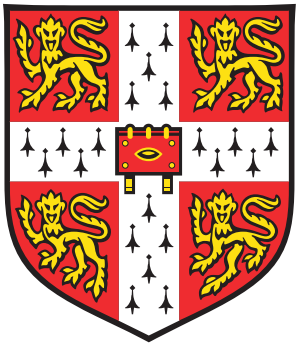
Sitewise error and constraint in mammalian comparative genomics



Gregory Jordan

European Bioinformatics Institute

University of Cambridge

A dissertation submitted for the degree of

Doctor of Philosophy

September 29, 2011

This thesis is dedicated to...

Acknowledgements

And I would like to acknowledge ...

I declare...

Abstract

I abstract...

Contents

Chapter 1

Introduction

Over the past decade, the comparative analysis of genomic sequences has immeasurably expanded our understanding of the evoluion, biology and diversity of mammals, the taxonomic class to which we belong Although the revolution in genomic medicine that was optimistically predicted during the unveiling of the draft human genome sequence is still far from being realized [Collins & McKusick, 2001; Varmus, 2010], the impact of comparative genomics on the study of human evolution, diversity and biology has been more immediate, far-reaching and deep [O'Brien *et al.*, 1999; Lander, 2011]. Many important questions in evolution have been asked—for example, what is the rate of mammalian speciation [Bininda-Emonds *et al.*, 2007; Venditti *et al.*, 2011], or what is the fraction of the genome under functional constraint [Boffelli *et al.*, 2003; Siepel *et al.*, 2005; Ponting & Hardison, 2011]—and, to some extent answered, using large amounts of genomic data.

The aim of this thesis is to show how the large-scale comparative analysis of genes and genomes can be used to identify genomic regions and biological features which have been subject to exceptional levels of selective constraint throughout mammalian evolution. When shared across many species, these evolutionary patterns can highlight genes and pathways involved in ongoing, universal mammalian genetic conflicts [Castillo-Davis *et al.*, 2004]; when observed in one or a few lineages, they may indicate more specific adaptations related to those species' unique evolutionary hsitory [Sawyer *et al.*, 2005; Nielsen *et al.*, 2007].

Furthermore, along with the increased use of high-throughput methods and datasets in biology has come a heightened awareness of the inescapable presence of noise and error within data. The study of genome sequences is no exception to this point; indeed, the many potential sources of error in any comparative genomic analysis combine to make it

difficult to assess accuracy or to identify anomalous results. This is especially problematic in comparative genomic analyses, where genome sequencing and sequence alignment errors can be mistaken for interesting biological signals [Mallick *et al.*, 2009; Schneider *et al.*, 2009a; Fletcher & Yang, 2010a; Markova-Raina & Petrov, 2011a]. Some of the difficulty of assessing results stems from a limited understanding of how various sources of error can impact downstream evolutionary analyses; thus, a secondary aim of this thesis is to better understand the impact of error on large-scale comparative analyses, and to further develop methods for appropriately predicting and handling such error.

1.1 The evolution of mammals and the mammalian genome

A major motivating factor behind the sequencing and study of mammalian genomes has been the desire to shed light on the human genome sequence through comparative study, leading to a better understanding of the diversity of genomic constraints under which our species has evolved [Mouse Genome Sequencing Consortium & Mouse Genome Analysis Group, 2002]. As the genome sequence of every animal is intertwined with all aspects of its biology, any comparison of genomes must be performed within the context of each species' phenotypic traits and evolutionary history; it will thus be useful to briefly review some salient features of the evolutionary history of mammals and their genomes which will serve as useful background for the analyses presented in this thesis.

Mammals are a diverse class of vertebrates, comprising roughly 5,400 species whose common ancestor lived ca. 165–170 million years (Myr) ago [Wilson & Reeder, 2005]. According to a comprehensive supertree constructed by Bininda-Emonds et al. using a combination of molecular data and fossil calibrations, the earliest major branching events were the split of Monotremata (containing the egg-laying mammals such as platypus and echidna) around 166 Myr ago and the divergence of the Marsupialia and Placentialia orders around 150 Myr ago. By 100 Myr ago the major placental superorders (e.g., Afrotheria, Euarchontoglires, Laurasiatheria and Xenarthra) had all diverged, and nearly all extant mammalian orders originated prior to 85 Myr ago [Bininda-Emonds *et al.*, 2007]. These dates were somewhat earlier than what had commonly been estimated based purely on fossil evidence [Archibald & Deutschman, 2001], but the early mammalian fossil record is sparse, which lends weight to the argument that the true date of origin is several Myr before the earliest discovered fossil. Taking this effect into account, an independent statistical analysis

of primate fossils provided corroborating evidence for the relatively early divergence of mammalian lineages [Martin *et al.*, 2007]. The Bininda-Emonds et al. phylogeny suggests that 43 placental lineages with extant descendants survived through the mass extinction at the K/T boundary, during which up to two-thirds of all mammalian species went extinct [Alroy, 1999]. After a 10 Myr period of overall decreased diversification levels (e.g. lowered speciation rates), most mammalian lineages continued to diversify at a relatively constant rate [Bininda-Emonds *et al.*, 2007; Martin *et al.*, 2007].

This evolutionary history has influenced the shape of the phylogenetic tree relating the extant mammalian species, a summarized version of which is shown in Figure 1.1. (Note that the dates of some of the earliest branches of the phylogeny in Figure 1.1, which was adapted from Haussler *et al.* [2009] using data from Hedges & Kumar [2009], disagree with the above description based on Bininda-Emonds *et al.* [2007], reflecting the large amount of uncertainty regarding more ancient events.) Deep but relatively short branches separate most of the ordinal groups, with the exception of Marsiupialia and Monotremata, which are separated from the other mammalian orders by much longer distances. Within each order, a fairly regular pattern of branching is seen (note, however that the tree in Figure 1.1 is truncated at the family level, omitting the relationships of individual species). Most orders are represented by several extant species, suggesting that the branch length separating any one species from its closest relative is fairly small, again with the exception of Monotremata which contains only 5 species spanning 45 Myr. These aspects of the mammalian phylogeny make it well-suited for large-scale comparative analysis, as long branches separating sequenced genomes (which are a major source of alignment error and uncertainty in evolutionary reconstruction) can continue to be shortened by sequencing additional species. Indeed, this was part of the motivation behind the Mammalian Genome Project (MGP) [Lindblad-Toh *et al.*, 2011], which generated much of the data used throughout this thesis and which I will introduce in more detail in Chapter 3.

Before the K/T boundary, ancestral mammal and primate species were likely smaller in size than they are today, as the ecological niches for larger animals were occupied by dinosaurs [**?**Smith *et al.*, 2010]. Their diet is assumed to have been largely insectivorous, as folivory in extant species is observed mainly in larger mammals [Smith *et al.*, 2010] (but see Martin *et al.* [2007] for an alternative perspective favoring a more folivorous primate ancestor). After the K/T extinction event around 65 Myr ago, mammals eventually diversified to occupy a wide range of the ecological roles left vacant by extinct species, with many lineages undergoing highly specialized morphological and behavioral adaptations and

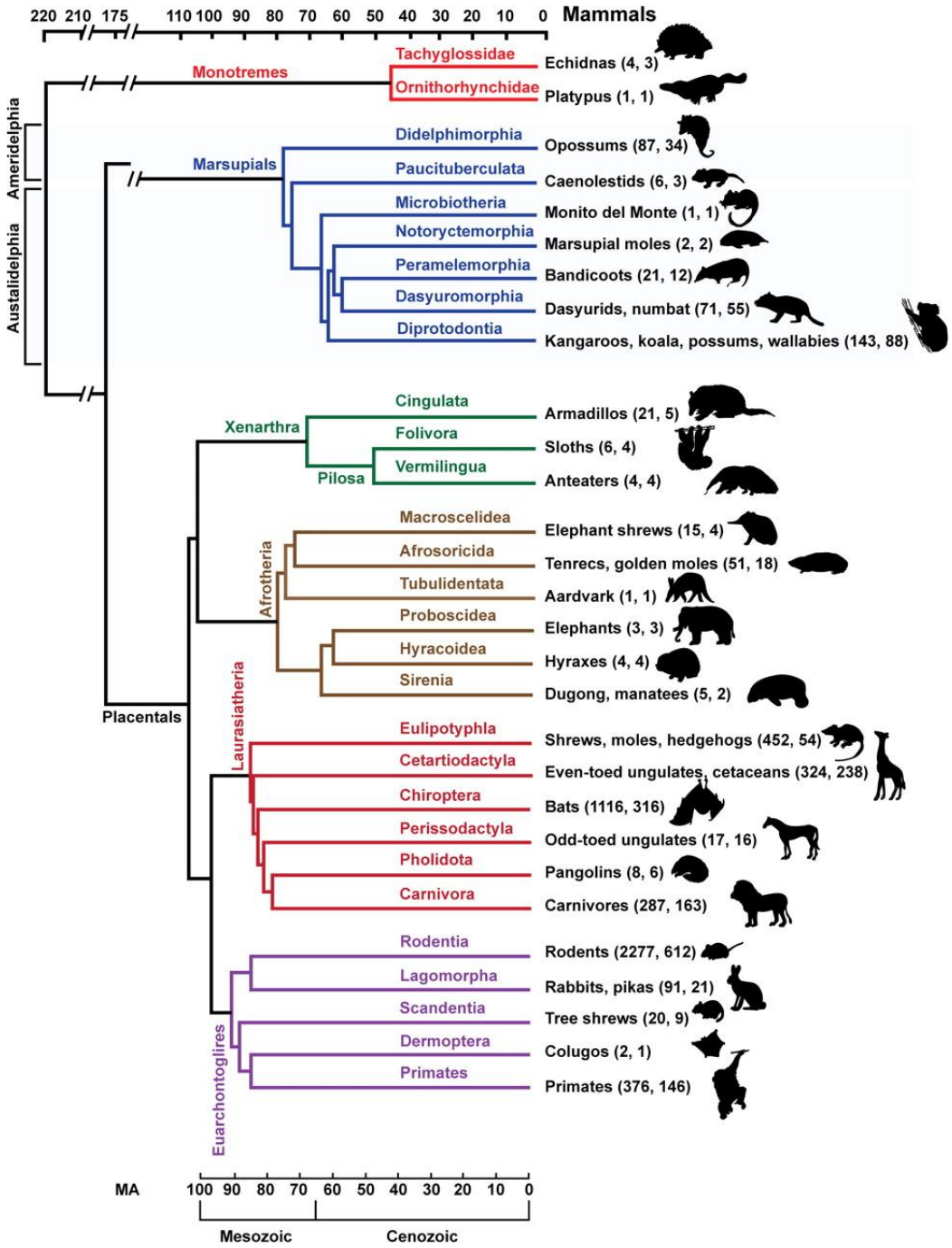


Figure 1.1: A time-resolved consensus phylogeny of the major mammalian lineages. Topologies and dates use data from Hedges and Kumar [2009]. Each terminal branch represents a mammalian family. The number of species contained in each family is included as the first number in parentheses after each family name (e.g., there are 2,227 species of rodents). Figure taken from Haussler *et al.* [2009].

the range of mammalian body sizes expanding by four orders of magnitude [Alroy, 1998]. A long-term trend towards larger body sizes has been observed in many lineages; the hypothesis that this is a general feature of mammalian evolution has been termed Cope's rule [Alroy, 1998], though its universality is controversial [Finarelli & Flynn, 2006; Monroe & Bokma, 2010].

The body size of mammals and their ancestors is an important consideration in sequence analyses, as body size has been shown to correlate with the overall rate of substitution in multicellular eukaryotes [Mouse Genome Sequencing Consortium & Mouse Genome Analysis Group, 2002; Hwang & Green, 2004; Welch *et al.*, 2008; Galtier *et al.*, 2009; Romiguier *et al.*, 2010; Bromham, 2011]. Other phenotypic features, such as metabolic rate and generation time, have been similarly linked to genomic evolutionary rates [Martin & Palumbi, 1993; Nabholz *et al.*, 2008], but all three of these characters are strongly cross-correlated in mammals, making it difficult to isolate the effect of each particular variable on the overall evolutionary rate or to identify the causative factor behind such variation. Regardless, it is clear that extant mammals exhibit a wide range of neutral evolutionary rates [Bininda-Emonds, 2007], with proposed explanatory factors including differences in the amount of mutagenic free radicals associated with an animal's metabolic rate, different rates of germ line cell divisions per year, and different DNA repair control mechanisms [Baer *et al.*, 2007].

The correlation between body size and neutral evolutionary rate has an important consequence for comparative genomic studies in mammals: extant species groups with smaller body sizes are expected to have experienced more DNA substitutions since their common ancestor than larger-bodied species groups, leading to increased branch lengths within smaller-bodied clades when branches are scaled by the neutral evolutionary rate. Figure 1.2 shows a phylogenetic tree for 29 mammals scaled by the genome-wide rate of observed substitutions, emphasizing the high observed substitution rates of most rodents and low rates of most hominids and some larger-bodied species from other mammalian orders. In comparative analyses where a larger number of substitutions increases the power of a method to detect a genomic feature or estimate an evolutionary rate (e.g., in detecting conserved regulatory elements or positive selection), the larger branch lengths of rodent species would be expected to result in improved power and statistical accuracy. This effect will be especially important in Chapter 4, where I compare sitewise estimates of selection pressures from groups of species from different mammalian orders.

A second biological characteristic showing significant variation between mammals, the effective population size (N_e), has important consequences for the study of genomic re-

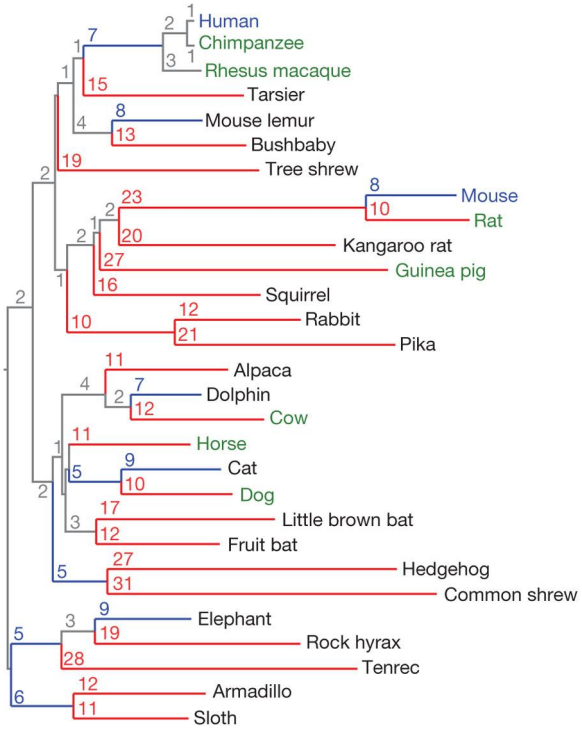


Figure 1.2: A phylogeny of 29 mammalian species, with branch lengths scaled with the neutral evolutionary rate estimated from genome-wide DNA alignments. Note the increased branch lengths of most rodent species (e.g., mouse, rat, pika) and various small members of Laurasiatheria (e.g., little brown bat, hedgehog, common shrew) relative to most primates (e.g., human, chimpanzee). Figure taken from Lindblad-Toh *et al.* [2011].

gions subject to natural selection [Charlesworth, 2009]. N_e is a fundamental parameter in population genetics, describing the size of an idealized population that exhibits the same amount of dispersion of allele frequencies due to genetic drift as the real population under study [Wright, 1931; Woolfit, 2009]. Many aspects of a population can influence its N_e, including the census count, breeding patterns, and geographical distribution of individuals [Caballero, 1994], and studies within mammals have consistently shown a much larger N_e for rodents than for primates and for small versus large mammals [Eyre-Walker *et al.*, 2002; Popadin *et al.*, 2007; Halligan *et al.*, 2010], suggesting that extant mammalian populations can significantly vary in this parameter. It is beyond the scope of this thesis to provide a comprehensive discussion of N_e and its importance within population genetics and molecular evolution, but Woolfit [2009] and Charlesworth [2009] provide focused reviews of the subject. The main predicted impact of N_e on the study of fixed substitutions between species is that slightly deleterious mutations are more likely to become fixed within a pop-

ulation having a small N_e versus a population with a large N_e. Closely tied to this effect is the prediction of the nearly neutral thoery of molecular evolution [Kimura, 1985] that many mutations in protein-coding regions are slightly deleterious and thus subject to this dependence on N_e [Kimura & Ohta, 1974a; Kimura, 1985; Ohta, 1992]. Several empirical studies have supported this hypothesis, showing that a different N_e leads to different rates of protein evolution in bacteria [Moran *et al.*, 2008; Warnecke & Rocha, 2011], birds [Axelsson & Ellegren, 2009] and mammals [Kosiol *et al.*, 2008a; Ellegren, 2009b] (but see Bachtrog [2008] for potentially contradictory evidence from *Drosophila*). Any analysis of comparative evolution in mammals should thus evaluated with respect to these well-established trends; in Chapters 4 and 5 I consider the possible effects of N_e on the observed patterns of positive selection within different groups of mammals, and in Chapter 6 I use genome-wide comparisons to estimate ancestral N_e in our closest primate relatives.

Key features of the mammalian genome itself are also worth highlighting. Mammals contain relatively large genomes (roughly 3 Gb of DNA, ranging from 2.5 to 4.5 Gb) with between 20 to 80 chromosomes [Bachmann, 1972]. The large range in chromosome count is likely a result of the high rate of chromosomal rearrangement in mammals [Eichler & Sankoff, 2003; Pevzner & Tesler, 2003]. Some regions of "rearrangement hotspots" show especially large amounts of large-scale genomic shuffling within mammals, and it has been speculated that these regions have contributed to the many lineage-specific gene family expansions in mammals [Eichler & Sankoff, 2003]. Breakpoints of mammalian chromosomal rearrangements tend to occur near transposable elements [Zhao & Bourque, 2009], which are small DNA sequences capable of replicating throughout the genome [Lander & International Human Genome Sequencing Consortium, 2001]. Transposable elements, a diverse class of sequence elements representing a variety of transposition mechanisms and sequence characteristics, together comprise roughly 45% of DNA in the human genome and have contributed signifciantly the ancient and ongoing evolution of mammalian genomes [Lander & International Human Genome Sequencing Consortium, 2001; Cordaux & Batzer, 2009].

In contrast to the rapid turnover of noncoding DNA and high rate of genomic rearrangement observed in mammalian genomes, the protein-coding gene complement appears to be less variable. Initial estimates of roughly 30,000 protein-coding [Lander & International Human Genome Sequencing Consortium, 2001; Mouse Genome Sequencing Consortium & Mouse Genome Analysis Group, 2002] genes in the human and mouse genomes have been lowered based on accumulating functional and phylogenetic evidence to roughly 21,000

genes [Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007]; the most recent gene annotations from Ensembl [Flicek *et al.*, 2011] contain 20,599 human and 21,873 mouse protein-coding genes. A majority of these genes are shared between all mammals" human and mouse are estimated to share 80% of genes in a "one-to-one" fashion, meaning no apparent gene duplications or deletions occurred since the common ancestor [Mouse Genome Sequencing Consortium & Mouse Genome Analysis Group, 2002], and a wider group of mammals including platypus show detectable orthologs (including genes with duplications or deletions in one or more lineages) in 82% of genes [Warren *et al.*, 2008]. Despite the relative consistency of the mammalian protein-coding catalogue, features such as alternative splicing and domain concatenations have been identified as potential contributors to mammalian phenotypic complexity and diversity [Lander & International Human Genome Sequencing Consortium, 2001].

In any study of vertebrate protein-coding genes, the two round (2R) hypothesis looms large. Originally proposed by Ohno [1970], the 2R hypothesis suggests that two polyplidization events occurred during the early evolution of the vertebrate common ancestor, explaining the observation that vertebrates often have up to four homologs of invertebrate genes [Hokamp *et al.*, 2003]. For three decades the veracity of the 2R hypothesis was hotly debated [McLysaght *et al.*, 2002; Dehal & Boore, 2005], but analyses based on comparisons between whole-genome sequences of several fish and basal chordates have repeatedly confirmed its predictions [Kasahara, 2007; Putnam *et al.*, 2008]. In addition to having interesting implications for the evolution of the immune system and of morphological diversity within vertebrates [Hughes & Yeager, 1997; Hoffmann *et al.*, 1999; Van de Peer *et al.*, 2009] the existence of ancient genomic duplications can cause problems in the inference of homology relationships between genes. Many of these aspects will be considered in more detail in Chapter 3 when a set of mammalian orthologs suitable for evolutionary analysis is identified.

1.2 Models of sequence evolution

The above section described several important genomic and evolutionary features of the mammalian clade. Many of those well-established observations resulted from the application of mathematical models of sequence evolution to comparative genomics data, which is now a widespread analytical appraoch. This section briefly introduces a selection of important methods and models for evolutionary analysis which will be applied throughout

the remaining chapters.

As the hereditary material of all free-living organisms, DNA represents a record of the history of life on earth. When an individual gives rise to offspring, some segments of DNA are replicated and passed on to all its descendants; importantly, the processes of DNA replication and repair are imperfect [Arnheim & Calabrese, 2009], and the resulting errors—called mutations—are passed on to successive generations. In addition to being a major source of the variation between individuals invoked in Darwin's theory of natural selection [Darwin, 1859], mutations in DNA leave a molecular record of evolutionary relationships and of the passage of time. The gradual accumulation of mutations in DNA, commonly observed as differences in DNA sequences between species at the same homologous location, can be reasonably modeled using phylogenetic trees and Markov models of sequence change [Yang, 2006].

The earliest observations that biological sequences tend to randomly change over time were made from the sequences of proteins, which are cellular molecules comprised of amino acid units whose arrangement is encoded in the DNA sequences of exons within genes. Zuckerkandl and Pauling, who were analyzing the amino acid sequences of hemoglobin genes from various species, noted in 1962 that the number of changes between sequences from different species corresponded well with the evolutionary distance those species based on fossil evidence; this led them to hypothesize that evolution at a molecular level may occur at a largely constant rate [Zuckerkandl & Pauling, 1962; Morgan, 1998]. Zuckerkandl and Pauling continued to explore the implications and applications of their "molecular evolutionary clock" hypothesis, using hemoglobin and cytochrome C sequences, for example, to estimate the date of human-gorilla divergence (at 11 million years) and to infer the ancestral protein sequences of mammalian ancestors [Zuckerkandl & Pauling, 1965]. Since those pioneering studies in proteins, a variety of evolutionary models have been developed to describe observed patterns of amino acid and DNA substitution. As this thesis is concerned largely with the application of such methods, I will only briefly summarize the key features of the more popular evolutionary models; Yang [2006] provides a comprehensive mathematical treatment.

The simplest model of DNA substitution, proposed by Jukes & Cantor [1969], assumes that every nucleotide has the same rate of changing into any other nucleotide. Although the assumption of equal rates may be reasonable for modeling a random process, the point mutation of a DNA base pair is a biochemical process (or rather, a set of potentially many unobserved biochemical processes which all produce the same class of observable result)

which may contain biases towards or against certain types of mutations. This was quickly discovered to be the case: analysis of the ever-increasing number of available biological DNA sequences showed that in many datasets *transitions*, defined as substitutions between two pyrimidine nucleotides (e.g., T→C or C→T) or between two purine nucleotides (e.g., A→G or G→A), are more common than *transversions*, defined as substitutions from a purine to a pyrimidine or vice-versa. Kimura [1980] thus proposed a more complex model, called K80 or Kimura's two-parameter model, which accounted for this bias. Specifically, Kimura's K80 model extends JC69 by incorporating an additional parameter, κ, referred to as the transition/transversion ratio, which represents the ratio of the rate of transition versus the rate of transversion substitutions. When κ is greater than one, transitions occur at a higher rate than transversions, providing a better fit to most biological datasets. The κ parameter of K80 a prototypical example of the parametric approach to building evolutionary models, where a parameter is introduced into the model which allows for a commonly-violated assumption of the simpler model to be relaxed. The value of the parameter is not specified in the model, however; it is usually estimated from the data on a case-by-case basis from the data using maximum likelihood (ML) [Whelan *et al.*, 2001].

Several nucleotide models were subsequently described which relax various assumptions of the JC69 and K80 models Whelan *et al.* [2001]; Yang [2006]. One especially unrealistic feature of K80 is its symmetric nature (meaning that the rate of substitution from one nucleotide to another is the same as the rate of the reverse substitution, e.g. G→C = C→G). A symmetric DNA Markov chain yields equal nucleotide frequencies when the substitution process reaches equilibrium, meaning that any starting DNA sequence, if left to evolve long enough under such a process, will end up with equal nucleotide frequencies. In reality, many biological sequences contain highly unequal nucleotide frequencies (owing to a variety of possible selective or mutationsal biases), making it inappropriate to assume an equal base composition [Yang, 2006]. Thus, models such as FEL [Felsenstein, 1981], HKY [Hasegawa *et al.*, 1985] and TN93 [Tamura & Nei, 1993] were developed to allow for various combinations of unequal base frequencies and an unequal transition/transversion ratio. The most general reversible nucleotide model, REV [Tavaré, 1986], includes four parameters describing the equilibrium nucleotide frequencies and six rate parameters, one for each possible pair of substitutions.

In contrast to the primarily parametric DNA models, evolution models for amino acid substitutions have generally been estimated empirically [Whelan *et al.*, 2001]. Developed during the early days of evolutionary sequence analysis, the JTT [Jones *et al.*, 1992] and

Dayhoff [Dayhoff & Schwartz, 1978] amino acid models were estimated using parsimony-based counting methods. The parsimony principle was first used to infer phylogenetic trees and ancestral protein sequences from each set of aligned proteins within a large database. Based on those phylogenetic trees and ancestral sequences, the entries of the 20x20 amino acid substitution matrix were populated with counts of inferred amino acid substitutions, and these counts were used to estimate a Markov substitution model. More recently, empirical amino acid models were estimated using ML methods, which improved upon a number of methodological issues with the parsimony approach [Adachi & Hasegawa, 1996; Whelan & Goldman, 2001].

A few important assumptions are shared by all of the models already described, namely that all sites within an alignment are (a) evolving independently of one another, (b) evolving under the same evolutionary process and at the same evolutionary rate (c) related by the same underlying phylogenetic tree. I will briefly review the development of models which relax these assumptions, all of which are clearly violated in many real datasets.

Independence between sites is generally a difficult assumption to relax for computational reasons [TOCITE], but the hypermutability of CpG dinucleotides within mammalian genomes (where CpG denotes a C nucleotide followed by a G nucleotide, the "p" representing the phosphodiester bond separating nucleotides on the same strand of a DNA molecule) has provided strong impetus to incorporate at least a dinucleotide context into models for estimating nucleotide substitution rates from large mammalian alignments [Blake *et al.*, 1992; Hwang & Green, 2004; Siepel & Haussler, 2004]. CpG hypermutability results from the methylation and subsequent deamination of the cytosine nucleotide at CpG sites in most mammalian genomic DNA; although all cytosine nucleotides are prone to deamination, cytosine deamination produces uracil, which is removed by uracil glycosylase allowing for DNA repair mechanisms to replace the original cytosine, while 5-methylcytosine produces thymidine, which results in frequent C→T transitions [Ehrlich *et al.*, 1982; Hwang & Green, 2004]. Context-dependent substitution models have shown that CpG mutations are by far the dominant form of mutation in mammalian genomes, and such models have also been essential for studying the evolution of GC content and mammalian isochores [Duret *et al.*, 2006; Duret & Arndt, 2008].

The major violation of the assumption of the same evolutionary rate and model between sites is the observation that the

[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 Detecting purifying and positive selection in proteins

[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.]

[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.]

1.4 Outline of the thesis

In this thesis, I describe... three studies concerning the use of phylogenetic codon models to identify

Chapter 2

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

2.1 Introduction

The decreasing cost of DNA sequencing has triggered a striking increase in the number of model and non-model organisms with planned genome sequencing projects, suggesting that the range and scale of comparative genomics applications will continue to expand [Green, 2007b; The ENCODE Project Consortium, 2007]. The existence of clusters of closely-related genome sequences across a wide taxonomic range has led to a better understanding of which aspects of molecular evolution are variable and which are constant [Wolf *et al.*, 2009], and an increased sampling of species should continue to boost the power and accuracy of individual analyses within a given clade.

The study of protein evolutionary rates and selective pressures in particular has flourished as a result of the growth in comparative genomic datasets. This is especially beneficial for the calculation of spatially precise evolutionary estimates, as additional species sampling has been shown to be an effective means of boosting the accuracy and power of sitewise detection of positive selection and evolutionary constraint [Anisimova *et al.*, 2001b; Massingham & Goldman, 2005a]. Site-specific evolutionary estimates have proved especially valuable when analyzed in conjunction with other protein-based datasets such as structural features [Lin *et al.*, 2007; Ramsey *et al.*, 2011], human population diversity [1000 Genomes Project Consortium, 2010] and human disease mutations [Arbiza *et al.*,

Table 2.1: Parameter Values Used in Simulations

Tree			Insertions and Deletions			ω Distribution		
Taxa	Source	MPL	Size Distribution	Mean Length (Std. Dev.)	Rate	Shape	Mean	$p(\omega > 1)$
6	Artificial		power law			lognormal		
17	β-globin	0.05–2.0	decay: 1.8	3.33 (5.51)	0–0.2	log mean: -1.864	0.277	0.06
44	Vertebrates		max length: 40			log SD: 1.201		

NOTE.—MPL is the mean path length of the tree in units of substitutions per synonymous site. Indel lengths are measured in units of codons, and the indel rate is defined as the number of insertion & deletion events per substitution.

2006].

A major concern in the detection of protein positive selection is that the effect of alignment error is not well characterized. Intuitively, one might expect alignment error to result mainly in an increased number of false positives, as the spurious alignment of non-homologous codons on average would result in a high number of apparent nonsynonymous substitutions and a low number of synonymous substitutions (since two randomly chosen codons are more likely to be nonsynonymous than synonymous). However, false negatives may also be introduced, either through the introduction of synonymous but non-homologous codons into a positively-selected site (thus reducing power due to an inflated synonymous substitution rate) or through the failure to align truly homologous codons at a positively-selected site (reducing power due to less evidence for positive selection at that site). Since different aligners employ a variety of algorithms, evolutionary models, and heuristic optimizations [Notredame, 2007], each program may be more or less prone to different types of alignment error, causing potentially large variations in the nature and magnitude of its impact on the detection of positive selection. Different aligners may also be designed for different downstream applications, such as phylogenetic inference or functional annotation [Morrison, 2009], making the optimal choice of aligner potentially dependent on the way in which the resulting alignment will be used. In this chapter, I focused on the sitewise detection of positive selection.

In addition, I expect the protein structure and the evolutionary divergence of a dataset to contribute to the effects of alignment error. Differently structured protein regions show variable tolerance to biological indels, with indels more common in extracellular and transmembrane proteins than in highly folded enzymes and housekeeping genes [de la Chaux *et al.*, 2007]. This suggests that well-folded protein regions will experience fewer biologi-

cal indels—and will therefore be less susceptible to alignment error—than less structured regions.

Furthermore, the evolutionary divergence of a dataset affects the power of sitewise inference and the prevalence of indels in multiple ways. As maximum-likelihood methods for detecting positive selection require data in the form of observed substitution events, they show little power at low divergence and their highest power at intermediate to high divergence levels [Anisimova *et al.*, 2001b]. However, alignment error should be greatest at high divergences, which may have the effect of reducing power. These two trends suggest that the overall power will be low at both extremes of divergence, with little inference power at low divergence (due to the scarcity of data in the form of observed substitutions) and an overwhelming amount of alignment error at high divergence (due to the large number of indel events).

Luckily the majority of genes in many biological clades of interest (such as mammals, vertebrates, fruit flies, and yeast) fall within the middle range of divergences where sitewise methods are at their most powerful and where multiple alignment is a difficult—but not hopeless—problem. As such, it is important to seek an understanding of the impact of alignment error on overall error rates within this important range of divergence levels.

A number of empirical analyses have established that errors in gene sequencing, annotation and alignment can contribute to errors in downstream evolutionary analyses such as phylogeny inference [Wong *et al.*, 2008] and estimates of positive selection [Schneider *et al.*, 2009b; Markova-Raina & Petrov, 2011b]. Most recently, Markova-Raina and Petrov [2011b] showed that the detection of positively-selected sites and genes in *Drosophila* genomes is highly sensitive to aligner choice, with PRANK's codon model [Löytynoja & Goldman, 2008] consistently producing alignments with the lowest amount of positive selection. Still, according to the authors' manual inspection of alignments, even positively-selected sites identified with PRANK alignments contained a sizable proportion of apparent false positives.

A limitation of the analysis of error in empirical datasets is the lack of a benchmark set of true alignments and positively-selected sites. Markova-Raina and Petrov [2011b] used their expected general effect of alignment error (an increase in false positives due to misalignment of non-homologous codons) as a proxy by which to compare different methods, allowing for the conclusion that PRANK was the least error-prone aligner in their analysis. However, the absolute number of false positives remained uncertain and there was the possibility of conflating multiple sources of error: in addition to alignment error, the authors noted

that gene mis-annotation was responsible for many apparent false positives, and there is also an expected error rate from the likelihood inference method itself. This limitation leaves important and interesting questions, regarding the nature of alignment error and its quantitative impact on the detection of positive selection, unanswered by empirical studies.

Controlled simulation experiments provide a natural framework for investigating error rates in detail, allowing one to pinpoint the sources of error in multi-step analyses such as alignment followed by evolutionary inference. This approach has been employed in assessing the robustness of phylogenetic inference methods to misalignment [Dwivedi & Gadagkar, 2009; Ogden & Rosenberg, 2006; Löytynoja & Goldman, 2008] but those results cannot be easily extrapolated to the analysis of sitewise selective pressures. More recently, Fletcher and Yang [2010b] performed a series of simulation experiments investigating alignment error in the use of the branch-site test to detect positive selection in genes. Their results showed that most aligners caused false positives by over-aligning codons and that datasets from mammalian and vertebrate gene families contain enough evolutionary divergence to make false positive errors resulting from misalignment a legitimate concern.

Reflecting a widespread awareness of the problem of misalignment, methods for identifying and removing uncertain or unreliable alignment regions have been commonly used in phylogenetic and molecular evolutionary analyses. The popular Gblocks program applies a set of heuristic criteria to identify conserved blocks deemed suitable for phylogenetic or evolutionary analysis [Castresana, 2000] while a number of aligners such as T-Coffee [Notredame *et al.*, 2000b] and PRANK [Löytynoja & Goldman, 2008] produce estimates of alignment confidence or reliability. GUIDANCE, which measures the robustness of alignment regions to perturbations in the guide tree used for progressive alignment, has also been proposed as an alignment confidence score [Penn *et al.*, 2010]. Unfortunately, despite their widespread use, the impact of the many available alignment scoring and filtering methods on phylogenetic and evolutionary analyses has not been well studied. Even for a single filtering program, Gblocks, results have been contradictory: one simulation-based study found that it improved the phylogenetic signal [Talavera & Castresana, 2007] while an empirical study across a wide range of taxa found that Gblocks-filtered alignments produced worse phylogenetic trees than unfiltered alignments [Dessimoz & Gil, 2010]. A recent study using a variety of filters suggested that the benefit of alignment filtering (in terms of improved accuracy) outweighs the cost (in terms of reduced power) when applied to detecting positive selection [Privman *et al.*, 2011], but this analysis was limited to a small range of possible evolutionary scenarios as discussed below. With the application of published

filtering methods to alignments before testing for positive selection becoming standard practice [Studer *et al.*, 2008a; Aguileta *et al.*, 2009], continued investigation of potential benefits of alignment filtering to the detection of positive selection seems well-warranted.

This paper aims to use a simulation framework to incorporate alignment error and alignment filtering into estimates of the error rate and power of sitewise evolutionary inference of positive selection. Our approach builds on those of Anisimova et al. [2002b], Fletcher and Yang [2010b], and Privman et al. [2011], using simulated protein alignments including insertions and deletions to evaluate methods for detecting sitewise positive selection. Furthermore, I incorporate a diverse sample of aligners and alignment filters into an experimental design that differs from previous ones in a number of important ways.

I focus on sitewise detection of positive selection occurring throughout a phylogeny and evaluate the impact of a number of alignment filtering methods on the sitewise analysis. Thus, the biological hypothesis being investigated is different from that studied by Fletcher and Yang [2010b], who focused on genewise selection acting at specific branches. Privman et al. [2011] recently published a related paper considering the evolutionary chracteristics of three HIV-1 genes. They concluded that alignment filtering improves the performance of positive selection inference by reducing false positive results. While this may be true for these three genes (and perhaps for HIV-1 in general), HIV-1 is known to evolve with widespread positive selection in the human host [Yang *et al.*, 2003]. Results valid for these genes may not be widely applicable to large-scale vertebrate and mammalian comparative datasets, which exhibit less adaptive evolution [Kosiol *et al.*, 2008b] and which comprise a larger diversity of protein structures and a wider range of species divergence levels.

I hypothesize that the divergence level and indel rate, two important evolutionary factors which are highly variable within and between different genomes, may strongly affect the performance of methods for alignment and detection of selection. Accordingly, our simulations encompass a wide range of biologically plausible indel rates and divergence levels while fixing other parameters at values typical to those encountered in the sitewise analysis of vertebrate gene families.

