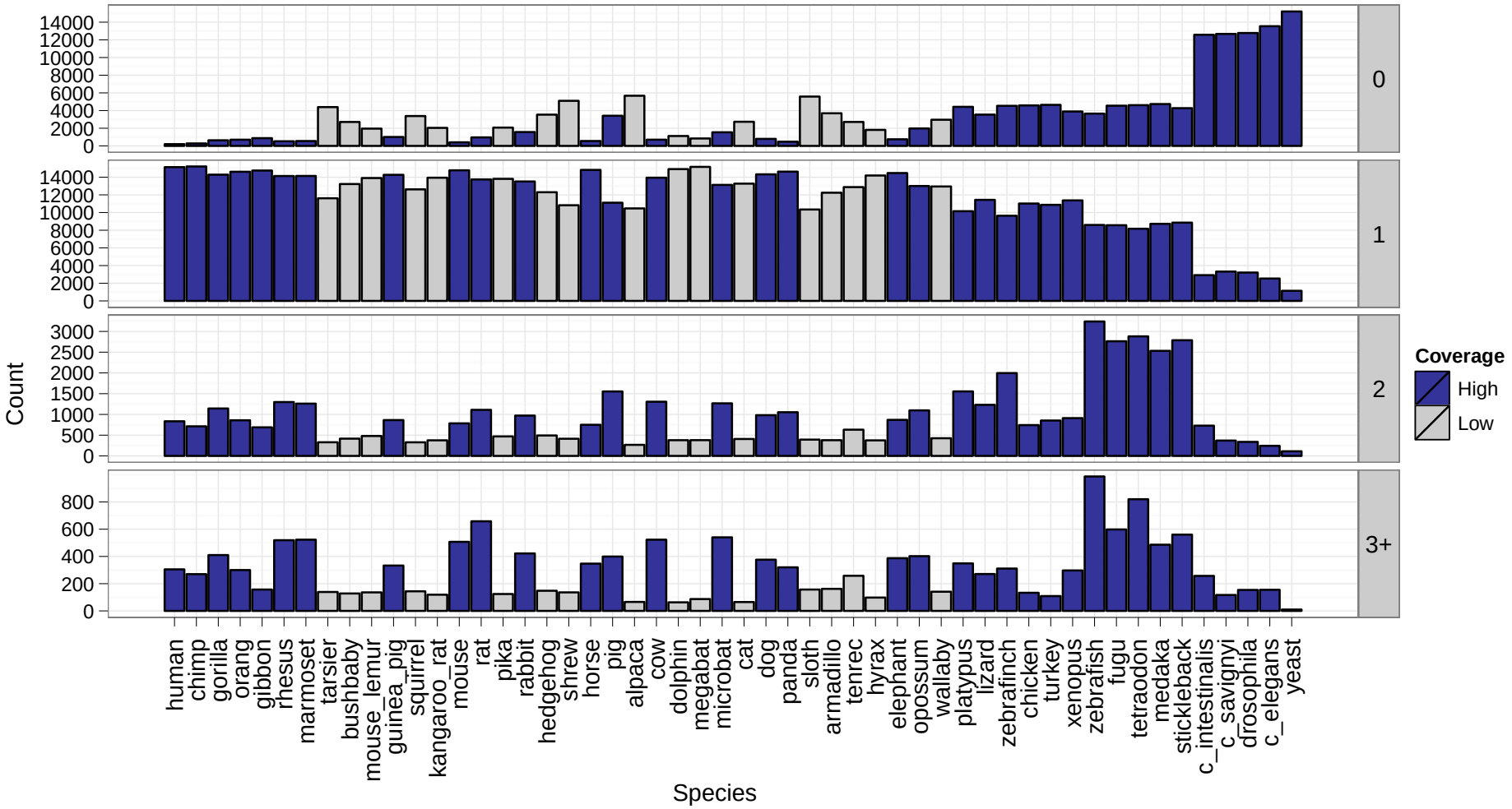


Figure 4.7: Taxonomic distribution of gene copy counts for the Eutheria subtrees defined by TC. The number of trees containing 0, 1, 2 or more than 3 sequences from each species is shown. Bars are colored blue and gray for species with high- and low-coverage genomes, respectively. Note that the y-axis scale is not the same for each panel.

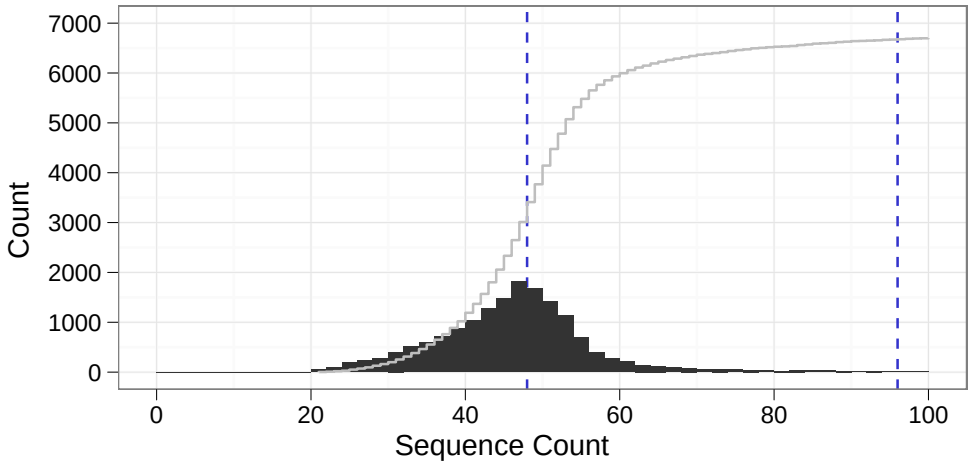


Figure 4.8: Sequence counts for the set of subtrees identified using the Eutheria clade taxonomic coverage constraint. Black bars show a histogram of sequence counts in bins of width 2, a gray line shows the cumulative fraction of sequences contained within trees of that size or smaller, and dashed blue lines are drawn at integral multiples of 48, the number of vertebrate species within Ensembl. The 230 trees with more than 100 sequences are not shown.

4.6 Conclusions

[Restate the goals: figure out which TC criteria could be used to isolate largely-orthologous mammalian gene trees]

Chapter 5

Patterns of sitewise selection in mammalian protein-coding genes

5.1 Introduction

5.1.1 The Mammalian Genome Project

A major goal of mammalian comparative genomics has been to quantify, identify and understand the fraction of the human genome that is under evolutionary constraint. The first non-human mammalian genomes showed at least 5% of the human genome to be under purifying selection [Lindblad-Toh *et al.*, 2005; Mouse Genome Sequencing Consortium & Mouse Genome Analysis Group, 2002; Rat Genome Sequencing Project Consortium, 2004], but the small number of genomes available limited the extent to which regions of evolutionary constraint could be identified. The Mammalian Genome Project, a coordinated set of genome sequencing projects initiated in 2005 and organised by the Broad Institute of MIT and Harvard, was designed with the primary purpose of increasing the accuracy and confidence with which regions of the human genome that have evolved under evolutionary constraint in mammals could be identified [Margulies *et al.*, 2007]. In line with this goal, 20 mammalian species were chosen for sequencing in order to maximise the amount of evolutionary divergence available for comparative analysis when combined with the 9 already available sequenced genomes [Margulies *et al.*, 2005]. To save on sequencing costs 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.

As the Mammalian Genome Project proceeded from its sequencing to analysis phase in late 2008, it became clear that the additional branch length afforded by the 29-species phylogeny would enable a number of improved evolutionary analyses beyond the identification of constrained non-coding regions. Among others, 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 Godlman group in analysing the ENCODE comparative sequencing data [ENCODE Project Consortium, 2007; Margulies *et al.*, 2007] and Massingham's work on a method and software program for sitewise evolutionary analysis [Massingham & Goldman, 2005], the group became involved in the protein-coding evolutionary analysis for the Mammalian Genome Project. This chapter describes my work on the project, which began in late 2008; all of the work described below was performed by me in consultation with members of the Goldman group (Nick Goldman and Tim Massingham), EnsEMBL team (Albert Vilella, Javier Herrero, Ewan Birney) and organisers and members of the Mammalian Genome Project (Manolis Kellis, Kerstin Lindblad-Toh, Mike Lin, Katie Pollard), and the major results from this analysis have been published [Lindblad-Toh *et al.*, 2011].

5.1.2 The Sitewise Likelihood Ratio test

As described in Section 1.3.1, differential survival of non-synonymous and synonymous mutations based on the degeneracy of the genetic code can be used as a source of information on the continued importance of mutations at a given protein-coding site over evolutionary time: a lower rate of non-synonymous substitution compared to synonymous substitution is indicative of purifying selection, or natural selection in favor of maintaining protein structure and function; equal rates of non-synonymous and synonymous substitutions is indicative of neutral selection, or no differential survival of protein-altering mutations; a greater rate of non-synonymous than synonymous substitution is indicative of positive selection, or natural selection in favor of protein-altering mutations.

Early evolutionary analyses of protein sequences showed large variation in the rates of amino acid change both within and between proteins resulting from the myriad structures and functions embodied by different proteins and protein domains [Kimura & Ohta, 1974]. Continued work suggested that the overall evolutionary rate Koonin & Wolf [2006] and the pattern of localised selective pressures Nielsen & Yang [1998]; Yang & Nielsen [1998] of a gene can reveal important insight into its role in the organism, establishing the study of

rates of non-synonymous and synonymous substitution in proteins as an effective method for using evolutionary information to investigate the functional characteristics of genes.

Maximum likelihood methods, introduced in Section 1.3.2, are commonly applied to biological sequence analysis due to their desirable statistical features... [[Finish out the paragraph with a quick recap of ML methods and their use in detecting selection]]

The Sitewise Likelihood Ratio (SLR) test is based on the mechanistic Goldman-Yang codon model of evolution, with an additional parameter for each site in the alignment representing the sitewise ω value. The inclusion of an additional parameter per alignment site makes the model extremely complex and difficult to optimise, but the dimensionality of the likelihood optimisation is reduced by making the assumption that the ω at each site does not contribute significantly to the overall likelihood, thus allowing for separate optimisation of the global parameters (including XYZ) across the whole alignment and the ω parameter at each site. In this way, SLR performs an approximate likelihood ratio test (LRT) for non-neutral evolution at each site...

5.1.3 Data quality concerns: alignment and sequencing error

The possibility that errors in the source alignments might cause false positives in the detection of sitewise positive selection was a major concern for this analysis. Although the SLR test and other sitewise maximum likelihood methods have been shown to be conservative for detecting positive selection 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.