2.2 Methods

Alignment Simulations

An overview of the simulation parameters used in this study can be found in Table ??. Three rooted trees were used to guide the simulation of protein-coding DNA alignments:

Table 2.2: Genome-wide Divergence Estimates for Commonly Analyzed Eukaryotes

Species	Pairwise dS	Root-to-tip dS	Reference
Human-Chimp	0.01	(0.005)	Nei *et al.*, 2010
Human-Mouse	0.43	(0.215)	Nei *et al.*, 2010
Human-Mouse	0.5 - 0.8	(0.25 - 0.4)	Ogurtsov *et al.*, 2004
Human-Chicken	0.9	(0.45)	Nei *et al.*, 2010
Human-Chicken	1.66	(0.83)	Hillier et al., 2004
Human-Zebrafish	1.38	(0.69)	Nei *et al.*, 2010
Vertebrates	–	0.75	Siepel et al., 2005
Drosophila	–	1.0	Siepel et al., 2005
Yeasts	–	1.25	Siepel et al., 2005

NOTE.—The root-to-tip dS is equivalent to the MPL (mean path length) used in our simulations. For two-species comparisons where the pairwise dS was given, the root-to-tip dS was calculated as half of the pairwise dS and is included in parentheses.

the artificial 6-taxon tree used by Anisimova et al. [2001b] and Massingham and Goldman [2005a] rooted at its midpoint, the 17-taxon vertebrate β-globin tree from Yang et al. [2000] and the 44-taxon vertebrate tree used by the ENCODE project [The ENCODE Project Consortium, 2007; Nikolaev *et al.*, 2007b]. Trees, shown with their original branch lengths in Figure 2.1, were scaled to comparable divergence levels by normalizing their mean path length (MPL), defined as the root-to-tip branch length averaged across all lineages in the tree. I simulated alignments with MPL divergence between 0.05 and 2.0 synonymous substitutions per synonymous site, spanning the range of evolutionary divergences observed in several clades of organisms with fully-sequenced genomes (Table **??**).

The INDELible program [Fletcher & Yang, 2009] was used to simulate codon sequences with indels along each phylogenetic tree. The length of the root sequence was set to 500 codons and $\boldsymbol{\kappa}$ (the ratio of transition to transversion substitutions) was fixed at 4. Indel lengths were drawn from a discretized power-law distribution with an exponential decay parameter of 1.8 and a maximum value of 40, yielding a mean indel length of 3.33 codons and standard deviation of 5.51 codons. The power-law model of indel lengths is well-supported by empirical studies [Benner *et al.*, 1993; Cartwright, 2009] and manual inspection of alignments from a range of parameter values identified the chosen model parameters as resulting in alignments most closely resembling those encountered in vertebrate alignments. The ratio of insertion to deletion events was set to 1, and the rate of indel formation was varied between 0 and 0.2 indel events per substitution per site.

The distribution of sitewise selective pressures (embodied by the parameter ω, the ratio of the rate of nonsynonymous substitution to the rate of synonymous substitution) was modeled with a log-normal distribution derived from a maximum-likelihood fit to a large

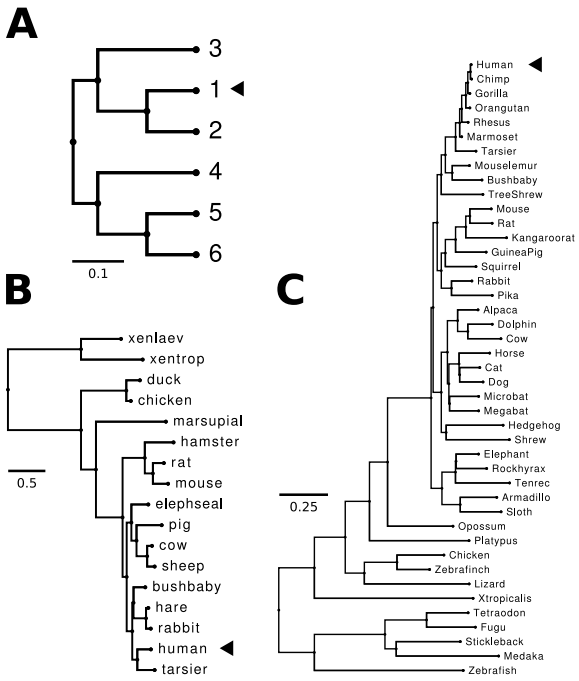


Figure 2.1: Phylogenetic trees used for simulation and analysis. The original scale for each tree is indicated by a scale bar, but trees were scaled to equal mean path length (MPL) divergence levels for simulation. (A) A 6-taxon artificial tree used in previous simulations [Anisimova *et al.*, 2001b; Massingham & Goldman, 2005a]. (B) A tree estimated from β-globin genes of 17 vertebrates and used in previous empirical analyses and simulation studies [Anisimova *et al.*, 2001b, 2002b]. (C) The 44-species tree used by the ENCODE project [The ENCODE Project Consortium, 2007; Nikolaev *et al.*, 2007b]. The nodes indicated by arrows were used as the reference species when comparing the true and inferred alignment (see Methods).

dataset of sitewise selective pressures estimated from mammalian gene trees (Lindblad-Toh et al. 2011; log-normal parameters shown in Table **??**). This distribution, with mean ω of 0.28 and 6% of sites having $\omega > 1$, is consistent with the structure-based expectation of many protein sites under purifying selection and few under neutral selection or positive selection [Smith, 1970; Kimura & Ohta, 1974b]. INDELible's general discrete model of sitewise ω variation was used to approximate the log-normal distribution by splitting the probability density into 50 equally-spaced bins between ω values of 0 and 3, with the highest bin containing the probability density for all values $\omega > 3$.

Branch lengths for each of the simulation trees were scaled before simulation to correct for the difference between our definition of branch lengths as the number of synonymous substitutions per synonymous site (dS) and INDELible's interpretation of branch length as the average number of substitutions per codon (t) [Fletcher & Yang, 2010b]. They are

related approximately by $t = 3(NdN + SdS) = 3dS(\bar{\omega}N + S)$, where N and S are the proportion of nonsynonymous and synonymous sites and $\bar{\omega}$ is the mean ω across all sites. S is approximately 0.3 when $\boldsymbol{\kappa} = 4$ [Yang & Nielsen, 1998] and the mean ω ratio for our chosen distribution is 0.277, yielding a dS-to-t conversion factor of 1.48 for all simulations performed.

Sequence Alignment and Filtering

Alignments were inferred using six alignment algorithms chosen for their widespread use or demonstrated accuracy: ClustalW v1.82 [Thompson *et al.*, 1994], MAFFT [Katoh *et al.*, 2005b], ProbCons [Do *et al.*, 2005], T-Coffee [Notredame *et al.*, 2000b] and two variants of PRANK [Löytynoja & Goldman, 2008] based on an amino acid model (subsequently referred to as PRANK_{AA}) or an empirical codon model (subsequently referred to as PRANK_{C}). Unaligned amino acid sequences were given as input to all alignment programs (except PRANK_{C}, which was provided the unaligned DNA sequences) and all software was run using default parameters with the true phylogenetic tree given as input where possible.

Alignments were filtered by masking out residues based on the output of three alignment scoring methods: Gblocks conserved blocks [Castresana, 2000], T-Coffee consistency scores [Notredame *et al.*, 2000b; Notredame & Abergel, 2003], and GUIDANCE alignment confidence scores [Penn *et al.*, 2010]. Gblocks, which identifies entire alignment columns as conserved or not conserved, was run using an increased gap tolerance and a reduced minimum block length in order to reduce the amount of each alignment removed (command-line parameters *b5=a* and *b4=3*), and all residues from any columns not within an identified conserved block were masked with *N*s.

GUIDANCE and T-Coffee filters produce scores for each residue, allowing individual residues to be masked instead of entire columns. Privman et al. [2011] found a residue-based filter to be more effective than its column-based equivalent, and I opted to filter residues instead of alignment columns where possible. GUIDANCE generates many replicate alignments, each using a slightly perturbed guide tree, with either MAFFT or PRANK_{AA} as the bootstrap aligner. The program then assigns to each residue from the input alignment a score from 0 to 1 based on how consistently it was placed in the replicate alignments. In order to maximize the similarity between the input aligner and the bootstrap aligner, I ran GUIDANCE with 100 MAFFT replicates when filtering ClustalW alignments and with 30 PRANK_{AA} replicates when filtering PRANK_{C} alignments. T-Coffee calculates

the residue-wise consistency between an input multiple alignment and independently calculated pairwise alignments [Notredame & Abergel, 2003], rounding and normalizing residue scores into integers between 0 and 9. T-Coffee was run using its default settings and the *evaluate_mode –output=score_ascii* command-line parameters to output alignment scores.

To filter alignments based on these residue-wise scores, a cutoff threshold was chosen for each method (0.5 for GUIDANCE and 5 for T-Coffee) and residues equal to or below that threshold were masked. On a per-alignment basis, if the default threshold caused greater than 50% of residues to be masked, then the threshold was relaxed to the highest value for which at least 50% of residues remained. I found this adjustment necessary because the scores from GUIDANCE and T-Coffee were strongly affected by the simulation conditions, with much lower average scores at higher indel rates and divergences. Requiring at least 50% of residues to remain unmasked ensured that enough data were available for meaningful evolutionary analysis, mimicking typical treatment of real data sets.

Two unrealistic but informative datasets were produced to serve as controls. First, the true simulated alignment was included in order to evaluate the sitewise performance without any alignment error. Second, an additional filtering method was constructed to represent an unattainable best-case scenario for sequence filtering, using knowledge of the true alignment to assign a score to each residue reflecting how correctly is has been placed in the inferred alignment. The approach taken was to calculate, for each residue, the branch length of the correct sub-tree (defined as the sub-tree connecting all sequences to which the current residue was correctly aligned) divided by the branch length of the total aligned sub-tree (defined as the sub-tree connecting all sequences with non-gap residues at the current alignment column). This residue-wise score ranges from 0 to 1 and reflects the expectation that correctly-aligned evolutionary branch length is the main source of information from which sitewise inference methods derive their power. I refer to this method as the 'optimal' filtering method. Scores were handled in a manner similar to GUIDANCE and T-Coffee, using a score threshold of 0.5.

Sitewise Evolutionary Analysis

Sitewise estimates of selective pressures were calculated using maximum-likelihood methods implemented in the Phylogenetic Analysis by Maximum Likelihood (PAML; Yang 2007a) and Sitewise Likelihood Ratio (SLR; Massingham & Goldman 2005a) software packages.

The *codeml* program from PAML implements a number of likelihood ratio tests (LRTs) for detecting the presence of positive selection in a gene while allowing the ω ratio to vary

among sites [Yang *et al.*, 2000]. These models, known as the sites or random sites models, use a variety of predefined statistical distributions to account for heterogeneous ω ratios among sites. After the likelihood optimization is performed, Bayesian methods can be used to estimate the posterior probability of each site being drawn from a given site class, where a high posterior probability of a site belonging to a class with $\omega > 1$ can be considered strong evidence that a site has evolved under positive selection [Yang *et al.*, 2005]. I used the two models for which the recommended Bayes Empirical Bayes method are implemented, M2a and M8.

SLR implements a method specifically designed for sitewise estimates which has been shown in simulations to perform as well as or better than PAML's sitewise random sites models [Massingham & Goldman, 2005a]. 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 ω ratios within the alignment. The value of ω is considered to be an independent parameter 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 ($\omega = 1$) and the alternative model is either purifying or positive selection ($\omega < 1$ or $\omega > 1$, respectively). The raw statistic measures the strength of evidence for non-neutral evolution at each site; following Massingham and Goldman [2005a] I use a signed version of the SLR statistic (created by negating the statistic for sites with $\omega < 1$) as the test statistic for positive selection.

Measuring Performance

In order to compare sitewise estimates from different alignments, a single sequence from each tree was chosen as the reference (arrows, Figure 2.1) and all sitewise statistics were mapped from alignment columns to sequence positions in the reference sequence. This approach corresponds to the process of mapping alignment-based evolutionary estimates onto a single member of the alignment for further analysis and integration with other genome-referenced data (as is often done, for example, using mammalian alignments and a human reference). As a result of this reference sequence based mapping, sites which were deleted in the reference sequence or inserted in a lineage not ancestral to the reference were not included in the final performance analysis.

To evaluate the power and error rates that might be achieved in real-world data anal-

ysis, the recommended cutoff thresholds for PAML's Bayesian posterior probabilities and the SLR statistic were used to identify positively selected sites. A posterior probability threshold of 0.95 was used for PAML [Yang *et al.*, 2005] and a threshold of 3.84, the 95% critical value of the χ^2 distribution with 1 degree of freedom, was used for SLR [Massingham & Goldman, 2005a]. Sites were compared to their true simulated state (e.g. positively-selected or non-positively selected) in order to identify correct and incorrect inferences, and from these classifications I calculated the false positive rate (FPR, defined as the proportion of all sites with true $\omega < 1$ falsely identified as positively selected) and true positive rate (TPR, defined as the proportion of all sites with true $\omega > 1$ correctly identified as positively selected).

As the addition of alignment error is expected to affect the power and error rates differently for each combination of simulation condition and aligner, I identified the score thresholds for each dataset that resulted in an actual FPR of 1% and calculated the TPR achieved at this actual error rate (hereafter referred to as $\text{TPR}_{1\%}$ to distinguish it from the TPR described above). Although this estimate of error-controlled power would be impossible to calculate in an empirical analysis where the error rate is unknown, it is useful in a simulation context for allowing a controlled comparison of the performance of sitewise analysis between different conditions. Specifically, it should be sensitive to changes in the numbers of both false positives and false negatives resulting from alignment error or alignment filtering; in both cases a lowered error-controlled power would result, as fewer true positives are identified at the constant 1% FPR.

I also evaluated the ability of each method to accurately infer the ω value at each site by collecting sitewise ω estimates from the output of each method and calculating the Pearson's correlation coefficient between the true and inferred ω values for each set of simulation conditions.

2.3 Results and Discussion

The Performance of Three Methods for Detecting Sitewise Positive Selection

I first evaluated the ability of three sitewise methods, PAML M2a, PAML M8 and SLR, to accurately estimate sitewise ω values and to detect positive selection under a range of tree lengths in the absence of alignment error. Figure 2.2 shows the TPR, FPR, $\text{TPR}_{1\%}$

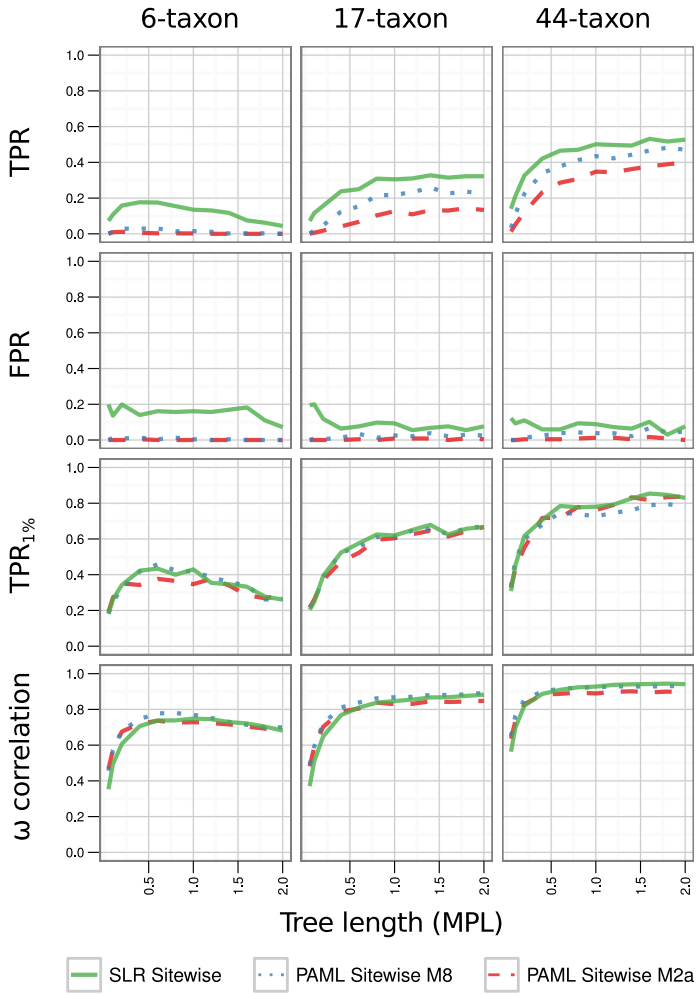


Figure 2.2: Alignments were simulated without indels for three tree shapes and analyzed with SLR, PAML M8, or PAML M2a. Fifty replicate alignments were simulated for each data point. The performance of each analysis method, as measured by four summary statistics, is plotted as a function of the mean path length (MPL) divergence. From top to bottom: true positive rate (TPR) at the recommended cutoff threshold (0.95 for PAML and 3.84 for SLR); false positive rate (FPR) at the recommended cutoff threshold; TPR at a 1% FPR threshold; Pearson's correlation coefficient between the true and inferred sitewise ω.

and sitewise ω correlation over a range of mean path lengths (MPL, defined as the mean root-to-tip branch length across all lineages) for each of the three simulation trees.

The detection power and ω correlation were weakest at low divergence levels for all methods and all trees due to the low amount of evolutionary information, as observed in previous simulations [Anisimova *et al.*, 2002b]. I found a positive correlation between tree size and detection power, with the highest performance in the 44-taxon tree. Power

generally increased monotonically with divergence, except for the 6-taxon tree which saw its maximum performance at moderate divergence levels (MPL 0.5–1.0) and began decreasing at higher values. The downward trend in the 6-taxon tree was likely due to the impact of saturation of synonymous sites in the very long branches present in such a sparse tree at high divergence levels. With lower average branch lengths at equivalent MPLs, the two larger trees showed no signs of decreased performance even at a MPL of 2 substitutions per site, which is greater than any of the divergence levels found in groups of commonly analyzed vertebrate, insect and fungal species (Table ??).

Comparing the three methods for detecting positively selected sites, I found that at the recommended cutoff threshold (Figure 2.2, top row) SLR showed the highest power to detect positive selection in all trees, followed by PAML M8 and PAML M2a. In the smallest tree, the power of the two PAML methods was virtually zero while SLR reached a maximum TPR of 18% (at MPL=0.5). At the same divergence, SLR yielded TPRs of 25% and 45% in the 17-taxon and 44-taxon trees, respectively, with PAML M2a ranging between 50–75% of SLR's power and PAML M8 falling between the two other methods.

The TPR measurements represent the power that might be achieved in real-world analysis using recommended cutoff thresholds, but the higher power from SLR may merely reflect a shifted balance between power and accuracy at the recommended cutoff threshold as opposed to an increased absolute ability to discriminate positive from neutral or purifying selection. The FPR and error-controlled $TPR_{1\%}$ results revealed that this was indeed the case: the FPR from SLR was higher than that from either of the PAML methods for all trees and divergence levels, suggesting that its higher power was the result of a less-conservative cutoff value. This was further verified by evaluating the TPR at a cutoff threshold that controlled for an actual FPR of 1% for each method ($TPR_{1\%}$, third row in Figure 2.2). The error-controlled $TPR_{1\%}$ values were virtually identical for all three methods, providing strong evidence that the three methods' sitewise statistics were nearly equally sensitive to positive selection under our chosen simulation conditions.

The conservative nature of the default thresholds for PAML and SLR has been previously noted [Anisimova *et al.*, 2002b; Yang *et al.*, 2005; Massingham & Goldman, 2005a], but the extremely low false positive rates in our simulations showed that in the absence of alignment error all three methods would yield very few false positives when analyzing genes with a typical mammalian-like distribution of ω values. The low FPRs were likely due to the large proportion of sites under moderately strong purifying selection in our ω distribution used for simulation. Such sites are less likely to yield false positives than sites

under neutral evolution (ω =1), the null model against which tests for positive selection are traditionally controlled.

For the purposes of our indel experiments, the observed similarity in error-controlled power levels indicated that the behavior of PAML M2a, PAML M8, and SLR was similar enough not to warrant separately evaluating all three methods in the subsequent indel simulation experiments. As the runtime for SLR was significantly lower than that of either PAML model, all subsequent results are presented only based on the SLR test.

The Effect of Alignment Error on Sitewise Power

When the indel rate was greater than zero, performance levels varied significantly for different tree sizes, alignment algorithms, and evolutionary divergences. Figure 2.3 shows the same performance measurements as Figure 2.2 for simulations without indels (gray lines, Figure 2.3) and with indels (black and textured lines, Figure 2.3) analyzed using three different aligners (ClustalW, MAFFT, and PRANK_C) and the true alignment. (Results for ProbCons, T-Coffee and PRANK_AA alignments are generally of intermediate quality, with ProbCons and T-Coffee showing slightly higher TPR, slightly higher FPR, and very similar $\text{TPR}_{1\%}$ compared to MAFFT, and PRANK_AA showing performance levels superior to these but inferior to PRANK_C. These results are omitted from Figure 2.3 in order to reduce visual clutter; $\text{TPR}_{1\%}$ results for PRANK_AA are shown in Figure 2.4C and discussed in the next section, and a comparison of results from all aligners tested can be found in Figure 2.6.) For the indel simulations, the indel rate here was held constant at 0.1 indel event per substitution.

Comparing the results without indels to those with indels under the true alignment I found a slight decrease in power and ω correlation and no noticeable increase in FPR. The decreased power was expected, since even in the absence of alignment error alignment columns containing gaps harbor less evolutionary information than columns with complete sequence data. The lack of increased FPR showed that SLR retained its conservative statistical performance even when analyzing gapped alignments. Surprisingly, at higher divergences (MPL $>$ 1.0) under the six-taxon tree, the FPR with indels was lower than the FPR without indels. This unexpected result may be attributed to the large number of alignment columns under such conditions that contained only a single non-gap sequence, as those columns were never inferred as positively-selected by SLR due to the complete lack of information. The two larger trees did not show a similar trend at high divergence levels, suggesting that this effect was indeed due to the highly sparse nature of the alignments in

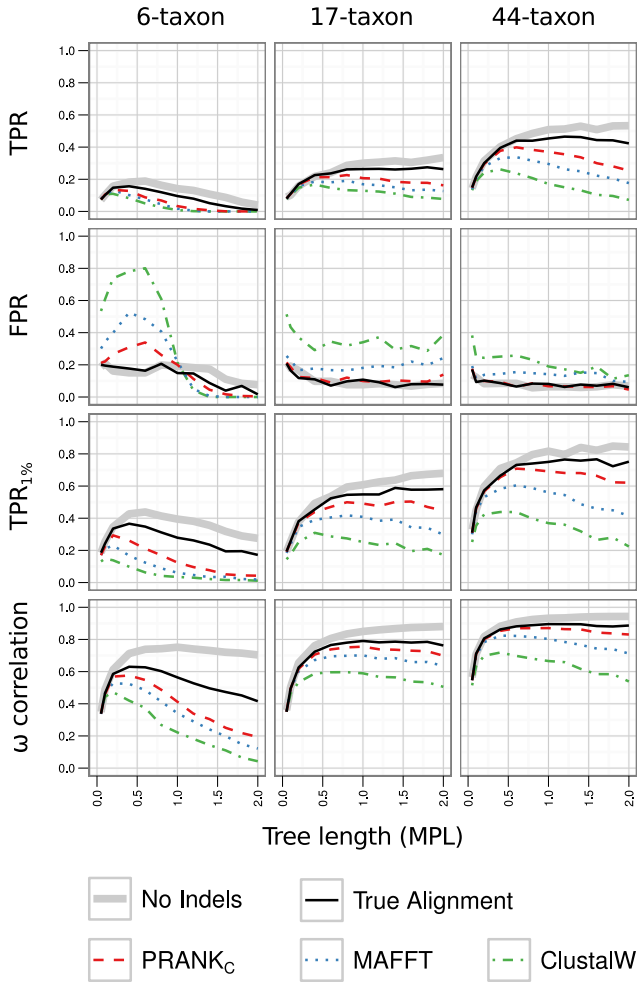


Figure 2.3: Sequences were simulated without indels (solid gray lines) or with indels (solid black and textured lines) using one of three tree shapes, aligned with one of three aligners, and analyzed with SLR; true alignments were separately analyzed with SLR (solid black lines). One hundred replicate alignments were simulated for each data point. The performance of each dataset, as measured by four summary statistics, is plotted as a function of the mean path length (MPL) divergence. From top to bottom: true positive rate (TPR) at the recommended cutoff threshold; false positive rate (FPR) at the recommended cutoff threshold; TPR at a 1% FPR threshold; Pearson's correlation coefficient between the true and inferred ω.

the 6-taxon tree. All other results from the 6-taxon tree at high divergences were similarly anomalous in this respect; I surmised that the sparseness of the true alignment, combined with the extreme difficulty of accurately aligning sequences along very long branches, made sitewise analysis with indels very unreliable at high divergences in the smallest tree.

When alignments were inferred using one of the three aligners tested, the TPR, $\text{TPR}_{1\%}$ and ω correlation were all reduced relative to the true alignment (dashed and dotted

lines, Figure 2.3). The degree of reduction varied depending on the aligner, simulation conditions, and performance measurement being analyzed. At low divergences (e.g. MPL < 0.2) the inferred alignments generally showed only a small decrease in performance. As divergence levels increased, so did the difference between the performance of the true alignment and the inferred alignments. The three aligners tested could be consistently and unambiguously ranked by all of the measured performance characteristics, with PRANK_C always performing best and ClustalW performing worst. The same ranking of aligners with respect to detecting positive selection has been observed in a number of studies [Fletcher & Yang, 2010b; Markova-Raina & Petrov, 2011b; Privman *et al.*, 2011]; our results corroborate these findings and provide evidence that this ranking may be consistent across a wide range of divergence levels and indel rates.

Looking at the TPR results for inferred alignments, I observed that in the 6-taxon tree the three aligners formed a cluster of lines well below the true alignment value, indicating similar tendencies among the different aligners to produce false negatives in the smaller tree. In larger trees the different aligners showed a wider spread of TPR values, but even PRANK_C showed a 5–10% reduction compared to the true alignment at MPL=1.0. These results show that the introduction of false negatives is a significant and seemingly unavoidable result of alignment error at medium to high divergence levels (MPL > 0.5), with even the most successful aligner producing a marked reduction in TPR compared to the true alignment. The $\text{TPR}_{1\%}$ and ω results in the larger two trees were qualitatively similar to the TPR results, showing that the aligners tested led to different levels of sitewise performance even when controlling for actual error rates or assessing the sitewise ω correlation.

The FPRs for infered alignments exhibited a very different trend from the other performance measures, with generally higher FPRs than the true alignment and the widest range of values occurring in the 6-taxon tree. In this tree at medium divergence levels (e.g. MPL 0.2–0.6) ClustalW showed up to a fourfold increase, and PRANK_C a nearly twofold increase, in FPR over the true alignment. As previously noted, the 6-taxon tree showed an anomalous FPR pattern at higher divergences, with lower FPRs for inferred alignments than the true alignment, likely due to the highly sparse true alignment under those conditions. In the two larger trees, FPRs from inferred alignments were less elevated compared to the true alignment, less variable between aligners, and relatively constant across the range of divergences. ClustalW's FPR ranged between 0.001 to 0.005, while PRANK_C's FPR was virtually identical to that of the true alignment in the 17-taxon and

44-taxon trees.

I found it useful to combine divergence estimates from Table ?? with the results from Figure 2.3 to characterize the combined effects of alignment error at different commonly analyzed levels of divergence. For example, at a human-mouse divergence level (MPL=0.2) misalignment had little impact on the TPR regardless of what aligner was used. However, ClustalW yielded a notably higher FPR than MAFFT or PRANK_{C}, and the error-controlled $\text{TPR}_{1\%}$ was correspondingly lower for ClustalW in all three trees. Thus, at low divergences I found that false positives were the main source of error from misalignment, and different aligners had highly variable tendencies to produce false positive results. At higher vertebrate and *Drosophila* divergence levels (MPL=0.8–1.0) false negatives became much more prevalent. The TPR for all inferred alignments was virtually zero in the 6-taxon tree, underscoring the necessity of including many species in the analysis of highly diverged sequences. In the two larger trees, PRANK_{C} resulted in very few additional false positives, but it suffered a 5–10% reduction in TPR relative to the true alignment. Meanwhile, ClustalW showed a 50% TPR reduction and maintained a strongly elevated FPR. At higher divergences and in larger trees, false negatives were thus the most persistent effect of alignment error, causing a marked reduction in sitewise power even with the best-performing aligner. Overall, the ranking of aligners was clear, with PRANK_{C} performing better than MAFFT and MAFFT performing better than ClustalW across all performance measures and MPL divergence levels.

Sitewise Power Under a Range of Indel Rates and Divergences

To explore the effects of alignment error across a wider range of simulation conditions, I extended the simulations of Figure 2.3 across multiple indel rates. Figure 2.4 shows heatmaps of the TPR and FPR for ClustalW, PRANK_{C} and the true alignment (Figure 2.4A, B) and a heatmap of the error-controlled $\text{TPR}_{1\%}$ for all aligners tested (Figure 2.4C). (MAFFT and PRANK_{AA} are omitted from Figure 2.4A, B and ProbCons and T-Coffee are entirely omitted from Figure 2.4 to save space. The performance of all these aligners fell between that of ClustalW and PRANK_{C} for all our measures. PRANK_{AA} slightly outperforms MAFFT, and ProbCons and T-Coffee show similar performance to MAFFT. A comprehensive set of TPR, FPR, and $\text{TPR}_{1\%}$ results can be found in Figure 2.6.) The results from Figure 2.3, which were simulated with an indel rate of 0.1, correspond to the middle row of each panel in Figure 2.4; rows above and below the middle row represent higher and lower indel rates, respectively. Similarly, the bottom row of each panel in Figure

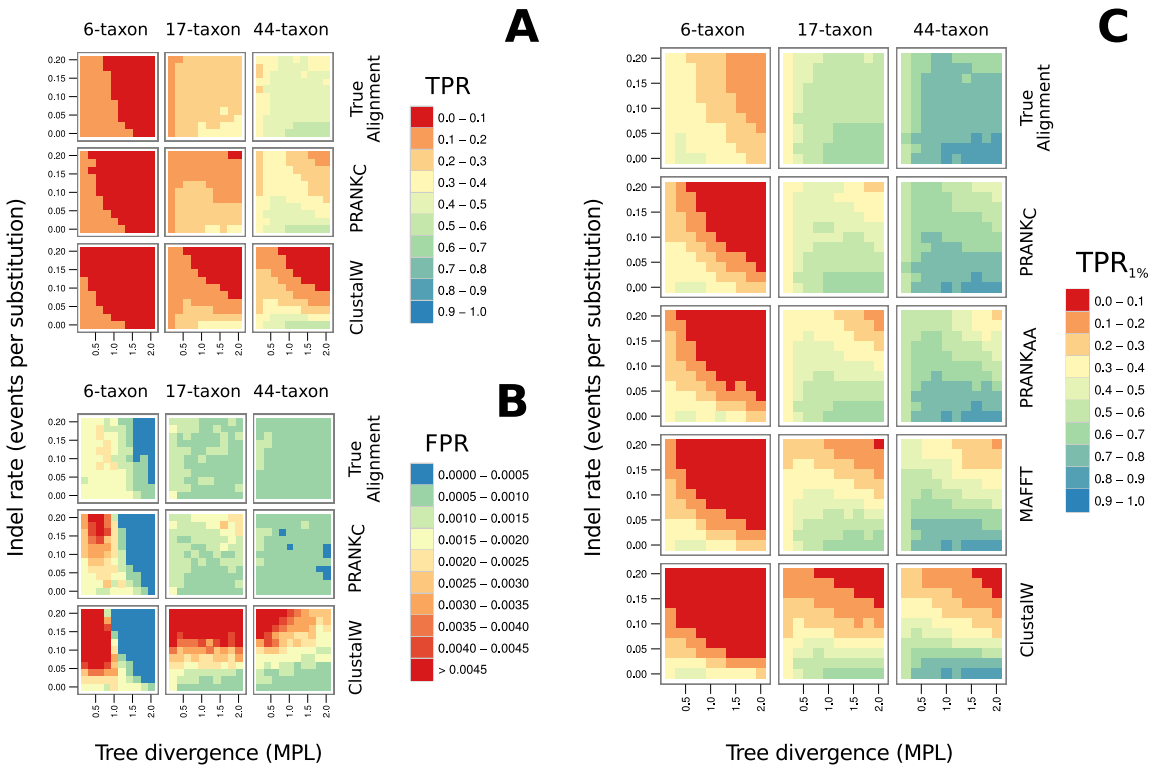


Figure 2.4: Sequences were simulated with indels using one of three tree shapes (6-, 17- or 44-taxon) and a range of indel rates and mean path length (MPL) divergence levels. Alignments are inferred with one of four aligners (ClustalW, MAFFT, PRANK_{AA}, PRANK_{C}) and analyzed with SLR; true alignments were separately analyzed with SLR. One hundred replicates were simulated for each set of conditions. Each cell is colored according to the performance at a given (indel rate, MPL) pair as measured by one of three summary statistics: (A) the true positive rate (TPR) at the recommended cutoff threshold, (B) the false positive rate (FPR) at the recommended cutoff threshold, or (C) the TPR at a 1% FPR threshold. Results for MAFFT and PRANK_{AA} are omitted from (A) and (B); as in (C) they show characteristics intermediate between ClustalW and PRANK_{C}.

2.4 was simulated with an indel rate of zero and corresponds to the 'No Indels' data in Figure 2.3.

The TPR values (Figure 2.4A) show a consistent pattern across the range of indel rates, with power decreasing as either the indel rate or the divergence level increases (except at the lowest divergence levels, where the lack of evolutionary information yielded slightly lower TPRs in the larger two trees). PRANK_{C} showed a greater ability than ClustalW to maintain a high TPR at higher indel rates, especially in the 17-taxon and 44-taxon trees. At lower indel rates, the TPR performance of both aligners and the true alignment were qualitatively similar.

PRANK$_C$ and ClustalW both showed a qualitatively similar pattern of elevated FPRs in the 6-taxon tree (Figure 2.4B), but their behavior diverged significantly in the 17-taxon and 44-taxon trees. In the 17-taxon tree, PRANK$_C$ only showed an elevated FPR compared to the true alignment at very high indel rates and divergence levels, but the ClustalW FPR increased steadily with the indel rate, quadrupling in value from the lowest to highest indel rate. Interestingly, for any given indel rate, the ClustalW FPR showed little variation across the range of divergence levels. This result was counter-intuitive, as we expected alignment errors to become more common as divergence increased and the number of observed indel events grew. Furthermore, PRANK$_C$ behaved according to our expectations, showing increased FPRs only at the highest divergences and indel rates in the 17-taxon tree. The FPR results in the 44-taxon tree confirmed the strange effect of ClustalW's alignments on the sitewise FPR: at the highest indel rates, ClustalW showed a negative relationship between FPR and divergence—exactly opposite to the trend I expected. PRANK$_C$'s FPR in the 44-taxon tree was equal to or below that of the true alignment under almost all conditions.

The error-controlled TPR$_{1\%}$ results (Figure 2.4C) provide a comprehensive picture of the effect of alignment error on the detection of sitewise positive selection. The two aligners not shown in the two other panels (MAFFT and PRANK$_{AA}$) exhibited TPR$_{1\%}$ values intermediate to those from ClustalW and PRANK$_C$ across the range of parameters tested, with PRANK$_{AA}$ performing better than MAFFT. As expected, performance was very similar between aligners at very low indel rates. At higher indel rates, most aligners yielded similar patterns of low TPR$_{1\%}$ in the 6-taxon tree, but in the larger two trees ClustalW and MAFFT alignments were unable to achieve high TPR$_{1\%}$ values, presumably due largely to their elevated FPR in those trees.

It is worth noting the exceptional ability of PRANK$_C$ to maintain a very low level of false positive sites even under extremely difficult alignment conditions. Although PRANK$_C$ showed slightly elevated FPRs at high indel rates in the 17-taxon tree, FPRs were nearly identical to the true alignment across all simulated conditions in the 44-taxon tree. This impressive performance suggests that, given a large enough number of taxa, PRANK$_C$ alignments would yield very few erroneous false positives in scans for positive selection in sequences with even very high divergence levels. Furthermore, these results showed that false negatives contributed more than false positives to PRANK$_C$'s reduction in sitewise performance—a novel observation which provides insight into the nature of PRANK$_C$ alignments and their application to sitewise evolutionary analysis.

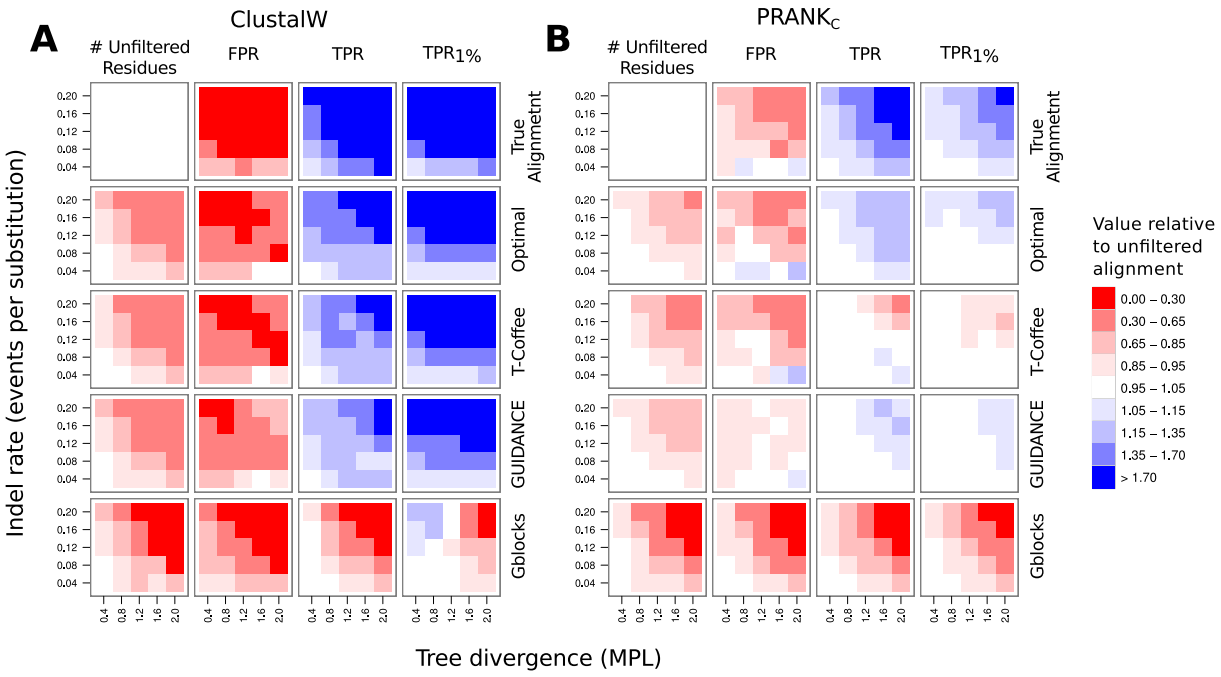


Figure 2.5: Sequences were simulated using the 17-taxon tree and a range of indel rates and mean path length (MPL) divergence levels. Alignments were inferred using (A) ClustalW or (B) PRANK$_C$, either left unfiltered or filtered with one of four alignment filters (Optimal, T-Coffee, GUIDANCE, Gblocks), and analyzed with SLR; true alignments were left unfiltered and separately analyzed with SLR. One hundred and fifty replicates were simulated for each set of conditions. Cells are colored according to the ratio of the performance of the indicated filter to the performance of the unfiltered ClustalW or PRANK$_C$ alignment as measured by one of four summary statistics. In columns from left to right: the number of unfiltered (i.e., non-N) residues remaining in the alignment; the false positive rate (FPR) at the recommended cutoff threshold; the true positive rate (TPR) at the recommended cutoff threshold; the TPR at a 1% FPR threshold (TPR$_{1\%}$). Note that the maximum percentage of residues removed by filtering was capped at 50% for all methods except Gblocks.