Alignment error results from the difficulty of reconstructing the evolutionary history of sequences evolving with indels and can cause non-homologous codons to be placed in the same alignment column. In Chapters 2 and 3 I explored the tendency of multiple aligners to produce such errors, showing that PRANK_C alignments would be expected to introduce few falsely identified positively-selected sites resulting from alignment errors at mammalian-like divergence levels.

Errors resulting from the inclusion of incorrect genomic sequence in alignments was 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 potential 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 error.

As most codon-based inference methods assume independence between amino acid sites, I first consider the effect—in isolation—of a single spuriously-assigned homologous codon on the maximum likelihood estimation of ω. Two cases can be considered: a single sequence error causing one spurious substitution within a codon, and one or multiple sequence or assembly errors causing multiple spurious substitutions within a codon. In the case of a single spurious substitution, if we assume no large difference between the natural mutational process and the process that caused the erroneous mutation, then the effect would be to shift the estimated ω in the branch containing the error towards 1. The sequence error would be incorporated into the maximum likelihood optimisation as an additional neutral substitution, inflating the estimated substitution rate but not affecting the relative non-synonymous and synonymous rates. This effect may be biased towards higher or lower ω values if a significant difference exists between the neutral biological mutational process and the pseudo-mutational process causing the erroneous substitution. On the other hand, a codon with multiple erroneous bases may cause greater elevation of the inferred substitution rate and ω, due to the necessity of maximum likelihood methods to infer a multi-step path of single substitutions between the two codons on either side of a given evolutionary branch. The path estimated between two completely non-homologous codons depends on the estimated codon frequencies, the genetic code, and the nature of the process causing misalignment of nonhomologous codons; while a detailed investigation of the expected effect on inferred ω values is beyond the scope of my analysis, it is not unreasonable to expect a greater number of false positive PSSs resulting from codons with multiple erroneous bases than from codons with single errors.

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 effect on the detection of positive selection. On its own, a miscalled base would obviously result in a single spurious substitution. However, low-quality bases tend not to be uniformly distributed among or within sequence reads, which makes for a larger 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 alignment and frameshift effects. Assembly errors, which result in larger-scale structural errors including missing, repeated, shuffled or inverted sequence regions, are most prone to produce codons with multiple erroneous errors due to the large amount of contiguous sequence data being misplaced.

I also note the impact of the inference method and terminal branch length on false positives resulting from sequence errors, which can be understood in terms of the information most directly affecting the inference of a positively selected site or a positively selected gene for a given detection method. Both the branch-site test and the sitewise tests (including SLR and PAML M8) are sensitive to substitutions at a subset of alignment sites, but the branch-site test is specifically sensitive to substitutions along the foreground branches of interest while the sitewise tests detect positive selection only throughout the entire tree. In the latter case, the effect of spurious synonymous and non-synonymous substitutions from sequence data depends on the ratio of the species' terminal branch length to the branch length of the entire tree: a longer terminal branch gives greater weight to the erroneous sequence data, making false positives more likely to result. In the former case of the branch-site test, the potential effect depends on the location and length of the foreground branches. If the terminal branch leading to the spurious sequence is within the foreground and the total foreground branch length is small, then false positives could easily result; if, however, the terminal branch is outside of the foreground then it should have little to no effect on the FPR of the branch-site test. Interestingly, this suggests that branch-site tests where the foreground only consists of internal branches may be less prone to false positives from sequencing error than tests that include terminal lineages in the foreground model.

To summarise, 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 phylogenetic branch length associated with sequencing-based errors when applied to detecting sitewise positive selection. In some cases even a large amount of sequencing error should not produce a strongly elevated FPR (e.g., when the total branch length is large, when analysing all mammals or vertebrates) but in other cases it could potentially bias results (e.g., when the branch length is small and/or many low-quality genomes are included, as in the major mammalian sub-clades).

Simulation studies similar to those I performed in Chapters 2 and 3 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 larger negative effects. For estimates of false positives resulting from these types of sequence errors, an empirical approach seems more appropriate.

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.

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 assembly contained a number of sequencing 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. The detailed identification and quantification of error sources performed by Mallick et al. [2009] is especially useful for designing filters to apply to an analysis based largely on low-coverage genomes; in particular, their observation that clusters of chimpanzee-specific mutations were responsible for many false positives motivated the window-based filter I developed and applied here and in the analysis of primate genomes in Chapters **??** and **??**.

[Possible figure summarizing the types of error and potential effects on the inference?]

5.2 Preparing mammalian alignments for sitewise analysis

The effects of sequencing, annotation and alignment error on the results of comparative evolutionary analyses can be severe, with a high potential for false positive results when using sensitive evolutionary models [Mallick *et al.*, 2009; Schneider *et al.*, 2009]. In order to minimize the potential for false positive results in this study, sequences were prepared for input to SLR with a series of filters and realignment steps designed to remove low-quality sequences prone to high error rates, realign sequences using an aligner with better performance for detecting sitewise positive slection, and mask out short alignment regions with dubious elevated rates of non-synonymous substitution.

5.2.1 Filtering out low-quality genome 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 included in Ensembl [Jaffe *et al.*, 2003], output a set of Phred quality scores alongside the genome sequence, with one Phred score per base ranging from 0 to 99. 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].

Unfortunately, Ensembl does not store quality scores from its source genome assemblies, so Phred quality scores had to be manually downloaded for all genomes with readily available Phred-like quality scores. 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 designed to process each quality score file to allow for quicker score retrieval and better memory performance.

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 and observed 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 error rates for scores below 25, indicating that the scores were accurate in this range. Error rates did not decrease significantly at scores above 25, however, suggesting that the value of using an extremely high Phred score threshold would be minimal. 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 considerations, 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 coding sequence (CDS) with quality scores available, a "minimum score" approach was used to filter 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 the cumulative distribution of quality scores found in Hubisz et al. [2011], approximately 5% of bases in low-coverage mammalian genomes contain Phred scores below 25. The worst case scenario, in terms of high-quality bases being masked as a result of using the minimum score, would be if only one base per codon had a score below the threshold. Were that the 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 5% total filtered bases if they always occurred in codon triplets. Thus, anywhere from 5% to 15% of nucleotides from low-coverage genomes were expected to be filtered by this approach.

5.2.2 Removing recent paralogs

To complement the subtree splitting process, which split apart ancient duplications to avoid paralogous comparisons in the sitewise analysis, a second paralog filtering step was applied to remove more recent paralogs. Some of these paralogs might have resulted from gene duplications that occurred subsequent to the two rounds of whole-genome duplication in the vertebrate ancestor, but it was also expected that some proportion of apparent paralogs in the Ensembl gene trees would be the result of errors in gene annotation or in the Compara pipeline.

A particular cause for concern in the current analysis was the possibility that stretches of missing or unassembled sequence in low-coverage genomes might produce gaps of missing data or assembly breakpoints between exons of a single gene, causing it to become annotated as two separate genes. 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 having been derived from the same single source gene. 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.

In the case of true recent paralogs, their inclusion in the sitewise analysis might skew the dataset towards increased levels of relaxed constraint or adaptive evolution, as has been hypothesized and observed to occur in recently-duplicated genes [Lynch & Conery, 2000]. Most models of evolution after gene duplication predict that one sequence will retain the ancestral function and diverge less from the common ancestor than the other, so the least-diverged copy would be the natural one to keep as the 'canonical' paralog for evolutionary analysis.

Both gene length and sequence divergence were used to identify which gene among a set of recent paralogs was most suitable to retain for sitewise analysis. It was expected that gene length would help discriminate spuriously shortened genes from true genes, while sequence divergence would distinguish between more and less diverged paralogs. The mean sequence divergence, estimated using the JC69 nucleotide model and the stock Compara gene tree alignments, was calculated between each putative paralog and the rest of the gene tree, and the ratio of the length of each putative paralog to the mean sequence length was also stored. Within each group of putative paralogs, the single gene to keep was chosen by the following rules, applied in 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 meaningful (i.e., non-zero and non-infinite) mean distance estimate, the sequence with the lowest distance was kept; (3) if no sequence yielded a meaningful distance estimate (or if all estimated distances were identical), the longest sequence was kept.

Figure 5.1 shows the distribution of gene lengths (relative to the mean across the alignment) for all putative paralogs, kept paralogs, and removed paralogs. The overall distribution of relative lengths shows that most putative paralogs have lengths similar to the alignment mean (with a peak at or slightly above 1), but the shape of the distribution is highly asymmetric with a strong bias towards shorter lengths. The length distribution of the kept paralogs shows that the bulk of highly-shortened genes were successfully removed. If anything, the distribution of kept genes is slightly biased towards lengths greater than

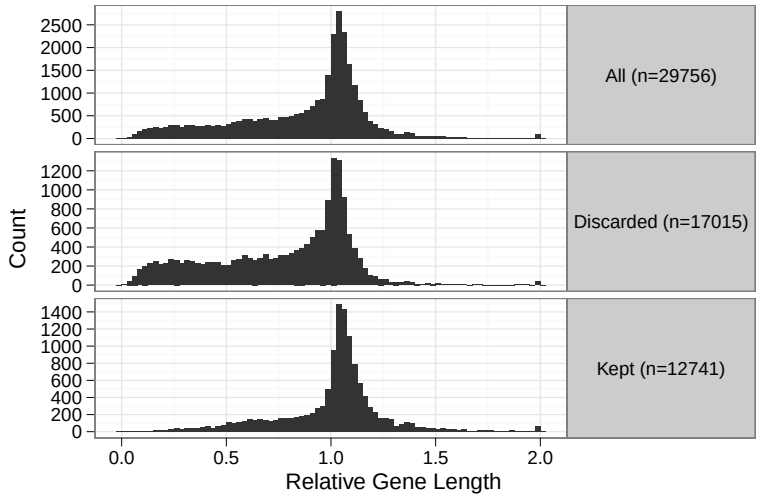


Figure 5.1: Gene lengths of all putative paralogs, normalized to the mean length all sequences in the enclosing tree of each paralog. Putatively paralogous genes (top panel) were either discarded (middle panel) or kept (bottom panel) according to rules based on their length and mean sequence divergence from other aligned sequences.

1, likely due to step 3 of the above process, where the longer gene was kept if putative paralogs yielded identical distance estimates.

An alternate view of the results of the paralog filtering process is shown in Figure 5.2, with the mean divergence of each discarded paralog compared that of the kept paralog, separated into panels according to the reason for discarding that gene. The spread of points above the diagonal in the first panel shows the difference in mean sequence divergence between the kept and discarded putative paralogs where divergence was used to choose between copies (TODO), and the middle panel represents paralogs whose mean divergences were identical (TODO). These two panels together show that although most recent paralogs within this set of gene trees contained indistinguishable levels of sequence divergence, around 40% showed a moderate difference that could be used for selecting the less-diverged copy. The rightmost panel shows the subset of apparent paralogs which were discarded due to their short gene length; a point worth noting here is that there was no bias towards higher or lower divergence levels in the discarded genes (the coefficient of the best-fit linear model for all non-negative values is TODO), suggesting (as expected) that many of the discarded short genes were in fact derived from a single orthologous gene.

n=XYZ

n=XYZ

0.983 ± 0.0

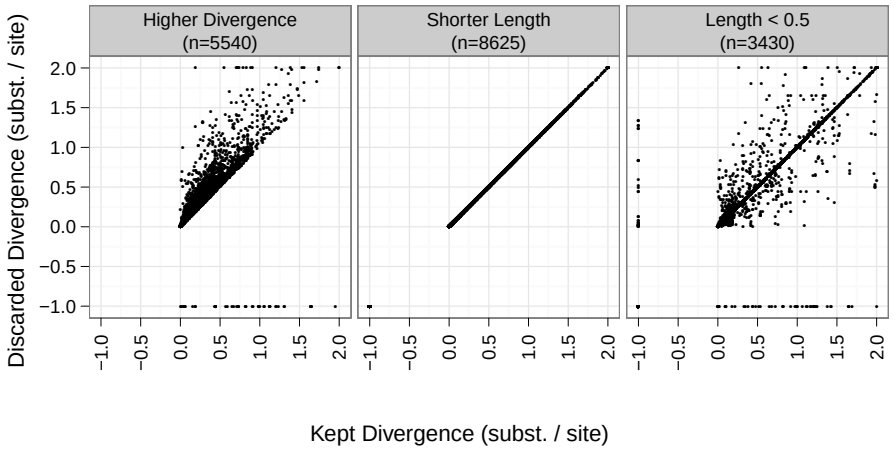


Figure 5.2: Sequence divergence of kept and discarded putative paralogs. Each point represents a gene which was discarded from the tree for one of three reasons: it had more sequence divergence than the kept gene (*Higher Divergence*; left panel), it had equal sequence divergence but shorter length than 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)

5.2.3 Realigning coding sequences

After filtering out codons with low quality scores and removing putative paralogs, sequences were aligned with PRANK [Löytynoja & Goldman, 2008] using its codon alignment model based on the empirical codon model [Kosiol *et al.*, 2007]. The simulation experiments described in Chapters 2 and 3 as well as numerous previously-reported empirical and simulation-based studies have shown PRANK's codon-based alignments to be superior for avoiding false positives in the detection of sitewise positive selection, strongly supporting the choice of PRANK for this analysis.

5.2.4 Filtering out clusters of non-synonymous substitutions

A final filtering step was evaluated for possible application to PRANK-aligned sequences in order to ensure that stretches of aligned but nonhomologous sequence, resulting either from misalignment or from exon mis-annotation artifacts, were not causing elevated rates of non-synonymous substitutions within specific regions of genes. This filtering step motivated

by the expectation that errors resulting from either misalignment or exon mis-annotation would both lead to clustered regions of nonhomologous aligned nucleotides and, correspondingly, clustered regions of elevated non-synonymous substitution rates. Clustering of non-synonymous substitutions was expected because in both cases, the existence of one misaligned column is not independent of nearby columns: for mis-annotation the misalignment would span the length of the erroneously-annotated exon, while the global nature of the progressive pairwise alignments performed by PRANK (and all other alignment algorithms) causes any misalignment error to be strongly non-independent with respect to errors in nearby alignment columns.

These arguments provided some hope that clustered non-synonymous substitutions could be used as a signal to detect potential misalignment and mis-annotated regions, but the utility of such a signal for filtering alignments must depend on the strength of error-caused non-synonymous clusters relative to both the frequency of true non-synonymous substitutions and their tendency to cluster together along the length of the protein sequence. Sequences separated by longer branch lengths will clearly show higher densities of true non-synonymous substitutions, possibly drowning out the error-caused signal and reducing the ability to use substitution clusters as a discriminating factor. Furthermore, non-synonymous substitutions were shown to be significantly more clustered than expected by chance in a number of genomic analyses of mammals and insects [Bazykin *et al.*, 2004; Callahan *et al.*, 2011; **?**], causing some concern that a filter based on detecting clusters of non-synonymous substitutions might attenuate the signal of true adaptive substitution that was one target of the present study.

[Write up the empirical investigation of non-synonymous substitution clustering.]

5.3 Genome-wide analysis of sitewise selective pressures in mammals

5.3.1 Mammalian species subsets for sitewise analysis

The SLR method was applied sequentially to several species subsets of each alignment of mammalian orthologs. For each subset, sequences corresponding to species within the subset were extracted from the alignment along with the corresponding subtree and input to SLR. If fewer than two sequences were available for a given subset, that subset was skipped and its absence from the dataset was recorded. Eight subsets in total were selected

for analysis; the species included in each subset and some phylogenetic measures of each subset are listed in Table 5.1.

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

Two larger species sets, Eutheria and Mammalia, were chosen for the purpose of measuring average sitewise selective pressures with high precision across a wider group of mammals. Using the Ensembl species tree as a guide, the estimated total synonymous branch lengths spanned by Ensembl species within Eutheria and Mammalia were 4.95 and 6.18, respectively. Simulations performed by Anisimova and Yang [**??**] and by myself in Chapter 2 predicted that the greater amount of branch length in the Eutherian and Mammalian trees—with two to three times the value of 1.71 for Laurasiatheria, the superorder with the largest total branch length—would result in significantly higher levels of power and accuracy for estimating sitewise ω and detecting sitewise positive selection. In this respect, Mammalia and Eutheria were more similar to each other than to any of the superorders.

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

Quantitatively, as measured by the MPL from the Ensembl species tree, the Eutheria subset (MPL =0.24) is far more similar to either of the three superorders (MPL from 0.13 to 0.27) than to the Mammalia subset (MPL =0.54). This is due to the striking

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

Finally, to further investigate the combined impact of evolutionary depth, biological variability and branch length on the results of large-scale sitewise analyses, two "sparse" subsets were created to act as controls relative to two existing species subsets. The species in the Sparse Glires group were chosen to approximate the total branch length of the Primate clade with species from the Glires clade, while the Sparse Mammals subset was constructed by selecting one species (preferably with a high-coverage genome) from each major mammalian branch, greatly reducing the total branch length covered but maintaining a similar evolutionary depth and distribution of branches within the species tree. The branch lengths in Table 5.1 show that the Sparse Glires group was only somewhat successful in its goal of approximating the Primates branch length (with total branch lengths of 0.99 and 0.68, respectively) while the Sparse Mammals group achieved a threefold lower total branch length compared to the full Mammalia group while maintaining a nearly identical MPL.

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

Sitewise data were collected from SLR and stored in a database for storage and further analysis. The Mammalia subset, containing the most branch length of all the datasets and representing the entire set of aligned sequences, and the Primate subset, containing the lowest overall branch length, were used to perform quality-control checks on the sitewise data. The point of these checks were to evaluate whether any additional filtering of the

Name	(Species Count) Species List	N_E	Ensembl MPL	Ensembl Total	Gene Median MPL	Gene Median Total
Primates	(10) Bushbaby, Chimpanzee, Gibbon, Gorilla, Human, Macaque, Marmoset, Mouse Lemur, Orangutan, Tarsier	20000	0.13	0.68	0.16	0.82
Glires	(7) Guinea Pig, Kangaroo rat, Mouse, Pika, Rabbit, Rat, Squirrel	230000	0.27	1.44	0.40	1.89
Laurasiatheria	(12) Alpaca, Cat, Cow, Dog, Dolphin, Hedgehog, Horse, Megabat, Microbat, Panda, Pig, Shrew	34410	0.19	1.71	0.26	2.14
Atlantogenata	(5) Armadillo, Elephant, Hyrax, Sloth, Tenrec	30000	0.20	0.83	0.26	0.96
Eutheria	(35) Alpaca, Armadillo, Bushbaby, Cat, Chimpanzee, Cow, Dog, Dolphin, Elephant, Gibbon, Gorilla, Guinea Pig, Hedgehog, Horse, Human, Hyrax, Kangaroo rat, Macaque, Marmoset, Megabat, Microbat, Mouse, Mouse Lemur, Orangutan, Panda, Pig, Pika, Rabbit, Rat, Shrew, Sloth, Squirrel, Tarsier, Tenrec, Tree Shrew	110000	0.24	4.95	0.35	6.39
Mammalia	(38) Alpaca, Armadillo, Bushbaby, Cat, Chimpanzee, Cow, Dog, Dolphin, Elephant, Gibbon, Gorilla, Guinea Pig, Hedgehog, Horse, Human, Hyrax, Kangaroo rat, Macaque, Marmoset, Megabat, Microbat, Mouse, Mouse Lemur, Opossum, Orangutan, Panda, Pig, Pika, Platypus, Rabbit, Rat, Shrew, Sloth, Squirrel, Tarsier, Tenrec, Tree Shrew, Wallaby	120000	0.54	6.18	0.66	8.11
Sparse Glires	(5) Guinea Pig, Kangaroo rat, Mouse, Rat, Squirrel	230000	0.25	0.99	0.36	1.31
Sparse Mammalia	(7) Armadillo, Dog, Elephant, Human, Mouse, Platypus, Wallaby	120000	0.51	2.18	0.60	2.80

Table 5.1: Species subsets used for sitewise analysis. Values under the "Ensembl" heading were calculated from subsets of the species tree used for evolutionary analyses by the Ensembl Compara pipeline, while values under the "Gene Median" heading were calculated as median values across the 15,XYZ gene trees analyzed (with branch lengths optimized by SLR). MPL – mean path length, Total – total branch length.

sitewise results was necessary before characterizing the global distribution of constraint in this and the other species subsets. Even if the sequence and alignment filters described above were successful at reducing the number of false positives due to incorrect input alignments, the behavior of SLR when applied to large datasets of heterogeneous alignments has not been well-studied, and a number of biases might have influenced the results. A particular point of concern was that columns with different patterns of gap and non-gap sequences, especially those with few non-gap sequences, might yield different performance characteristics. Although the SLR method was sensibly designed to account for uncertainty in the estimation of ω and detection of positive selection, one might reasonably expect less-desirable statistical properties from sites with 2 non-gap sequences compared to sites with 20.