Effect of Alignment Filtering on Sitewise Error Rates

Having established that alignment error can lead to reduced sitewise performance through the introduction of false negatives and false positives, I tested whether alignment filtering methods could reduce error rates and improve the power of sitewise detection of positive selection. Using sequences simulated from the 17-taxon tree and a range of indel rates and divergence levels, I calculated inferred alignments using ClustalW and PRANK$_C$ and applied four filtering methods before performing the sitewise analysis. Since I wished to determine whether alignment filters either improved or worsened the error rates and power of sitewise analysis, I measured the ratio of each performance measure to the value obtained

from the equivalent unfiltered alignments. These relative values are presented in Figure 2.5. (Filtering results for PRANK_{AA}, ProbCons and T-Coffee were not calculated, and results for MAFFT are omitted from Figure 2.5 to save space. The gain or loss in performance resulting from filtering MAFFT alignments was generally intermediate to that resulting from filtering ClustalW or PRANK_{C} alignments. A comprehensive set of filtering results, including MAFFT alignments and $\text{TPR}_{5\%}$ values, can be found in Figure 2.7.)

As alignment filters act through the removal of alignment residues or columns, a certain amount of reduction in both the FPR and TPR was expected purely from the decreased amount of information available. For example, a filter that randomly removes a fraction of residues of each alignment would be expected to produce equal reductions in FPR and TPR. A more effective filter may also yield a reduced TPR, but the FPR reduction would be larger in magnitude, making the detection of positive selection more powerful for a given error rate. Thus, a reduced FPR is not necessarily indicative of good filtering performance, nor is a reduced TPR necessarily indicative of poor filtering performance. Additionally, the prevalence of false negatives resulting from misalignment suggested the interesting possibility that alignment filters may also improve power by removing false negatives, perhaps by masking out residues that were preventing positive sites from being identified. The removal of false negatives would result in an increased TPR, further complicating the assessment of filtering results based on FPR or TPR alone. As a result, I focused on the change in error-controlled $\text{TPR}_{1\%}$ as the best single measure of whether a filter had successfully improved the sitewise power of a dataset since this value is sensitive to changes in both the FPR and TPR. Note that the $\text{TPR}_{1\%}$ controls the FPR post-filtering, accounting for the tendency of filtering to reduce the FPR at a given cutoff threshold.

I first examined the two controls, the unfiltered true alignment and the inferred alignments filtered with the optimal filter (top two rows, Figure 2.5). The true alignment nearly always showed smaller FPR (red cells) and greater TPR and $\text{TPR}_{1\%}$ (blue cells) compared to the inferred alignments, with a greater magnitude of change relative to the ClustalW alignments than to the PRANK_{C} alignments (darker shades cf. lighter shades). These scores represented the direction and an upper limit on the magnitude of change that might be achieved by a perfect alignment filter. One exception to the general trend of lower FPR in the true alignment was the observation of two simulation conditions with slightly elevated FPRs in the true alignment compared to PRANK_{C} alignments (at an indel rate of 0.04 and MPL of 0.8 and 2.0). This small inconsistency may be explained by stochastic variation in false positive counts, as the absolute value of the FPR was very low in both

datasets under those conditions. Figure 2.4B shows the FPR to be on the order of 5×10^{-4} under those conditions; in total, I observed 63 false positives for the true alignment and 67 false positives for the PRANK_C alignment at the indel rate of 0.04 and MPL of 0.8 across all 150 replicates (comprising ca. 75,000 analyzed sites). A similar slight FPR elevation was also observed at the same indel rate for the optimal, GUIDANCE and T-Coffee filters.

Our hypothesis was that the optimal filter would show the same direction of change in FPR and TPR as the true alignment, but with slightly lower magnitudes. Indeed, improved sitewise performance was achieved in nearly all simulation conditions by the optimal filter, with the magnitude of $\text{TPR}_{1\%}$ change slightly lower than for true alignment. For ClustalW alignments the amount of improvement was quite large, with >70% increase in $\text{TPR}_{1\%}$ for nearly all conditions with an indel rate above 0.1. The improvement was more modest for PRANK_C alignments with a maximum of 15–35% $\text{TPR}_{1\%}$ increase.

Looking at the reduction in the number of non-masked residues remaining after filtering, I found that the optimal filter reached the maximum of 50% filtered residues for all ClustalW alignments with MPL>1 and an indel rate>0.1. This meant that more than 50% of residues were correctly aligned across less than 50% of the tree in those alignments. By contrast, the optimal filter applied to PRANK_C alignments only reached the maximum of 50% filtered residues at the highest tested divergence level and indel rate combination.

The TPR improvements achieved by the optimal filter provided some insight into the nature of sitewise false negatives resulting from alignment error. Two different types of alignment error might cause a false negative at a positively-selected site: either misalignment of one or more non-homologous codons causing the positive signal to be masked, or non-alignment of homologous codons causing the amount of evolutionary information to be reduced. The former type of error would be recoverable by alignment filtering (through removal of the codon(s) masking the positive signal), but the latter would not. Thus, the ability of the optimal filter to improve TPR levels across the board provided evidence that a sizeable portion of false negatives from both ClustalW and PRANK_C alignments were due to misaligned codons and thus amenable to recovery by filtering. Although the optimal filter was unrealistic in that it was based on perfect knowledge of which codons were misaligned, this result provided hope that one of the other filters might show a similar ability to recover false negative errors from PRANK_C alignments.

Turning to the three filters under investigation, I found T-Coffee and GUIDANCE both to be highly effective at improving ClustalW alignments, with magnitudes of improvement near those of the optimal filter. When applied to PRANK_C alignments, however, the

two filters' behavior diverged: T-Coffee only showed unchanged or reduced $TPR_{1\%}$, but GUIDANCE yielded slightly improved $TPR_{1\%}$ at high divergence levels and indel rates, with values 5–15% greater than the unfiltered $PRANK_C$ alignments. Both filters removed similar amounts of sequence information and resulted in similarly reduced FPR levels, but GUIDANCE showed a unique ability to recover false negatives from $PRANK_C$ alignments at the highest divergence levels and indel rates, and the resulting TPR elevation appears to have been responsible for the increased $TPR_{1\%}$ performance.

Gblocks behaved very differently from the other filters tested, resulting in reduced FPR, TPR, and $TPR_{1\%}$ under nearly all simulation conditions. Only at high indel rates and low divergence levels in the ClustalW alignments did Gblocks show increased $TPR_{1\%}$ relative to the unfiltered alignments. This poor performance was likely due to overly-aggressive removal of alignment columns. I could not limit the amount of sequence masked by Gblocks, so many alignments saw more than 70% of residues removed, resulting in the loss of a large number of correctly-aligned true positive sites. Dessimoz and Gil [2010] found Gblocks filtering to have a negative effetc on the accuracy of phylogenetic inference; our results provide additional evidence in support of their finding, suggesting that Gblocks filtering tends to reduce, rather than increase, the power and accuracy of alignments when applied to a number of evolutionary analyses.

There was some evidence that the column-wise nature of Gblocks filtering was partly responsible for its poor performance in our experiments. Since false negative errors cannot be recovered through the removal of entire alignment columns, it made sense that the $PRANK_C$ alignments—which resulted in varying numbers of false negatives but always very few false positives—would not see improved performance after column-wise filtering. On the other hand, ClustalW alignments showed a relatively constant level of false positives and an increasing number of false negatives as divergence levels increased. The application of a column-wise filter like Gblocks would thus be expected to show good improvement at low divergences where false positives dominated, but less improvement at higher divergences where false negatives became more prominent. Indeed, this was the pattern observed when applying Gblocks to ClustalW alignments.

Overall, our alignment filtering simulations found that Gblocks rarely improves alignments for sitewise detection of positive selection, but filtering methods based on GUIDANCE and T-Coffee scores have a good ability to mask out misaligned residues that cause false positives and false negatives in sitewise inference. This beneficial effect is highly dependent on simulation conditions and the input aligner. For ClustalW alignments (which,

left unfiltered, led to many false positives and false negatives) both GUIDANCE and T-Coffee showed good ability to improve sitewise performance, behaving qualitatively similarly to the optimal filter.

Since the performance measurements shown in Figure 2.5 are expressed relative to values obtained with unfiltered alignments, they do not allow for easy comparison of the absolute performance of any combination of aligner and filter. To more directly compare the absolute TPR, FPR and $TPR_{1\%}$ values obtained with filtered ClustalW and MAFFT alignments to those obtained with other unfiltered alignments, I simulated additional datasets with ClustalW and MAFFT alignments and GUIDANCE filtering using all three trees and the same range of indel rates and MPLs as used for Figure 2.4. These results are included in Figure 2.6. Comparing absolute performance, I find that GUIDANCE filtering generally improved the $TPR_{1\%}$ for ClustalW and MAFFT alignments, largely due to strong FPR reductions in regions of high indel rates and low divergence levels. The resulting $TPR_{1\%}$ performance for filtered ClustalW alignments was comparable to (but not better than) unfiltered MAFFT alignments, and the $TPR_{1\%}$ for filtered MAFFT alignments was slightly better than unfiltered MAFFT alignments and substantially worse than unfiltered $PRANK_{AA}$ alignments.

Filtering was less beneficial when applied to the more accurate $PRANK_C$ alignments, with T-Coffee filtering reducing performance and GUIDANCE yielding only mild $TPR_{1\%}$ improvements. Importantly, GUIDANCE only showed improved performance at high divergence levels (e.g., MPL>1.6), well above those found in commonly-analyzed groups of species. Thus, the use of unfiltered $PRANK_C$ alignments would yield largely equivalent performance to GUIDANCE-filtered alignments for detecting sitewise positive selection when analyzing protein-coding sequences at most commonly encountered divergence levels. Equally important is the observation that GUIDANCE (when judiciously applied, including an upper limit on the amount of sequence data removed) did not significantly reduce performance compared to the unfiltered alignments. Put simply, filtering neither significantly hurt nor significantly improved performance. Finally, I note that the performance of the 'optimal' filter on $PRANK_C$ alignments suggests that mild further improvements to filtering strategies may be possible, but the potential for improvement is small and may be of little value.

2.4 Conclusions

In this paper, I investigated the performance of sitewise detection of positive selection under a range of tree sizes, indel rates, and divergence levels, using simulation parameters designed to approximate the analysis of typical mammalian protein-coding genes. We evaluated the ability of six alignment methods and three alignment filtering methods to produce alignments for detecting positively selected sites, using the FPR, TPR, and error-controlled $TPR_{1\%}$ of the sitewise detection of positive selection as our main measures of performance.

The simulation results showed that alignment error can have a measurable impact on the error rates and power of the sitewise detection of positive selection under all but the least difficult alignment conditions. I confirmed and extended the findings of Fletcher and Yang [2010b], Markova-Raina et al. [2011b], and Privman et al. [2011] regarding the relative accuracy of different aligners, showing that $PRANK_C$ had the best performance and ClustalW had the worst performance for subsequent sitewise analysis. Notably, our simulations found that ClustalW produced more sitewise false positives than any other aligner tested even at low divergence levels, suggesting that its use should be avoided even when analyzing closely-related sequences. $PRANK_C$, on the other hand, resulted in very low FPRs even at higher divergences. In particular, when the number of sequences in the tree was large, $PRANK_C$'s sitewise FPRs were virtually indistinguishable from those of the true alignment.

An important observation regarding the size of tree analyzed was that the 6-taxon tree caused qualitatively similar problems (e.g., elevated FPRs and reduced TPRs) for all aligners, suggesting that poor performance is inevitable when analyzing a small number of moderately divergent sequences. The small amount of evolutionary information combined with the longer branch lengths makes alignment difficult and increases the tendency of misalignment to cause sitewise false positives. Thus, I reiterate the well-established recommendation to use large numbers of sequences when inferring sitewise positive selection [Anisimova *et al.*, 2001b, 2002b]. We additionally point out that when analyzing sequences with indels, the shape of the tree may matter as well: trees with long internal branches may be especially prone to false positives, as longer branches are more difficult to align.

The very low FPRs observed for $PRANK_C$ alignments conflict somewhat with the results of Fletcher and Yang [2010b], who found that the FPRs for the branch-site test were not under control even with $PRANK_C$ alignments. This apparent discrepancy can be explained by different sensitivities to alignment error: the branch-site test would yield false

positives when misalignment causes apparent positive selection along only the foreground branch, while the SLR sitewise test would produce false positives only when misalignment causes a signal of positive selection strong enough to overpower the non-positive signal throughout the tree. This effect stems from the different biological hypotheses tested by the two methods; their differential sensitivity to misalignment underscores the necessity of considering the biological sensitivity and robustness to alignment error when applying either of these tests to detect positive selection within an alignment. For the detection of positive selection in highly divergent or indel-prone sequences, the use of sitewise models instead of the branch-site test may be a sensible alternative, sacrificing some sensitivity for better error control.

Despite producing very low FPRs, PRANK_C alignments still resulted in an increased number of false negatives compared to the true alignment. I showed that some of these false negatives were possible to recover as true positives through alignment filtering, and I found that both the 'optimal' filter and GUIDANCE were able to successfully recover these false negatives at high divergence levels, resulting in small but measurable performance improvements over the unfiltered PRANK_C alignments.

The manual or automated adjustment of alignments has been thought by many to be an important step in evolutionary analyses due to fear of a high prevalence of misalignment-induced false positives. While this is true for some aligners, I find that more accurate alignment algorithms result in significantly fewer false positives in the subsequent detection of sitewise positive selection. This strongly reduces the beneficial effect of alignment filtering, so much so that current filtering methods are scarcely able to improve the performance of PRANK_C alignments when analyzed with SLR.

As a result, I cannot unequivocally recommend the use of alignment filtering in the detection of sitewise positive selection, except perhaps for the most divergent and indel-prone sequences, which are unlikely to be encountered in analyses focusing on mammalian or vertebrate genes. Although GUIDANCE showed some ability to improve the error-controlled power under difficult alignment conditions at high divergence levels (whereas T-Coffee and Gblocks filtering failed to improve upon any PRANK_C alignments), this improvement was modest in magnitude and was achieved largely through the recovery of false negatives as opposed to the elimination of false positives. For an analysis where the control of false positives is the primary concern, the added computational expense of running many bootstrap alignment replicates (as performed by GUIDANCE) may not be offset by the possibility of a slight increase in power. However, GUIDANCE never

significantly reduced power, so its use would not be expected to yield worse results.

Some of our conclusions differ from those of Privman et al. [2011], who found strong improvements in error-controlled power by filtering alignments in simulations focused on three HIV-1 genes. Although the phylogenetic trees they used to guide their simulations contained divergence levels at the low end of our tested range (MPLs of 0.38, 0.34 and 0.33 for gag, pol and env, respectively; E. Privman, personal communication) and were roughly comparable in size and shape to our 17-taxon tree, I failed to find a significant benefit for alignment filters at any MPL below 1.2 when using PRANK_C alignments. Interestingly, the authors also found Gblocks to be roughly comparable in performance to GUIDANCE, while I found Gblocks' performance to be very poor. Some of these discrepancies may be due to differences in the details of our simulations or filtering procedures, but in the end our results are largely complementary: Privman et al. showed that filtering can be beneficial for detecting positive selection, especially in the case of fast-evolving (but fairly closely-related) sequences, while I have shown that divergence levels and indel rates have a significant impact on the performance of different aligners and filters.

Our simulations did not include fully biologically realistic models of spatial or temporal variability in the rate of indel formation or in the distribution of selective pressures (e.g.; Whelan 2008a). We do not expect that such heterogeneity would affect our main conclusions regarding the relative performance of different aligners and filters: the trends I observed were consistent across a wide range of parameter values and tree sizes, suggesting that they reflect fundamental differences in each method's ability to align or filter sequences as opposed to artifacts due to the relative simplicity of our simulations.

However, such heterogeneity is clearly important to the evolution of mammalian proteins [Fay & Wu, 2003]. Many proteins contain combinations of structured domains and unstructured regions along their length, resulting in a discontinuous mix of different structure- and function-related evolutionary pressures. False positives may be more prominent in small unconstrained regions, raising FPR levels above those predicted by simulations with a uniform pattern of evolution. Functional differences between genes may also influence the ω distribution, with some genes or domains showing fewer or more neutrally-evolving sites than modeled in our simulations, making false positive results either less or more likely, respectively. As such, the appropriateness of our simulation scheme should be critically considered when evaluating our specific power and error rate estimates in the context of real-world data analysis.

In the case of a protein evolving with a mix of insertion and deletion rates, our results

based on uniform rates could be used to identify tentative upper and lower bounds for the overall error rate. For example, if a small region of a protein is evolving with higher divergence and indel rates than the rest, then the error rates within the less-constrained region should be comparable to our results based on proteins with uniformly high divergence and indel rate. Correspondingly, the overall error rate for the protein would fall somewhere between those observed in our homogeneous simulations corresponding to the least difficult and most difficult regions of the alignment. Critical to such an analysis is the accurate estimation of the local indel rate, for which a number of methods are currently available or under development [Holmes, 2005; Cartwright, 2009].

Species-level differences may also have an effect on error rates in detecting positive selection, as the efficacy of natural selection is highly dependent on effective population size [Ellegren, 2009a]. For example, proteins evolving in *Drosophila* species with a high effective population size should experience stronger positive and purifying selection than in mammals, potentially leading to increased power and reduced error rates when compared to our simulations based on a mammalian-like ω distribution.

A tangential but interesting observation from our simulations was that all three methods I tested for the sitewise detection of positive selection were highly conservative when analyzing alignments with a mammalian-like ω distribution. SLR, for example, yielded FPRs well below the nominal 5% level. This was due to a mismatch between SLR's null model of neutral evolution ($\omega = 1$) and our more realistic distribution of non-positive sites (where $\bar{\omega} = 0.277$). I note that while the ability of SLR and PAML's sitewise models to distinguish between neutral evolution and positive selection in a well-controlled manner is important, the majority of protein sites evolve under moderate purifying selection. Future work on methods to adaptively adjust the cutoff thresholds to achieve better statistical control under non-neutral (and possibly unknown) ω distributions could yield much greater power to detect positive selection while maintaining good control of error rates.

I showed here that even relatively simple evolutionary simulation experiments could sensitively assess the performance characteristics of different aligners, provide quantitative insight into the practical effects of alignment error, and suggest areas for future development of alignment and filtering methods. In the future, I expect the development of more realistic simulations for protein evolution—perhaps incorporating structurally-motivated and empirically validated models of mutation, indel formation and constraint—to further increase the applicability and accuracy of such experiments, and I believe that flexible and accessible simulation programs such as INDELible [Fletcher & Yang, 2009] and PhyloSim

[Sipos *et al.*, 2011] will play an important role in the quantitative assessment of alignment algorithms and alignment-dependent comparative analyses.

As genomes rapidly accumulate in the databases and large-scale analyses become the norm, I hope that the development and application of alignment methods, which are arguably the most important step in any evolutionary analysis, will be based on a rigorous understanding of their behavior and performance when applied to a wide variety of evolutionary analyses.

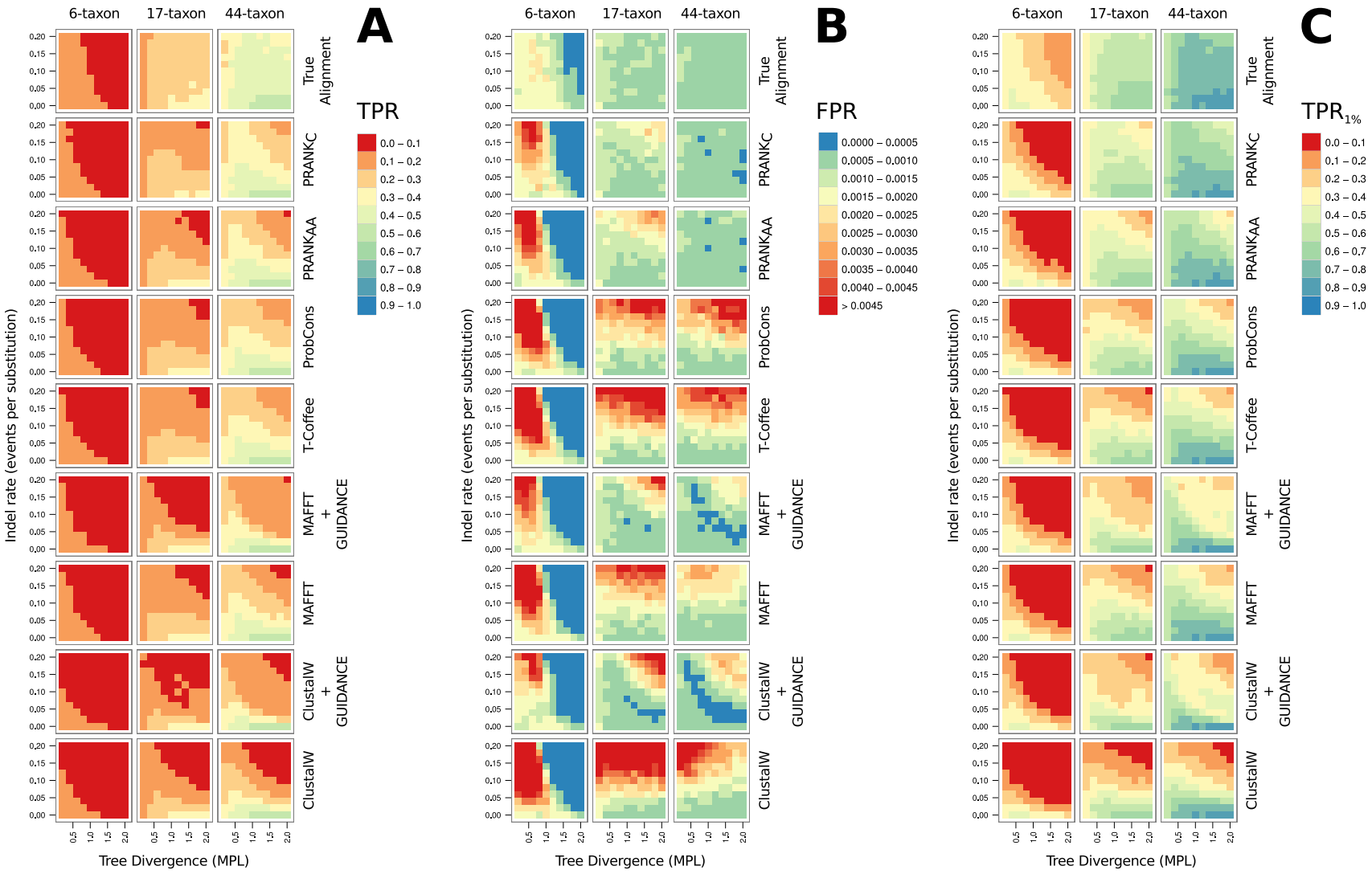


Figure 2.6: This figure depicts the same simulations and uses the same formatting as Figure 4, except that results for MAFFT and PRANK_{AA} have been added to sections (A) and (B) and results using ProbCons, T-Coffee, ClustalW + GUIDANCE, and MAFFT + GUIDANCE have been added to all sections.

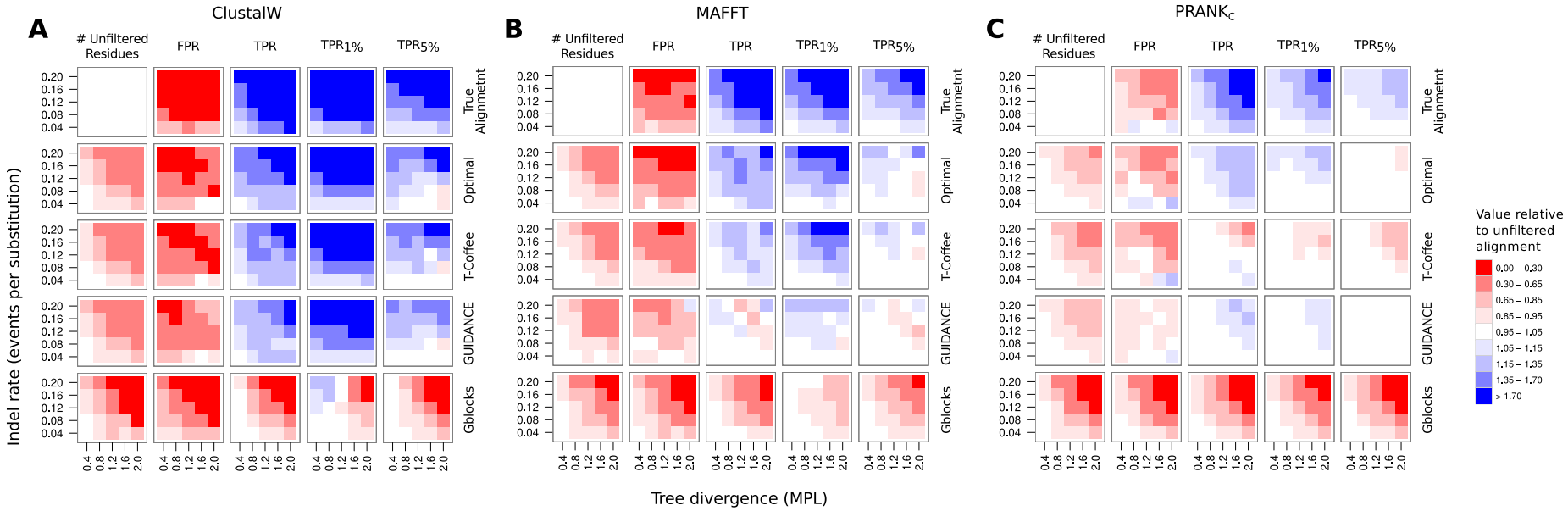


Figure 2.7: This figure depicts largely the same simulations and uses the same formatting as Figure 5 except for two changes. First, results for MAFFT have been added to section (B) and results for PRANK$_C$ have been moved to section (C). Second, the rightmost column has been added, showing the true positive rate (TPR) at a 5% false positive rate (FPR) threshold (labeled TPR$_{5\%}$).

Chapter 3

Curating a set of orthologous mammalian gene trees

3.1 Introduction

This chapter considers the identification of groups of orthologous genes in vertebrates and mammals. I begin with a discussion of the detection of orthologs within groups of vertebrate and other eukaryotic genomes, focusing on the tree-based approach to ortholog and paralog identification as used by the Optic, Compara, and Treefam databases. In preparation for the sitewise analysis of mammalian orthologs to be presented in the subsequent two chapters, I construct a simple method for isolating subtrees with largely orthologous characteristics from the set of Compara gene trees and analyze the taxonomic distribution of gene duplication and loss within those largely orthologous trees (LOTs), demonstrating how strongly the history of genome duplications has affected the structure of orthologous and paralogous relationships between and within vertebrate and invertebrate genomes. The set of gene trees resulting from a simple tree-splitting process based on taxonomic coverage is selected for use in the subsequent sitewise analysis.

3.2 Methods for ortholog identification

Central to any evolutionary sequence analysis is the assumption that the sequences being analyzed, or some parts therein, share a common evolutionary origin. Thus, the first step in any such analysis is the collection of homologous sequence data. Starting with a source sequence, homologs are usually identified by searching a sequence database for sequences

with a minimum overall similarity according to some evolutionary model. In some cases, such as when sequences are highly divergent or subject to domain shuffling or horizontal gene transfer, a more localized or more specialized measure of similarity may be desirable [Koonin *et al.*, 2001; Sjölander *et al.*, 2011], but within closely related groups of organisms such as vertebrates and mammals, the process of gene duplication and divergence dominates patterns of relatedness between protein sequences [Ohno, 1970], making overall sequence similarity a useful method by which to identify homologous sequences in the organisms of interest.

Heuristic algorithms have been developed for performing quick and sensitive sequence homology searches within databases of protein and nucleic acid sequences [Altschul *et al.*, 1997; Eddy, 2009]. The power of these methods is high enough that homology within vertebrates, even for fast-evolving genes, can be readily detected. However, the prevalence of historical and ongoing gene duplication and loss in vertebrates complicates the problem, as orthologous genes (e.g., homologs between species, sharing a common ancestral population and related through a speciation event) and paralogous genes (e.g., homologs within a single species, sharing a common ancestral genomic sequence and related through a duplication event) are important, yet difficult, to distinguish from one another [Jun *et al.*, 2009]. In other words, sets of genes that may be confidently identified as sharing homology may still contain paralogy and orthology relationships that are more difficult to resolve. Since the evolutionary trajectories of paralogous versus orthologous genes are expected to be quite distinct [Lynch & Conery, 2000], the correct identification of orthologous versus paralogous relationships in vertebrates is critical for any detailed molecular evolutionary analysis.

Methods for distinguishing orthologous from paralogous relationships within vertebrate genomes have been in development for over a decade [Remm *et al.*, 2001; Yuan *et al.*, 1998], but the amount of overlap between vertebrate orthologous groups identified by different methods has historically been disappointingly small [Jun *et al.*, 2009; Chen *et al.*, 2007], suggesting that the problem of orthology inference is far from resolved. Still, one might expect the accuracy and power of orthology inference methods to improve with time, given the steady increase in available computing power and the sequencing of many complete vertebrate genomes over the past decade. Indeed, the growing number of available sequenced genomes, greater available computational power, and improved understanding of patterns of gene duplication and loss have led to the growing popularity of phylogeny-based approaches, which were once considered computationally impractical and too difficult to

automate [Remm *et al.*, 2001]. In general, the phylogenetic approach involves estimating a phylogenetic tree from an entire cluster of homologous genes and inferring duplication events based on the discordance between the gene tree and the species tree. Several variants of this approach have been developed and applied to gene tree reconstruction in insects, fungi and vertebrates [Muller *et al.*, 2010; Cepas *et al.*, 2007; Datta *et al.*, 2009; Vilella *et al.*, 2009; Ruan *et al.*, 2008; Hahn *et al.*, 2007], and validation against manually-curated gene trees has shown these phylogeny-based methods to be more sensitive and accurate than pairwise or graph-based methods [Datta *et al.*, 2009].

One particular phylogeny-based approach, which uses a known species tree to guide the resolution of orthologs and paralogs within a gene tree in a Bayesian framework, has been implemented in a software package called TreeBeST and used extensively to infer genome-wide vertebrate and eukaryotic gene trees for the Optic, TreeFam and Compara databases [Heger & Ponting, 2008; Ruan *et al.*, 2008; Vilella *et al.*, 2009].

I describe in this chapter an initial analysis of the gene trees within Compara which I undertook in preparation for the sitewise evolutionary analysis presented in Chapters 4 and 5. The major goals of this analysis were to better understand the distribution of orthologous genes within vertebrate genomes, identify which taxonomic constraints would best define largely-orthologous groups of mammalian genes, and evaluate whether low-coverage genomes, which are unique to the Compara database of orthologs, showed high enough annotation quality and consistency to be included in the mammalian evolutionary analysis. Throughout the chapter, attention will be paid to those aspects of the gene tree identification pipeline which may contribute to errors in the downstream evolutionary analysis.

3.3 Low-coverage genomes in the Ensembl database

A major feature that distinguishes Compara from the Treefam and Optic databases is the inclusion in Compara of several mammalian genomes which have been sequenced to low coverage as a part of the Mammalian Genome Project (MGP). The inclusion of these additional genomes was a major reason why I performed the analysis described in the following chapters using Ensembl as a source of gene trees and alignments, representing a major advantage over otherwise similar orthology databases due to the greater species sampling density. With respect to the actual sequencing of the low-coverage mammalian genomes, the motivation and goals of the MGP itself will be outlined in the next chapter.

For the purpose of the present discussion, I will outline the way low-coverage genomes are handled within the Ensembl annotation pipeline, as certain features of the low-coverage genomes cause characteristic anomalies in the orthology analyses, as will be shown in the ensuing analysis.

The prevalence of missing sequence data and fragmented contigs in low-coverage genomes presented a unique set of problems for the generation of transcript annotations in Ensembl. In recognition of these differences, the procedure used by Compara 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 genome, and gene models are projected from human to the target genome. Small frame-disrupting insertions or deletions within orthologous exons are corrected, and missing or incomplete exons are padded with Ns in order to produce a transcript with a length equal to that of the human reference transcript. The inclusion of these error-correcting features allows for a set of intact, if not complete, coding transcripts to be generated for low-coverage genomes, leveraging the high level of sequence similarity between human and other Eutherian mammals to project genes and transcripts from the high-quality human genome to the unannotated, highly fragmented low-coverage genome assemblies. Still, in many cases the Ensembl pipeline could not map complete genes or transcripts from human to the target genome, causing difficulty in identifying duplications. On one hand, the lack of completely assembled chromosomes meant that segmental duplications in low-coverage genomes were often unresolved or unidentifiable, making it difficult or impossible to confidently identify recently duplicated genes. On the other hand, the shorter length of assembled fragments caused genes to occasionally be split between two sequence fragments; the Ensembl pipeline currently annotates such genes as two separate shortened gene fragments, resulting in an excess of shortened apparent paralogs in resulting gene trees (see Section **??** for more detail on this artifact).

The Compara gene family pipeline, which I describe in more detail below, uses the set of annotated transcripts from each species as its input [Vilella *et al.*, 2009], so the quality of gene annotation from each source genome has a direct impact on the overall quality and accuracy of the resulting 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 less necessary.

3.4 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 [Vilella *et al.*, 2009; Flicek *et al.*, 2011]. Although a complete description of the design, implementation, and validation of the pipeline behind the Ensembl database would be beyond the scope of this chapter, I will briefly outline the major aspects of the approach used by the Compara pipeline, focusing on a few details which are relevant to the current analysis and 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, however, 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].

In terms of organizing biological data, however, pervasive alternative splicing is somewhat burdensome. Roughly 34% of human genes contain 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. Within the context of the Compara database, the first problem in maintaining consistency across many source genome sequences is the fact that primary data on alternative transcript structures (e.g., data from expressed sequence tags, RNA-seq, or proteomics experiments) are largely absent from most organisms with sequenced genomes. Furthermore, the task of incorporating multiple transcripts per gene

into an evolutionary framework 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 definition of orthologous groups, whose main goals are to understand the evolutionary relationships of gene families within some set of species' genomes.

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. The OPTIC database, which contains orthologous groups identified within a wide range of animal and fungal clades, uses the transcript with the maximum length as the canonical transcript [Heger & Ponting, 2008].

Once the set of canonical transcripts is chosen, an all-against-all protein BLAST search is performed using the Washington University variant of BLAST and genes are clustered into homologous groups 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.*, 2005a], MUSCLE [Edgar, 2004], KAlign [Lassmann *et al.*, 2009], and T-Coffee [Notredame *et al.*, 2000a]. 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 contrast to the equivalent task in sets of insect or fungal genomes where divergence levels between extant species with sequenced genomes 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. However, it is worth noting that the recent development of a new Bayesian method for gene tree reconciliation, based on a generative model of gene family evolution as a birth-death process and incorporating species-specific and gene-specific rate variation, showed good performance in resolving fungal and invertebrate gene trees, and could easily be adapted to the vertebrate genomes in the Compara database . Additional validation of the overall approach taken by Compara comes in the form of Treefam and Optic [Ruan *et al.*, 2008], databases of animal and fungal gene trees which applied a similar set of tools to infer gene trees from a more diverse set of genomes, producing qualitatively similar results [Vilella *et al.*, 2009].

3.5 Quantifying paralogous relationships within Ensembl gene trees

The first task in preparing the Compara data for subsequent sitewise analysis was to identify and extract a set of gene trees or gene subtrees comprising largely of orthologous relationships within mammals (i.e. LOTs), avoiding as much as possible the inclusion of paralogous relationships. It was necessary to postprocess the Compara gene trees in order to achieve this goal, because many of the Compara trees contained multiple sets of complete mammalian orthologous trees linked by ancient gene duplication events, while I wished to study the evolution of each mammalian LOT in isolation. In other words, the Compara gene trees could be charcterized as being over-clustered with respect to the core set of mammalian orthologous trees. This over-clustering was not necessarily inaccurate with respect to the evolutionary history of vertebrate genes, but it was not a desirable feature for my intended use of the data. This section is concerned with identifying and quantifying the level of such over-clustering within the Compara database.

I first collected the set of 18,607 Compara gene trees and analyzed their overall size and the number of human genes contained within each tree. The results of this analysis are presented in Table 3.2 and Figure 3.3; a complete description of these data will follow in Section 3.7, but the aspects relevant to characterizing the amount of paralogy contained within these trees will be discussed here.

The first row of Table 3.2 shows various summary statistics from the full set of Compara gene trees, with the columns under the "Human Content" heading showing the fraction of

all gene trees containing zero, one, or two or more human genes. Evidence for the existence of large numbers of paralogs within Compara trees came from the observation that 20% of Compara trees contained 2 or more human genes. If each Compara tree contained only one set of mammalian orthologs, then the 20% of trees with multiple human gene copies could only be explained by an unrealistically high rate of gene duplication in the lineage leading towards human. Instead, a more parsimonious explanation was that many of the Ensembl trees contained not one group of mammalian LOTs, but two or more sets of mammalian LOTs joined by one or more ancient duplication events. This explanation was further supported by Figure 3.3, which shows a histogram of total gene counts in the Ensembl root trees. A large number of trees contained more than 48 sequences (the number of vertebrate genomes in Ensembl), with clear peaks in the histogram of trees with 2, 3, and 4 times the number of vertebrate genomes in Ensembl. These patterns were highly consistent with a high number of Compara gene trees containing multiple mammalian LOTs.