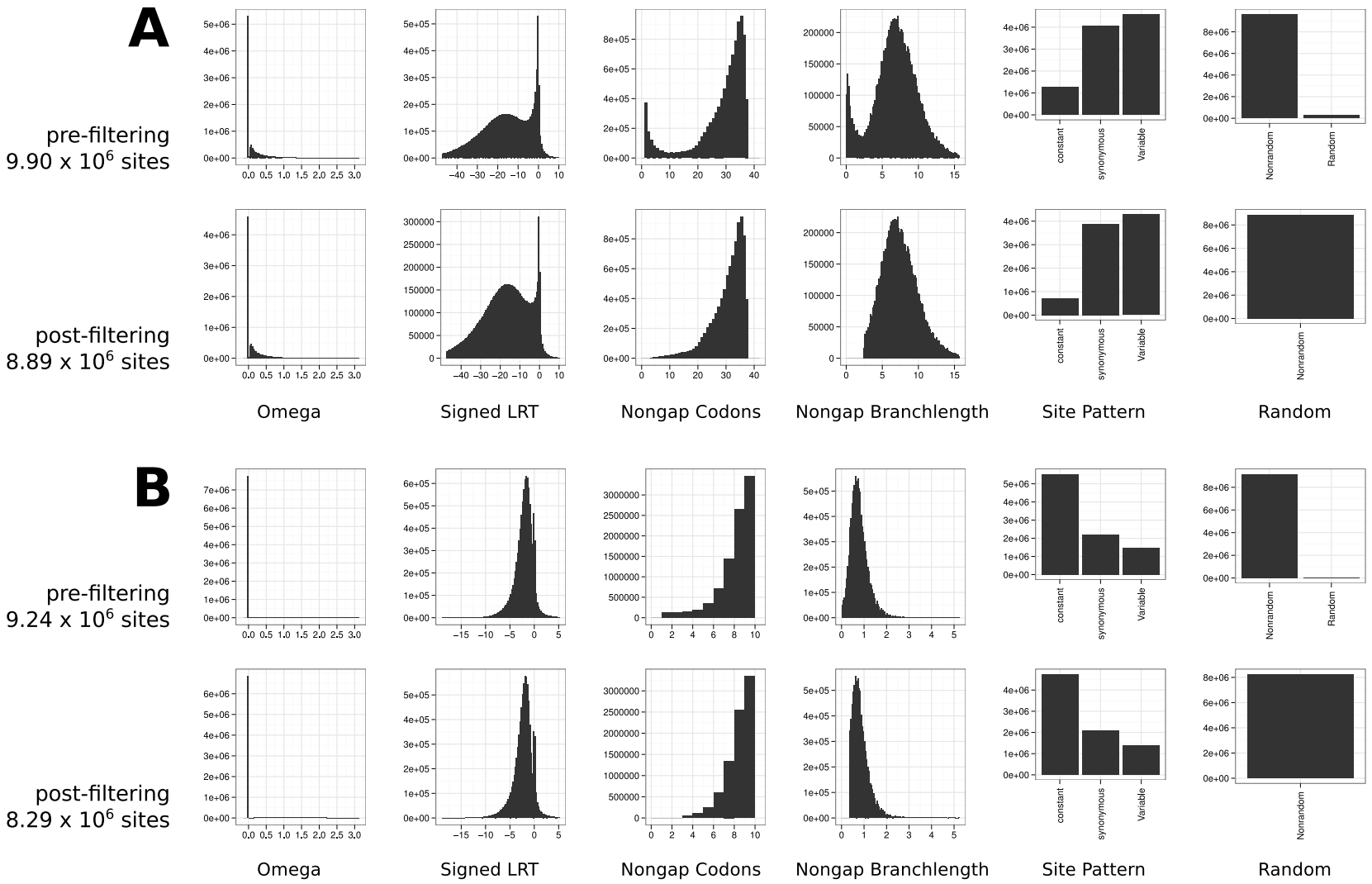


Figure 5.3: Distributions of sitewise values for Mammalia (A) and (B) Primates, before (top row) and after (bottom row) removing sites based on the filtering scheme (see text). Note: the y-axis scale varies between rows, and the x-axis scale varies between (A) and (B) for the Signed LRT, Nongap Codons and Nongap Branch Length values.

Figure 5.3 shows the distributions of six sitewise values: two continuous values output by SLR (Omega and Signed LRT), two categorical values from SLR (Site Pattern and Random), and two values calculated from the codon alignment (Nongap Codons and Nongap Branch Length). The Nongap Codons value measures the number of non-gap codons in the given alignment columnn, while the Nongap Branch Length represents the total branch length connecting all non-gap sequences (using the gene tree with branch lengths optimized by SLR).

A prominent feature of the distribution of ω values for the unfiltered Mammalian data, shown in the top panel of Figure 5.3A, was the large number of sites with a maximum-likelihood estimated ω of zero. Further inspection of the data revealed that all zero-ω sites contained either synonymous or constant site patterns, and all sites with constant patterns (and nearly all sites with synonymous patterns) yielded a maximum likelihood ω estimate of zero. An estimate of zero for synonymous sites is intuitively appropriate, as the lack of any non-synonymous substitutions throughout the tree would seem to provide no evidence for a non-synonymous substitution rate of greater than zero. For constant sites the case is less clear, because no data regarding the rate of either synonymous or non-synonymous substitutions exists. However, given SLR's assumption of a constant synonymous substitution rate throughout each gene [Massingham & Goldman, 2005], the ω value which maximizes the likelihood of observing zero substitutions is zero, since that value minimizes the non-synonymous (and total) substitution rate.

Use the sit
and 'seq' t
do a more
tive analys

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

the result of simultaneous multiple nucleotide changes or successive single changes) would be observed [Kosiol *et al.*, 2007].

The distributions of Nongap Codons and Nongap Branch Length values in the top row of Figure 5.3A showed that many alignment columns contained only a few non-gap sequences. Both distributions were bimodal with a larger peak at the upper end of the range and a smaller peak at the lower end of the range. If the sites with low non-gap codon counts represented accurate evolutionary histories, the observed excess of mostly gapped sites might be taken as an indication that insertion events in terminal lineages or recent ancestral lineages were prominent enough in mammalian evolution to leave a noticeable signature of sites with low non-gap codon counts. This would be a very interesting observation, but unfortunately it is not likely a correct one. Given the many possible sources of error in the creation of Ensembl gene trees, however, a more likely scenario was that sites with low codon counts and low branch lengths came from stretches of sequence which only exist in a few species as a result of annotation or alignment error, with a higher probability of being nonhomologous and showing spurious signals of positive selection. This would make such sites prime candidates for filtering out prior to a large-scale analysis.

To test the hypothesis that sites with few non-gap sequences are less reliable than other sites, the Mammals and Primates data were split into deciles by nongap branch length and sites within each decile were summarized by the proportion of sites showing evidence for purifying and positive selection; the results of this analysis are presented in Table 5.2. The lowest decile appeared to be a clear outlier in the Mammalia dataset, with nearly 17% of sites having an estimated ω of greater than 1 and 2% of sites showing significant evidence for positive selection at a nominal 5% error rate. Deciles with greater nongap branch lengths showed lower proportions of sites with $\omega > 1$ and less evidence for positive selection, with a gradual increase in both values at progressively higher deciles. The gradual increase in evidence for positive selection with increasing nongap branch length could be explained by genes with higher overall dN/dS ratios (and perhaps more positive selection) producing, on average, higher estimated branch lengths due to the increased non-synonymous substitution rate. Overall, these patterns were consistent with the expectation that sites with few non-gap sequences were not consistent with the bulk of the dataset, and Table 5.2 showed that removing sites with the lowest 10% of nongap branch length would remove most of the apparently anomalous sites.

The breakdown of Primates data in Table 5.2 showed a trend similar the Mammalia dataset, although the distinction between the lowest decile and the rest of the dataset was

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

Table 5.2: Proportions of sites with evidence for purifying and positive selection in the Mammalia and Primates datasets broken down by nongap branch length. Sites were separated into 10 equally-sized bins of nongap branch length and summarized by the 25^{th}, 50^{th} and 75^{th} percentiles of nongap branch length and nongap codons, the percentage of sites with ω estimated below or above 1, and the percentage of sites with significant evidence of positive selection at a nominal 5% FPR.

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

Turning back to the bulk distributions in Figure 5.3, the rightmost panel depicts the small set of sites designated as "random" by SLR. These sites were flagged by SLR as having a site pattern not significantly different from random [Massingham & Goldman, 2005], and they were also targeted for removal before analysis of the global distribution.

The final filtering protocol applied to each sitewise dataset included three steps. First, all sites within the bottom 10% of nongap branch length values were removed; second, sites

flagged by SLR as "random" were removed; third, all sites with fewer than four non-gap sequences were removed.

The most prominent effect of the filter on the bulk distributions in Figures 5.3A and 5.3B was, as expected, the removal of the excess of sites with low non-gap branch lengths and non-gap codon counts. The distribution of ω estimates and Signed LRT statistics were largely unchanged, indicating that the overall characteristics of each dataset were not significantly altered by the filter. The lack of large-scale change was a somewhat reassuring result, given that the filter only removed roughly 10% of sites from each dataset.

A more detailed comparison of various summary statistics for the filtered and unfiltered datasets showed that filtering had a noticeable impact on three quantities of interest: it reduced the proportion of constant sites, lowered the mean ω, and slightly decreased the percentage of positively-selected sites. Tables 5.3 and 5.4 contain summary statistics and calculations performed on the filtered and unfiltered Mammalia and Primates data. Most of the data contained in these tables will be more fully described in the next subsection, but a comparison of the filtered and unfiltered rows for Primates and Mammalia provided a means by which to quantitatively assess the effect of filtering on particular aspects of the dataset. First, the percentage of constant sites was reduced in the post-filtering data, moving from 60.04% to 57.62% in Primates and from 12.62% to 8.17% in Mammalia. This was expected, as the sites removed by filtering were enriched in constant site patterns due to their lower non-gap branch lengths. Second, the mean ω value was slightly reduced in Primates (from 0.32 to 0.27) and greatly reduced in Mammalia (from 0.49 to 0.20), likely due to the removal of sites containing a small number of nonhomologous codons which might have produced abnormally high sitewise maximum likelihood ω estimates. Third, the proportion of positively-selected sites (shown for a range of significance thresholds in Table 5.4) was moderately reduced in both Primates (from 0.56% to 0.53% at $P_{\chi^2_1} < 0.05$) and Mammalia (from 0.72% to 0.56%). These three effects of filtering, each showing a shift indicating more useful data (e.g., a lower percentage of constant sites) or less evidence for positive selection (e.g., lower mean ω and proportion of positively-selected sites) in the post-filtering datasets, together provided evidence supporting the inclusion of such a filtering step prior to the analysis and comparison of sitewise estimates in different species sets. Although most quantities of interest were not noticeably changed, those that were affected by filtering shifted to more conservative values, which was taken to be a positive sign given the persistent concern regarding the presence of false positives in detecting positive selection.