The prevalence of over-clustered Eutherian orthologs in the Compara database could be 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 Section 3.4, 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 (TC) 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 were tuned to perform well on the dataset as a whole (i.e., to successfully cluster a majority of vertebrate orthologs), 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. The rate of protein evolution varies widely within a genome, as evidenced by a study of amino-acid substitution rates of roughly 6,000 eukaryotic orthologous genes by Koonin et al. [Koonin *et al.* 2004], who found that the middle 90% of genes show a

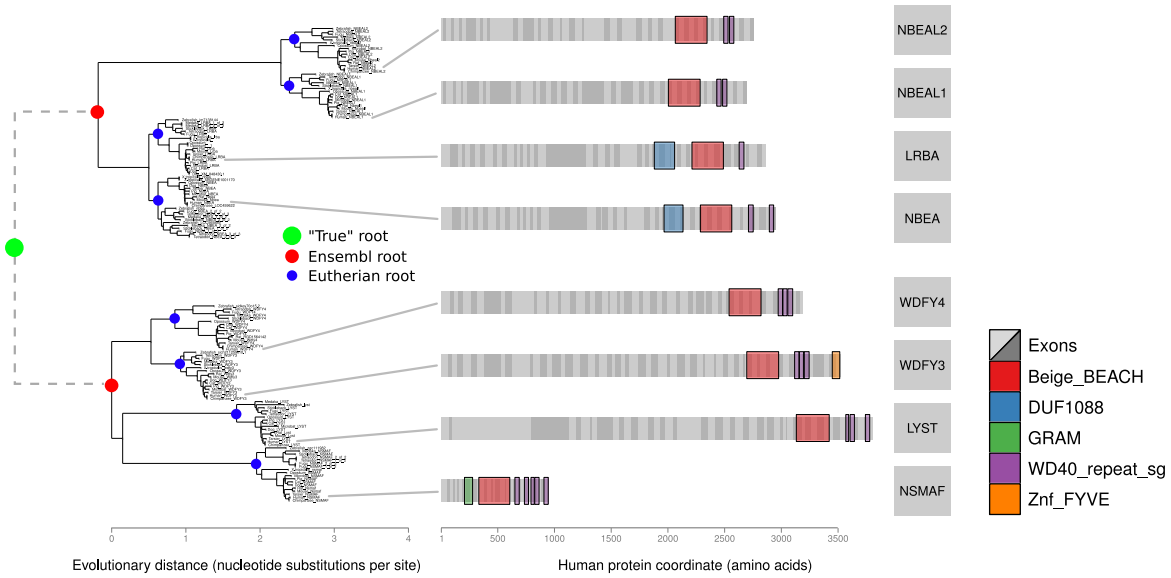


Figure 3.1: The evolutionary history of the human *neurobeachin-like 2* gene (*NBEAL2*) and its paralogs. Left, two phylogenetic trees from Ensembl Compara (release 60) are shown, summarizing the evolution of *NBEAL2* and its three paralogs (top) and *LYST*, a presumed distant paralog of *NBEAL2*, and 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. Adapted from [Albers *et al.*, 2011]

nearly fourfold variation in evolutionary rate. Given this wide range of protein evolutionary rates, the excess of over-clustered orthologs in the Compara database is understandable and even somewhat expected.

Before continuing with a description of a method to identify LOTs within Compara gene trees, 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 mammals. Certainly these large "over-clustered" trees, which represent a more distant evolutionary history than a single mammalian 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.

As an example, take the gene *NBEAL2* and its human paralogs, whose gene trees,

exon structures and domain classifications were extracted from Ensembl v62 and summarized in Figure 3.1. 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]. I performed, with Botond Sipos, an evolutionary analysis of *NBEAL2* for the article describing the discovery, and 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 evolutionary distance between paralogs. Had each mammalian ortholog been displayed independently in Ensembl (i.e., using the blue "Eutherian root" nodes in Figure 3.1), 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 root connecting these two sets of trees (represented by the green node in Figure 3.1), the connection between these eight genes would have been even more immediately apparent.

For the purposes of this project, 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 TC criteria; the remainder of this chapter presentes design and results of this process when applied to the Compara gene trees.

3.6 Using taxonomic coverage to extract largely orthologous mammalian subtrees

Based on the observation that the clustering step of the Compara pipeline did not make use of any taxonomic information, I hypothesized that a relatively simple set of rules based on TC would be sufficient to identify most mammalian LOTs. Two well-established observations in mammalian genomes supported the decision to use TC in this context. 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 which occurred in the ancestral vertebrate genome) 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 TC to identify mammalian subtrees should be largely unaffected by continued (i.e., post whole-genome duplication) gene duplication or loss events. The potential utility of TC was further bolstered by the star-like shape of the mammalian tree [Bininda-Emonds *et al.*, 2007]: a star-like trees contains more branch length within terminal lineages than a ladder-like tree 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 TC of the gene tree.

The TC-based tree splitting process worked as follows. For every internal node N of each Compara gene tree, the TC was calculated for several vertebrate clades. The TC for node N and clade C was calculated as $TC(N, C) = species(N)/species(C)$, 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 was then traversed from root to tip, and if a given set of taxonomic coverage constraints (TCCs) were satisfied by both subtrees below node i, then the tree was 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 continued recursively until every node was tested against the TCCs. The smallest subtrees which satisfied the TCCs were included in the resulting subtree set. If only the entire gene tree satisfied the TCCs then the entire tree was included; if the entire gene tree failed to satisfy the TCCs, it was excluded altogether from the resulting subtree set.

I chose a variety of TCCs to apply to the set of Compara trees, all of which were run against the 18,607 gene trees within the Compara database to generate several genome-wide subtree sets. Table 3.1 shows the details of the various TCCs used. The clade names (e.g., $TC(Primates)$) refer to the set of species in the Ensembl database that

are contained within the given subtree of the NCBI taxonomy; the NCBI classification of species contained within Ensembl is shown in Figure 3.2, including labels on internal nodes corresponding to the clade names given in Table 3.1.

For the Ingroup and Outgroup categories of TCCs, a TC value of greater than 0.6 was required for a single taxonomic clade. This value was not arbitrarily chosen; rather, it was important to use a TC value slightly above 0.5 to achieve the desired result of identifying orthologous subtrees. A value much higher would be too restrictive: if, for example, the required TC value were set to 1, then all subtrees containing a deletion in any species within the clade of interest would not satisfy the TCC. On the other hand, a required TC value of less than 0.5 would allow a single LOT to be split into two subtrees, with one subtree having TC$<$ 0.5, and the other subtree, containing the other half of the species within the clade of interest, also having TC$<$ 0.5. Thus, 0.6 was deemed a sufficient TC requirement for isolating subtrees with reasonably high TC while allowing for some amount of gene deletion.

Two additional types of TCCs were designed for use in the MammalSubgroups and MammalSubgroupsPlusOutgoup methods. Inspired by the alignment filtering method from Pollard et al. [2010], which required at least one sequence to be present from each of the three major mammalian superorders (Primates, Glires, and Laurasiatheria) for a column to pass through the filter, the TC_{all} TCC required that the TC for all of the included clades was above a given minimum value. To complement the TC_{all} constraint, the TC_{any} constraint was designed to require that the TC for any of the included clades was above a given value. These more complicated TCCs were included in the analysis in order to determine whether combinations of more specialized constraints would perform as well or better than the simplest approach at isolating LOTs from the Compara gene trees.

The methods within the Orthologs category of subtree sets were implemented separately from the rest. Instead of splitting Compara trees based on TCCs, subtrees in the Orthologs category were defined from the set 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 Ensembl-annoated orthologs was extracted and stored; this was guaranteed to yield exactly one subtree for every gene in the source species. Human, mouse, zebrafish, and drosophila were chosen as source species for testing. This approach differed from the tree-splitting strategy in two ways: first, it made use of the orthology annotations resulting from Ensembl's orthology pipeline, which uses the Compara gene trees as its source. Second, this approach did not guarantee that each subtree would contain

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 3.1: 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

a completely unique set of genes. For example, a gene which was recently duplicated in the human terminal lineage 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 annotations would serve as a useful reference against which to compare the methods based purely on TCCs.

The subtree splitting scheme was applied to the 18,607 gene trees from the Compara database, producing a set of subtrees for each of the TCCs shown in Table 3.1. In the next two sections I will describe the resulting sets of trees and subtrees and discuss what they reveal about the evolutionary history of vertebrates and the feasibility of using TCCs to isolate mammalian LOTs for sitewise analysis.

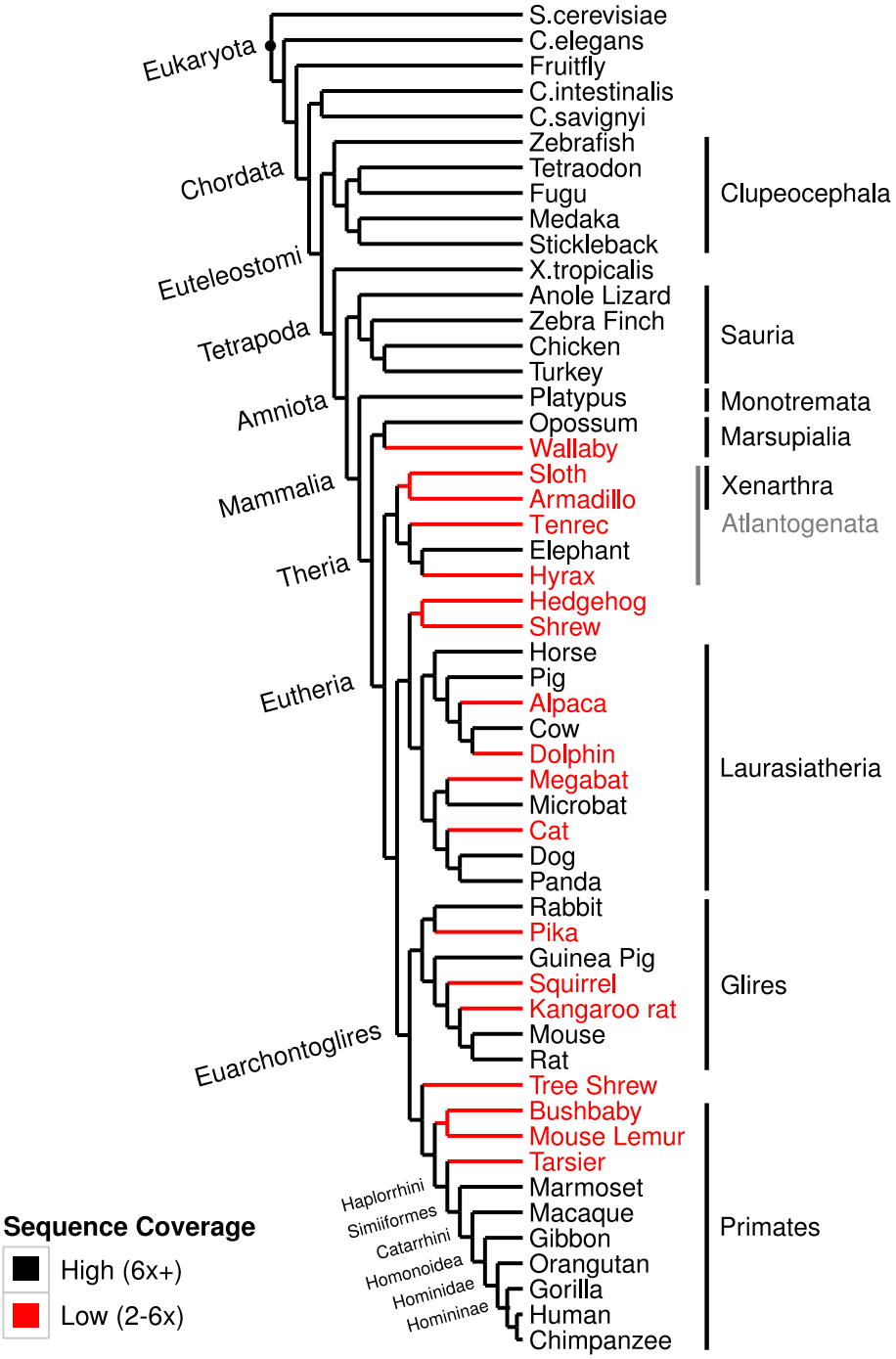


Figure 3.2: The NCBI taxonomy of species within the Ensembl Compara database. Note that branch lengths are not drawn to scale. Low-coverage genomes (e.g., those with 2-6x mean sequence coverage) are labeled in red, and high-coverage genomes are in black. Clade names are included on the left and on the right side of the tree.

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 3.2: 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

3.7 Analysis of the set of root Compara gene trees

Table 3.2 presents a summary of the set of root Compara gene trees and the subsets of trees with greater or fewer than 15 sequences.

A first observation was that, somewhat surprisingly, nearly half of all Compara gene trees contained few sequences: 9,378 out of 18,607 root trees constituted 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 gene trees might have resulted from extensive lineage-specific gene duplications or from pseudogenes being mis-annotated as genes, in either case causing tight clusters of very closely-related transcripts that would have been identified by *hcluster_sg* as independent clusters of homologs. Some evidence for the latter scenario could be found in the median species counts and mean path lengths of the smaller (fewer than 15 sequences) versus larger (greater than 15 sequences) trees, shown in the second and third rows of Table 3.2. The set of small root trees had a median species count of 2, compared to 47 for the large trees, indicating that the smaller trees encompassed sequences from a very small taxonomic range. Furthermore, the median mean path length (MPL) for small trees was 0.04 compared to 1.04 for the large subset, revealing a much smaller level of sequence divergence within each tree. These two summary statistics together provided evidence that the smallest gene trees in the Compara database consist of highly species-specific, closely-related proteins; it was likely that many of these small trees represented artifactual gene annotations and would be most appropriately removed prior to any downstream analysis.

Despite the existence of many small gene trees in the Compara database, they comprised only a small fraction of all protein-coding sequences. Only 809 human genes, or 4% of the

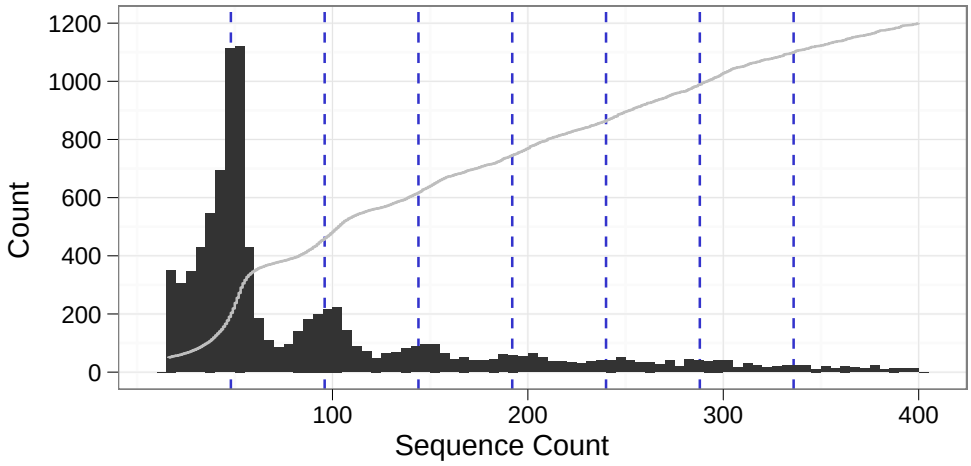


Figure 3.3: Sequence counts for the set of root Compara trees. Black bars show a histogram of sequence counts in bins of width 5, and the gray line shows the cumulative fraction of sequences contained within trees of that size or smaller. For clarity, 9,378 trees with 15 or fewer sequences are not shown. Dashed blue lines are drawn at integral multiples of 48, the number of vertebrate species within Ensembl.

total human gene set—a gene set which I expected 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—was contained within the subset of small trees. This indicated that whatever process was causing the Compara pipeline to yield a high number of small gene trees did not have much of an impact on the overall placement of human protein-coding genes within the database of root Compara gene trees.

As mentioned in Section 3.5, a closer examination of the distribution of tree sizes in the set of root Compara trees presented a clear view of the over-clustering of mammalian LOTs. The black bars in Figure 3.3 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 exist at larger tree sizes, with the location of these wave-shaped peaks corresponding closely to the second, third and fourth multiples of 48. The pattern of recurring peaks is 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 provided good support for the situation described in Section 3.5, with the Compara pipeline often clustering together two or more mammalian LOTs related

by a more distant duplication event.

I found it interesting to characterize the set of root Compara gene trees by the proportion of all protein sequences which are covered by trees of a given size or smaller. This value is plotted in Figure 3.3 as a gray line. First, this plot showed that trees with fewer than 15 sequences (which were excluded from the plot but inlcuded in the calculation of the cumulative distribution shown) represented a trifling fraction of the sequences within the Compara database; this observation was similar, but not identical, to the statistic noted above, that only 4% of human genes were contained within small trees. Second, the steady upward slope of the cumulative curve contrasted with the declining height of the histogram bars at higher sequence counts. This was a result of the increasing number of sequences encompassed by each of the larger trees: although relatively few trees contained more than 300 sequences, together they comprised around 10% of all protein-coding genes in Compara. Two points along this cumulative plot were of particular interest. First, one could use the height of the cumulative line at the beginning of the second hump in the histogram (at around 75 sequences) to identify the fraction of vertebrate proteins which exist in the Compara database as members of a duplicated mammalian LOT. The line showed that around 30% of proteins are covered by trees of 75 sequences or fewer, meaning that 70% of vertebrate proteins are contained within large gene trees containing sequence-based evidence of ancient paralogy. Second, using the cumulative line in the reverse direction could identify the tree size above which 50% of all sequences were 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 a set of gene trees than the median tree size. A similar calculation is often performed to characterize the size distribution of contiguous sequence blocks (contigs) 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 Compara gene trees, the N50 tree size was 139, slightly less than three times the number of vertebrate species (Table 3.3). Thus, the average protein in the set of Compara gene trees belonged to a tree with 139 sequences, corresponding roughly to three mammalian LOTs related by ancient duplication events, likely resulting from the two rounds of vertebrate genome duplication.

Another way to characterize the distribution of Compara gene trees was across the taxonomic space. Given the wide range of quality in genome assemblies and annotation sets contained within Ensembl, a question of particular interest was whether levels of gene presence and absence were consistent across different species groups and different levels of

assembly quality. To investigate this question in the context of the root Compara gene trees, data were collected by counting the number of sequences from each species contained within each gene tree. Results of this analysis are presented in Figure 3.4, showing the number of trees containing 0, 1, 2, or more than 3 genes from each of the 53 species within Ensembl. Comparing the ranges of values for each copy count (labeled 0, 1, 2 and 3+), I found that the largest number of trees contained zero copies from any one species (between 8,000–11,000 within vertebrate species), a smaller but large number of trees contained one copy from a species (between 4,000-6,000) and several thousand trees contained two, three or more copies (between 1,000-1,500 for 2 copies and 1,500-2,000 for 3+). Note that these numbers were tabulated independently between species; for example, the 10,000 zero-copy trees in human were counted independently of the 10,000 zero-copy trees in chimpanzee, so nothing could be said about how many trees contained zero copies in both human *and* chimpanzee. As noted in the analysis of tree sizes, the large number of zero-copy trees reflected the large number of small, species-specific trees within the set of Compara gene trees. Similarly, the large number of trees with many copies from each species reflected the trees with multiple mammalian LOTs clustered together.

Comparing numbers horizontally across the range of species in Figure 3.4 revealed that the zero-copy count tended to increase with increased evolutionary distance from human, while the 1, 2 and 3+ copy counts tended to decrease as the distance from human increased. Both trends were most striking at the distant end of the tree, where the five non-vertebrate species are shown. 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 long branch lengths connecting those species to each other and to the more densely sampled 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 large distances, and heterogeneity in evolutionary rates across lineages [Whelan, 2008b]. However, the reduced numbers of 2 and 3+ copy count trees in non-vertebrate species were most likely due to the whole-genome duplications at the basal vertebrate lineage, resulting in the non-vertebrate species, which underwent no such genome duplications, being strongly depleted of multi-copy duplicates compared to their vertebrate relatives.

Knowing that many of the low-coverage genomes were annotated using projections based on the human set of transcripts, it was slightly concerning that human and its close primate relatives contained fewer zero-copy genes and more one-copy and two-copy genes

than any other group of vertebrates in the Compara set of trees. There was no *ab initio* biological reason to expect this to be the case, so I suspect that this pattern, which was fairly small in effect, was due to the widespread reliance on human annotation and protein experimental data in the annotation of non-human genomes. There was, however, one region of Figure 3.4 where human and primates did not contain the largest number of one-copy or 2-copy trees: the 3+ copy count for the fish species was greater than that for human or any primates, which was most likely a result of gene duplicates retained after the third round of genome duplication in the teleost ancestor [Jaillon *et al.*, 2004]. The signal resulting from the teleost genome duplication event was clearer in the Fish set of TC-defined subtrees, so I will defer further discussion of this effect to Section 3.8 where that set of trees is described.

Finally, the differences in copy counts between species with low- and high-coverage genome sequences showed the tendency of low-coverage genome sequences to be absent from gene trees or to show reduced copy counts. Low-coverage species contain more zero-copy, roughly the same number of one-copy, and noticeably fewer multi-copy genes than closely-related high-coverage species. These clear differences showed 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 impact of low-coverage genomes on gene duplication inference in the Compara database [Milinkovitch *et al.*, 2010], but the latter point was largely ignored. Again, this signal was also stronger in the sets of TC-defined mammalian subtrees and will be revisited in the next section.

The preceding analysis of the set of root Compara gene trees, in which I characterized the distribution of trees with respect to size and across the taxonomic space, showed that despite the over-representation of small, species-specific trees, most sequences were contained in trees with biologically plausible sizes given the history of vertebrate whole-genome duplications. A gene tree-based equivalent of the N50 statistic was developed for summarizing the distribution of protein sequences within differently-sized gene trees, and two visual summaries of this distribution were introduced (in Figures 3.4 and 3.3), providing evidence for the clustering of paralogous mammalian sub-trees and for species-based and genome quality-based trends in the breakdown of gene copy counts within the root Compara gene trees.

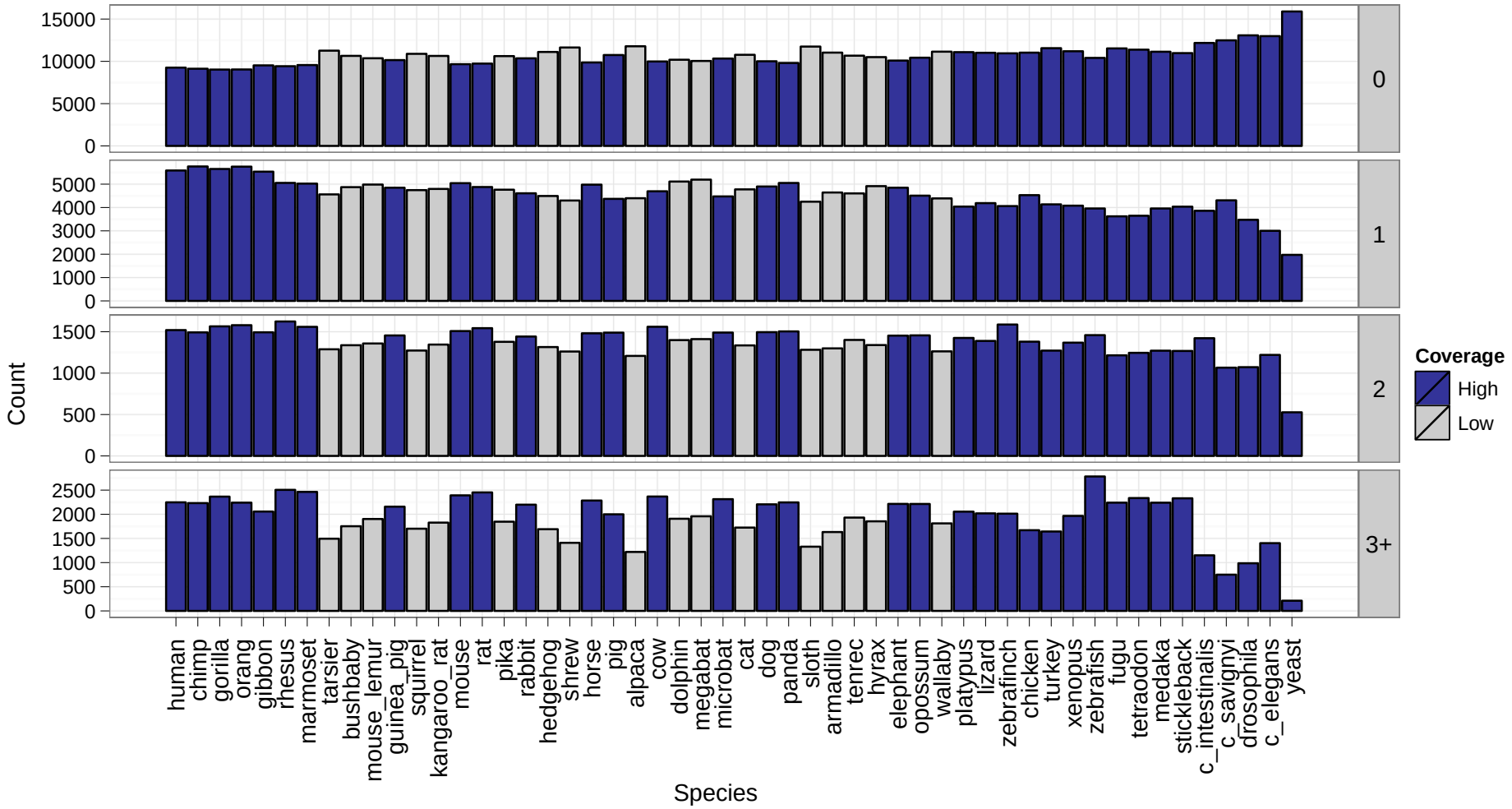


Figure 3.4: Taxonomic distribution of gene copy counts for the root Ensembl trees. The number of trees containing 0, 1, 2 or more than 3 sequences from each species is shown. Bars are colored blue and gray for species with high- and low-coverage genomes, respectively. Note that the y-axis scale is not the same for each panel.

3.8 Analysis of sets of subtrees defined by taxonomic coverage and orthology annotation

The sets of trees resulting from applying the subtree splitting scheme with various TCCs to the Compara gene trees are summarised in Table 3.3, with a summary of the root Compara gene trees and a summary of the set of seven-species amniote gene trees from the Optic database [Heger & Ponting, 2008] included at the bottom for comparison.

The Ensembl Roots and Drosophila Orthologs sets were two clear outliers among the subtree sets shown in Table 3.3, with much higher N50 values (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) than any other subtree set. In fact, these two subtree sets were very similar except for the excess of small species-limited trees in the Ensembl Roots set: the Drosophila Orthologs set contained fewer trees than the Ensembl Roots (9,210 vs. 18,607) and a larger average tree size (60 vs. 15), and the summary values closely resembled the set of Ensembl Roots with small trees removed (Table 3.2).

Within the Ingroups category, methods based on mammalian TCCs (Primates, Glires and Laurasiatheria) produced largely similar sets of trees, with the Primates set containing around 2,000 more trees and covering around 1,000 more human genes than the Glires and Laurasiatheria sets. There was no readily apparent reason for the higher number and human gene coverage of Primate trees, although it may speculatively be due to an excess of primate-specific gene trees that were not captured by non-primate TCCs. Further investigation of the trees unique to this set would be required to reveal the root cause of this minor discrepancy.

The Sauria set of subtrees was noticeably different from the mammal-based TCCs sets from the Ingroups category. The Sauria clade was represented by only four species in Ensembl and diverged from the mammalian ancestral population at an early point in the evolution of amniotes; it is plausible that the lower clade size and long branch length separating Sauria from the other vertebrate clades caused 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).

The Fish clade TCC produced a strikingly different set of trees, resulting from the impact of the teleost-specific whole-genome duplication on the structure of gene trees in fish. Although the Fish subtree set yielded a N50 value of 49, which was no different from the N50 of the other Ingroups sets, Table 3.3 highlights three major differences in the Fish

set: it contained many more trees, a higher proportion of trees with zero human copies, and a lower total human gene count than the other Ingroups sets.

The reason for the drastically different Fish tree set was that the tree splitting procedure was designed to identify the largest non-overlapping set of subtrees that satisfied the given TCCs. Genes where one or both of the duplicate copies from the teleost-specific whole-genome duplication were lost would appear as one-to-one orthologs or deletions with respect to the other vertebrate lineages. Genes that were retained in duplicate form, however, would result in a gene tree with two teleost-specific subtrees, each containing a high TC value (i.e., near or equal to 1.0) for the Clupeocephala clade that contains the fish species within Ensembl. In this case, the splitting procedure would produce two small fish-specific subtrees, "ignoring" the surrounding set of mammalain orthologs because two smaller 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 3.5, 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 TCCs 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 species composition. Each subtree 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 strong evidence that the tree-splitting method was accurately isolating mammalian LOTs. Some slight differences between subtree sets 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 potentially be explained by the minimum required tree size increasing along with the clade size, as a result of the increased number of species required

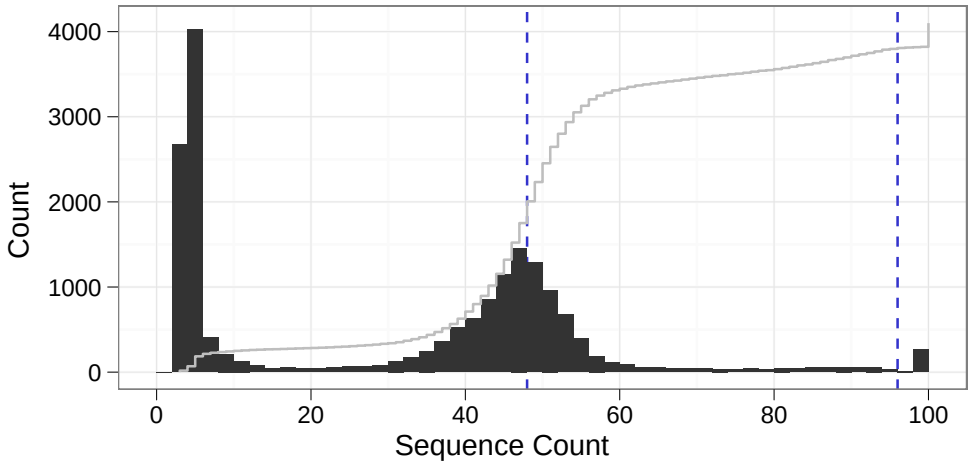


Figure 3.5: Sequence counts for the set of subtrees identified using the Fish clade taxonomic coverage constraint, showing an excess of small subtrees resulting from the teleost genome duplication. Black bars show a histogram of sequence counts in bins of width 2, a gray line shows the cumulative fraction 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. Trees with more than 100 sequences are included in the topmost bin.

to produce a TC value of 0.6. The minimum subtree size ranged 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 Outgroups methods. The MammalsSubgroups set produced more trees than the Outgoups sets, but the N50 was slightly lower (46 vs. 49) and the proportion of zero-copy human trees was higher (0.18 vs. 0.01), suggesting that the additional trees in the MammalsSubgroups set were spurious subtrees containing limited 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 the three different ortholog sources and between the overlapping and non-overlapping sets of subtrees. The *Drosophila* ortholog set was highly contrasted with the vertebrate sets due to the two rounds of whole genome duplication, while there was minimal variation among the other ortholog sets. It is interesting to note that the protein-coding transcripts used by the Compara pipeline included 21,873 mouse protein-coding genes and

only 19,991 human genes, indicating either a larger number of true protein-coding genes in mouse or a higher tolerance for false positive gene predictions in the mouse genome compared to the human genome. Zebrafish, on the other hand, contained 24,540 genes; this number agreed well with the 17.5% proportion of retained duplicate genes that I estimated earlier in this section. Overall, 76% and 81% of mouse and zebrafish genes contained an apparent orthologous relationship with one human gene, which was only slightly lower than the 92% of Eutheria subtrees containing one human sequence.

Subtree Method			Med. Size		Human Content			Human	Med.	Med.
Category	Name	Count	(Min / Max)	N50	0	1	2+	Total	MPL	Species
Ingroups	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
Outgroups	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
Default Trees	Ensembl Trees	18607	15 (2 / 400)	139	0.50	0.30	0.20	19995	0.55	8
	Optic Trees	17372	9 (2 / 789)	9	0.12	0.79	0.09	18477	0.00	8

Table 3.3: Summary of Ensembl subtrees identified using taxonomic criteria or Ensembl ortholog annotations. The set of Compara gene trees from Table 3.2 and the set of trees from the Optic database [Heger & Ponting, 2008] are included at the bottom 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

In order to easily compare patterns between subtree sets and across the taxnomic space, I tabulated gene copy counts within each Ensembl species for each generated subtree set; the results of this analysis are shown in Figure 3.6. By way of reference, the values shown in each of the separate panels of Figure 3.4 appear in the bottom panel of Figure 3.6 as stacked bars of different colors. Although the various summary characteristics of each of the subtree methods have already been discussed at length, this more taxonomic-oriented view revealed some salient features of the patterns of gene deletion and duplication within the tree sets and showed the pervasive impact of genome-wide duplications on the evolution of vertebrate genes. The large fraction of species with multiple copies in the Drosophila Orthologs subtree set (Figure 3.6, blue and purple bars) showed the effect of two rounds of vertebrate genome evolution, producing a large number of trees with multiple Drosophila orthologs in vertebrate species; similarly, the elevated fraction of multi-copy fish trees in the Outgroups subtree sets (e.g., the blue bars in the Eutheria panel of Figure 3.6) showed the impact of the teleost-specific duplication event on the number of mammalian LOTs with multiple fish copies.

3.9 Gene duplication and loss in the set of Eutherian largely orthologous trees

The subtrees defined by the Eutheria TCC were chosen as the final set of gene trees for use in the downstream sitewise analysis. This set was chosen due to its slightly larger number of trees and better coverage of human genes compared to the other subtree sets from the Outgroups category (Table 3.3).

The distribution of tree sizes for the set of Eutherian subtrees is shown in Figure 3.8. The histogram of tree sizes showed a single peak at exactly the number of vertebrate species in Ensembl, with no trees containing fewer than 20 sequences and a few hundred trees with more than 100 sequences. The distribution of tree sizes was consistent with the set of 16,477 gene trees representing an accurate set of genome-wide mammalian LOTs, with variations in sequence counts resulting from sporadic gene duplication or loss events or unannotated genes in low-coverage genomes.

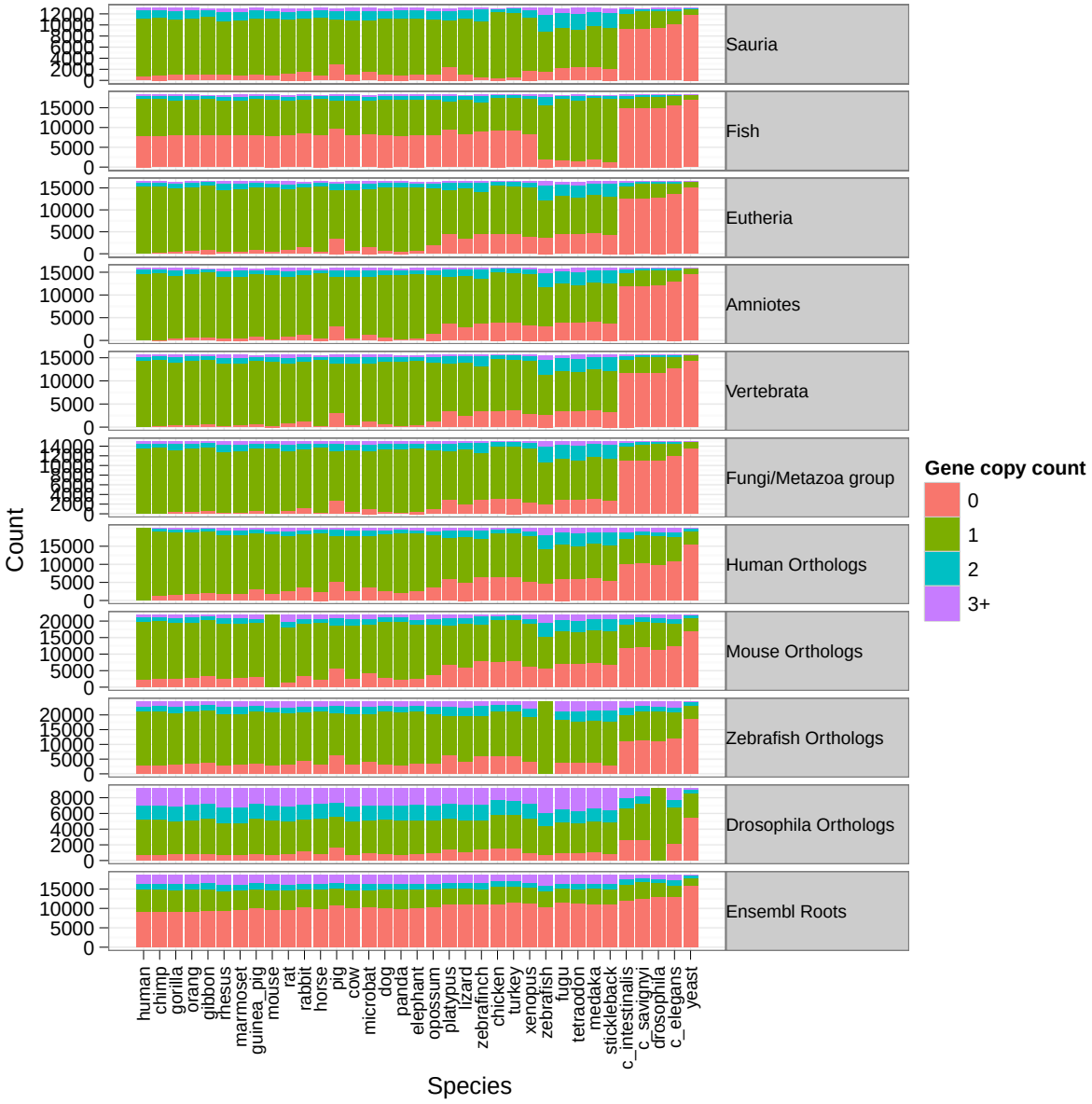


Figure 3.6: Taxonomic distribution of gene copy counts for different subtree methods. The numbers of trees containing 0 (red), 1 (green), 2 (blue) or more than 3 (purple) sequences from each species are shown as stacked colored bars. The Ingroup and Subgroups methods were omitted for clarity, as were species with low-coverage genomes. Note that the y-axis scale is not the same for each panel.

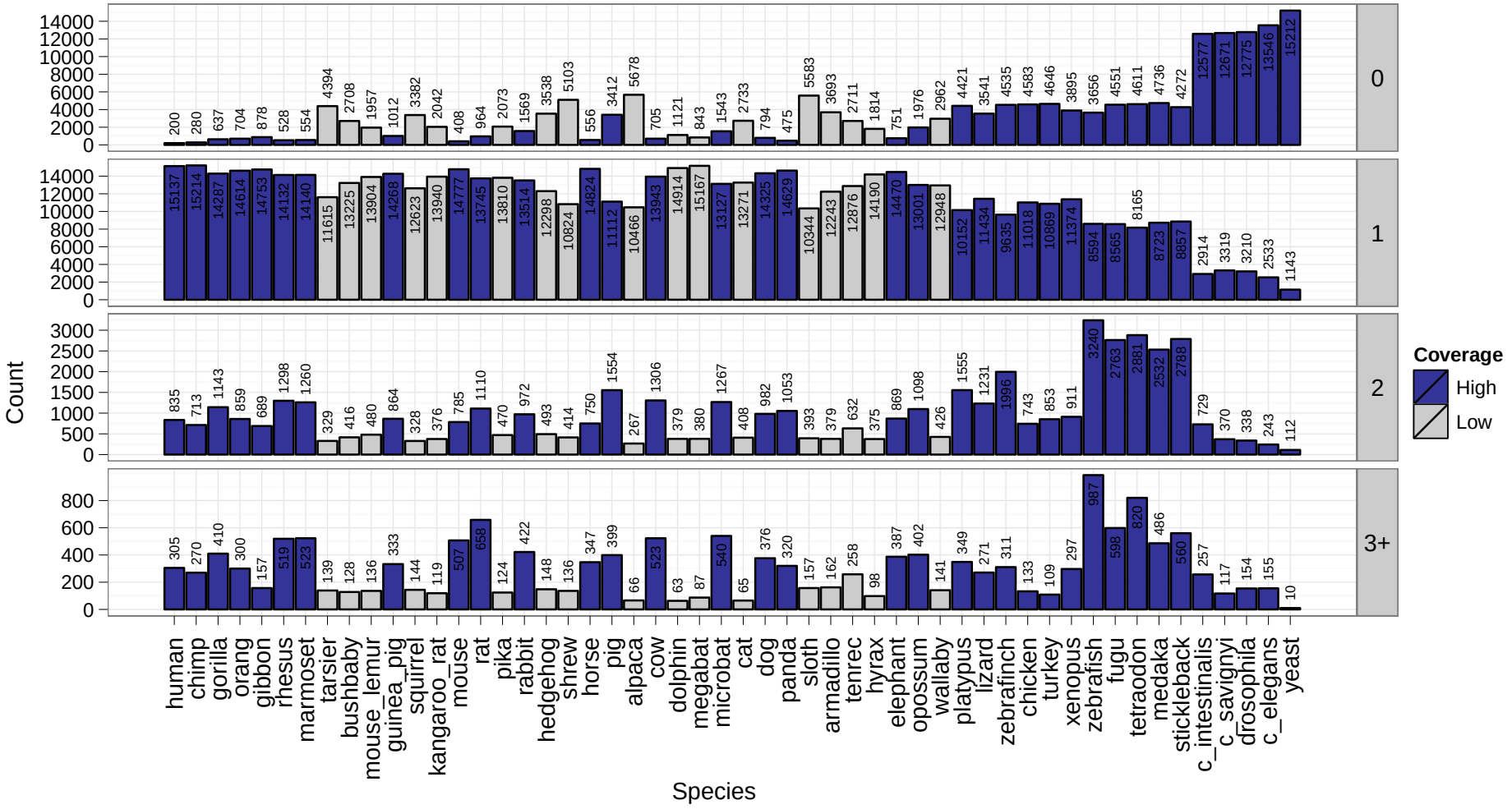


Figure 3.7: Taxonomic distribution of gene copy counts for the Eutheria subtrees defined by taxonomic coverage constraints. Each panel from top to bottom shows the number of trees containing 0, 1, 2 or more than 3 sequences from each species. Bars are colored blue and gray for species with high- and low-coverage genomes, respectively. Note that the y-axis scale is not the same for each panel.

I also analyzed the detailed taxonomic distribution of gene duplications and losses implied by the set of Eutherian subtrees, as the relative prevalence of zero-copy and multi-copy trees in this set of LOTs might provide some indication of whether gene deletion or gene duplication has been more prevalent in the evolution of vertebrate genomes. Figure 3.7 shows the gene copy counts for the set of Eutherian subtrees across all Ensembl species, similar to what Figure 3.4 showed for the set of root Compara gene trees. The excess of zero-copy trees and deficit of duplications in low-coverage genomes was immediately apparent from Figure 3.7, confirming the trend observed in the set of root Compara trees. Similarly, the trend of increased zero-copy trees in non-mammalian and non-vertebrate species was stronger than in Figure 3.4, as was the excess of trees with multiple copies in fish genomes.

A quantitative comparison of the number of multi-copy versus zero-copy trees in each species showed that gene duplication has had a greater apparent impact on gene copy counts than gene deletion, at least within primates and most mammals with high-coverage genomes. For example, human contained 200 zero-copy trees and 1,140 combined two- or three-copy trees within the set of Eutherian subtrees, showing evidence for a greater prevalence of gene duplication than gene deletion in LOTs since the common Eutherian ancestor. On the other end of the spectrum in primates was gibbon, which had the most zero-copy trees (878) and the fewest multi-copy trees (846) of all the primates, showing roughly equal tendencies towards gene deletion and duplication across the set of Eutherian LOTs. This pattern was consistent across the mammalian tree, save for a few exceptions:n guinea pig showed roughly the same number of zero-copy and multi-copy trees (1012 vs. 1197, respectively), rabbit showed slightly more zero-copy trees (1569 vs. 1394), and pig showed a much higher number of zero-copy trees (3412 vs. 1953). Beginning with opossum, vertebrates more distantly related to the Eutherian common ancestor showed greater numbers of zero-copy trees, leveling off at ca. 4,000 zero-copy trees, and higher variation in the number of multi-copy genes; at the low end were chicken and turkey with 876 and 972 multi-copy genes, respectively, and at the high end (excluding the fish species) were zebrafinch and platypus with 1904 and 2307 multi-copy genes, respectively.

In the analysis of vertebrate genomes, one must be vigilantly aware of potential biases arising from the frequent reliance on homology with the well-annotated human and mouse genomes in the annotation of newly sequenced genomes. Such a bias could have plausibly led to anomalous results in the present analysis of gene copy counts, for example by reducing the chance of correctly identifying gene trees containing deletions in human or mouse, thus

over-inflating the prevalence of duplications versus deletions. The level of consistency seen in the relative numbers of zero-copy and multi-copy trees across the range of mammals provided some evidence against such a bias, although it did not entirely rule out the possibility, as all mammalian genomes may have been similarly affected by the use of human or mouse proteins in the Ensembl gene annotation pipeline.

An unexpected result of the compraison between mammalian species was the identification of certain genomes with uncharacteristically high numbers of zero-copy or multi-copy trees. The most striking example was pig, which contained nearly double the number of zero-copy trees of any other high-coverage mammalian genome and a noticeably elevated number of multi-copy trees, but rabbit and guinea pig also deviated from the normal patterns. Given the otherwise consistently low number of zero-copy trees throughout the range of Eutherian mammals, I would expect the number of zero-copy trees in these species to decrease as finished-quality genome sequences ad produced and the gene annotation pipelines are further optimized to work with each species. In particular the anomalous nature of the pig gene trees may be related to the draft quality of the genome sequence and assembly; I would expect the number of zero-copy trees to be substantially reduced once a finished-quality genome sequence and annotation set is made available [Archibald *et al.*, 2010].

3.10 Comparison to gene trees from the Optic database of amniote orthologs

Given the tendency of the root Compara gene trees to contain multiple mammalian LOTs, I wished to evalute whether a different phylogenetic tree-based orthology pipeline produced a similar distribution of gene tree sizes. The Optic database, a project from the Ponting group which used an independently designed gene-building and comparative genomics pipeline combined with the TreeBeST software to infer duplication-resolved gene trees within a variety of vertebrate and invertebrate clades, was ideal for such a comparison [Heger & Ponting, 2008]. I downloaded the set of Optic gene trees inferred from a group of eight vertebrate genomes (human, mouse, dog, opossum, platypus, chicken, zebrafinch, and tetraodon), characterized the set of gene trees using a variety of summary statistics, which are included in the bottom row of Table 3.3, and plotted the distribution of gene tree sizes in Figure 3.9.

After accounting for differences in the number of sampled species, the set of Optic

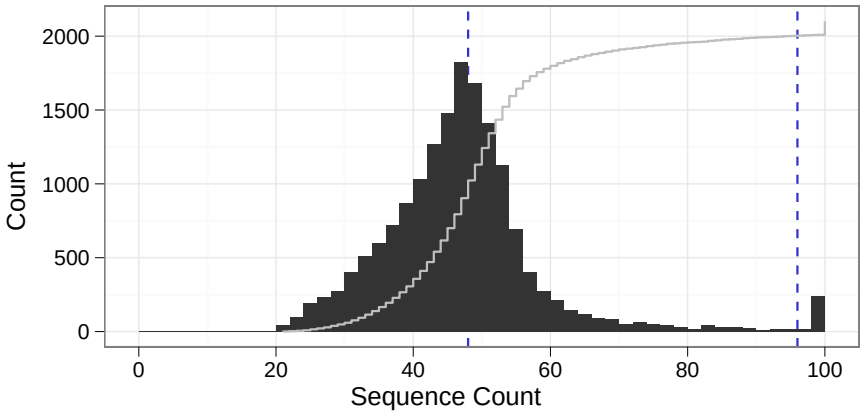


Figure 3.8: Sequence counts for the set of subtrees identified using the Eutheria clade taxonomic coverage constraint. Black bars show a histogram of sequence counts in bins of width 2, a gray line shows the cumulative fraction of sequences contained within trees of that size or smaller, and dashed blue lines are drawn at integral multiples of 48, the number of vertebrate species within Ensembl. Trees with more than 100 sequences are included in the topmost bin.

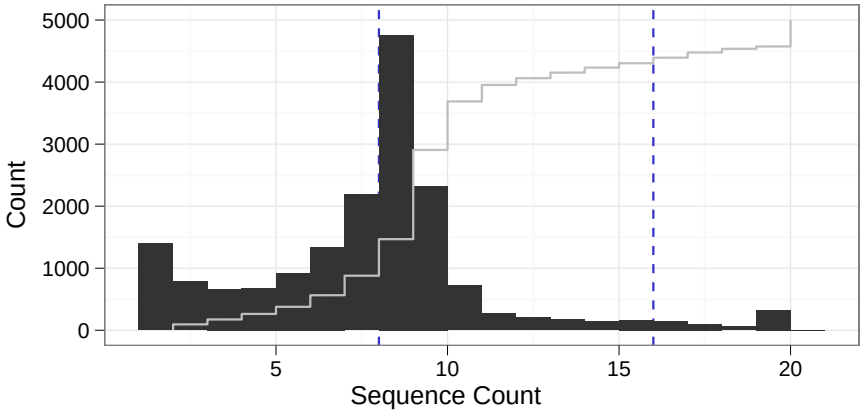


Figure 3.9: Sequence counts for gene trees in eight amniote species from the Optic orthology database [Heger & Ponting, 2008]. Black bars show a histogram of sequence counts, a gray line shows the cumulative fraction of sequences contained within trees of that size or smaller, and dashed blue lines are drawn at integral multiples of 8, the number of amniote species analyzed by Optic. Trees with more than 20 sequences are included in the topmost bin.

gene trees more closely resembled the set of Eutherian subtrees than the root Compara gene trees. Nearly 80% of the Optic gene trees contained exactly one human sequence and only 9% contained two or more human sequences; the large proportion of single-copy trees suggested that the Optic trees did not contain many "over-clustered" mammalian LOTs, and the 9% of trees with multiple human genes was close to the 7% seen in the

set of Eutherian subtrees. Figures 3.8 and 3.9 clearly show the similar sequence count distributions between the Optic and Eutherian trees; each histogram has a distinct peak at a sequence count corresponding to the number of species included in the database, with no sign of the long tail of large gene trees seen for the distribution of root Compara gene trees in Figure 3.3.

3.11 Conclusions

This chapter was concerned with identifying a set of gene trees for subsequent sitewise evolutionary analysis. Three characteristics were desired in the ideal set of trees: maximal coverage of available mammalian genomes, minimal inclusion of paralogous relationships, and consistent taxonomic representation throughout the set of trees. I chose the Ensembl Compara database as a source of root gene trees due to its well-established methodology [Heger & Ponting, 2008; Vilella *et al.*, 2009] and its increased species sampling compared to equivalent vertebrate databases such as Treefam and Optic.

Using the set of root Compara gene trees as a basis for further analysis, I characterized the distribution of gene trees in a varity of ways, including a tree-based analogue of the N50 statistic commonly used in evaluating genome assemblies. This analysis showed that the "over-clustering" of multiple Eutherian LOTs within single Compara gene trees had a major impact on the composition of the gene tree set. I then developed a simple scheme for isolating LOTs from within larger gene trees using flexible TCCs and applied the system to the set of Compara trees to generate several sets of genome-wide TC-defined subtrees.

These sets of TC-defined subtrees provided a number of insights into the orthology and paralogy relationships within and between mammalian and vertebrate genomes, including a quantification of the proportion of duplicate genes that were retained after the teleost whole-genome duplication that matched the prediction resulting from a detailed analysis of fish genomes [Brunet *et al.*, 2006]. The comparison between subtree sets resulting from different TCCs showed that a simple threshold based on TC in the Eutheria clade produced a set of subtrees that contained a high percentage of human genes and satisfied the three desired characteristics for subsequent sitewise analysis.

An analysis of the number of trees with zero, one, two or more sequences for a given species revealed patterns of variation in gene copy counts across vertebrate genomes. Importantly, low-coverage genomes showed a large increase in zero-copy trees (e.g. trees with no sequences from the low-coverage genome) and a notable decrease in multi-copy trees.

Other, more subtle trends were identified, including individual genomes, such as pig, with apparently anomalous numbers of zero- or multi-copy trees, and more widespread trends, such as the relatively few multi-copy trees in avian genomes and the higher prevalence of multi-copy trees versus zero-copy trees within mammals.

The set of Eutherian LOTs identified in this chapter represented, to my knowledge, the largest and most accurate set of mammalian orthologous trees identified to date. Although the set of Eutherian subtrees were isolated from the Compara root trees and not independently derived, the separation of the "over-clustered" Compara trees into biologically realistic sets of Eutherian genes sharing mostly orthologous relationships was an important and fundamental part of preparing these gene trees for subsequent evolutionary analysis.

One might argue that a gene tree with fully-resolved duplication and speciation events, as provided by the Compara database, should be enough to allow for identification of orthologous versus paralogous subtrees. In theory this is true, but such an approach to identifying LOTs would have failed on two grounds: first, it would not distinguish between ancient paralogs resulting from the two rounds of whole-genome duplication in vertebrates and more recent paralogs resulting from the gradual accrual of gene duplications over time, causing recent duplication events to be handled in the same way as ancient duplications absent the incorporation of taxonomic information. Second, in practice the resolution of gene deletions within the Compara gene trees was far from perfect, a point raised by Milinkovitch et al. [2010] in their criticism of the inclusion of low-coverage genomes in the Ensembl Compara orthology pipeline. A lack of data, caused either by missing genes from low-coverage genomes or by the limited amount of sequence data available for phylogenetic reconstruction, may often contribute to errors in the accurate identification of duplication and speciation nodes within the Compara gene trees; the use of TC information sidestepped this source of uncertainty by not relying on accurate resolution of duplication nodes.

The performance of the TC-based scheme described in this chapter showed that simple taxonomic criteria could be used to identify largely orthologous subtrees. The utility of a taxonomic-based approach to identifying meaningful gene tree subtrees was indirectly validated by comparison to gene trees from the Optic orthology database, which showed many similar characteristics to the set of Eutherian subtrees; the pipeline used to create the Optic database includes an explicit step for splitting trees based on taxonomic criteria [Heger & Ponting, 2008], suggesting that a similar technique was found to be important in identifying LOTs within the context of an independently designed orthology pipeline.

Chapter 4

Patterns of sitewise selection in mammalian genomes

4.1 Introduction

This chapter describes the use of sitewise evolutionary estimates to characterize the global distribution of selective constraint across 38 mammalian genomes and within the major mammalian superorders. I will apply the Sitewise Likelihood Ratio (SLR) test, evaluated in Chapter 2, to the set of mammalian orthologous gene trees from Chapter 3 to generate genome-wide sets of sitewise statistics measuring selective constraint in several groups of mammalian species. Both this chapter and the following one are concerned with the analysis of these data: here I will consider the overall distribution of constraint observed in several groups of mammalian genomes, while Chapter 5 will look at the use of these sitewise data to identify trends in the evolution of genes and protein domains.

The first section of this chapter introduces the context in which this project was performed—namely, the sequencing and analysis of several mammalian genomes for the Mammalian Genome Project (MGP)—and outlines the main biological questions underpinning the sitewise analysis I performed.

The next section describes the preparation and alignment of the mammalian gene tree from Chapter 3 and introduce a protocol for filtering genome-wide sitewise estimates. Although the simulations from Chapter 2 showed that sequences with divergence levels above that of most mammalian proteins can be aligned without introducing many false positives due to misaligning biological insertions and deletions, the analysis of empirical sequence data involves many potential non-evolutionary sources of alignment error. A

sequenced and annotated genome is not a piece of observed data; rather, it is the result of a succession of inferences (based ultimately on the observation of a pool of genomic DNA by some sequencing technology), each step along the way involving potential errors and biases. Chapter 3 looked at the identification of mammalian orthologs, showing that the inference of correct gene tree structures is fraught with difficulty and that low-coverage genomes are under-represented in gene duplications. Other sources of error, including those occurring while reading DNA bases [TOCITE], assembling genomic fragments [TOCITE], and annotating gene-coding regions [TOCITE] have all been previously highlighted as being important in the large-scale analysis of genomic data. As such, care was taken to design and evaluate a variety of filters to reduce the probability of yielding misleading results.

The third section looks at the global distribution of mammalian selective constraint in three ways: first, using sitewise estimates from SLR to identify sites evolving under purifying and positive selection at various confidence levels; second, fitting parametric distributions to the set of sitewise estimates to infer the distribution of selective pressures; and third, evaluating the impact of genomic variation in GC content and recombination rate on the distribution of sitewise estimates. The consistency of these results with previous work will be explored.

4.2 The Mammalian Genome Project

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

twofold coverage, meaning each genomic base pair would be covered on average by two sequence reads and roughly 85% of genomic sequence would be covered by at least one read. The decision to sequence many genomes at low coverage was a deliberate choice, designed to maximize the average amount of branch length available for the identification of constrained sequence [Margulies *et al.*, 2007].

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

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

2006].

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

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

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

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

4.3 Data quality concerns: sequencing, assembly and annotation error

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

The possibility that erroneously-aligned sequences might cause false positives in the detection of sitewise positive selection was a major concern for this analysis, especially given the low-coverage nature of the 20 newly-sequenced genomes. Although the SLR test and other sitewise maximum likelihood methods have been shown to be conservative in their identification of positively selected sites under most conditions, even when the amount of data is low or the null model is violated [Anisimova *et al.*, 2002a, 2003; Massingham &

Goldman, 2005b], most evolutionary analyses are based on the assumption that all sites within an alignment column are truly homologous. This assumption can be violated in a number of ways, some of which are described below.

Of course, alignment error can result from errors in reconstructing the evolutionary history of sequences evolving with indels, causing non-homologous codons to be placed in the same alignment column. In Chapter 2 I explored the tendency of a number of progressive multiple alignment programs to produce such errors, showing that PRANK_C alignments introduce few falsely identified positively-selected sites resulting from alignment errors at mammalian-like divergence levels. Thus, PRANK_C was used to align all coding sequences, and the number of false positives resulting from misalignment of biological insertions and deletions was expected to be low.

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

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

estimated substitution rate but not affecting the relative nonsynonymous and synonymous rates.

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

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

For detecting positive selection, the nature of the model used for inferring positive selection and the branch lengths separating the species being tested may also have an impact on the prevalence false positives resulting from sequence errors. Sequence errors should only substantially affect the estimation of nonsynonymous and synonymous substitution rates along the terminal lineage leading to the erroneous sequence data; thus, the potential impact of sequencing error on the inference of a positively selected site or gene can be estimated by considering the potential impact of an inflated rate of nonsynonymous substitution along the terminal branch on the inference of positive selection with a given test.

Both the branch-site test for positive selection (which is not used in this analysis) and the sitewise tests for positive selection (including PAML M8 and SLR, first described in Chapter 2) are sensitive to erroneous substitutions occurring at individual alignment columns, with the major difference between the two types of test being that the branch-site test is highly sensitive to substitutions along the foreground branch(es) being tested for positive selection, while sitewise tests only measure the signal for positive selection across the entire evolutionary tree.

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

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

To summarize, the expected effect of alignment errors on the sitewise detection of

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

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

Schneider et al. [2009a] 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.*, 2009a]. 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; Studer *et al.*, 2008b; Casola & Hahn, 2009], high GC content [Ratnakumar *et al.*, 2010] or functional shifts [Storz *et al.*, 2008; Wang & Gu, 2001] might have had an error-independent effect resulting in a higher proportion of PSGs in low-scoring categories. The heads-or-tails method has also been shown to be inappropriate for estimating alignment uncertainty [Fletcher & Yang, 2010a], so results based on this measurement should be taken with caution. Despite these criticisms, the analysis did provide good evidence that some of the obvious sources of error may be contributing to excessive estimates of branch-specific positive selection in mammals.

Mallick et al. [2009] took a different approach to the same problem by performing a careful resequencing and reassembly of the chimpanzee genome (the initial assembly of which had lower coverage and lower quality than the human genome) and re-analysing the evidence for positive selection along the chimpanzee linegae in 59 genes which had previously been identified as chimpanzee PSGs. The authors, who were motivated by a concern that previous reports of a larger proportion of PSGs in chimpanzee than in human [Bakewell *et al.*, 2007] were the result of its lower-quality genome rather than a biologically significant difference in levels of adaptation, found that the vast majority of PSGs identified in two previous studies showed no evidence for positive selection when using their reassembled and higher-coverage version of the chimpanzee genome [Mallick *et al.*, 2009]. This suggested that the original 4x coverage chimpanzee genome contained a number of sequencing and assembly errors leading to false inferences of positive selection. A detailed analysis of 302 codons with multiple spurious nonsynonymous 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. [2009a] and Mallick et al. [2009] provide strong evidence in support of the hypothesis that errors in sequencing, assembly, annotation and alignment can result in strongly elevated inferred ω values when using sensitive tests for detecting positive selection. Furthermore, the detailed identification and quantification of error sources performed by Mallick et al. [2009] provided an empirical estimate of how important each potential source of error would be in the detection of positive selection. Although both of these studies used the branch-site test for detecting positive selection, their results could be expected to generalize well enough to guide the design of filtering methods for the present sitewise analysis. With this work in mind, I implemented three filtering steps to help identify and remove sequences and alignment regions potentially subject to the errors noted above: filtering out low-quality sequence, removing gene fragments

and recent paralogs, and identifying alignment regions with extremely high numbers of clustered substitutions.

Filtering out low-quality sequence

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

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

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

A suitable score threshold for filtering coding regions was chosen based on a study by Hubisz et al. [2011], who performed a detailed analysis of Phred quality scores, which are a probabilistic prediction of the error rate, and actual error rates in low-coverage mammalian genome assemblies by comparing the low-coverage assemblies to matched regions of high-quality sequence from the ENCODE comparative genomics dataset [ENCODE Project Consortium, 2007]. The authors identified a strong correlation between Phred scores and actual error rates for scores below 25, indicating that the scores were accurate predictors of the true error rate in this range. Error rates did not decrease significantly at scores above

25, however, suggesting that the use of an extremely high Phred score threshold would only minimally reduce error levels below those obtained with a moderate threshold. Furthermore, Hubisz et al. noted that 85% of bases in the low-coverage mammalian genomes contain very high Phred scores (> 45) and only 4% have low scores (< 20).

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

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

The above filtering scheme was applied to all coding sequences from each species for which quality scores were available, which included all of the species with low-coverage genomes as well as five with high-coverage genomes: chimpanzee, guinea pig, dog, horse, cow, and elephant. (Note that guinea pig and elephant genomes were originally sequenced at low 2x coverage for the MGP, but they have since undergone additional sequencing to produce high-coverage 7x assemblies. These assemblies were used in Ensembl version 63 and thus in the analysis described below.) The overall percentage of nucleotides filtered from each genome is shown in Figure 4.1. As expected, genomes with high-coverage se-

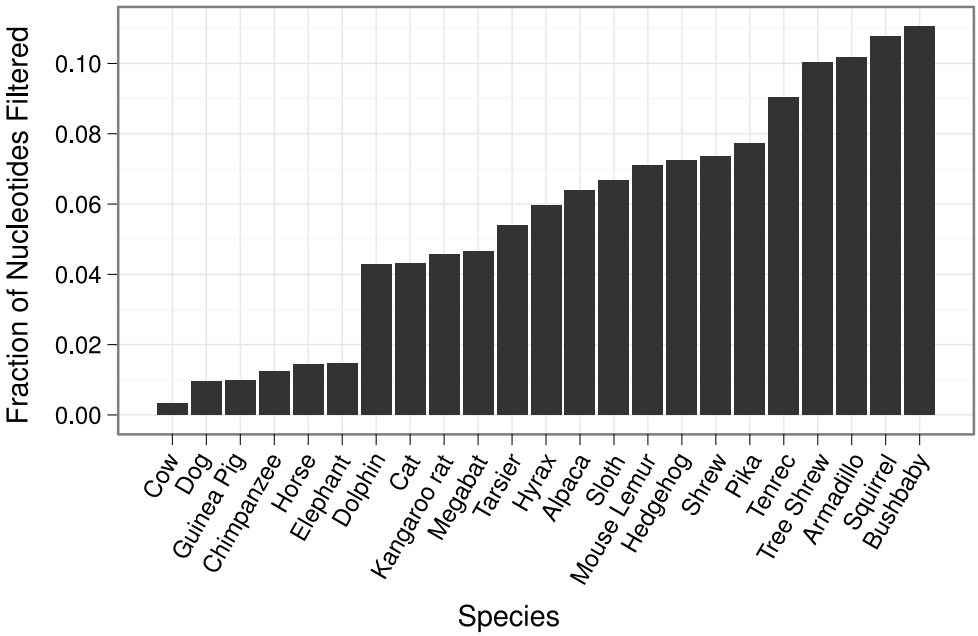


Figure 4.1

quences contained fewer bases with low Phred scores, resulting in 1-2% of nucleotides being filtered. The bulk of low-coverage genomes resulted in 4-8% of nucleotides being filtered, while five genomes (bushbaby, squirrel, tree shrew, armadillo and tenrec) showed a noticeably higher proportion of low-quality bases, with 9-11% nucleotides being filtered out. The distribution of filtered nucleotide proportions confirmed the expectation that 5-15% of nucleotides would be filtered using a Phred score threshold of 25, and the variation in filtered nucleotide proportions between different species showed that despite the uniform 2x coverage of the low-coverage mammalian genomes, different assemblies varied widely in their distributions of sequence quality scores within coding regions.

Removing recent paralogs

As discussed in Section 3.5, the inclusion of paralogous gene relationships in a large-scale analysis of orthologous gene evolution may produce misleading signals of adaptive evolution [Lynch & Conery, 2000], artifacts resulting from gene conversion [Casola & Hahn, 2009], and produce biases due to lineage-specific family expansion, a process which is relatively common in mammalian gene families [Gu *et al.*, 2002]. As a result, it has traditionally been considered important to filter out recently-duplicated genes (e.g., genes duplicated after the whole-genome duplication event in the vertebrate ancestor) in large-scale evolu-

tionary analyses. Previous genome-wide scans for positive selection involving six or fewer mammalian genomes have either required strict one-to-one orthology [Clark *et al.*, 2003; Nielsen *et al.*, 2005] or allowed very limited numbers of recent duplications in specific lineages [Kosiol *et al.*, 2008a]. With larger mammalian trees, however, the requirement of strict one-to-one orthology becomes increasingly untenable: if gene duplications and deletions occur randomly in time, then the probability of observing at least one such event in a given gene family should increase linearly with the amount of branch length covered by the tree. The requirement of one-to-one orthology would result in fewer genes being available for analysis as more species are incorporated into the analysis, which is clearly an undesirable trend. As an alternative to ignoring genes which do not satisfy the requirement of strict orthology, I developed an approach, described below, for handling recently duplicated genes by removing the more-divergent paralogous copy from the the gene tree.

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

minimal.

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

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

The protocol I implemented for filtering apparent paralogs used both gene length and sequence divergence to identify which gene among a set of apparent paralogous copies was most suitable to retain for sitewise analysis. Gene length was used primarily to discriminate spuriously shortened genes from true genes, and sequence divergence was used to distinguish between more- and less-diverged paralogs. First, the mean pairwise sequence distance was calculated between each putative paralog and all other sequences in the gene

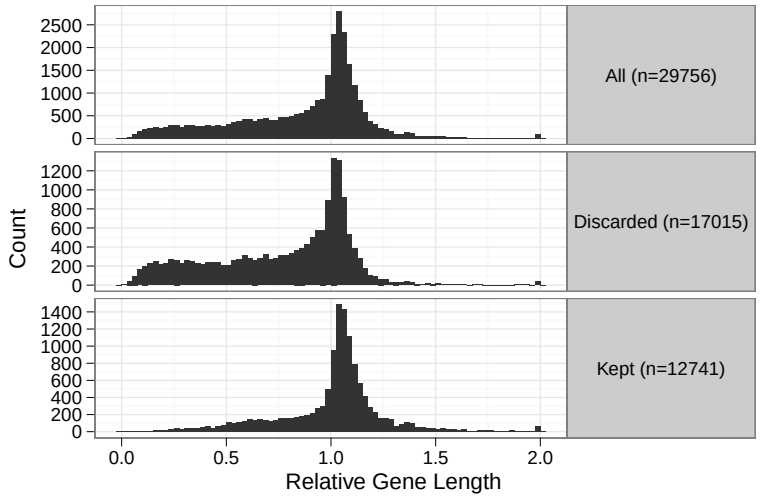


Figure 4.2: Length ratios of putative paralogs. The length ratio was calculated as the length of a putative paralogous copy divided by the mean length all sequences its corresponding gene tree. Putatively paralogous genes (top panel) were either discarded (middle panel) or kept (bottom panel) according to rules based on their length and mean sequence divergence from other aligned sequences, as described in the text.

tree, resulting in one mean pairwise distance estimate per putative paralog (hereafter referred to as the mean distance). For these distance calculations, the stock Compara codon alignments and the JC69 nucleotide model to estimate distances. Second, the ratio of the sequence length of each putative paralog to the mean sequence length across the tree (hereafter referred to as the length ratio) was also calculated.

Genes were grouped by species within each gene tree, and any group of 2 or more genes was considered to be a set of putative paralogs .Within each set of putative paralogs, a single gene was chosen to be retained for evolutionary analysis based on three rules applied in the following order: (1) if only one sequence had a length ratio above 0.5 and all others had a length ratio below 0.5, the longest sequence was kept; (2) if at least one sequence yielded a mean distance below the others, that sequence was kept; (3) if all mean distances were identical then the longest sequence was kept, or if all mean distances and length ratios were equal, an arbitrary choice was made.

These rules were applied to each of the 29,756 putative paralogs contained within the 16,XYZ largely orthologous gene trees from the previous chapter. Figure 4.2 shows the distributions of length ratios separately for the set of all putative paralogs, those discarded from the alignments, and those kept for subsequent analysis. The overall distribution of length ratios shows that most putative paralogs had lengths similar to the mean length

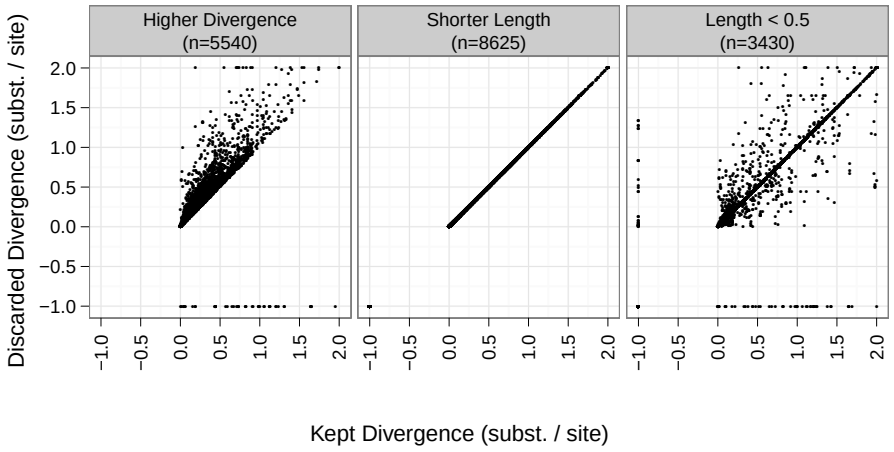


Figure 4.3: Sequence divergence of kept and discarded putative paralogs. Each point represents a gene which was discarded from the tree for one of three reasons: it had more sequence divergence than the kept gene (*Higher Divergence*; left panel), it had equal sequence divergence but shorter length than the kept gene (*Shorter Length*; middle panel), or it had a gene length (relative to the mean across all sequences) of less than 0.5 while the kept copy had a relative length greater than 0.5 (*Length < 0.5*; right panel). Divergence was measured as the mean pairwise divergence between the gene and all other sequences in the tree, and a value of -1 was assigned to genes for which no reliable divergence estimate could be attained due to a lack of sufficient data)

across the gene tree (with a peak at or slightly above 1), but the shape of the distribution was asymmetric, with a strong bias towards shorter lengths. The filtering protocol effectively removed these shortened genes, as evidenced by the strong enrichment of lower length ratios in the distribution of discarded genes and the less skewed distribution of length ratios in the set of XYZ kept paralogs.

To better compare the characteristics of the discarded and kept genes, a dmore detailed view of the results of the paralog filter is presented in Figure 4.3, showing a scatter plot of the mean distance and length ratio of each discarded paralog compared to that of the corresponding kept paralog. Figure 4.3 is separated into panels according to the rule used to discard the paralogous copy: the first panel corresponds to rule (1), where genes with a length ratio below 0.5 were discarded; the second panel corresponds to rule (2), where genes with higher mean distances were removed; the third panel corresponds to rule (3), where all genes had equal mean distances and the longest gene was kept (or, if all lengths were equal, an arbitrary choice was made).

The first panel of Figure 4.3 shows that genes discarded on the basis of having a very

short length contained sequence distances similar to the kept copies, as the highest density is along the diagonal and there is no apparent bias for genes to lie above or below the diagonal. This is in line with the expectationq that these discarded genes were not truly paralogous copies, but rather fragments of split genes resulting from unassembled sequence segments. The second panel shows that when paralogous copies could be differentiated by their mean distances, they tended to have low average distances (<0.5 substitutions per nucleotide site) and only a small difference between the kept and discarded copy (e.g., most of the distribution is just above the diagonal, and few points are above the dashed line with a slope of 2). Finally, the distribution of length ratios and mean distances in the set of genes where length was the discriminating factor (or where an arbitrary decision was made) shows that most of these genes were mostly identical whether measured by sequence distance or sequence length.

These results provided evidence that a sizeable fraction of recently duplicated mammalian genes are identical or very similar to each other: for roughly 30% of putative paralogs, not enough time has elapsed since the duplication event for a detectable amount of sequence change to have occurred, and the choice between retaining one copy or the other was essentially arbitrary. For the roughly 40% of putative paralogs where differences in mean distance could be identified, these differences tended to be small, suggesting that massive functional divergence of recent gene duplicates has not been a common phenomenon in mammalian evolution. Nonetheless, this protocol was designed to identify the least-diverged copy of a recently duplicated gene, and for 40% of putative paralogs the mean distance to other sequences in the gene family allowed a sensible decision to be made.

This was obviously not the most conservative approach to dealing with recent duplications—one could remove all copies from a set of putative paralogs, creating an apparent gene deletion, or one could simply ignore all gene families with any recent duplications (e.g., require one-to-one orthology allowing for gene deletions). The latter option is almost certainly too conservative, but the former option may be appropriate for a more conservative approach. As the main concern over the handling duplicated genes has been that they may introduce a bias towards elevated evolutionary rates, I marked the XYZ genes containing at least XYZ sets of putative paralogs for further evaluation. Sitewise estimates from these genes were excluded from the most conservatively-filtered sitewise dataset and examined separately for excess signal of positive selection (see Section 4.4), and in the next chapter I examine whether using the more conservative approach of removing all paralogous copies

from genes removed the signal of positive selection from a subset of genes (see Section **??**).

Identifying clusters of nonsynonymous substitutions

After filtering for sequence quality and removing paralogous genes and shortened gene fragments, PRANK was used to align the codon sequences of each of the 16XYZ mammalian gene trees. Manual analysis of a number of these alignments revealed many short stretches of clearly nonhomologous sequence in one species, often flanked by stretches of perfect homology and often lying on the borders of exon junctions. These obviously erroneous stretches were likely due to mis-assembly of a genomic region or mis-identification of exon boundaries within the gene of one species. These errors were particularly concerning with respect to the detection of positive selection, as the incorporation of a stretch of apparently nonhomologous material into a sequence alignment would produce many alignment columns with multiple nucleotide differences per codon. As discussed in Section 4.3, this type of error is particuarly prone to cause false positives in the detection of positive selection.