Name	Sites	Site Pattern, %			Med.	Nongap BL			ω_{ML}		ω_{ML} Above / Below, %			
		Const.	Syn.	Nsyn.	Codons	Med.	Mean	SD	Mean	SD	< 0.5	< 1	> 1	> 1.5
Primates (raw)	9.24e+06	60.04	24.08	15.88	9	0.73	0.85	1.36	0.19	0.50	86.21	91.08	8.92	5.79
Primates	8.24e+06	57.62	25.49	16.89	9	0.78	0.92	1.24	0.20	0.51	85.45	90.86	9.14	5.72
Glires	7.70e+06	33.99	41.95	24.06	7	1.81	2.00	1.24	0.15	0.35	89.33	96.01	3.99	1.76
Laurasiatheria	8.16e+06	32.52	38.93	28.54	10	2.03	2.24	1.32	0.19	0.41	86.46	94.08	5.92	2.95
Atlantogenata	5.15e+06	55.95	29.75	14.30	5	0.98	1.09	0.62	0.13	0.38	89.03	94.91	5.09	2.55
Eutheria	8.79e+06	11.54	44.49	43.97	30	5.83	6.17	2.76	0.19	0.36	87.01	95.27	4.73	1.91
Mammalia (raw)	9.90e+06	12.62	40.96	46.43	32	6.89	6.88	3.74	0.21	0.40	86.26	94.29	5.71	2.95
Mammalia	8.89e+06	8.17	43.56	48.26	33	7.23	7.55	3.32	0.19	0.35	87.52	95.74	4.26	1.63
Sparse Glires	6.72e+06	44.22	38.40	17.38	5	1.28	1.43	1.14	0.12	0.34	90.07	96.35	3.65	1.69
Sparse Mammalia	7.52e+06	26.32	44.99	28.69	6	2.55	2.82	1.81	0.14	0.32	90.69	96.85	3.15	1.33

Table 5.3: Summary statistics and maximum likelihood ω estimates for sitewise data in eight species groups. Rows labeled "Primates (raw)" and "Mammalia (raw)" were computed based on unfiltered data and are included for reference. Columns under the "ω_{ML} Above / Below, %" heading measure the cumulative percentage of sites with ω_{ML} below or above the indicated value. Med.—median, Const.—constant, Syn.—synonymous, Nsyn.—non-synonymous, ML—maximum likelihood.

Name	Positively Selected Sites (%)								$P_{\chi^2_1} < 0.1$, %			$P_{\chi^2_1} < 0.05$, %		
	$P_{\chi^2_1} < 0.1$		$P_{\chi^2_1} < 0.05$		$P_{\chi^2_1} < 0.01$		FDR< 0.05		Pos	Neg	Neut.	Pos	Neg	Neut.
Primates (raw)	89734	(0.97)	52005	(0.56)	15973	(0.17)	357	(0.00)	0.97	29.86	69.16	0.56	14.31	85.13
Primates	75629	(0.92)	43382	(0.53)	12985	(0.16)	158	(0.00)	0.92	33.17	65.92	0.53	15.97	83.50
Glires	17388	(0.23)	8735	(0.11)	1953	(0.03)	0	(0.00)	0.23	76.73	23.04	0.11	65.58	34.31
Laurasiatheria	57494	(0.70)	34693	(0.43)	11589	(0.14)	471	(0.01)	0.70	74.49	24.81	0.43	64.65	34.92
Atlantogenata	11987	(0.23)	5766	(0.11)	1166	(0.02)	0	(0.00)	0.23	46.82	52.94	0.11	24.24	75.64
Eutheria	80406	(0.91)	55063	(0.63)	24804	(0.28)	9378	(0.11)	0.91	85.09	14.00	0.63	82.10	17.27
Mammalia (raw)	107595	(1.09)	71167	(0.72)	28894	(0.29)	7822	(0.08)	1.09	80.15	18.77	0.72	76.98	22.30
Mammalia	72278	(0.81)	49311	(0.55)	22123	(0.25)	8335	(0.09)	0.81	86.51	12.68	0.55	83.87	15.57
Sparse Glires	10582	(0.16)	4928	(0.07)	933	(0.01)	0	(0.00)	0.16	69.33	30.51	0.07	48.67	51.26
Sparse Mammalia	12737	(0.17)	6347	(0.08)	1396	(0.02)	0	(0.00)	0.17	79.85	19.98	0.08	72.43	27.49

Table 5.4: Proportions of sites subject to positive, purifying and neutral selection at various LRT$_{SLR}$ thresholds. The Benjamini-Hochberg method [Benjamini & Hochberg, 1995] was used to identify the LRT$_{SLR}$ threshold at which FDR$<$0.05. For columns under the headings "$P_{\chi^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.

5.3.3 The global distribution of sitewise selective pressures in mammals

Each set of sitewise data was filtered as described above. The resulting global distributions of site patterns, sitewise ω estimates, and 95% confidence intervals are shown in Figure 5.4; the left panel in each row plots the number of sites with constant, synonymous, and non-synonymous patterns. All sites with $\omega_{ML} = 0$ had constant or synonymous patterns, and all sites with $\omega_{ML} > 0$ had non-synonymous patterns; the distributions of ω_{ML} for these non-synonymous sites are shown as histograms on the right panel in each row.

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

The site pattern counts in Figure 5.4 showed that the branch length of each species group had a strong effect on the overall composition of the sitewise data. Species groups covering little branch length, such as Primates and Atlantogenata, contained a majority of constant sites, while groups comprising lots of branch length, such as Eutheria and Mammalia, contained few constant sites and roughly equal proportions of synonymous and non-synonymous sites. Comparing the Glires and Mammalia data with their corresponding "sparse" datasets confirmed that this trend was largely due to branch length as opposed to biological factors: the Sparse Glires data yielded a smaller proportion of non-synonymous sites and a greater proportion of constant sites than the Glires data (17.41% versus 24.08% for non-synonymous sites, 44.21% versus 33.98% for constant sites; numbers from Table 5.3), and the pattern for Sparse Mammalia and Mammalia was qualitatively the same.

Turning to the distribution of these ω_{ML} estimates, shown in Figure 5.4 as a series of histograms representing the ω_{ML} density (for nonzero values only) and a series of solid lines representing the cumulative density (for all values), it is clear that the vast majority of protein-coding sites have evolved under purifying selection in mammals. The Mammalia group, which contained a very small proportion of potentially uninformative constant sites (8.17%), showed a maximum density of nonzero ω_{ML} estimates at $\omega \approx 0.1$, and the vast majority of sites showed some evidence of purifying selection, with ω_{ML} estimates below 1. The height of the ω_{ML} cumulative distribution at $\omega = 1$ corresponds to the proportion of such sites; the exact value, included in Table 5.3 under the "< 1" column, is 95.74%. The nonzero ω_{ML} values were more evenly spread in the other species groups, with Glires showing a maximum nonzero ω_{ML} density at around $\omega \approx 0.25$ and Primates at $\omega \approx 0.7$.

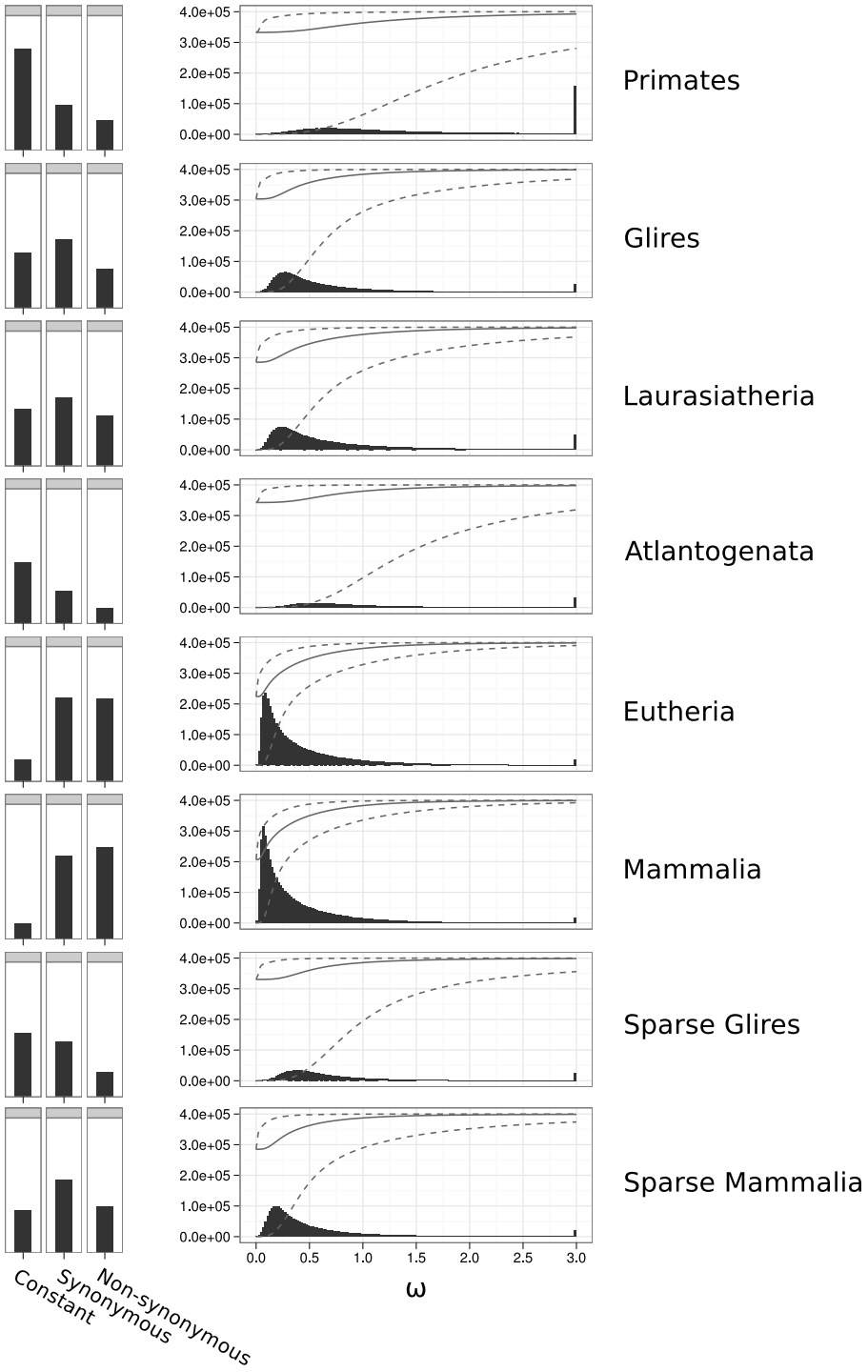


Figure 5.4: Global distributions of site patterns and ω estimates for eight species groups. Left, bars represent the number of sites showing constant, synonymous, and non-synonymous patterns. Note, the y-axis is held constant between rows. Right, bars represent a histogram of maximum likelihood ω estimates where $\omega > 0$. Sites where $\omega > 3$ are counted in the bin at $\omega = 3$. A solid line is drawn showing the cumulative proportion of sites with ω below the current value, and dashed lines are drawn above and below the solid line, showing the cumulative proportion of sites with the lower or upper range, respectively, of their 95% confidence interval below the current value.