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

Two factors could confound the effectiveness of using clustered substitutions to identify regions of non-homologous aligned sequence. First, the length of the terminal branch leading to each species determines how many lineage-specific substitutions would be expected to occur within a window of a certain size. The terminal human branch, for example, is very short (as it shares a very recent common ancestor with chimpanzee), while the platypus branch is very long (sharing a most recent common ancestor only with the entire eutherian clade). Thus, one would expect to observe many more lineage-specific substitutions in platypus than in human for a given alignment window. In contrast, a stretch of non-homologous aligned sequence should introduce, on average, a constant number of nonsynonymous and synonymous substitutions into the branch ancestral to the sequence

in which it exists. The end result is that it should be more difficult to distinguish homologous from non-homologous stretches in species with long terminal lineages, as species with long terminal lineages will have higher numbers of substitutions in truly homologous regions. On the other hand, this trend should also serve to limit the negative impact of non-homologous stretches in those species on the detection of positive selection, because the resulting elevation in nonsynonymous or synonymous substitutions rates would be less severe.

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

I investigated the distributions of nonsynonymous and synonymous substitutions within windows of mammalian alignments by using *codeml* [Yang, 2007b] under the M0 model (e.g., assuming one ω for all sites and all branches in the tree) to perform the marginal reconstruction of ancestral sequences at internal nodes [Yang *et al.*, 1995] and to identify the substitution events implied by the reconstructed ancestral sequences of each gene alignemnt. Only substitution events occurring between codons with high posterior probabilities in the marginal ancestral reconstruction (> 0.9) were analyzed, and the location of each substitution event along the alignment and within the gene tree was stored. This analysis was performed on all gene trees, yielding a large database of substitution events along internal and terminal branches of the phylogenetic tree confidently inferred from the codon-based PRANK alignments of mammalian gene trees.

Using this set of inferred substitutions, counts of synonymous and nonsynonymous substitutions within non-overlapping 15-codon alignment windows for all terminal and internal nodes were collected; the results for a selection of species and internal nodes are shown in Figure 4.4, which plots the number of 15-codon windows containing a given number of nonsynonymous and synonymous substitutions for a selection of terminal and internal nodes. The mean length of the branch ancestral to the given node, indicated in

parentheses after each node name, was calculated from the set of branch lengths estimated by *codeml*.

Figure 4.4 shows that the vast majority of 15-codon windows in these alignments contained few substitutions (note that the y-axis uses a logarithmic scale), but a long tail of nonsynonymous and synonymous substitutions were observed for some nodes. Comparing the counts of nonsynonymous vs. synonymous substitutions within the terminal nodes (Figure 4.4, top panel), a pattern is seen where the nonsynonymous counts (red bars) are higher than synonymous counts at 0 substitutions, lower than synonymous counts in the middle range of substitutions (1–5 substitutiosn), and higher again in the higher range of substitutions (>5 substitutions). The pattern in the lower range is consistent with the action of purifying selection on protein-coding regions, causing a reduced number of windows with multiple nonsynonymous substitutions compared to synonymous substitutions. The excess of windows with large numbers of nonsynonymous substitutions, on the other hand, runs against the pattern of purifying selection; instead, it shows unexpectedly long clusters of nonsynonymous substitutions to be a widespread feature of these mammalian alignments. The red and blue triangles drawn in each plot mark the number of substitutions below which 99.9% of windows are contained; the shift of the nonsynonymous markers to the right emphasizes the excess of highly clustered nonsynonymous substitutions. Interestingly, human—which has the highest quality and best annotated genome—does not show the same level of excess seen in the other genomes analyzed.

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

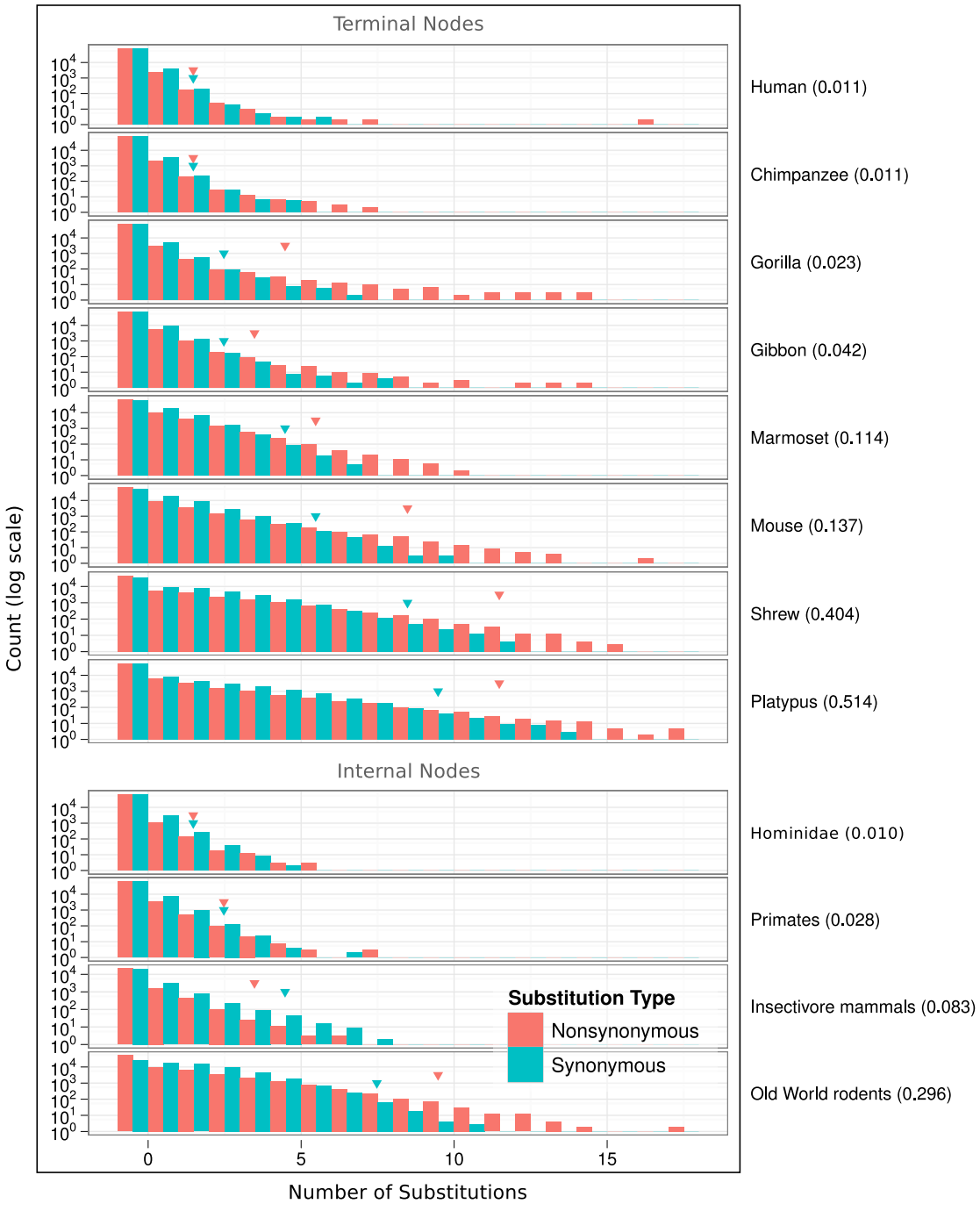


Figure 4.4: Counts of inferred nonsynonymous (red bars) and synonymous (blue bars) substitutions in 15-codon windows along terminal and internal branches of the mammalian tree. The leftmost two bars correspond to windows with 0 substitutions, the next two bars correspond to windows with 1 substitution, and so on. Red and blue arrows indicate the number of nonsynonymous and synonymous substitutions, respectively, corresponding to the 99.9% percentile across all windows in that node.

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

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

4.4 Genome-wide analysis of sitewise selective pressures in mammals

Species groups for sitewise analysis

For each alignment of mammalin orthologs, SLR was run separately on 10 different sets of mammalian species to obtain sitewise estimates in a variety of species groups. For each species group, sequences corresponding to species within the group were extracted from the whole mammalian alignment (along with the corresponding subtree) and input to SLR, which was run with its default parameters. If fewer than two sequences were available for a given gene and species group, the sitewise analysis was skipped for that group. The species included in each group are listed in Table 4.1 alongside the MPL and total branch length of their subtrees estimated as the median value across all 16xyz gene-wise dS branch length estimates from SLR.

Three species groups (Glires, Primates, and Laurasiatheria) were chosen because they represent the three mammalian superorders with the greatest taxonomic representation in Ensembl, providing an opportunity to compare the molecular evolutionary dynamics of three monophyletic mammalian groups containing varying levels of divergence, diverse

	Species		Median dS	
Name	Count	List	MPL	Total
Primates	10	Bushbaby, Chimpanzee, Gibbon, Gorilla, Human, Macaque, Marmoset, Mouse Lemur, Orangutan, Tarsier	0.16	0.83
Atlantogenata	5	Armadillo, Elephant, Hyrax, Sloth, Tenrec	0.26	0.97
HMRD	4	Dog, Human, Mouse, Rat	0.34	1.01
Sparse Glires	5	Guinea Pig, Kangaroo rat, Mouse, Rat, Squirrel	0.36	1.32
HQ Mammals	9	Chimpanzee, Cow, Dog, Horse, Human, Macaque, Mouse, Pig, Rat	0.31	1.61
Glires	7	Guinea Pig, Kangaroo rat, Mouse, Pika, Rabbit, Rat, Squirrel	0.40	1.90
Laurasiatheria	12	Alpaca, Cat, Cow, Dog, Dolphin, Hedgehog, Horse, Megabat, Microbat, Panda, Pig, Shrew	0.26	2.16
Sparse Mammals	7	Armadillo, Dog, Elephant, Human, Mouse, Platypus, Wallaby	0.61	2.86
Eutheria	35	Alpaca, Armadillo, Bushbaby, Cat, Chimpanzee, Cow, Dog, Dolphin, Elephant, Gibbon, Gorilla, Guinea Pig, Hedgehog, Horse, Human, Hyrax, Kangaroo rat, Macaque, Marmoset, Megabat, Microbat, Mouse, Mouse Lemur, Orangutan, Panda, Pig, Pika, Rabbit, Rat, Shrew, Sloth, Squirrel, Tarsier, Tenrec, Tree Shrew	0.35	6.43
Mammals	38	Alpaca, Armadillo, Bushbaby, Cat, Chimpanzee, Cow, Dog, Dolphin, Elephant, Gibbon, Gorilla, Guinea Pig, Hedgehog, Horse, Human, Hyrax, Kangaroo rat, Macaque, Marmoset, Megabat, Microbat, Mouse, Mouse Lemur, Opossum, Orangutan, Panda, Pig, Pika, Platypus, Rabbit, Rat, Shrew, Sloth, Squirrel, Tarsier, Tenrec, Tree Shrew, Wallaby	0.67	8.21

Table 4.1: Species groups used for sitewise analysis by SLR. The median MPLs and the median total branch length are shown for each species group, taken from the 15,XYZ branch lengths estimated by SLR for each gene. MPL – mean path length.

biological characteristics, and a number of high-quality reference genomes. A fourth parallel mammalian subclade, Atlantogenata, consisting of sloth, armadillo, tenrec, elephant and hyrax, was also included, but the monophyly of this group is still under debate [Murphy *et al.*, 2007; Churakov *et al.*, 2009] and it contains only one high-coverage genome. As such, it was not considered a primary target for the mammalian superorder analysis. The different mammalian superorders contained a wide range of total branch lengths, with 0.83 for Primates, 0.97 for Atlantogenata, 1.90 for Glires, and 2.16 for Laurasiatheria. A slightly different ordering was found when measuring the trees by MPL, with Glires having a significantly higher MPL (0.40) than the other groups despite having fewer species and a lower total branch length than Laurasiatheria. This reflected the higher neutral evolutionary rate in the Glires group, a well-documented feature of rodent evolution likely resulting from their long-term shorter generation time, which has been strongly correlated with higher neutral evolutionary rates [Nikolaev *et al.*, 2007a; Smith & Donoghue, 2008].

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

median total branch lengths for Mammalia and Eutheria were 8.21 and 6.43, respectively, and the MPLs were 0.67 and 0.35.

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

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

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

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

for potential errors or biases.

I first examined the overall distributions of ω estimates and sitewise LRT statistics from SLR. Figures 4.5 and 4.6 show the distributions of six sitewise values for each group of species: two continuous values and two categorical values output by SLR (Omega, Signed LRT, Site Pattern and Random) and two values calculated from the codon alignment (Non-gap Codons and Non-gap Branch Length). Non-gap Codons is a count of the number of non-gap codons in each alignment columnn, and the Non-gap Branch Length value represents the total branch length connecting all non-gap sequences (using the gene-wide branch lengths optimized by SLR).

A prominent feature of the distribution of ω values for the unfiltered Mammals data, shown in the top row of Figure 4.6, was the large number of sites with a zero value for ω_{ML}, the maximum likelihood (ML) point estimate of ω. Further inspection of the data revealed that all $\omega_{ML} = 0$ sites contained either synonymous or constant site patterns. Furthermore, all sites with constant patterns (and nearly all sites with synonymous patterns) yielded a ω_{ML} estimate of zero. Intuitively, an estimate of zero for synonymous sites is appropriate, as the lack of any nonsynonymous substitutions throughout the tree would provide no evidence for a nonsynonymous substitution rate of greater than zero. For constant sites the case is less clear, because no data regarding the rate of either synonymous or nonsynonymous substitutions exists in the alignment column. However, given SLR's assumption of a constant synonymous substitution rate throughout each gene [Massingham & Goldman, 2005b], the ω value which maximizes the likelihood of observing zero substitutions is zero, since that value minimizes both the nonsynonymous and the total substitution rate.

It is not evident from Figure 4.6, but a small proportion (ca. 0.2%) of sites containing synonymous site patterns resulted in maximum likelihood estimates greater than zero. Analysis of the alignment columns corresponding to these sites showed them all to include synonymous codons coding for serine or arginine which are separated by multiple nucleotide differences. Under the mechanistic codon model implemented by SLR, which does not allow for multiple simultaneous nucleotide changes, inferring an evolutionary path between these multiply-substituted codons required the inference of multiple nonsynonymous substitutions to reach one codon from the other. This produced a nonsynonymous substitution rate of greater than zero for a site with a synonymous site pattern. The existence of multiply-substituted codons in alignments has been previously reported [Averof *et al.*, 2000; Whelan & Goldman, 2004], and empirical results have supported the notion that codon models that allow for multiple simultaneous nucleotide changes better describe

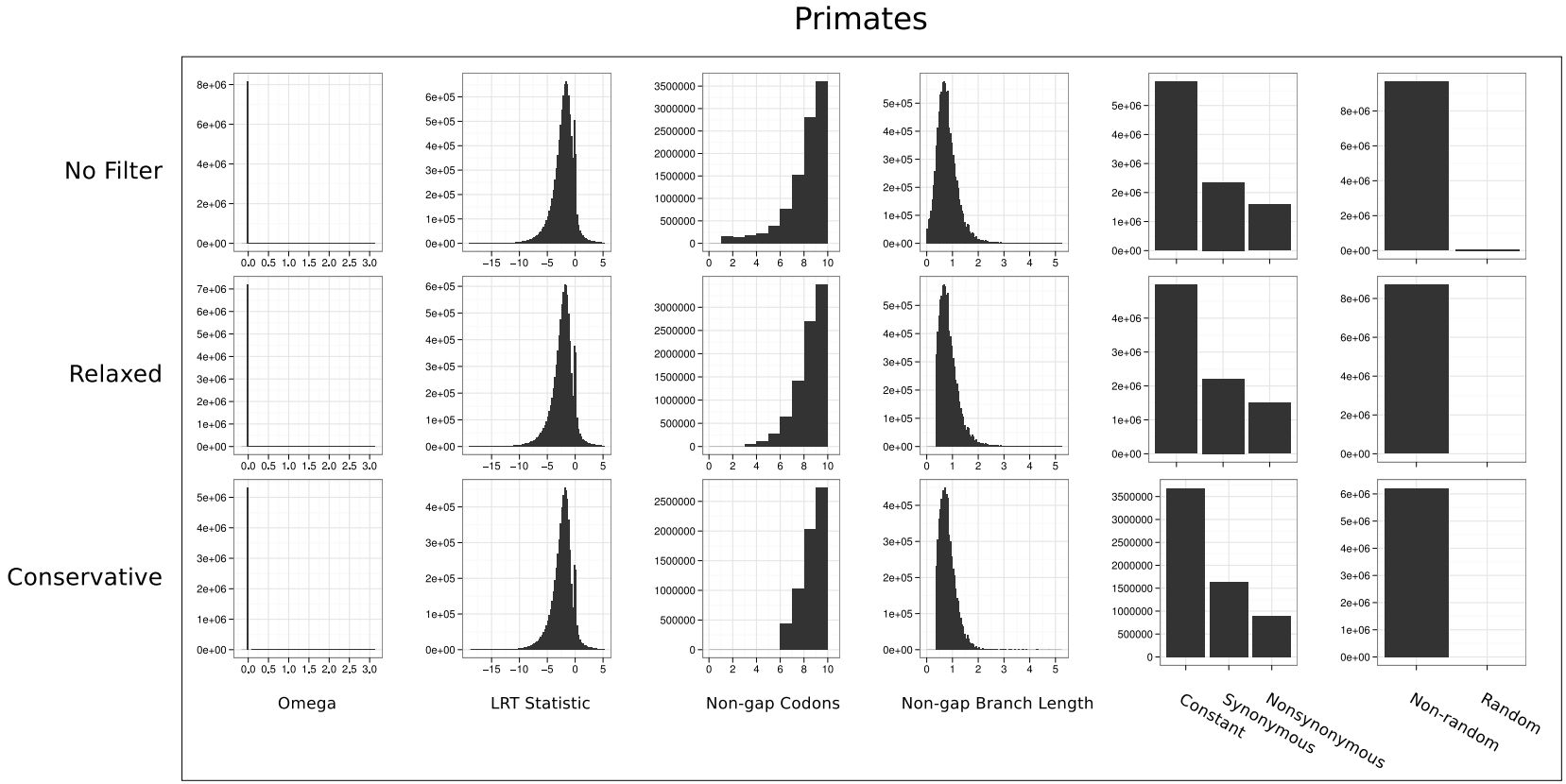


Figure 4.5: Distributions of sitewise values for the Primates species group, showing the raw data (top row) and the result of applying the relaxed (middle row) and conservative (bottom row) filters.

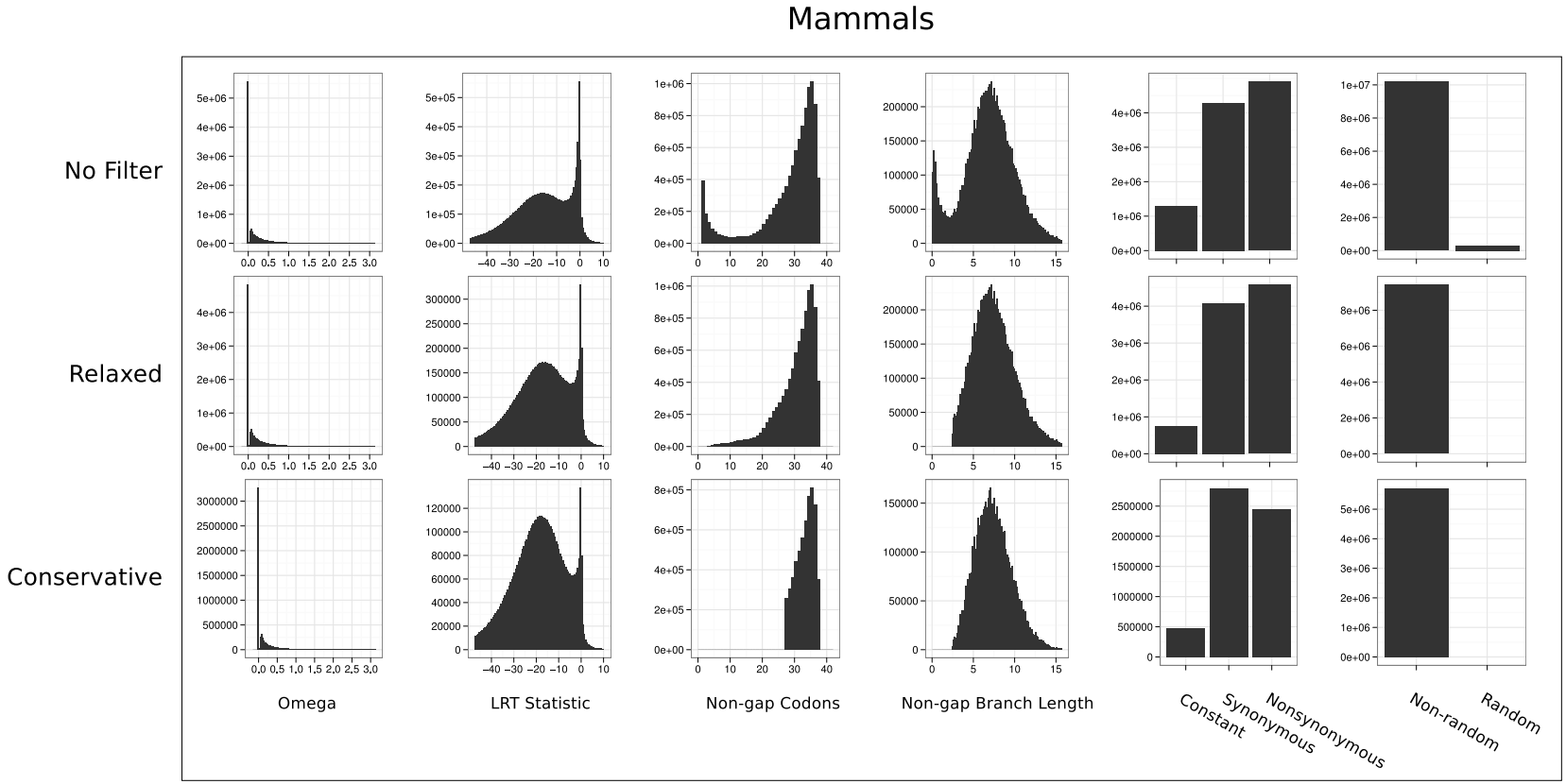


Figure 4.6: Distributions of sitewise values for the Mammals species group, showing the raw data (top row) and the result of applying the relaxed (middle row) and conservative (bottom row) filters.

evolution than those that do not [Kosiol *et al.*, 2007]. However, the very low proportion of synonymous sites requiring nonzero nonsynonymous substitution rates suggested that the impact of these effects on the current dataset was minimal; this is likely due to the relatively short branch lengths separating the nodes of the mammalian tree, making it less probable that codons with multiple substitutions (whether the result of simultaneous multiple nucleotide changes or successive single changes) would be observed [Kosiol *et al.*, 2007].

The distributions of the Non-gap Codons and Non-gap Branch Length values in the unfiltered row of Figure 4.6 showed that most alignment columns contained sequence data from many species (with Non-gap Codons peaking at 36 and Non-gap Branch Length peaking at around 8 substitutions per site), but a noticeable portion contained only a few non-gap sequences. If the alignment columns with low non-gap codon counts represented accurate evolutionary histories, then the observed excess of highly-gapped sites might be taken as an indication that insertion events in terminal lineages or recent ancestral lineages were prominent enough throughout mammalian evolution to leave a noticeable signature of sites with very low non-gap codon counts. Given the many possible sources of error in the annotation and alignment of these sequences, however, a more likely scenario was that sites with low codon counts and low branch lengths came from stretches of sequence which only exist in a few species as a result of annotation or alignment error. As a result, these sites might be expected to show a higher probability of being nonhomologous and showing spurious signals of positive selection. This would make such sites prime candidates for filtering out prior to analysis.

To test the hypothesis that sites with few non-gap sequences would be less reliable for analysis than other sites, I split the sitewise estimates from the Mammals and Primates groups into ten equally-sized bins of non-gap branch length. Sites within each bin were summarized by calculating the percentage of sites with ω_{ML} less than or greater than 1, as well as the percentage of sites showing evidence for positive selection at a nominal 5% false positive rate (FPR), hereafter referred to as PSCs. The results of this analysis are presented in Table 4.2. The lowest bin was a clear outlier in the Mammals data, with nearly 17% of sites having $\omega_{ML} > 1$ and 2% of sites being PSCs. The other 9 bins with greater non-gap branch lengths showed fewer sites with $\omega > 1$ and less evidence for positive selection; within those 9 bins, a pattern of gradual increase in the proportion of sites with $\omega_{ML} > 1$ and PSCs was observed at progressively higher non-gap branch lengths. The increase in evidence for positive selection with increasing non-gap branch length could be

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

Table 4.2: Proportions of sites with evidence for purifying and positive selection in the Mammalia and Primates datasets broken down by non-gap branch length. Sites were separated into 10 equally-sized bins of non-gap branch length and the sites within each bin were summarized by the 25^{th}, 50^{th} and 75^{th} percentiles of non-gap branch length (BL) and non-gap codons, the percentage of sites with ω estimated below or above 1, and the percentage of sites classified as positively-selected codons (PSCs) at a nominal 5% FPR. BL–branch length; PSC–positively selected codons.

explained by genes with higher overall dN/dS ratios (and presumably more PSCs) having higher branch lengths due to the increased rate of nonsynonymous substitution. Overall, the pattern observed for the Mammals data was consistent with the prediction that sites with few non-gap sequences were not consistent with the general pattern of sitewise data. In terms of choosing an appropriate threshold on which to filter, Table 4.2 indicated that removing sites with the lowest 10% of non-gap branch length would remove most of the apparently anomalous sites.

Table 4.2 showed a similar trend for the Primates dataset, although the distinction between the lowest bin and the rest of the dataset was less obvious. The percentage of PSCs in the lowest decile was only slightly higher than in the next-highest decile, and the proportion of sites with $\omega_{ML} > 1$ was lower than in all other bins. Thus, despite

weaker evidence in the Primates data for the anomalous nature of sites with few non-gap sequences, it still appeared that filtering sites in the bottom 10% bin would improve the overall quality and consistency of the data.

Turning back to the bulk distributions in Figure **??**, two other criteria were used to target sites for removal before analysis. First, the rightmost panels of Figures 4.6 and 4.5 depict a small set of sites designated as "random". These sites were flagged by SLR as having a site pattern not significantly different from random [Massingham & Goldman, 2005b], and they were also targeted for removal before analysis of the global distribution. Second, all sites with fewer than four non-gap sequences were removed. This was done to avoid analyzing sites with very few sequences which were not within the bottom 10% of sites by non-gap branchlength.

At this point, all of the criteria used to define the relaxed filter have been described: non-gap branch lengths, random flags, and the number of non-gap sequences at each site. The middle rows of Figures 4.5 and 4.6 show the summary distributions resulting from applying the relaxed filter to the Mammals and Primates sitewise data.

Three additional criteria were added to create the more conservative filtered dataset. First, the threshold on non-gap sequence counts was increased: all sites with a non-gap codon count below 75% of the maximum non-gap count for that species group were removed. Second, sites and genes containing windows of clsutered nonsynonymous substitutions (as identified in Section 4.3) were removed: all sites overlapping the 23,116 15-codon windows with excess nonsynonymous substitutions (using the 99.9% quantile based definition of excess substitutions from Section 4.3) were masked out, and 819 genes with greater than 10% of sites covered by windows with excess nonsynonymous substitutions were removed. Finally, the 3,333 genes which contained more than 2 sets of putative paralogs were excluded.

As with the relaxed filter the result of applying the conservative filter to the Primates and Mammals datasets is shown in the bottom rows of Figures 4.5 and 4.6. Comparing between the distributions in the three rows of Figure 4.6, the most prominent effect of the two filters on the bulk distributions in was the removal of the excess of sites with low non-gap branch lengths and non-gap codon counts. The distributions of ω_{ML} estimates and LRT statistics were qualitatively unchanged unchanged, indicating that the overall characteristics of the dataset were not significantly altered by this filter.

Tables 4.3 and 4.4 provide a quantitative summary of the Mammals and Primates datasets before and after applying the two filters. Also shown is the subset of sites over-

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

The columns in Table 4.3 show various summary statistics of each sitewise dataset including the number of sites, the proportions of different site patterns, and the proportions of purifying and positive selection based on ω_{ML} estimates from SLR. Table 4.4 provides the number and proportion of identified PSCs (columns under the heading "Positively Selected Sites") as well as the breakdown of sites into purifying, neutral, and positively-selected at two different FPR thresholds (columns under the headings "$P_{\chi^2_1} < 0.1$" and "$P_{\chi^2_1} < 0.05$").

These views make clear the impact of extensive filtering on the genome-wide levels of positive and purifying selection observed in the data. The unfiltered data from the Primates group contained 9.07% of sites with $\omega_{ML} > 1$, and 0.59% of sites were PSCs at a nominal 5% FPR; the evidence for positive selection was reduced in the conservatively-filtered data, showing 7.87% sites with $\omega_{ML} > 1$ and 0.41% PSCs. An even stronger effect of filtering was seen for the Mammals data, with $\omega_{ML} > 1.5$ being reduced from 5.71 to 2.73 between the unfiltered and conservatively-filtered datasets, and the percentage of PSCs reduced from 0.72% to 0.35%. The rows representing two sets of sites which were removed during the conservative filtering process showed higher signals of positive selection than the unfiltered data, suggesting that these two filtering steps were at least somewhat effective in removing potentially anomalous or untrustworthy sites from the dataset. For sites removed from being within clusters of nonsynonymous substitutions, the enrichment for signals of positive selection was clear: in Primates, 18.28% of sites yielded $\omega_{ML} > 1$, and 1.47% of sites were PSCs at a 5% FPR threshold, more than three times the proportion of PSCs seen in the conservatively-filtered dataset. Sites removed as a result of being within genes containing recent duplications showed less of a signal for positive selection, but the proportions of PSCs and sites with $\omega_{ML} > 1$ were still above those seen in either the relaxed or conservatively filtered datasets for both Mammals and Primates. Thus, genes that have experienced many recent duplications in mammals contained higher levels of positive selection even after the most-divergent paralogous copies were removed.

			Site Pattern, %			Med.	Nongap BL			ω_{ML}		ω_{ML} Below / Above, %			
Name	Filter	Sites	Const.	Syn.	Nsyn.	Codons	Med.	Mean	SD	Mean	SD	< 0.5	< 1	> 1	> 1.5
Primates	None	9.86e+06	59.78	24.08	16.14	9	0.74	0.86	1.19	0.25	0.77	86.03	90.97	9.03	5.86
	Relaxed	8.81e+06	57.34	25.51	17.15	9	0.78	0.93	1.24	0.25	0.74	85.28	90.76	9.24	5.78
	Conservative	6.35e+06	59.31	26.24	14.45	9	0.76	0.82	0.51	0.21	0.67	87.28	92.13	7.87	4.81
	Pfam	3.53e+06	59.45	27.69	12.86	9	0.79	0.97	1.54	0.16	0.58	89.73	94.10	5.90	3.52
	(Clusters)	9.60e+05	44.48	21.57	33.95	9	1.17	1.42	1.44	0.51	1.02	71.68	81.70	18.30	12.01
	(Paralogs)	1.61e+06	52.59	25.12	22.29	9	0.87	1.31	2.61	0.30	0.79	82.64	89.27	10.73	6.73
Mammals	None	1.05e+07	12.41	40.89	46.70	32	6.95	6.95	4.12	0.24	0.58	86.25	94.32	5.68	2.92
	Relaxed	9.46e+06	8.03	43.44	48.53	33	7.30	7.62	3.79	0.19	0.38	87.53	95.77	4.23	1.61
	Conservative	5.78e+06	8.29	48.98	42.72	34	7.28	7.48	2.42	0.15	0.31	90.86	97.27	2.73	0.91
	Pfam	3.73e+06	9.04	52.11	38.85	33	7.44	7.84	4.44	0.12	0.29	92.93	97.83	2.17	0.80
	(Clusters)	9.75e+05	4.30	22.07	73.63	29	9.51	9.69	4.85	0.40	0.56	71.90	88.81	11.19	4.71
	(Paralogs)	1.79e+06	7.18	38.66	54.16	32	7.68	8.30	6.84	0.22	0.42	85.33	94.85	5.15	2.03

Table 4.3: Summary statistics of sitewise estimates for Mammals and Primates data with various filters applied. Rows labeled (Clusters) and (Paralogs) contain sites excluded by the Conservative filter. Columns under the "ω_{ML} Below / Above" heading measure the percentage of sites with ω_{ML} below or above the indicated value. Med.—median, Const.—constant, Syn.—synonymous, Nsyn.—nonsynonymous, BL—branch length.

		Positively Selected Sites (%)								$P_{\chi^2_1} < 0.1$, %			$P_{\chi^2_1} < 0.05$, %		
Name	Filter	$P_{\chi^2_1} < 0.1$		$P_{\chi^2_1} < 0.05$		$P_{\chi^2_1} < 0.01$		FDR< 0.05		Neg.	Neut.	Pos.	Neg.	Neut.	Pos.
Primates	None	99487	(1.01)	58188	(0.59)	18328	(0.19)	244	(0.002)	29.89	69.11	1.01	14.26	85.15	0.59
	Relaxed	83239	(0.95)	48176	(0.55)	14704	(0.17)	104	(0.001)	33.20	65.86	0.95	15.91	83.54	0.55
	Conservative	46140	(0.73)	26261	(0.41)	7801	(0.12)	50	(0.001)	33.85	65.42	0.73	15.87	83.71	0.41
	Pfam	19563	(0.55)	11353	(0.32)	3561	(0.10)	31	(0.001)	38.89	60.55	0.55	19.88	79.80	0.32
	(Clusters)	23543	(2.45)	14173	(1.48)	4663	(0.49)	40	(0.004)	30.52	67.03	2.45	16.36	82.16	1.48
	(Paralogs)	18821	(1.17)	11058	(0.69)	3437	(0.21)	30	(0.002)	33.62	65.21	1.17	17.45	81.86	0.69
Mammals	None	114105	(1.08)	75536	(0.72)	30735	(0.29)	2063	(0.020)	80.30	18.62	1.08	77.13	22.15	0.72
	Relaxed	76450	(0.81)	52166	(0.55)	23382	(0.25)	1890	(0.020)	86.57	12.62	0.81	83.95	15.50	0.55
	Conservative	29432	(0.51)	20150	(0.35)	9140	(0.16)	795	(0.014)	90.61	8.88	0.51	88.54	11.11	0.35
	Pfam	17253	(0.46)	12320	(0.33)	6159	(0.17)	706	(0.019)	92.69	6.85	0.46	91.01	8.66	0.33
	(Clusters)	23443	(2.40)	16355	(1.68)	7592	(0.78)	656	(0.067)	70.08	27.51	2.40	65.71	32.61	1.68
	(Paralogs)	17735	(0.99)	12174	(0.68)	5471	(0.31)	426	(0.024)	83.92	15.08	0.99	80.88	18.44	0.68

Table 4.4: Proportions of sites subject to positive, purifying and neutral selection at various LRT$_{SLR}$ thresholds for Mammals and Primates data with varous filters applied. The Benjamini-Hochberg method [Benjamini & Hochberg, 1995] was used to identify the LRT$_{SLR}$ threshold at which FDR$<$0.05. For columns under the headings "$P_{\chi^2_1} < 0.1$, %" and "$P_{\chi^2_1} < 0.05$, %", Pos. and Neg. are the percentage of sites with significant evidence for positive and negative selection, respectively, and Neut. is the percentage of "neutral" sites not showing significant evidence for non-neutral selection.

The global distribution of sitewise selective pressures in mammals

To produce high-confidence sitewise estimates across the ten chosen species groups, sitewise data from each species group were processed with the conservative filter as described above. The resulting global distributions of site patterns, sitewise ω_{ML} estimates, and 95% confidence intervals are shown in Figure 4.7. The left panel in each row plots the number of sites with constant, synonymous, and nonsynonymous patterns; all sites with $\omega_{ML} = 0$ had constant or synonymous patterns, and all sites with $\omega_{ML} > 0$ had nonsynonymous patterns. The right panel in each row shows the distributions of ω_{ML} for sites which contained a nonsynonymous site pattern.

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

The site pattern counts in Figure 4.7 showed that the branch length of each species group had a strong effect on the overall composition of the sitewise data. Groups covering little branch length, such as Primates and Atlantogenata, contained mostly constant sites, while groups covering a large amount of branch length, such as Eutheria and Mammals, contained few constant sites and roughly equal proportions of sites with synonymous and nonsynonymous site patterns. Comparing the Glires and Mammals data with their corresponding "sparse" datasets confirmed that this trend was largely due to branch length as opposed to biological factors: the Sparse Glires data, for example, yielded a smaller proportion of nonsynonymous sites and a greater proportion of constant sites than the Glires data (17.41% versus 24.08% for nonsynonymous sites, 44.21% versus 33.98% for constant sites).

The distributions of ω_{ML} estimates are shown in Figure **??** as a series of histograms showing the ω_{ML} density (for nonzero values of ω_{ML} only) and a series of solid lines showing the cumulative ω_{ML} density (representing all values); the lower and upper dashed lines show the cumulative density of the lower and upper 95% confidence interval resulting from each sitewise estimate. From these distributions, it is clear that the majority of protein-coding sites have evolved under purifying selection in mammals, a fact which is most easily seen in the larger species groups. The Mammalia group, which contained only a small proportion of uninformative constant sites (8.17%), showed a maximum density of nonzero ω_{ML} estimates at $\omega \approx 0.1$, and the vast majority of sites showed some evidence of purifying selection with ω_{ML} estimates below 1. The height of the ω_{ML} cumulative distribution at $\omega = 1$ corresponds to the proportion of sites with some evidence for purifying selection.

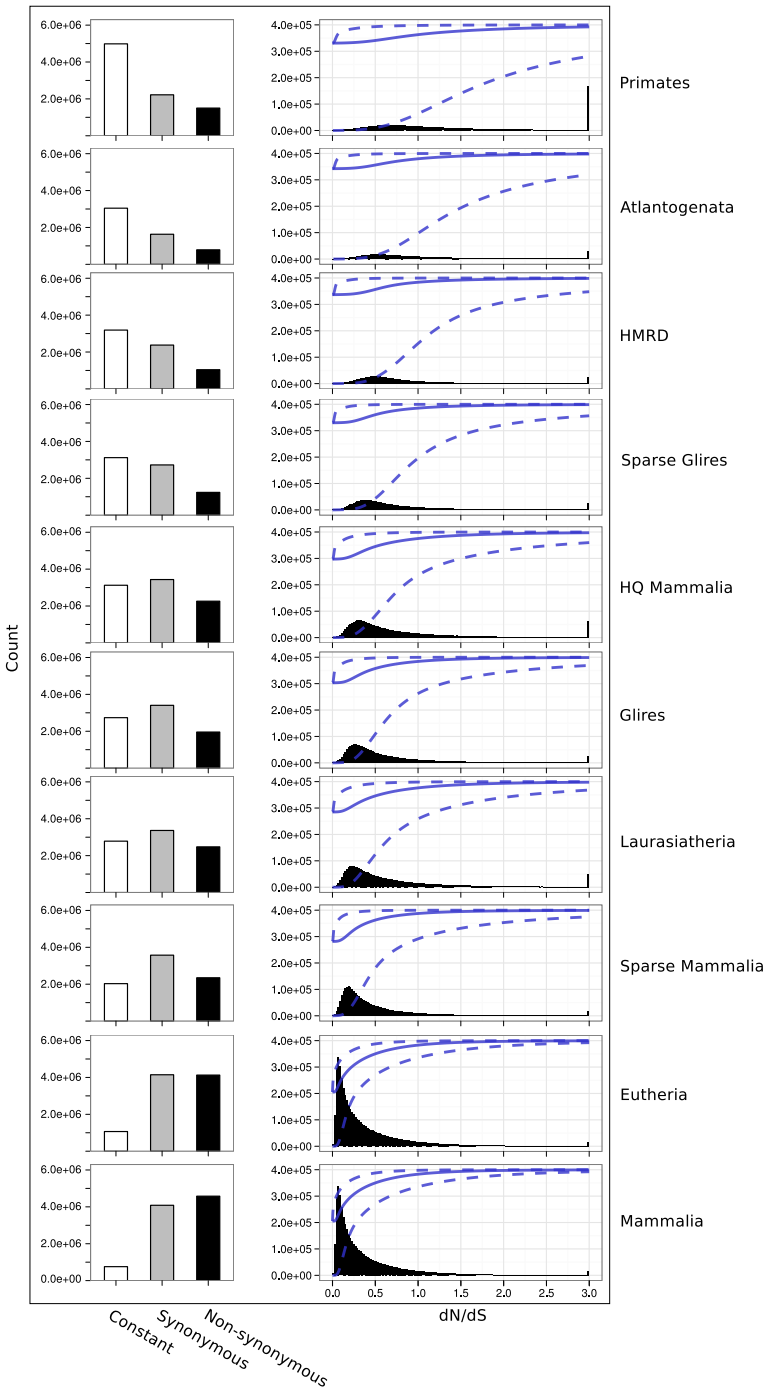


Figure 4.7: Global distributions of site patterns and ω estimates for ten species groups. Left panels: bars represent the number of sites showing constant, synonymous, and non-synonymous patterns. Note, the y-axis is held constant between rows. Right panels: bars represent histograms of ω_{ML} estimates for sites where $\omega_{ML} > 0$. Sites with $\omega_{ML} > 3$ are counted in the bin at $\omega_{ML} = 3$. A solid line is drawn showing the cumulative distribution of ω_{ML}, and dashed lines are drawn above and below the solid line showing the cumulative distributions of the lower and upper bounds, respectively, of the 95% confidence interval associated with each sitewise estimate.

The nonzero ω_{ML} values were more evenly spread in the other species groups: Glires contained a maximum nonzero ω_{ML} density at around $\omega \approx 0.25$ and Primates at $\omega \approx 0.7$. This upwards shift in nonzero ω_{ML} estimates relative to Mammalia was likely due to the greater proportion of constant and synonymous sites in datasets with lower overall branch lengths: sites which were truly evolving with $0 < \omega < 1$, but where no nonsynonymous or synonymous substitutions were observed, would have their ω_{ML} estimate "pushed" towards zero, presumably causing a concomitant upwards shift in the distribution of the remaining nonzero ω_{ML} values.

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 [2005b], I used a signed version of the LRT statistic (hereafter LRT_{SLR}), formed by negating the LRT statistic for sites where $\omega_{ML} < 1$, as a way to sort sites according to their evidence for purifying and positive selection. Thus, sites with $\text{LRT}_{SLR} < 0$ showed at least some evidence for purifying selection, and sites with $\text{LRT}_{SLR} > 0$ showed at least some evidence for positive selection. It should be noted that the LRT_{SLR} is a measure of the strength of evidence for purifying or positive selection, but not necessarily the actual strength of that selection. For example, an alignment covering a very large branch length might yield a strongly negative LRT_{SLR} for a site with ω_{ML} only moderately below 1, because the evidence for purifying selection at that site was highly statistically significant; on the other hand, a strongly-purifying site in an alignment covering less branch length might produce a much less-negative LRT_{SLR}, even with an estimated ω_{ML} near zero.

To further explore this point, Figure 4.8A shows the relationship between LRT_{SLR}, ω_{ML} and the $\text{CI}_{95\%}$ width for sites from the Mammals group. The left panel, comparing the LRT_{SLR} to nonzero ω_{ML} estimates, shows that the two values are highly correlated, with the greatest number of low ω_{ML} estimates occurring at sites with strongly negative LRT_{SLR}s. Correspondingly, the middle panel shows an even stronger relationship between the LRT_{SLR} magnitude and the $\text{CI}_{95\%}$ width, with the tightest confidence intervals at sites with very strong evidence for purifying selection. The rightmost panel compares the

ω_{ML} of each site with the width of its $CI_{95\%}$, revealing a more linear and diffuse positive relationship between ω_{ML} and the size of the $CI_{95\%}$. The equivalent plots for Primates, shown in Figure 4.8B, reveal similar patterns, but with generally less-negative LRT_{SLR} values, higher ω_{ML}, and larger $CI_{95\%}$. These differences highlight the impact of branch length on the amount of confidence with which ω can be estimated on a per-site basis. The low branch length of the Primates clade rarely yields ω_{ML} estimates with $CI_{95\%}$ intervals smaller than 1, while the bulk of sites from the Mammalia dataset have relatively small $CI_{95\%}$s. Thus, the distribution of ω_{ML} estimates from datasets with low branch lengths (e.g., the histogram densities seen in Figure 4.7) should be interpreted with caution, as any comparison between ω_{ML} from different sites or datasets may be more sensitive to the amount of statistical confidence placed on each estimate than to any meaningful biological difference between the two sets of data.

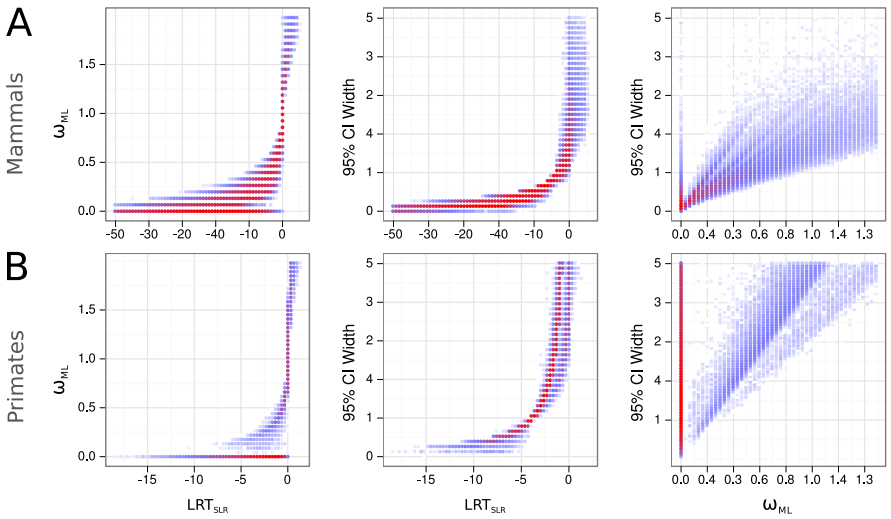


Figure 4.8: The relationship between LRT_{SLR}, ω_{ML}, and $CI_{95\%}$ width in (A) Mammalia and (B) Primates datasets. Each point represents the binned density of sites; no points are drawn where no density exists, while blue and red points are drawn at areas of low and high density, respectively. The left panel shows sites where $\omega_{ML} < 0$, the middle panel shows all sites, and the right panel shows sites where $0 < \omega_{ML} < 1$. Note the change in x-axis scales between plots in (A) and (B), reflecting the paucity of sites in Primates with strong evidence (LRT_{SLR} <-12) for purifying selection.

			Site Pattern, %			Med.	Nongap BL			ω_{ML}		ω_{ML} Below / Above, %			
Name	Filter	Sites	Const.	Syn.	Nsyn.	Codons	Med.	Mean	SD	Mean	SD	< 0.5	< 1	> 1	> 1.5
Primates	Conservative	6.22e+06	59.29	26.26	14.45	9	0.76	0.82	0.50	0.19	0.57	87.27	92.13	7.87	4.80
Atlantogenata		4.07e+06	57.23	30.07	12.70	5	0.94	1.01	0.37	0.13	0.41	90.01	95.38	4.62	2.29
HMRD		5.02e+06	49.75	36.41	13.84	4	0.96	1.01	0.36	0.12	0.37	90.06	96.37	3.63	1.73
Sparse Glires		5.35e+06	45.36	39.11	15.53	5	1.24	1.32	0.68	0.12	0.36	90.91	96.71	3.29	1.51
HQ Mammals		6.38e+06	37.09	40.68	22.23	8	1.46	1.55	0.64	0.17	0.43	88.31	94.97	5.03	2.50
Glires		5.79e+06	34.70	43.68	21.62	7	1.77	1.87	0.84	0.13	0.36	90.54	96.53	3.47	1.50
Laurasiatheria		5.35e+06	33.36	41.99	24.66	11	2.03	2.15	0.87	0.16	0.41	88.88	95.36	4.64	2.22
Sparse Mammals		5.65e+06	25.81	46.75	27.44	6	2.55	2.75	1.45	0.13	0.32	91.65	97.28	2.72	1.10
Eutheria		5.72e+06	11.96	49.78	38.26	32	5.80	6.01	1.96	0.15	0.33	90.17	96.76	3.24	1.18
Mammals		5.72e+06	8.30	48.98	42.72	34	7.28	7.48	2.42	0.15	0.30	90.86	97.27	2.73	0.91

Table 4.5: Summary statistics of sitewise estimates for all species groups with the conservative filter applied. Columns under the "ω_{ML} Below / Above" heading measure the percentage of sites with ω_{ML} below or above the indicated value. Med.—median, Const.—constant, Syn.—synonymous, Nsyn.—nonsynonymous, BL—branch length.

		Positively Selected Sites (%)								$P_{\chi^2_1} < 0.1$, %			$P_{\chi^2_1} < 0.05$, %		
Name	Filter	$P_{\chi^2_1} < 0.1$		$P_{\chi^2_1} < 0.05$		$P_{\chi^2_1} < 0.01$		FDR< 0.05		Neg.	Neut.	Pos.	Neg.	Neut.	Pos.
Primates	Conservative	45179	(0.73)	25710	(0.41)	7661	(0.12)	50	(0.001)	33.89	65.38	0.73	15.88	83.70	0.41
Atlantogenata		8143	(0.20)	3852	(0.09)	757	(0.02)	0	(0.000)	46.96	52.84	0.20	23.75	76.15	0.09
HMRD		6538	(0.13)	3040	(0.06)	534	(0.01)	0	(0.000)	63.74	36.13	0.13	37.42	62.52	0.06
Sparse Glires		7233	(0.14)	3316	(0.06)	644	(0.01)	0	(0.000)	70.34	29.52	0.14	49.07	50.87	0.06
HQ Mammals		29344	(0.46)	16374	(0.26)	4548	(0.07)	0	(0.000)	74.87	24.67	0.46	61.55	38.19	0.26
Glires		11155	(0.19)	5581	(0.10)	1221	(0.02)	0	(0.000)	78.93	20.88	0.19	67.92	31.98	0.10
Laurasiatheria		29058	(0.54)	17617	(0.33)	5944	(0.11)	41	(0.001)	78.31	21.15	0.54	68.74	30.93	0.33
Sparse Mammals		7953	(0.14)	3913	(0.07)	857	(0.02)	0	(0.000)	81.99	17.87	0.14	75.28	24.65	0.07
Eutheria		35270	(0.62)	24234	(0.42)	11006	(0.19)	999	(0.017)	89.00	10.38	0.62	86.54	13.04	0.42
Mammals		29075	(0.51)	19900	(0.35)	9025	(0.16)	781	(0.014)	90.61	8.88	0.51	88.54	11.11	0.35

Table 4.6: Proportions of sites subject to positive, purifying and neutral selection at various LRT_{SLR} thresholds. The Benjamini-Hochberg method [Benjamini & Hochberg, 1995] was used to identify the LRT_{SLR} threshold at which FDR<0.05. For columns under the headings "$P_{\chi^2_1} < 0.1$, %" and "$P_{\chi^2_1} < 0.05$, %", Pos. and Neg. are the percentage of sites with significant evidence for positive and negative selection, respectively, and Neut. is the percentage of "neutral" sites not showing significant evidence for non-neutral selection.

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

Tables 4.5 and 4.6 provide summaries of the sitewise estimates obtained for each of the 10 mammalian species groups, showing the same values provided earlier in Tables 4.3 and 4.4 for the different filters.

Table 4.6 presents the proportions of PSCs identified at a variety of LRT_{SLR} thresholds, demonstrating that anywhere between 0.01% to 0.73% of sites could be confidently identified as under positive selection in mammals at nominal FPR thresholds between 1% and 10%. Interestingly, however, different species groups yielded strikingly different estimates of the proportion of PSCs. At a 5% FPR threshold, the Primates, HQ Mammals, Laurasiatheria, Eutheria, and Mammals groups produced broadly comparable proportions of positively-selected sites, ranging from 0.33% to 0.42%. The proportions of PSCs in these groups were higher using a 10% FPR threshold (ranging from 0.46% to 0.73%) and lower using a 1% FPR threshold (ranging from 0.07% to 0.19%). When the FDR was controlled using the Benjamini-Hochberg method, however, far fewer PSCs were identified. Only the Eutheria and Mammalia groups yielded a substantial number of positively-selected sites at this level of control; the Primates and Laurasiatheria data yielded non-zero numbers of PSCs as well, but these species groups were likely limited in their power to yield positively-selected sites after FDR control due to their lower total branch lengths.

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

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

In Mammalia, the breakdown of sites into positive, negative and neutral categories at 10% and 5% significance thresholds produced a pattern similar to that seen in the ω_{ML} distributions from Figure 4.7, with a large amount of purifying constraint (83.87% of sites at 5% FPR), a small proportion of neutrally-evolving sites (15.57%), and a diminishing number of positively-selected sites (0.55%). As expected given the use of a fixed LRT_{SLR} threshold to identify purifying sites, the fraction of sites confidently identified as under purifying selection showed a strong dependency on the branch length of the species set, with a much higher power in Mammalia than in Primates to confidently detect purifying selection (83.87% vs. 15.97%).

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

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

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

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

Thus, the relaxed constraint argument tempers the interesting observation of strong differences in the numbers of PSCs between different species groups. A lower historical effective population size for the Primates and Laurasiatheria species groups may explain some of the increase in the number of PSCs detected, even in the absence of true variation in the prevalence of positive selection between the species groups investigated here. Still, the argument may be made that statistical methods for controlling error rates, such as the Benjamini-Hochberg method for FDR control used to identify PSCs at an expected FDR< 0.05 in Table 4.6 [Benjamini & Hochberg, 1995], should account for the potential

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

4.5 Conclusions

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

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

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

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

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

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

[...]

Chapter 5

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

5.1 Introduction

This chapter describes the use of sitewise data to identify trends in the evolution of protein-coding genes and domains, focusing on the detection of PSGs. I will first develop a number of methods for using sitewise estimates to identify signals of positive selection within genes and domains and apply these methods to the sitewise data generated in Chapter 4. Next, to provide a higher-level interpretation of these results I will use functional gene annotations to identify categories enriched for genes with evidence of positive selection in different species groups. Lastly, I will place these results within the context of the literature by circectly comparing the sets of PSGs identified by this and previously-published studies.

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

Analysis Consortium, 2007]; analyses in other mammalian genomes have revealed largely similar patterns [Kosiol *et al.*, 2008a; Li *et al.*, 2009]. To explain the increased $d\mathrm{N}/d\mathrm{S}$ values observed within PSGs, three distinct evolutionary dynamics have commonly been invoked: an evolutionary arms race between host and parasite interacting genes [Yang, 2005; Meyerson & Sawyer, 2011], sexual selection or genetic conflict between the sexes [Wyckoff *et al.*, 2000; Clark & Civetta, 2000], and functional adaptation following gene duplication [Zhang *et al.*, 2002].

As the power of phylogenetic analysis using codon models depends strongly on the amount of branch length encompassed by the species being compared [Anisimova *et al.*, 2001a, 2002a], there was some reason to believe *a priori* that the detection of PSGs using mammalian alignments incorporating low-coverage genomes would be more powerful than in previous whole-genome analyses, which typically included 12 or fewer species across mammals and lower total branch length [ELLEGREN, 2008]. However, differences in the specific models used to detect positive selection are expected to affect the sensitivities of one study compared to another [Anisimova & Kosiol, 2009], so the set of genes identified using the current methodology would necessarily be expected to be a superset of those identified in previous studies. Most large-scale studies have used the branch-site test for positive selection [Zhang *et al.*, 2005], while the results described in this chapter were generated using SLR. I showed in Chapter 2 that SLR has similar power to the site-based test implemented in PAML for detecting sitewise positive selection, but no analysis has yet compared the differences in PSGs identified by site-specific and branch-site methods on a large scale. For this reason, I hoped that a quantitative comparison between PSGs identified using the current methodology and those found in previously-published studies may improve our understanding of how similar or different the PSGs identified by different methods can be.

5.2 Combining sitewise estimates to identify positive selection

In Chapter 5 I covered the generation and analysis of several highly filtered sets of genome-wide sitewise selective pressures within different groups of mammalian species. These sitewise estimates were used to characterize the global distribution of evolutionary constraint and to compare overall levels of purifying and positive selection between groups of mammalian species. The focus on individual codons as an evolutionary unit of investigation

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

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

A major issue in combining sitewise estimates to identify PSGs is that of correcting for performing multiple sitewise tests per gene. The SLR method performs an independent statistical test at each site, producing a sitewise statistic which can be compared to a χ^2_1 distribution to yield a p-value representing the strength of evidence against strict neutral evolution [Massingham & Goldman, 2005b]. When combining these p-values to decide whether a gene contains significant evidence for positive seletcion, one must take into account the number of tests performed. For example, a 100-codon gene evolving under the null model ($\omega = 1$) would be expected to produce 5 sites with p-values at a nominal FPR of 0.05; correspondingly, the chance that at least one site within the gene would have $p < 0.05$ is 99.4%. This is calculated as the complement of the probability that no sites out of n have $p < x$, which is $(1 - x)^n$. Thus, if the set of genes containing at least one site with nominal

$p < 0.05$ were called PSGs, nearly all genes evolving under the true null model would be selected. In contrast, the LRTs for positive selection implemented in PAML only perform one statistical test per gene and do not suffer from the same multiple testing problem. Clearly, some procedure for correcting or combining the results from multiple tests must be applied in order to identify PSGs using sitewise data in a statistically controlled manner.

I tested 3 types of methods which are capable of correcting for multiple sitewise tests within genes to identify PSGs: first, adjusting significance thresholds to control the family-wise error rate (FWER); second, combining p-values from multiple tests to produce a single p-value summarizing the overall evidence against the null hypothesis; third, estimating empirical gene-wise p-values based on the genome-wide distribution of sitewise estimates. Each approach makes different use of the sitewise data from each gene to identify a set of significant PSGs and thus had the potential to yield a unique set of PSGs. The remainder of this section provides some background on each approach and describes how it was applied to the current problem.

Controlling the FWER

The FWER is defined as the probability, for a given set of tests performed, of one or more tests producing a false positive result. In the example of a 100-codon gene evolving under the null model, the FWER at a nominal p-value of 0.05 was 0.994. Assuming an appropriate uniform null distribution of p-values and independence between tests, the Šidàk equation (to which the more popular Bonferroni correction is an easily computed approximation) identifies the p-value threshold x which is necessary to control the FWER at the desired level α. The FWER expected for a family of n tests thresholded at a nominal p-value of x is $\alpha = 1 - (1 - x)^n$, so the p-value threshold necessary to control for a desired FWER can be found by rearranging the equation: $x = 1 - (1 - \alpha)^{1/n}$. A similar but more powerful approach to controlling the FWER is the step-up method from Hochberg; this method is implemented internally by SLR for reporting the number of positively- and negatively-selected sites after multiple testing correction [Hochberg, 1988; Massingham & Goldman, 2005b].

I used the *p.adjust* method from the R statistical project to apply the Hochberg procedure to the set of sitewise p-values from each gene; this produced a new set of p-values representing the FWER expected if all sites with p-values equally or more extreme than the given site were called significant. The overall p-value for each gene was taken as the minimum FWER-adjusted p-value across all sites.

One weakness of this approach is that the evidence for assigning positive selection comes only from the site with the most extreme LRT_{SLR}, ignoring any signal of positive selection from sites with weaker p-values. As it has been previously observed that PSGs often contain multiple sites subject to similar elevated dN/dS levels [Sawyer *et al.*, 2005; Kosiol *et al.*, 2008a], the gene-wise p-values resulting from the above approach to controlling the FWER may lack power to detect positive selection in genes with many sites showing moderate to strong evidence for positive selection. The next two methods described are both sensitive to more than just the most significant site, making them potentially more powerful for identifying PSG in a statistically-controlled manner.

Combining p-values

The second approach to multiple testing directly addresses this problem by combining p-values from a series of independent tests, producing an overall p-value for the null hypothesis given the set of tests performed. The motivation behind such methods is that moderately significant results from independent tests of a common null hypothesis should be considered as good or better evidence than one strongly significant test. Many different specific techniques of this type have been discussed in the literature (see Cousins [2007] for an extensive annotated bibliography). Two of the most popular methods are Fisher's combined probability test and Stouffer's method (Fisher, 1932; Stouffer *et al.*, 1949; reviewed in Whitlock, 2005). Briefly, each method combines p-values from independent tests in some way (Fisher's test takes the product of all p-values, while Stouffer's method transforms p-values into normal quantiles and sums the resulting z-scores) and compares the resulting statistic to the expected distribution given a null distribution of the same number of input p-values. Comparisons of both tests suggested that they provide similar power overall, but Stouffer's method generally yields smaller p-values when the input p-values are more similar and Fisher's test yields smaller p-values when the input p-values vary widely [Darlington & Hayes, 2000]. When the distribution of input p-values is nonuniform or the number of tests is large, however, performance can be reduced. It has been noted that a relatively small number of large p-values can limit the power of Fisher's test [Zaykin *et al.*, 2002], and the Stouffer method should be equally sensitive to small and large p-values. Since the majority of mammalian protein-coding sites showed moderately strong signals of purifying selection in the global distribution, the distribution of one-sided p-values for positive selection would be heavily weighted towards 1 for the set of sitewise estimates in most genes. As a result, both the Fisher and Stouffer methods were expected to lack

power to identify PSGs, as the dominant signal of purifying selection in most genes would tend to produce non-significant combined p-values for positive selection even when strong evidence for positive selection exists.

The variants of Fisher's and Stouffer's methods which incorporate a truncation step (i.e., including only p-values below a pre-specified threshold to calculate the combined statistic) provided a potentially more powerful approach to combining sitewise p-values within genes [Darlington & Hayes, 2000; Zaykin *et al.*, 2002, 2007]. Zaykin et al. [2002; 2007] showed that the truncated product method (TPM), a truncated version of Fisher's product method, is well-suited for large-scale genomics experiments where the number of tests is large and the standard methods lack power. The authors suggest a truncation threshold of $p < 0.05$ provides a good balance of sensitivity and power, and they note that the method is asymptotically equivalent to Fisher's combined test as the p-value trunctation is increased to 1. Thus, the truncation threshold determines the extent to which the method focuses on more significant test results. The test statistic is calculated as the product of all p-values below the truncation threshold, and in the implementation provided by Zaykin get al. [2002] the significance of the statistic is determined by simulation based on the null model. As an example, for a gene with 100 sites and 5 where $p < 0.05$, the test statistic would be the product of those 5 p-values and its significance would be tested by generating 5,000 replicates under the null model (i.e., uniformly distributed p-values) using the same $p < 0.05$ criterion to calculate the truncated product of p-values.

To explore the behavior of the TPM at various p-value truncation thresholds, I used the implementation provided by Zaykin et al. [2002] to calculate combined p-values at truncation thresholds corresponding to a nominal 5%, 10%, 20%, and 50% sitewise FPRs. I also calculated a combined p-value using Fisher's standard combined method to test the hypothesis that the method lacked power to detect PSGs in protein-coding genes due to the presence of many purifying sites.

Assigning empirical p-values based on the global sitewise distribution

The previous two approaches are fairly generic statistical methods, with formulas whose accuracy depends on the assumption of uniformly-distributed p-values under the null hypothesis. Since the overall distribution of one-tailed p-values from sitewise estimates is far from uniform, however, the large proportion of sites with strong evidence for purifying selection may cause problems when a uniform distribution of p-values is assumed.

This problem is alleviated somewhat by the fact that FWER control mainly uses only the most significant test result to identify a PSGs. The sensitivity of the TPM method to largely non-uniform p-values should also be reduced, as sites with p-values above a certain threshold from the calculation of the combined statistic, thus avoiding undue influence from non-significant p-values. Still, the apparent mismatch between the neutral null model tested by SLR and the large majority of sites evolving under purifying selection suggested that tests based on the theoretical distribution of LRT statistics may be overly conservative. For the confident identification of PSGs this may be desirable, but for the global analysis of functional trends in genes subject to PSGs, a less conservative approach should provide more signal. Given the large set of sitewise estimates available for each species group, the identification of PSGs based on empirical p-values was an attractive alternative approach, with potentially more power to detect genes with significant deviations from the observed genome-wide distribution of LRT_{SLR} statistics within each species group [Noble, 2009].

I implemented a randomization method to assign an empirical p-value to each gene based on the length of the gene and the number of sites with p-values below a certain pre-specified significance threshold. This design shares some characteristics with the TPM method, as the test statistic comes from the subset of sites exceeding a certain significance threshold. The test statistic here, however, was a simple count of the significant sites. To assess the significance of the observed count for a given gene, a set of pseudo-replicate genes (each with the same number of sites as the real gene) was generated by sampling with replacement from the genome-wide set of sitewise estimates. Using the pre-specified significance threshold, the number of significant sites from each replicate was counted. Given n, the number of replicates, and r, the number of replicates with as many or more significant sites than the observed count, the empirical p-value was calculated as $(r+1)/(n+1)$ [North *et al.*, 2002]. This method was applied to each gene using 100,00 replicates; as with the TPM method, the effect of different truncation thresholds was assessed by separately calculating empirical p-values using nominal 0.05%, 1%, 5%, and 10% FPR thresholds.

5.3 Analysis of PSGs identified using sitewise selective pressures

The methods described above was applied to sitewise estimates from each of the 10 species groups 3 levels of sitewise filtering from Chapter 4.

To assess the overall behavior of each method, I first looked at the distribution of p-values for different species sets using the conservatively-filtered sitewise data. Figure 5.1 shows the distribution of gene-wise p-values for 10 species groups using the Hochberg, Fisher, TPM, and empirical methods described above. (Note that for the TPM and empirical methods, only one truncation threshold is shown for simplicity; the distributions of p-values for the other truncation thresholds were qualitatively similar.) As expected, Fisher's product method produced very few p-values below 1, showing little to no power to detect positive selection in any species group. The TPM was slightly more sensitive than Fisher's product method with roughtly 10% of genes yielding p-values below 1 for the Mammals species group; the comparison between Fisher's method and the TPM showed that the truncation slightly increased the sensitivity of the method, but the overall sensitivity remained low with very few genes producing low p-values.

The Hochberg and empirical methods both showed much greater sensitivity and revealed strong differences in the distributions of p-values between species groups. For the Hochberg method, Eutheria and Mammals groups showed a large proportion of p-values in the realm of significance, with roughly 10% of genes having $p < 0.05$. Primates and Laurasiatheria clustered together with the next highest proportion of low p-values (roughly 5% with $p < 0.05$), followed by the HQ Mammals group with roughtly 2% of genes with $p < 0.05$. The other species groups all showed no visible enrichment for low p-values, with a largely uniform distribution of p-values in the range of $0 < p < 1$ and less than 5% of genes with a p-value below 1. The empirical method produced two tight clusters of species groups: the first cluster, with roughly 5-7% of genes with $p < 0.05$, contained Eutheria, Mammals, Primates, Laurasiatheria and HQ Mammals; the second cluster, with roughly 2% of genes having $p < 0.05$, contained the other 5 species groups. Note that the cumulative curve for species groups in the lower cluster levels off at around $p = 0.2$. This leveling off occurred at the maximum p-value given to genes with at least one site below the truncation threshold; the substantial fraction of genes with zero sites below the truncation threshold yielded p-values near to 1.

The major differences between the Hochberg and empirical methods were the tighter

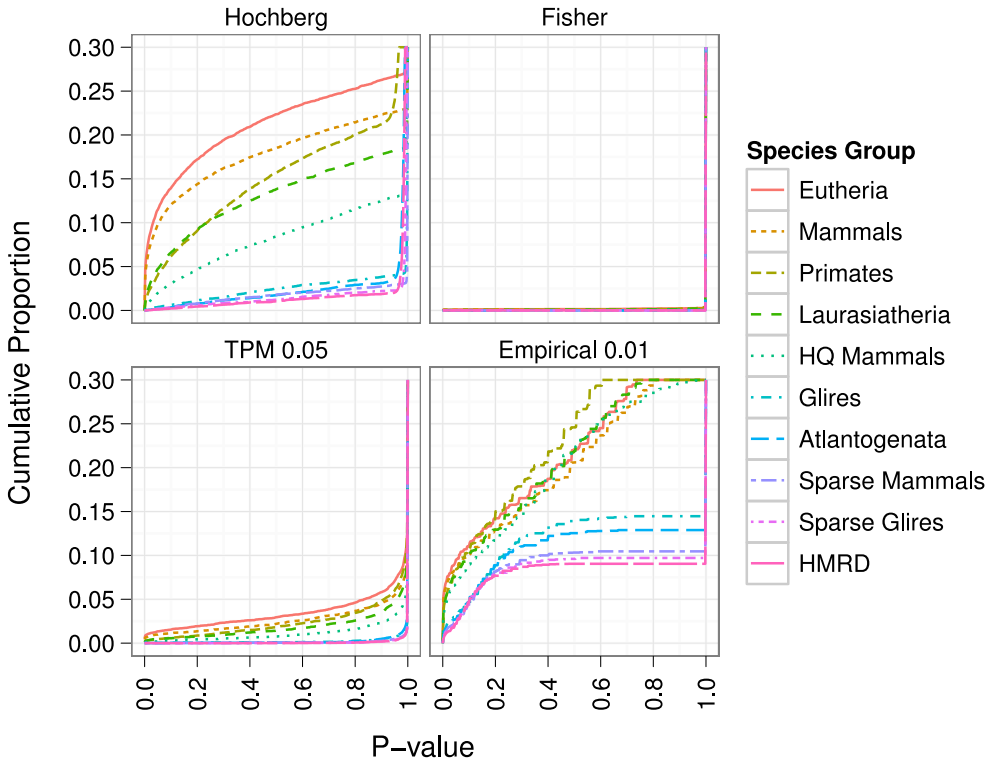


Figure 5.1: Cumulative distributions of gene-wise p-values for positive selection resulting from 4 different methods for combining sitewise estimates within genes. Note that the species groups are listed in order of their cumulative proportions at a p-value of 0.5 for the Hochberg method. To more clearly show the separation between species groups at lower y-values, cumulative proportions above 0.3 are not shown.

clustering of the empirical p-values for the 5 species groups with greater evidence for positive selection and the greater proportion of low empirical p-values for the 5 species groups with less evidence for positive selection. Both differences could be explained by the fact that the Hochberg method assessed significance based on the absolute magnitude of the LRT statistic for positive selection, while the empirical method assessed significance based on the magnitude of evidence for positive selection *relative* to all sitewise estimates for a given species group. This had the effect of increasing the proportion of genes with low p-values for species sets with less branch length (e.g., Primates, Laurasiatheria, and HQ Mammals) or less overall evidence for positive selection (e.g., the five species groups from the lower cluster). As a result, although the overall pattern for each method was somewhat similar, it appeared that the empirical method provided greater sensitivity to detect signals of positive selection while accounting for differences in branch lengths and the background distribution of sitewise selective pressures.

In order to identify a set of confident PSGs for each method it was important to control for multiple testing across genes, since several thousand genes were independently tested for positive selection. This multiple testing issue, resulting from performing many tests across a genome, was distinct from the previously discussed issue of multiple testing across *sites* within a gene. In the case of testing many sites within a gene, the driving question was an overall hypothesis about the gene (e.g., does the gene contain any positively-selected sites or not) and the appropriate error rate to control was the FWER. In contrast, the goal of testing many genes across a genome was not to answer a specific global question (e.g., are *any* genes under positive selection), but rather to identify candidates with a reasonably low number or proportion of likely false positive results. For this purpose, the false discovery rate (FDR), defined as the expected proportion of rejections of the null hypothesis that are false, is a powerful and easily interpreted type of statistical control [Benjamini & Hochberg, 1995]. Thus, the PSGs reported in Table 5.1 are those genes which remained significant after controlling for an expected FDR< 0.1 using the Benjamini Hochberg method [Benjamini & Hochberg, 1995].

Filter	Species Group	Genes	$\bar{\omega}_A$	$\bar{\omega}_G$	Hochberg	*Fis.*	$TPM_{0.5}$	$TPM_{0.2}$	$TPM_{0.1}$	$TPM_{0.05}$	$Emp_{0.005}$	$Emp_{0.01}$	$Emp_{0.05}$	$Emp_{0.1}$
Relaxed	Primates	15197	0.20	0.12	89	16	18	40	65	102	1455	887	1854	2466
	Atlantogenata	11181	0.17	0.11	0	0	0	0	0	1	155	58	430	761
	HMRD	13173	0.14	0.09	0	0	0	0	0	0	103	53	363	795
	Sparse Glires	13625	0.14	0.08	0	0	0	0	0	0	119	38	361	801
	HQ Mammals	15348	0.16	0.10	0	1	2	8	18	36	1096	745	1533	2065
	Glires	14761	0.14	0.08	0	0	0	0	0	1	279	122	624	1157
	Laurasiatheria	15146	0.17	0.10	74	6	12	35	51	80	1248	857	1658	2131
	Sparse Mammals	14988	0.14	0.09	0	0	0	1	1	1	208	100	559	1001
	Eutheria	15856	0.17	0.11	1443	55	57	131	223	363	1733	1360	1987	2352
	Mammals	15946	0.16	0.11	1247	50	49	107	189	312	1605	1233	1898	2308
Conservative	Primates	10683	0.17	0.12	36	1	2	4	6	10	804	459	1141	1523
	Atlantogenata	7714	0.15	0.11	0	0	0	0	0	1	90	37	274	470
	HMRD	9243	0.12	0.09	0	0	0	0	0	0	47	28	201	502
	Sparse Glires	9547	0.12	0.08	0	0	0	0	0	0	72	20	210	507
	HQ Mammals	10905	0.14	0.10	0	0	0	1	5	8	633	345	923	1340
	Glires	10125	0.12	0.08	0	0	0	0	0	0	151	59	409	728
	Laurasiatheria	9667	0.15	0.10	46	2	2	7	10	14	661	400	885	1176
	Sparse Mammals	10237	0.12	0.09	0	0	0	0	0	0	90	54	299	558
	Eutheria	10189	0.14	0.11	611	8	10	15	32	72	823	629	1034	1253
	Mammals	10192	0.14	0.11	474	8	10	13	25	55	728	585	934	1132
Pfam	Primates	9957	0.17	0.12	14	1	1	4	6	14	286	146	434	804
	Atlantogenata	7004	0.15	0.11	0	0	0	0	0	0	9	4	62	165
	HMRD	8408	0.12	0.09	0	0	0	0	0	0	14	6	40	134
	Sparse Glires	8758	0.12	0.08	0	0	0	0	0	0	9	4	32	134
	HQ Mammals	10090	0.14	0.10	0	0	0	1	4	5	205	134	416	636
	Glires	9565	0.12	0.08	0	0	0	0	0	0	23	11	82	230
	Laurasiatheria	9900	0.14	0.10	15	2	2	7	9	15	257	189	420	653
	Sparse Mammals	9624	0.11	0.09	0	0	0	0	0	0	26	12	78	167
	Eutheria	10501	0.14	0.11	352	15	18	32	49	93	454	369	608	777
	Mammals	10587	0.14	0.11	297	14	15	27	42	78	413	349	573	746

Table 5.1: PSGs identified using sitewise data with 3 sitewise filters, 10 species groups and different methods to combine p-values across sites. The columns $\bar{\omega}_A$ and $\bar{\omega}_G$ present the arithmetic and geometric means, respectively, of the gene-wide ω values estimated by SLR. To identify PSGs, only genes with at least 50 sitewise estimates from the given species group and filter were tested. The Benjamini-Hochberg method was used to identify PSGs significant at FDR$<$ 0.1 for all methods. Hoch.—Hochberg's method for FWER control; Fis.—Fisher's combined p-value test; TPM—truncated product method using 50%, 20%, 10% and 5% FPR thresholds; E—empirical p-values using 0.5%, 1%, 5% and 10% FPR thresholds.