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

5.3.3.2 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, or lack thereof, for purifying and positive selection. Thus, sites with $\text{LRT}_{SLR} < 0$ showed at least some evidence for purifying selection and sites with $\text{LRT}_{SLR} > 0$ showed at least some evidence for positive selection. It should be noted that the LRT_{SLR} is a measure of the strength of evidence for purifying or positive selection, but not necessarily the actual strength of that selection. For example, an alignment covering a very large branch length might yield a strongly negative LRT_{SLR} for a site with ω_{ML} only moderately below 1, because the difference between dN and dS was highly statistically significant; on the other hand, a strongly-purifying site in an alignment covering less branch length might produce a much less-negative LRT_{SLR}, even with ω_{ML} near zero.

Figure 5.5A shows the empirical relationship between LRT_{SLR}, ω_{ML} and the $\text{CI}_{95\%}$ width for sites from the Mammali dataset. The left panel, comparing the LRT_{SLR} to nonzero ω_{ML} estimates, shows that the two values are highly correlated, with the greatest number of low ω_{ML} estimates occurring at sites with strongly negative LRT_{SLR}s. Correspondingly, the middle panel shows an even stronger relationship between the LRT_{SLR} magnitude and the $\text{CI}_{95\%}$ width, with the tighest windows at sites with very strong evidence for purifying selection. The rightmost panel compares the ω_{ML} of each site with the width of its $\text{CI}_{95\%}$, revealing a more linear and diffuse positive relationship between ω_{ML} and the size of the $\text{CI}_{95\%}$. The equivalent plots for Primates, shown in Figure 5.5B, reveal

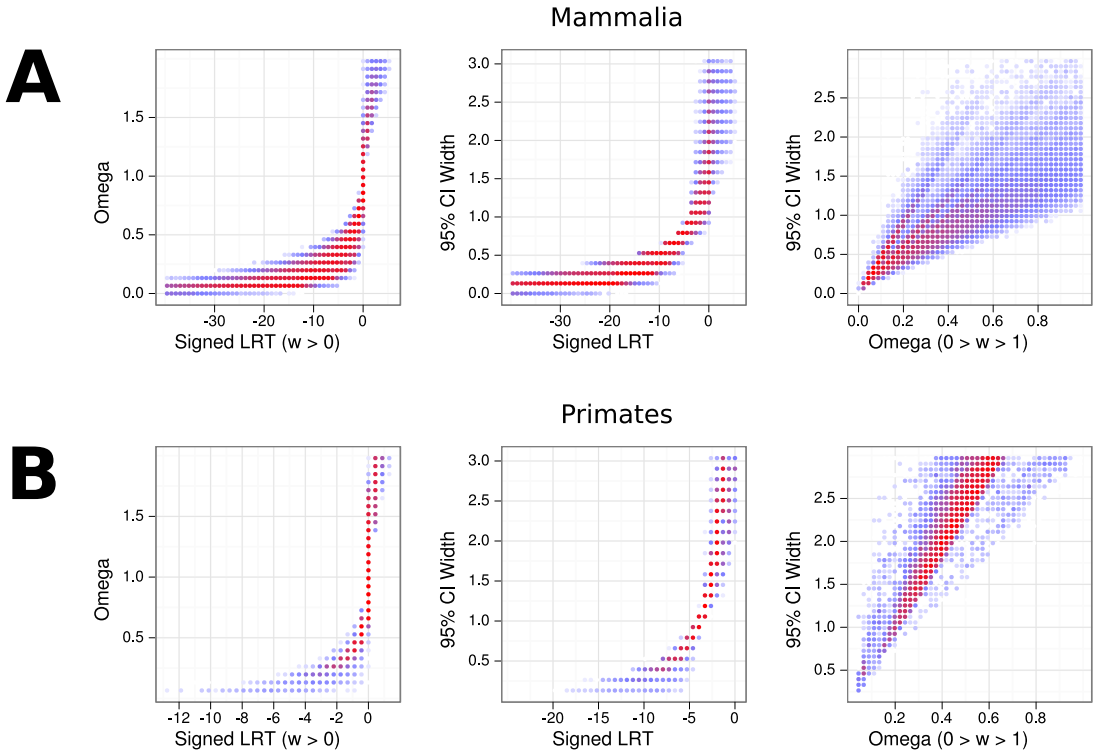


Figure 5.5: The relationship between LRT_{SLR}, ω_{ML}, and $\text{CI}_{95\%}$ width in (a) Mammalia and (b) Primates datasets. Each point represents the binned density of sites; no points are drawn where no density exists, while blue and red points are drawn at areas of low and high density, respectively. 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 scale between plots in (a) and (b), reflecting the paucity of sites in Primates with strong evidence (LRT_{SLR} <-12) for purifying selection.

similar patterns, but with more weight towards less-negative LRT_{SLR} values, higher ω_{ML}, and larger $\text{CI}_{95\%}$. These differences highlight the impact of branch length on the amount of confidence with which ω can be estimated, showing that the low branch length of the Primates clade rarely yields ω_{ML} estimates with $\text{CI}_{95\%}$ intervals smaller than 1, while the bulk of sites from the Mammalia dataset have a relatively small $\text{CI}_{95\%}$. Thus, the ω_{ML} distribution from datasets with low branch lengths (e.g., Figure 5.4) should be interpreted with caution, and any comparison between different sites or datasets must be sensitive to the amount of statistical confidence placed on each estimate.

The statistical information at each site could be used to identify sites evolving under purifying or positive selection with statistical confidence. Sites with a CI_{upper}, the upper bound of the $\text{CI}_{95\%}$ interval, below $\omega = 1$ showed evidence of purifying selection with

an expected 5% FPR, and sites with a CI_{lower} above $\omega = 1$ showed evidence of positive selection with an expected 5% FPR. In both cases, the 5% FPR was expected under SLR's null model of neutral evolution. There was a strong relationship between CI_{upper} and the χ^2_1 approximation to the LRT_{SLR} distribution, whereby the set of sites with $\text{CI}_{upper} < 1$ 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 instead of $\text{CI}_{95\%}$ intervals when discussing sites with significant evidence for purifying or positive selection. In some cases, however, use of the full $\text{CI}_{95\%}$ will be preferable, as the LRT_{SLR} critical values only correspond to one end of the $\text{CI}_{95\%}$, depending on whether the site shows greater evidence for purifying or positive selection.

Table 5.4 summarizes the results from using the χ^2_1 approximation to the LRT_{SLR} distribution to identify sites subject to purifying selection and positive selection at various FPR thresholds. The left group of columns show the number and proportion of sites with evidence for positive selection at nominal 10%, 5%, and 1% FPR thresholds, respectively, as well as an expected 5% FDR calculated using the Benjamini Hochberg method for FDR control [Benjamini & Hochberg, 1995]. The two groups of columns on the right show the result of breaking sites into three groups (positive, negative, and neutral) based on the result of a χ^2_1 test at a given FPR threshold.

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

In Mammalia, the breakdown of sites into positive, negative and neutral categories at

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

Even for the Mammalia dataset, which encapsulated roughly 7.5 synonymous substitutions per site on average, one might reasonably expect that the power to confidently identify sites under purifying selection, though higher than in Primates, was still less than 100%. If this is the case, then the proportion of confidently identified purifying sites must be an underestimate of the true proportion of negatively-selected sites (and by symmetry, absent any methodological bias, the same should be true for positively-selected sites). As a result, the fractions of positively- and negatively-selected sites in Table 5.4 should not be taken as best estimates of the actual proportions of such sites, but more appropriately as lower bounds on those proportions. In fact, In the next two sub-sections, I will separately consider the issue of using the sets of sitewise estimates in different species groups to estimate the proportion of sites subject to purifying and positive selection in mammals.

5.3.3.3 Estimating the proportion of negatively-selected sites

For sites under purifying selection,

In , the value of ≈95% based on $\omega_{ML} < 1$ (Table 5.3) is likely closer to the true number; despite the caveats involved in ignoring the uncertainty involved in ω_{ML} estimates, the 95% number was surprisingly consistent across different species groups and branch lengths, ranging from ≈91% in Primates to ≈97% in Sparse Mammalia.

5.3.3.4 Estimating the proportion of positively-selected sites

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

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

identifying purifying selection, but there was no obvious correlation between higher branch lengths and higher numbers of confidently identified PSCs.

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

To quantify this observation, Spearman's rank correlation coefficients between the median branch length of each species set and each of several summary statistics were calculated (Table 5.5). Although the number of samples was small with only eight species groups, these correlations should be able to provide some indication as to which aspects of the sitewise data might be easily attributed to the effects of branch length and which aspects suggested an alternative (e.g., biological or artefactual) cause for the differences between species groups. The results were quite striking: the site classifications (constant, synonymous and non-synonymous) and the proportion of negatively-selected sites at $P_{\chi^2_1} < 0.05$ were strongly correlated with branch length, while the other factors, including mean ω_{ML}, the proportion of sites with $\omega_{ML} < 1$ and $\omega_{ML} < 0.5$, and the proportion of positively-selected sites, were weakly correlated or largely uncorrelated with branch length.

The results in Table 5.5 can be interpreted in a number of ways. First, they emphasized the unambiguous correlation between the branch length in a tree and the power to detect sitewise purifying selection. However, they also showed that various measures based on sitewise estimates contained variation between species groups that was not well-explained by branch length.

Refer to C
Pollard 20
showed tha
mates and
vores have
BDS (bias
gent subst
while glire

5.3.4 Modeling the global distribution of sitewise selective pressures

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

Table 5.5: Correlations between median non-gap branch length, N_E, and various summary statistics of the sitewise data across eight species sets. The magnitude (Rho) and significance (P-value) of Spearman's rank correlations between variables were calculated using N_E values from Table 5.1 and branch lengths and summary statistics from Tables 5.3 and 5.4. All summary statistics except for mean ω_{ML} were measured as a fraction of total sites. More highly significant correlations are shaded in darker blue. N_E – estimated effective population size, BL – non-gap branch length, $PSC_{5\%}$– positively-selected codons at $P_{\chi^2_1} < 0.95$.

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

Table 5.6

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

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

5.3.5 Simulations to evaluate the fit of empirical data to a simulated poisson process at varying branch lengths and ω ratios

The large difference in branch lengths covered by the species sets under investigation letd to some uncertainty in whether differences in the observed summary statistics, such as the mean ω_{ML} or the proportion of sites with $\omega_{ML} < 1$, were due to biological differences between species groups (i.e., differences in the efficacy of purifying selection resulting from population size differences) or to branch length effects. For example, two

I performed a small simulation study to further investigate the interactions between branch length, effective population size, and the summary statistics presented in Table 5.3,

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

Previous simulations on the power and accuracy of maximum-likelihood methods for detecting sitewise positive selection have provided strong evidence for increased power with increased branch length and number of taxa [Anisimova and Yang 2002 PMID:12032251, Massingham and Goldman 2005 PMID:15654091]. Other major effects observed have been a reduction in power when branch lengths are very short, due to the scarcity of data in the form of observed substitutions, and a reduction in accuracy when branches are very long and the ancestral reconstruction procedure becomes inaccurate due to saturation of substitutions at synonymous sites [Anisimova and Yang 2002 PMID:12032251]. These results have provided general guidance to empirical analyses, but the potential effects of tree shape, divergence level, distribution of dN/dS levels, and misalignment error make it difficult to extrapolate expected power levels or error rates from generic simulations to specific empirical analyses and real-world datasets. In order to estimate the power of the SLR method when applied to the present set of species and to quantify the power gained from the additional 20 mammalian genomes, we ran a series of simulation experiments with parameters tuned specifically to the analysis of mammalian gene families. Based on the prevoius studies described above we hypothesized that the addition of 20 mammalian genomes to the available dataset would significantly improve SLRs sensitivity for detecting positive selection at a reasonable error rate, with some portion of that improvement coming from a reduction in alignment error due to the shorter average length of branches in the phylogenetic tree being aligned.

We used the Indelible program [Fletcher and Yang 2010 PMID:19423664] to simulate 100 replicate codon alignments with a root sequence length of 500 codons for each of three trees: all 29 Eutherian mammals, the 9 mammals with high-coverage genomes, and the four-species Human-Mouse-Rat-Dog quartet used in a number of previous comparative analyses. The dN/dS value at each simulated site was drawn from a discretized lognormal distribution with log(mean)=-1.864 and log(sd)=1.201 with the maximum dN/dS capped at 3. This distribution yielded a mean dN/dS of 0.277 and 6% of sites with dN/dS ¿ 1, which is consistent with our estimates from the global mammalian distribution. We ran each set of simulations twice: once with an insertion and deletion (indel) rate of zero, and once with an indel rate of 0.05 indel events per substitution event. The length of each indel event was drawn from a discretized power-law distribution with a parameter of 1.8, a maximum insertion or deletion length of 40 codons, and equal insertion and deletion lengths

and probabilities. Each simulated alignment was aligned using PRANKs codon model of evolution and analyzed with SLR. The resulting SLR score at each human sequence position was compared to the true dN/dS at the equivalent site in the true alignment and used to calculate the power and accuracy of the detection of positive selection for each set of simulated alignments. A summary of the results is provided in Supplementary Table S17.12.

The simulation results without indels show a dramatic increase in the ability to detect sitewise selective pressures and sitewise positive selection in larger mammalian trees. We used ROC curves based on the true and inferred dN/dS values to calculate a number of statistics summarizing the performance of sitewise inference under each tree tested. The Spearmans rank correlation coefficient between inferred and true dN/dS was 0.749, 0.849, and 0.942 in the 4-taxon, 9-taxon, and 29-taxon trees, respectively, representing a 10% increase in the accuracy of inferred maximum-likelihood dN/dS values as a result of the added 20 mammalian species. The number of true positives recovered when controlling for a false discovery rate (FDR) below 0.1 was 50, 782, and 3760 for the three trees; for FDR ¡ 0.05, the numbers of true positives were 10, 429, and 2990. This represents a 4.8-fold increase for FDR¡0.1 and a 6.9-fold increase for FDR¡0.05 resulting from the additional 20 species in the tree. It should be noted that the two previous power estimates re

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

Multiple lines of evidence have lent support to the hypothesis that GC-biased gene conversion (BGC) has been a major force in the evolution of mammalian genomes [Galtier et al 2001 PMID:11693127, Galtier 2003 PMID:XYZ, Dreszer et al 2007 PMID:17785536]. Both empirical and theoretical results have shown that BGC can significantly affect patterns of observed substitutions in both selectively neutral and functionally constrained sites [Galtier et al 2009 PMID:19027980, Berglund et al 2009 PMID:19175294]. Recently, Ratnakumar et al. [2010, PMID:20643747] re-analyzed the dataset of positively-selected genes from Kosiol et al. [2008, PMID:18670650] for signatures of BGC and found that up to 20% of cases of identified elevated dN/dS ratios could be due to BGC rather than adaptive evolution. However, the strongest signals of BGC were found only in genes showing signals of positive selection along short branches in the phylogenetic tree using so-called

branch-site models of evolution; when the authors looked for similar BGC signatures in genes with evidence for positive selection at specific sites throughout the mammalian tree (e.g., genes with significant LRTs for PAMLs sites model) they found no evidence for a strong BGC influence [Ratnakumar et al. 2010 PMID:20643747].

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

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

that if our sitewise data do contain a strong BGC influence, then the positively-selected sites we detected would be expected to be associated with regions of high male-specific recombination.

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

Theoretical and empirical evidence pointed towards an increased sensitivity of dN/dS estimates to BGC influence in regions of high GC content, so we separated out the top 10% of sites by GC content and analyzed them according to quantiles of male recombination rate (Supplementary Table S17.13, High GC, Male Recombination. The middle four

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

5.4 Conclusions

[...]

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

Table 5.7

Chapter 6

The use of sitewise selective pressures to characterise the evolution of genes and domains in mammals

6.1 Introduction

A clear definition of what constitutes a positively selected gene (PSG) is elusive and depends heavily on the model the data. Neutral theory, however, does provide a precise definition: the dN/dS ratio gives a natural threshold for distinguishing purifying from positive selection at dN/dS = 1, above which point positive selection can be inferred. This leads logically to the most stringent definition of a PSG as a gene whose overall mean dN/dS ratio is greater than 1. However, as noted by Yang [5], this definition of a PSG is extremely conservative and is likely to be met by very few genes. To examine this hypothesis in the context of mammalian proteins, we averaged the site-wise dN/dS estimates from each gene tree tested to obtain an average dN/dS value, mean-dN/dS. At a threshold of mean-dN/dS = 1, 126 out of 15,451 gene trees analyzed showed evidence for positive selection (0.82%). At a threshold of mean-dN/dS = 0.5, 1,438 genes were PSGs (9.3%).

More powerful scans for positive selection relax the requirement that the overall dN/dS of a gene is above 1 under the assumption that the force of positive selection is unlikely to act along an entire protein sequence and throughout the entire evolutionary history of a gene tree. This led to models that allow for variation in the dN/dS ratio along the sequence or along the phylogeny. If the data shows statistically significant evidence for positive selection under a given model, then it is considered a PSG thus, the definition of

a PSG is tightly linked to the evolutionary model being applied and the evolutionary clade or lineage being analyzed. In the case of the SLR method (which allows for variation of dN/dS along the sequence but not along the phylogeny), the most appropriate definition of a PSG is a gene where one or more sites show statistical evidence for positive selection after correcting for multiple tests. With this definition, we find 2,865 positively selected gene trees in mammals (18.5%). We may also formulate a more conservative definition in an attempt to minimize the number of false positives resulting from a background of reduced purifying constraint, considering genes with mean-dN/dS of less than 0.5 and at least 2 positively-selected codons to be PSGs. This definition yields 739 positively selected gene trees (4.8%).

Although the mean-dN/dS ¿ 1 definition of a PSG is unreasonably conservative (and, upon manual inspection, picks out many proteins with especially severe alignment or gene annotation errors), the numbers of PSGs resulting from both codon-based definitions applied to the current data (4.8% at the low end and 18.5% at the high end) are similar to the range of other recently published comparative scans for positive selection, which identified between 3.3% [6] and 16.6% [7] of genes as PSGs in mammals or primates.

6.2 Comparison of sitewise results to previously described sets of positively selected genes

6.3 Using sitewise selective pressures to characterise the evolution of genes

6.3.1 Identifying genes subject to positive selection

6.3.2 Identifying genes subject to strong or weak purifying selection

6.4 Using sitewise selective pressures to characterise the evolution of protein domains

[To cite: Moses & Durbin [2009] — They use residue probabilities to generate "preferred residue" predictions, and look at human polymorphism to assign whether a SNP is moving

to or from a preferred residue.]

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.

The domain annotations were collated in a variety of ways: (1) taking the mean omega value across all annotated sites [omega column in the table below], (2) counting the number of PSCs within all annotated sites [psc], and (3) counting the number of PSCs per annotated site [psc_corr]. The len column represents the number of alignment sites containing the given Pfam annotation. When the list of domains is sorted by the total number of annotated PSCs we see the more familiar positively-selected domains at the top of the list (immunoglobulin, collagen), followed by more novel positively-selected domains such as the 7 transmembrane receptor, protein kinase domain, and ion transport protein.

6.4.1 Identifying protein domains subject to positive selection

6.4.2 Identifying protein domains subject to strong or weak purifying selection

6.5 Identifying genes under unusual selective pressures in mammalian superorders

Chapter 7

Evolution of protein-coding genes in gorilla and the African apes

7.1 Introduction

7.1.1 The gorilla and other primate genome projects

7.1.2 Incomplete lineage sorting

7.1.3 Effective population sizes of extant and ancestral primate populations

7.1.4 Measuring shifts in selective pressures using branch-specific likelihood ratio tests

7.1.5 Data quality concerns: sequencing, assembly and alignment error

7.2 Constructing codon alignments of one-to-one orthologous genes in six primate species

7.2.1 Identification of genes with one-to-one homology

7.2.2 Collection of homologous DNA sequences from genome- or transcript-based multiple alignments

7.2.3 Filtering sequence regions with low sequence quality

7.2.4 Filtering sequence regions with high substitution counts

7.2.5 Filtering sequence regions with evidence of incomplete lineage sorting

Chapter 8

Gorilla part 2

8.1 Analysis of incomplete lineage sorting in the African great apes within and nearby protein-coding genes

8.2 Analysis of dN/dS levels in six primate genomes

8.2.1 Genome-wide dN/dS in six primates and their ancestors

8.2.2 Genome-wide dN/dS in regions of differing sitewise constraint

8.2.3 Analysis of the impact of sequence and alignment filtering on primate dN/dS estimates

8.3 Conclusions and future work

Bibliography

Albers, C., Cvejic, A., Favier, R., Bouwmans, E., Alessi, M., Bertone, P., Jordan, G., Kettleborough, R., Kiddle, G., Kostadima, M., Read, R., Sipos, B., Sivapalaratnam, S., Smethurst, P., Stephens, J., Voss, K., Nurden, A., Rendon, A., Nurden, P. & Ouwehand, W. (2011). Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. *Nat Genet*, **43**, 735–7. 14

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. 61

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. 36, 61

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. 36

Archibald, A., Bolund, L., Churcher, C., Fredholm, M., Groenen, M., Harlizius, B., Lee, K., Milan, D., Rogers, J., Rothschild, M., Uenishi, H., Wang, J., Schook, L. & Swine Genome Sequencing Consortium (2010). Pig genome sequence–analysis and publication strategy. *BMC Genomics*, **11**, 438. 30

Archibald, J.D. (1999). Divergence times of eutherian mammals. *Science*, **285**, 2031. 48

Averof, M., Rokas, A., Wolfe, K. & Sharp, P. (2000). Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science*, **287**, 1283–6. 51

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. 40

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. 46

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. 39

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. 55, 60

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L. & Purvis, A. (2007). The delayed rise of present-day mammals. *Nature*, **446**, 507–512. 48

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X., Flicek, P., Gräf, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kähäri, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlic, A., Rae, M., Rios, D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C. & Hubbard, T. (2006). Ensembl 2006. *Nucleic Acids Res*, **34**, D556–61. 12

Brunet, F., Roest Crollius, H., Paris, M., Aury, J., Gibert, P., Jaillon, O., Laudet, V. & Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol*, **23**, 1808–16. 27

Callahan, B., Neher, R., Bachtrog, D., Andolfatto, P. & Shraiman, B. (2011). Correlated evolution of nearby residues in drosophilid proteins. *PLoS Genet*, **7**, e1001315. 46

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. 39

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. 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. 41

TOCITE (2011). Citation will be inserted at a later point in time. 8

Csuros, M., Rogozin, I. & Koonin, E. (2011). A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. *PLoS Comput Biol*, **7**, e1002150. 10

Dehal, P. & Boore, J. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, **3**, e314. 15

Demuth, J., De Bie, T., Stajich, J., Cristianini, N. & Hahn, M. (2006). The evolution of mammalian gene families. *PLoS One*, **1**, e85. 16

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–7. 12

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. 35, 41

Flicek, P., Amode, M., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H., Rios, D., Ritchie, G., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y., Trevanion, S., Vandrovcova, J., Vilella, A., White, S., Wilder, S., Zadissa, A., Zamora, J., Aken, B., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X.,

Herrero, J., Hubbard, T., Parker, A., Proctor, G., Vogel, J. & Searle, S. (2011). Ensembl 2011. *Nucleic Acids Res*, **39**, D800–6. 9

Green, P. (2007). 2x genomes–does depth matter? *Genome Res*, **17**, 1547–9. 37

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. 9, 41

Hubisz, M., Lin, M., Kellis, M. & Siepel, A. (2011). Error and error mitigation in low-coverage genome assemblies. *PLoS One*, **6**, e17034. 9, 41, 42

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. 41

Jaillon, O., Aury, J., Brunet, F., Petit, J., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z., Cattolico, L., Poulain, J., De Berardinis, V., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K., McEwan, P., Bosak, S., Kellis, M., Volff, J., Guigó, R., Zody, M., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E., Weissenbach, J. & Roest Crollius, H. (2004). Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, **431**, 946–57. 23

Jun, J., Mandoiu, I. & Nelson, C. (2009). Identification of mammalian orthologs using local synteny. *BMC Genomics*, **10**, 630. 8

Katoh, K., Kuma, K., Toh, H. & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, **33**, 511–8. 12

Kellis, M., Birren, B. & Lander, E. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, **428**, 617–24. 12

Kimura, M. & Ohta, T. (1974). On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences*, **71**, 2848–2852. 35

Koonin, E. & Wolf, Y. (2006). Evolutionary systems biology: links between gene evolution and function. *Curr Opin Biotechnol*, **17**, 481–7. 10, 35

Koonin, E., Fedorova, N., Jackson, J., Jacobs, A., Krylov, D., Makarova, K., Mazumder, R., Mekhedov, S., Nikolskaya, A., Rao, B., Rogozin, I., Smirnov, S., Sorokin, A., Sverdlov, A., Vasudevan, S., Wolf, Y., Yin, J. & Natale, D. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol*, **5**, R7. 14

Kosiol, C., Holmes, I. & Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Mol Biol Evol*, **24**, 1464–79. 45, 51, 52

Lassmann, T., Frings, O. & Sonnhammer, E. (2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res*, **37**, 858–65. 12

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. 34

Lindblad-Toh, K., Garber, M., Zuk, O. & ,et al. (64 co-authors) (2011). A high-resolution map of evolutionary constraint in the human genome based on 29 eutherian mammals (in press). *Nature*, **0**, 0. 35

Löytynoja, A. & Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–5. 45

Lynch, M. & Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–5. 43

MacKenzie, D., Defernez, M., Dunn, W., Brown, M., Fuller, L., de Herrera, S., Günther, A., James, S., Eagles, J., Philo, M., Goodacre, R. & Roberts, I. (2008). Relatedness of medically important strains of saccharomyces cerevisiae as revealed by phylogenetics and metabolomics. *Yeast*, **25**, 501–12. 12

Mallick, S., Gnerre, S., Muller, P. & Reich, D. (2009). The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res*, **19**, 922–33. 39, 40

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. 34

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. 34, 35

Massingham, T. & Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–62. 35, 36, 51, 53, 58, 61

Milinkovitch, M., Helaers, R., Depiereux, E., Tzika, A. & Gabaldón, T. (2010). 2x genomes–depth does matter. *Genome Biol*, **11**, R16. 12, 24

Miller, J., Koren, S. & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315–27. 22

Mironov, A., Fickett, J. & Gelfand, M. (1999). Frequent alternative splicing of human genes. *Genome Res*, **9**, 1288–93. 10

Moses, A. & Durbin, R. (2009). Inferring selection on amino acid preference in protein domains. *Mol Biol Evol*, **26**, 527–36. 73

Mouse Genome Sequencing Consortium & Mouse Genome Analysis Group (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–62. 34

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. 47

Nielsen, R. & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–36. 35

Notredame, C., Higgins, D. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**, 205–17. 12

Parmley, J., Chamary, J. & Hurst, L. (2006). Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol*, **23**, 301–9. 10

Pollard, K., Hubisz, M., Rosenbloom, K. & Siepel, A. (2010). Detection of non-neutral substitution rates on mammalian phylogenies. *Genome Res*, **20**, 110–21. 17

Pruitt, K., Harrow, J., Harte, R., Wallin, C., Diekhans, M., Maglott, D., Searle, S., Farrell, C., Loveland, J., Ruef, B., Hart, E., Suner, M., Landrum, M., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J., Curwen, V., Dicuccio, M., Kellis, M., Lee, J., Lin, M., Schuster, M., Shkeda, A., Amid, C., Brown, G., Dukhanina, O., Frankish, A., Hart, J., Maidak, B., Mudge, J., Murphy, M., Murphy, T., Rajan, J., Rajput, B., Riddick, L., Snow, C., Steward, C., Webb, D., Weber, J., Wilming, L., Wu, W., Birney, E., Haussler, D., Hubbard, T., Ostell, J., Durbin, R. & Lipman, D. (2009). The consensus coding sequence (ccds) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*, **19**, 1316–23. 11

Rasmussen, M. & Kellis, M. (2007). Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res*, **17**, 1932–42. 12

Rat Genome Sequencing Project Consortium (2004). Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521. 34

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. 39

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L., Guo, Y., Hériché, J., Hu, Y., Kristiansen, K., Li, R., Liu, T., Moses, A., Qin, J., Vang, S., Vilella, A., Ureta-Vidal, A., Bolund, L., Wang, J. & Durbin, R. (2008). TreeFam: 2008 update. *Nucleic Acids Res*, **36**, D735–40. 8, 12

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. 39, 40

Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L., Richards, S., Weinstock, G., Wilson, R., Gibbs, R., Kent, W., Miller, W. & Haussler, D. (2005). Evolutionarily

conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034–50. 12

Storz, J., Hoffmann, F., Opazo, J. & Moriyama, H. (2008). Adaptive functional divergence among triplicated alpha-globin genes in rodents. *Genetics*, **178**, 1623–38. 39

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. 39

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. 42

Vilella, A., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. & Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, **19**, 327–35. 9, 12

Vilella, A., Birney, E., Flicek, P. & Herrero, J. (2011). Considerations for the inclusion of 2x mammalian genomes in phylogenetic analyses. *Genome Biol*, **12**, 401. 12

Wallace, I., O'Sullivan, O., Higgins, D. & Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*, **34**, 1692–9. 11

Wang, Y. & Gu, X. (2001). Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics*, **158**, 1311–20. 39

Whelan, S. (2008). Spatial and temporal heterogeneity in nucleotide sequence evolution. *Mol Biol Evol*, **25**, 1683–94. 23

Whelan, S. & Goldman, N. (2004). Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*, **167**, 2027–43. 51

Yang, Z. & Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol*, **46**, 409–18. 35

Zhu, J., He, F., Hu, S. & Yu, J. (2008). On the nature of human housekeeping genes. *Trends Genet*, **24**, 481–4. 10