Table 5.1 provides a summary of PSGs identified by each method for each sitewise filter and species group. Only genes with at least 50 sitewise estimates were tested, resulting in different numbers of genes for different species groups and sitewise filters. Groups containing fewer species, such as Atlantogenata and HMRD, tended to contain slightly fewer genes than larger groups; this mirrored differences between species groups in the genome-wide number of sitewise estimates seen in Chapter 4 (see Table 4.5).

The pattern of PSG counts was qualitatively similar between different sitewise filters, with fewer PSGs found using more stringent filters. For each combination of species group and method, the greatest number of PSGs was generally found using the relaxed filter, fewer were found using the conservative filter, and the fewest were found using only sites within Pfam domains. This was partially due to the lower total number of genes retained for analysis with the two more conservative filters: for the Mammals species group, 15,946 genes contained at least 50 sites for analysis using the default filter, while the conservative and Pfam filters resulted in only 10,192 and 10,587 genes, respectively. Even after accounting for the different total gene counts in different filters, the number of PSGs as a proportion of all genes was still reduced in the more conservative filters: as an example, for PSGs identified in the Mammals group using Hochberg FWER, 7.8% of genes were PSGs using the relaxed filter, 4.7% using the conservative filter, and 2.8% using only sites within annotated Pfam domains. A similar trend was observed for the other PSGs identification methods, showing that the conservative and Pfam filtered datasets contained progressively lower proportions of genes subject to positive selection. This corresponded well with the pattern seen in Chapter 4 for the prevalence of positively-selected sites.

Comparing between the different methods for identifying PSGs, the Hochberg FWER control and empirical p-value methods were much more sensitive than the Fisher and TPM methods, as expected from the p-value distributions in Figure 5.1. The Fisher method was the most conservative, identifying a vanishingly small number of PSGs in all species groups. Comparing results from the TPM method at different truncation thresholds, the method provded to be increasingly more sensitive as the truncation threshold was decreased; in the Mammals group using the conservative filter, 55 PSGs were identified with a truncation threshold of $p < 0.05$. The empirical method was the least sensitive with a truncation threshold of $p < 0.01$, with increased sensitivity using the lowest threshold ($p < 0.005$) and the two higher thresholds ($p < 0.05$ and $p < 0.1$). The Hochberg method and the most conservative empirical method yielded 474 and 585 PSGs in the Mammals group, respectively.

Although the Hochberg and empirical methods resulted in similar numbers of PSGs for the Mammals species group, the empirical method identified the greater number of PSGs in the smaller species groups. The pattern of Hochberg PSG counts across species groups was reminiscent of the pattern of significant PSCs identified after controlling the FDR (Table 4.6): Mammals and Eutheria yielded several hundred PSCs and PSGs, Primate and Laurasiatheria yielded a much smaller but still non-zero number, and the other species groups yielded none. The consistency of this pattern between PSCs and PSGs reflected the fact that the Hochberg method for identifying PSGs was sensitive largely to the existence of any one site within a gene having a very strong signal of positive selection. Thus, only the species groups with a large total branch length and a high prevalence of positive selection produced a large number of Hochberg PSGs.

In contrast, PSGs from empirical p-values reflected a significant clustering of less extreme PSCs. As a result, the empirical method identified some PSGs in species groups where the Hochberg method identified none. The qualitative pattern between species groups was largely similar to that seen for the Hochberg PSGs: using the conservative filter and the empirical method with a truncation threshold of $p < 0.01$, Mammals and Eutheria yielded around 600 PSGs, Primates, Laurasiatheria and HQ Mammals produced around 400, and most other species groups had 50 or fewer PSGs. The species group with the most striking difference between the Hochbeg PSGs and the empirical PSGs was the HQ Mammals group, which had zero Hochberg PSGs but several hundred empirical PSGs. This was consistent with the intermediate location of the cumulative curve for HQ Mammals under the Hochberg method in Figure 5.1; although this species group showed a greater enrichment of low p-values than the lowest cluster of curves, it was not strong enough to produce any significant genes at FDR< 0.1.

In summary, the 3 types of methods for combining sitewise estimates to identify PSGs showed very different performance patterns across the different species groups. While the TPM and Fisher's method have been extensively used in large-scale studies, they appeared to lack power in this application. Control of the FWER or the use of empirical p-values yielded greater numbers of PSGs. Using these methods to identify PSGs, the 10 species groups fell into 2 clusters, each with a very different proportion of identified PSGs.

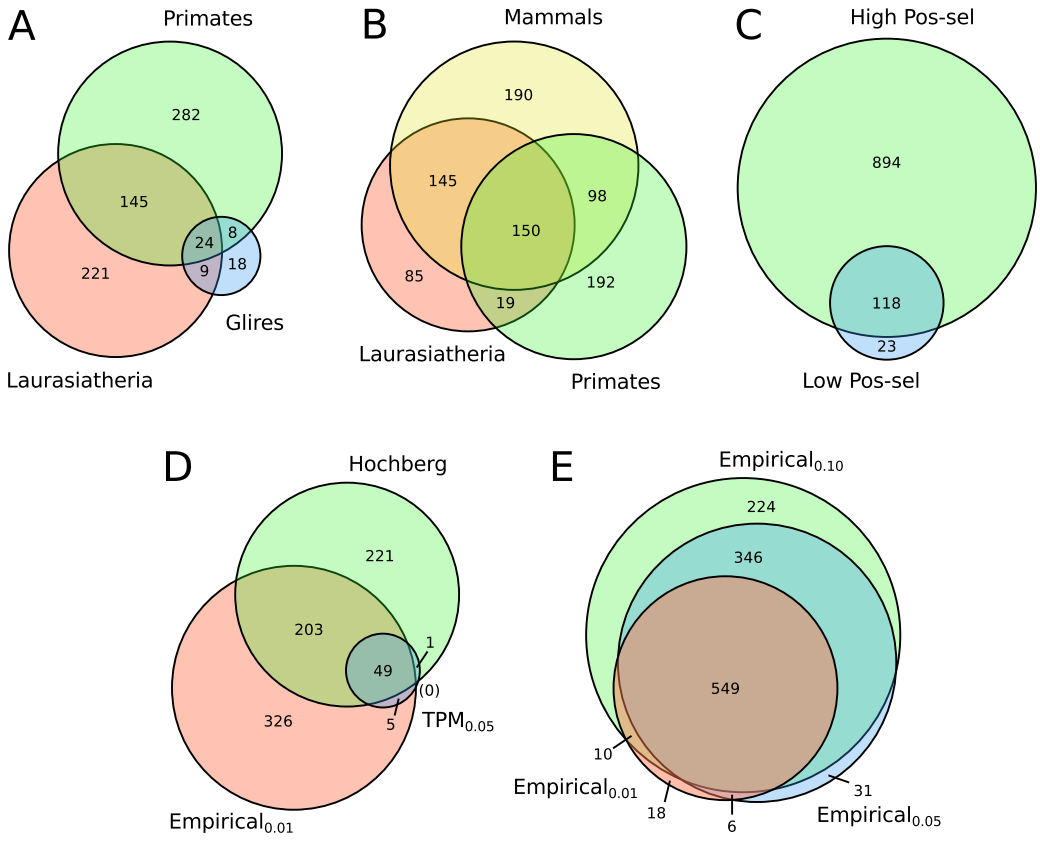


Figure 5.2: Venn diagrams of PSGs identified in different species groups and using different methods. (A) PSGs identified using empirical p-values in Primates, Glires, and Laurasiatheria. (B) PSGs identified using empirical p-values in Mammals. Laurasiatheria and Primates. (C) PSGs identified using empirical p-values in species groups with high and low levels of positive selection. (D) PSGs identified in the Mammals group using three different methods for combining sitewise estimates within genes. (E) PSGs identified in the Mammals group using the empirical p-value method with 3 different truncation thresholds: $p < 0.01$ (smallest circle, left), $p < 0.05$ (middle circle, right), $p < 0.10$ (largest circle, top).

Overlaps between positively-selected genes in different species groups

Using the sets of significant PSGs from Table 5.1, it was possible to identify how many PSGs were shared between, or unique to, different species groups or methods. Unless otherwise specified, all future analyses in this chapter will be derived from the conservatively-filtered dataset.

I first looked at the distribution of PSGs from the empirical method with a $p < 0.01$ truncation threshold across species groups. Overall, a total of 1035 out of 11520 genes, or

8.9% of those investiated, were identified as a PSG in at least one of the species groups. Figure 5.2 shows a more detailed breakdown of how many PSGs were shared between various species groups. Figure 5.2A compares genes from the three major mammalian superorders, showing that Primates and Laurasiatheria share roughly a third of their PSGs and that around two-thirds of PSGs in Glires are also significant in Primates, Glires, or both. Figure 5.2B looked at PSGs shared between Primates, Laurasiatheria, and Mammals (which contained all of the species within the Primates and Laurasiatheria groups), showing roughly equal mixtures of shared and unique genes. Finally, I split the 10 species groups into 2 clusters based on the prevalence of PSGs: Mammals, HQ Mammals, Eutheria, Primates and Eutheria were considered "High Pos-sel" groups, and the rest were considered "Low Pos-sel" groups. Figure 5.2C shows the overlap between the union of PSGs identified in each group; PSGs from the High Pos-sel cluster of species groups is largely a superset of those from the Low Pos-sel cluster, with only 23 PSGs unique to the species groups which showed less overall positive selection.

Expanding the count to include PSGs identified by the Hochberg, Fisher, and truncated product methods (at a truncation threshold of $p < 0.05$), the total number of PSGs identified was 1300, or 11.3% of all genes tested. Compared to the 1035 from the empirical method alone, the additional 265 genes came from the Hochberg method in the Eutheria and Mammals groups. Figure 5.2D shows the overlap of PSGs identified in the Mammals group by different methods; while the TPM yielded no unique PSGs, a large number of the Hochberg genes were unique, indicating that the Hochberg and empirical p-value methods were sensitive to different patterns of positively-selected sites within genes. In contrast, Figure 5.2E shows that the different variants of the empirical method using different truncation thresholds yielded largely the same set of PSGs, but with increasing sensitivity as the truncation threshold was relaxed from $p < 0.01$ to $p < 0.10$.

5.4 Functional analysis of PSGs and comparison to previous studies

I used Gene Ontology (GO) term annotations from the Ensembl database to identify functional categories enriched for PSGs. GO annotations for all human genes were downloaded from version 64 of Ensembl and were applied to the mammalian alignment containing each human gene. As the GO ontology contains links between terms forming a directed acyclic graph, I followed the common practice of applying the set of all ancestral, and thus less-

specific, terms to each gene as well [Rivals *et al.*, 2007]. Only terms within the Biological Process ontology were included in this analysis, as the Molecular Function and Cellular Component hierarchies contain less information on the types of processes generally associated with the presence of positive selection in mammalian genes [Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007].

Two methods were employed to identify GO terms enriched for PSGs. First, a simple test for independent association was performed for each term: a 2x2 contingency table was filled with the counts of PSGs and non-PSGs which were annotated and not annotated with the current term (each combination of which filled one cell of the table), and Fisher's Exact Test (FET) was used to perform a one-sided test for independence of rows and columns. A highly significant FET p-value thus represented strong evidence for a positive association between a gene being positively-selected and being annotated with the given term [Rivals *et al.*, 2007]. To control for multiple tests being performed, I excluded all terms containing fewer than 5 PSGs (to reduce the number of tests performed and to avoid including highly specific and less biologically-informative GO terms) and used the Benjamini-Hochberg method to identify the FET p-value needed to control for an expected FDR$<$ 0.1 within each set of PSGs. The second method I used to assess significance was the `weight` algorithm from the `topGO` program [Alexa *et al.*, 2006]. The `weight` algorithm also uses FET to identify significant associations between terms and genes of interest, but it accounts for the fact that gene annotations for nearby terms in the GO graph structure are highly correlated by reducing the significance of terms which have more specific, significantly-enriched descendant terms. The result of this weighting is that clusters of closely-related and highly significant terms, which may otherwise clutter the list of top FET results with an uninformative set of very similar terms, are thinned out by reducing the p-values of the less-specific ancestors. Only terms which were significant by both FET (FDR$<$ 0.1) and the `weight` algorithm ($p < 0.1$) were included in the top and bottom sections of Table 5.2. Terms with more than 300 or fewer than 30 annotated genes were also excluded from inclusion in the top or bottom sections of Table 5.2 for clarity.

These two methods were applied to several sets of PSGs in order to assess the consistency of enriched terms between different methods for identifying PSGs and different species groups. The conservatively-filtered dataset was used for all tests. Each set of enriched terms was assigned a letter for identification in Table 5.2; those letters are included here in parentheses for reference. From the Mammals group, I tested for enriched GO terms in the 474 Hochberg PSGs (`H`), the 585 $Emp_{0.01}$ PSGs (`M`), the 934 $Emp_{0.05}$ PSGs

(m), and the 202 genes in the top 2% genome-wide by overall dN/dS value (D). The latter group was defined using gene-wide dN/dS values output by SLR, based on fitting a M0-like codon model to the mammalian alignment. To evaluate PSGs identified in the mammalian superorders, I tested the 459 $Emp_{0.01}$ PSGs from Primates (P), the 409 $Emp_{0.05}$ PSGs from Glires (g), and the 400 $Emp_{0.01}$ PSGs from Laurasiatheria (L). Finally, the set of 273 genes with independent evidence for positive selection in each of the Primates, Glires, and Laurasiatheria groups was obtained by taking the least significant $Emp_{0.05}$ p-value for each gene from each species group and identifying genes which remained significant (i). Note that groups indicated by lowercase letters correspond to those using the less conservative $Emp_{0.05}$ PSG definition.

In order to facilitate a comparison with functional associations reported in previously-published studies, I also collected the lists of terms enriched for PSGs from Clark et al. [2003] (C), the Rhesus Macaque Genome Sequencing and Analysis Consortium [2007] (R), and Kosiol et al. [2008a] (K).

GO Term		Enriched in		Values for Mammals $Emp_{0.05}$ (label M in "This Study" column)					
ID	Description	This Study	Lit.	FET	topGO	Ann.	Sig.	Exp.	Top 5 Genes
Top 10 Enriched Terms									
GO:0006954	inflammatory response	P LMmHD	RK	**9.4e-11**	**2.0e-06**	202	35	10.2	TFRC, TLR1, ITGAL, TLR4, A2M
GO:0045087	innate immune response	P LMmHDi	RK	**2.4e-09**	**2.7e-04**	144	27	7.3	SAMHD1, TLR1, TLR4, A2M, C8A
GO:0051607	defense response to virus	P LMmH i	K	**4.2e-05**	**5.4e-03**	43	10	2.2	SAMHD1, CD4, ZC3HAV1, MAVS, DDX58
GO:0042742	defense response to bacterium	PgLMmHDi	K	**8.8e-05**	**1.2e-04**	38	9	1.9	TLR1, TLR4, LTF, MAVS, MBL2
GO:0000236	mitotic prometaphase	PgLMm Di		**3.7e-04**	**3.7e-04**	55	10	2.8	**CENPT, CENPI, CENPQ, REC8, DSN1**
GO:0019221	cytokine-mediated signaling	Mm	K	**1.3e-03**	**4.1e-02**	98	13	5.0	MAVS, IL1RL1, NLRC5, STAT2, IFNGR2
GO:0050900	leukocyte migration	P LMm		**1.5e-03**	**1.4e-03**	112	14	5.7	ITGAL, ITGAM, CD84, COL1A2, CD34
GO:0007067	mitosis	Pg Mm Di		**3.4e-03**	**4.7e-02**	218	21	11.0	**HAUS6, CENPT, CENPI, SPAG5, KIAA1009**
GO:0002576	platelet degranulation	Mm		**4.7e-03**	**4.7e-03**	42	7	2.1	A2M, KNG1, HRG, SELP, ITGA2B
GO:0051297	centrosome organization	gLMm		**5.6e-03**	**1.2e-02**	33	6	1.7	**HAUS6, CEP152, CEP250, BRCA2, HAUS5**
Other Terms Commonly Identified in the Literature									
GO:0006952	defense response	P LMmHDi	RK	**2.0e-15**	1.5e-01	376	59	19.0	TFRC, SAMHD1, TLR1, ITGAL, TLR4
GO:0006955	immune response	PgLMmHDi	RK	**2.2e-12**	**3.3e-03**	415	57	21.0	SAMHD1, TLR1, ITGAL, TLR4, CD164
GO:0009611	response to wounding	P LMmHD	RK	**3.2e-07**	5.8e-01	553	56	28.0	TFRC, TLR1, VCAN, ITGAL, TLR4
GO:0050896	response to stimulus	P LMmHD	RK	**4.0e-06**	6.0e-01	2343	160	118.7	TERF2, TFRC, SAMHD1, GGH, TLR1
GO:0009607	response to biotic stimulus	P LMmHDi	RK	**2.7e-05**	1.0e+00	265	30	13.4	SAMHD1, TLR1, TLR4, LTF, CD4
GO:0050909	sensory perception of taste		RK	5.7e-01	5.7e-01	16	1	0.8	RTP4
GO:0007600	sensory perception		C K	7.4e-01	1.0e+00	274	12	13.9	TLR4, FAM161A, RP1, RTP4, COL4A3
GO:0007606	sensory perception of chemical stimulus		RK	8.5e-01	1.0e+00	36	1	1.8	RTP4
GO:0007166	cell surface receptor linked signaling		CR	1.0e+00	8.2e-01	1107	37	56.1	ITGAL, TLR4, ITGAM, HRH4, COL16A1
GO:0007608	sensory perception of smell		C K	1.0e+00	1.0e+00	17	0	0.9	
Other Terms Identified in This Study but Not in the Literature									
GO:0006302	double-strand break repair	Mm i		**8.4e-03**	**2.6e-02**	58	8	2.9	**SETX, XRCC5, BRCA2, UIMC1, APLF**
GO:0051301	cell division	g Mm		**1.5e-02**	**7.2e-03**	264	22	13.4	**HAUS6, SPAG5, KIAA1009, DCLRE1A, SYCP1**
GO:0031295	T cell costimulation	LMm D		**2.1e-02**	**2.1e-02**	32	5	1.6	CD4, CD3G, DPP4, CD86, CD274
GO:0007059	chromosome segregation	Mm		**2.4e-02**	**1.6e-02**	83	9	4.2	REC8, **BRCA2, DSN1, CENPH, SETDB2**
GO:0015711	organic anion transport	mH		7.6e-02	9.1e-02	45	5	2.3	ABCC2, SLC26A8, SLC16A7, SLC4A1, SLC13A2
GO:0071706	TNF superfamily cytokine production	L mH		7.6e-02	9.8e-02	32	4	1.6	TLR1, TLR4, MAVS, CD86
GO:0007283	spermatogenesis	P L Di		8.3e-02	5.4e-02	184	14	9.3	NLRP14, SLC9A10, CYLC1, REC8, SYCP1

Table 5.2: Example GO terms enriched for PSGs in this study and in the literature. Top section: the 10 terms most significantly enriched for $Emp_{0.05}$ PSGs in the Mammals species group. Middle section: other terms found in at least 2 of 3 published genome-wide scans. Bottom section: other terms enriched for PSGs in this study but not in the literature. The presence or absence of characters under the columns "This Study" and "Lit." indicates which sets of genes from this or previously-published studies showed enrichment for PSGs for that term (see text for definitions). The last 6 columns show values from the Mammals $Emp_{0.05}$ set, corresponding to the 'M' flag; bold P-values indicate significance (FDR$<$ 0.1 for FET and p$<$ 0.05 for topGO). Genes discussed in the text are presented in bold face. Lit.—literature; FET—Fisher's Exact Test; Sig.—Significant; Exp.—Expected.

Table 5.2 summarizes the results of the GO term enrichment tests, showing for three sets of terms which groups of genes from this study, and which previously-published studies, identified a significant enrichment of PSGs.

The top section shows 10 of the GO terms most strongly enriched for Mammalian $Emp_{0.05}$ PSGs according to FET. The top few terms, including *inflammatory response*, *innate immune response*, *defense response to virus* and *defense response to bacterium*, represented genes involved in host defense and immune response—two of the functions most commonly associated with positive selection in mammals [Nielsen, 2005]. Accordingly, the top four terms were identified in one or two previously-published studies and in most or all of the species groups and PSGs identification methods evluated in this study. Interestingly, the term *mitotic prometaphase* was associated with PSGs in almost all gene sets from this study, but it was not found by any of the sets of enriched terms from the literature. The next several terms, most of which were not found in the literature, showed a more mixed pattern of enrichment across gene sets from this study. Some of these terms, including *mitosis* and *centrosome organization* were connected to the more strongly-enriched *mitotic prometaphase* term and showed many of the same significant genes; others, such as *platelet degranulation* and *leukocyte migration* were distinct in function and composition of Mammalian $Emp_{0.05}$ PSGs.

The middle section of Table 5.2 focuses on GO terms commonly associated with PSGs in the literature which were not included in the 10 top terms, showing all terms identified in at least 2 of the 3 previously published studies. The first 5 terms largely recapitulated those included in the top section relating to defense response and inflammation, all of which were identified in most gene sets from the current study. The next several terms, including *sensory perception of taste* and *cell surface receptor linked signaling*, were more related to sensory perception and were uniformly not associated with PSGs in this study. The lack of an association for these terms in the current study was surprising, as olfaction and sensory perception have been among the most consistently identified functional categories in large-scale scans for positive selection [Nielsen *et al.*, 2005; Nielsen, 2005]. One explanation for this difference may be that the removal of highly-duplicated genes from the conservatively-filtered dataset has reduced the number of olfactory and sensory genes available for analysis. While there was some evidence that the current dataset was depleted of olfactory genes compared to previous analyses (according to Table 5.2 only 17 genes were annotated with *sensory perception of smell*, while Kosiol et al. [**?**] analyzed 229 such genes), the number of genes annotated with *sensory perception* (274) and *sensory perception of chemical stimulus*

(36) were still large enough to produce a significant enrichment if one existed. Other possible explanations included the possibility that positively-selected sensory genes were more prone to exclusion from the current analysis for other reasons (for example, if their alignments contained more clustered nonsynonymous substitutions) or less sensitivity in the current study to the patterns of positive selection occurring in sensory perception genes.

The bottom section of Table 5.2 shows the remainder of terms which were identified in the current study, but not in previous studies providing GO term enrichments, as associated with PSGs. The first term, *double-stranded break repair*, was identified in 3 of the 8 gene sets, with the association with PSGs driven by genes such as *SETX*, a RNA helicase which causes ataxia and lateral sclerosis when defective [Suraweera *et al.*, 2007], and *BRCA2*, a tumor suppressor gene for which a common allele is associated with an increased risk of breast cancer and whose close relative, *BRCA1*, has been shown to be positively selected in mammals [Huttley *et al.*, 2000]. Some of the next terms, including *cell division* and *T cell costimulation* and *TNF superfamily cytokine production*, were similar to other terms in the first two sections and contained similar sets of PSGs, but the terms *organic anion transport* and *spermatogenesis* were quite distinct in their function and set of associated PSGs. The anion transport term contained largely members of the solute carrier (SLC) gene superfamily, a 300-strong group of membrane-bound transporter genes [He *et al.*, 2009], while the *spermatogenesis* category has been widely reported in other studies of mammalian positive selection not included in Table 5.2 [Torgerson *et al.*, 2002; Swanson *et al.*, 2003; Clark & Swanson, 2005; Nielsen *et al.*, 2005].

The GO term enrichments indicated a strong prevalence of positive selection in genes related to core cellular processes such as cell division and DNA repair. Many of these associations were noted and discussed by Nielsen et al. [2005], who hypothesized an interesting connection between PSGs and cancer-related genes in these functional categories. Nielsen et al. suggested that cancer-related genes, which are often involved in cell proliferation and apoptosis pathways, may be likely targets of positive selection resulting from genetic conflict due to their involvement in processes known to lead to positive selection, such as the proliferation of immune cells [Sawyer *et al.*, 2005] or sperm competition [Torgerson *et al.*, 2002; Clark & Swanson, 2005]. This hypothesis was developed and expanded by Crespi and Summers [2006], who analyzed the results of several scans for positively-selected genes through the lens of cancer risk. Crespi and Summers argued that positive selection resulting from "antagonistic coevolution" between various entities (e.g., hosts and parasites, parents and offspring, or sperm cells and eggs) has been the driving force behind the evo-

lution of increased cancer risk. Although similar trends were observed by Nielsen et al., the current study provided additional support for an association between positive selection and cancer-related genes, expanding the list of PSGs in functional categories related to cancer progression and containing known tumor suppressor genes.

A more surprising result from the GO term analysis was the strong enrichment of PSGs in terms related to mitosis and chromosome segregation. None of these terms were identified in the previous studies analyzed, but I found strong enrichments for terms such as *mitosis*, *centrosome organization*, and *chromosome segregation*. All of these terms were identified as enriched for $Emp_{0.01}$ PSGs in Mammals, while *centrosome organization* was enriched for $Emp_{0.01}$ in Laurasiatheria and for $Emp_{0.05}$ in Glires. Among the top PSGs within these terms were *HAUS6*, a member of the HAUS microtubule-binding complex which is vital to the mitotic spindle assembly and maintenance of the centrosome, centrosomal proteins *CEP152* and *CEP250*, and several centromere proteins including *CENPT*, *CENPI*, *CENPQ*, and *CENPH*. There has been great interest surrounding the evolution of centromeric DNA and proteins ever since Henikoff, Ahmad and Malik proposed the "centromere paradox" [2001]. Based on the observation that both centromeric DNA and centromere-related proteins were rapidly evolving in animals, the authors postulated an ongoing genetic conflict between centromeric DNA and proteins resulting from the unequal transmission of chromosomes during female meiosis [Henikoff *et al.*, 2001; Malik & Henikoff, 2002, 2009]. Initial comparative analysis of the major centromeric protein *CENPA* gene showed it to be positively-selected in *Drosophila* and *Arabidopsis* but not in mammals, while a more recent study in primates identified positively-selected residues in *CENPA* and three other centromeric proteins [Schueler *et al.*, 2010]. The current identification of several positively-selected centromere proteins provided large-scale corroboration of the result from primates, showing that positive selection in centrosomal and centromeric proteins is a major component of the overall set of PSGs throughout mammals. In all, 12 out of the 17 centromeric proteins included in this analysis showed evidence of positive selection in either the relaxed or conservatively-filtered datasets.

5.5 Comparing PSGs identified by different studies

Somewhat surprisingly, no direct comparison between PSGs identified in large-scale scans for positive selection has been published, despite the observation that many similar terms and genes tend to occur in studies including different species and using different methods

[Nielsen *et al.*, 2005; Kosiol *et al.*, 2008a]. To gain a better understanding of the amount of similarity between the results from this analysis and from previously-published studies, I performed a gene-by-gene comparison with the sets of PSGs described by Clark et al. [2003], Nielsen et al. [2005], the Rhesus Macaque Genome Sequencing and Analysis Consortium [2007], and Kosiol et al. [2008a]. The goals of this analyses were conceptually similar to those of the previous section: to identify trends in shared and unique signatures of positive selection from this and previous genome-wide scans.

I first mapped the sets of genes described by each of the above studies to the set of genes included in this analysis using the supplementary data tables provided alongside each publication. The process was slightly different for each study due to the different formats provided.

Clark et al. [2003] provided NCBI RefSeq gene IDs, which I converted to Ensembl gene IDs using index files downloaded from the NCBI Entrez gene database [Maglott *et al.*, 2005]; this resulted in 5,636 of the original 6,145 genes being successfully mapped. Following Clark et al. [2003], genes with $p < 0.01$ for the M2 test in either the human or chimpanzee branch were taken as PSGs, yielding 272 successfully mapped PSGs. Nielsen et al. [2005] provided a table including Ensembl gene IDs, NCBI RefSeq gene IDs, and gene names for each of the 20,362 genes included in their study. I used all three pieces of information to attempt to match those genes to the current dataset, but only 11,402 of the original genes were successfully matched. Still, these genes appeared to contain most of the 50 top **pst!**s (**pst!**s) reported in their analysis [Nielsen *et al.*, 2005]. Although the authors did not provide specific criteria by which PSGs were defined, I took the lowest LRT value from the 50 genes described, 1.67, and identified 142 successfully mapped genes with LRT values greater than that value. The Rhesus Genome Sequencing and Analysis Consortium [2007] provided the names of 179 PSGs identified using a branch-site test along any branch of the primate tree. Of those, 123 genes were matched by name to genes included in the current study. Finally, Kosiol et al. [2008a] provided a UCSC browser track with chromosomal coordinates and scores based on a test across the entire mammalian phylogeny for 16,529 genes, of which 544 were positively-selected at FDR< 0.1. Using chromosomal coordinates to match genes in the current dataset, 14,460 genes and 395 PSGs were identified.

Figure 5.3 shows the overlap between PSGs identified in this and previously-published studies. Note that the total numbers of PSGs are smaller than those noted in the previous paragraph, as genes which were absent from the conservatively-filtered dataset were removed. (Results from a comparison using the relaxed filter were qualitatively similar to

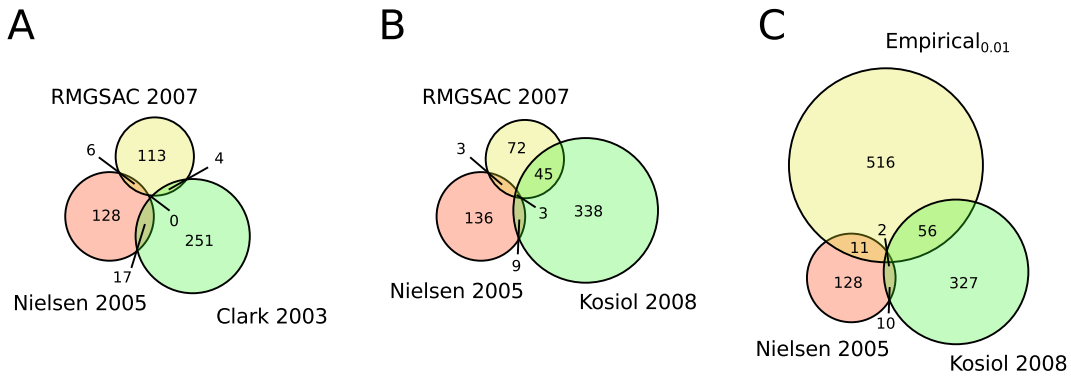


Figure 5.3: Venn diagrams of PSGs identified in different studies. (A) PSGs identified in primates by Clark et al. [2003], Nielsen et al. [2005] and the Rhesus Macaque Genome Sequencing and Analysis Consortium [2007]. (B) PSGs identified in primates and mammals by Nielsen et al. [2005], the Rhesus Macaque Genome Sequencing and Anaalysis Consortium [2007] and Kosiol et al. [2008a]. (C) PSGs identified in primates and mammals by Nielsen et al. [2005], Kosiol et al. [2008a] and this study using the Mammals species group, conservative filter, and the $Emp_{0.01}$ method.

those in Figure 5.3.) Overall, the lack of overlap in identified PSGs was striking: Figure 5.3A shows the overlap between the three studies in primates, with zero genes shared by all 3 studies and from 4 to 17 genes shared between any pair. Although each analyses identified similar numbers of PSGs, very few of the actual genes identified were in common. This result did not appear to be an artifact of genes lost during the mapping process, as Nielsen et al. also noted that only 1 of their top 50 genes was also identified by Clark et al. [?] as evolving under positive selection. Figure 5.3B shows slightly more overlap between the two most recent studies, with 45 PSGs shared between Kosiol et al. and the Rhesus genome analysis. The comparison between PSGs from Nielsen et al., Kosiol et al., and the set of $Emp_{0.01}$ PSGs from the Mammals species group shown in Figure 5.3C revealed a similar number of overlapping genes, despite the larger overall number of PSGs identified in the current study.

The comparison of overlapping PSGs was somewhat limited, as it required the use of a cutoff threshold to identify each set of PSGs and did not easily allow for a comparison between the different methods. For example, although Figure 5.3C showed a greater number of overlapping genes between Kosiol et al. and the current study than between Nielsen et al. and the current study (154 vs. 31), it was unclear whether this was due to the greater overall number of PSGs identified by Kosiol et al., or to a greater tendency for this study and Kosiol et al. to identify common PSGs. By eye, it seemed as if both Kosiol et al. and

Nielsen et al. shared a similar proportion of PSGs with the current study.

As an alternative approach to comparing between the current results and previous studies, I constructed a series of receiver operator characteristic (ROC) curves for each published study. For each study, the set of PSGs was used as the binary classifier (or "truth" value), and a set of 4 gene-wide p-values or dN/dS estimates from the current study were evaluated as test statistics. Curves were constructed by sorting the list of matched genes by each test statistic and counting the cumulative number of PSGs identified as the test statistic increased in value. To test whether the choice of species group affected the proportion of shared PSGs, I included gene-wide dN/dS estimates for Primates and Mammals (where the test statistic was the negative dN/dS value, so the genes with highest dN/dS were sorted first), and to test whether the method used to combine sitewise estimates within genes had an effect, I included $Emp_{0.01}$ and Hochberg p-values as test statistics.

Figure 5.4 shows the ROC curves comparing the current dataset to each of the four previously published studies. The vertical dashed lines correspond to the FDR< 0.1 threshold of the $Emp_{0.01}$ p-values in Mammals, making the intersection of the $Emp_{0.01}$ ROC curve at the vertical lines equivalent to the numbers of overlapping PSGs seen in Figure 5.3C for the Nielsen et al. [2005] and Kosiol et al. [2008a] sets of PSGs.

The Clark et al. [2003] curves hardly strayed from the diagonal line, showing little ability beyond random chance to identify PSGs from that study. This was not necessarily unexpected, as that study tested for positive selection only along the very short human and chimpanzee branches of the primate tree, while even the Primates species group from the current study contained sequences from species as distant as tarsier, covering much more branch length and a much more diverse set of primate species. The Nielsen et al. [2005] study showed a noticeably stronger enrichment for PSGs in genes with low p-values or high dN/dS values in the current study, with each ROC curve rising well above the diagonal, and the Primates dN/dS curve showing the greatest performance. The difference between the curves for Nielsen et al. and Clark et al. was interesting, as both studies used the same set of sequences and alignments. Presumably the analytical method used by Nielsen et al. was more similar to the current study in its sensitivity to patterns of positive selection than that used by Clark and colleagues. Within the Nielsen et al. panel, the difference between the two curves based on overall dN/dS ratios and the two curves based on p-values from sitewise estimates was noticeable, with the dN/dS curves showing greater performance throughout the range of cutoff values. This may be explained by the small amount of branch length included in that study providing only enough power to

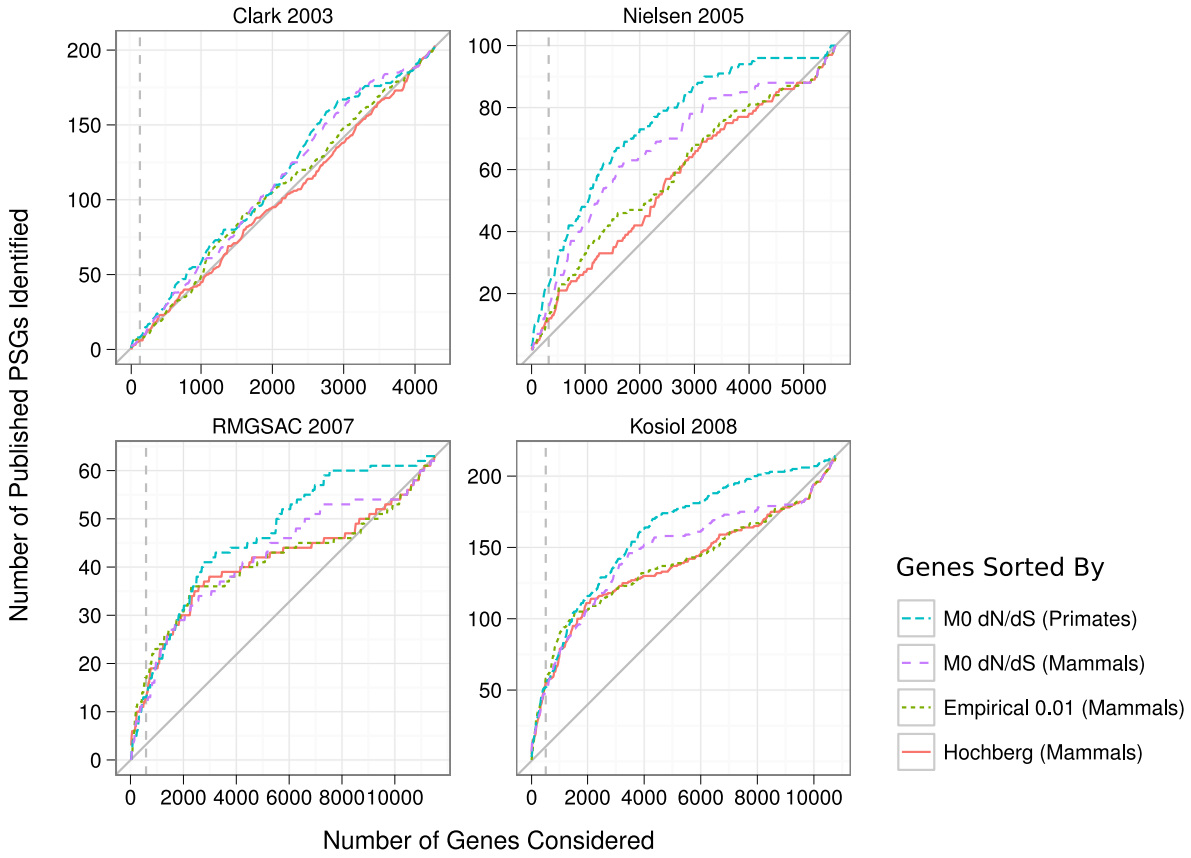


Figure 5.4: ROC curves for using dN/dS estimates and different PSG identification methods to identify PSGs in previously-published studies. Within each panel, the x-axis represents all genes successfully matched between the published study and the current analysis and the y-axis represents the number of PSGs within the matching genes. Each curve traces the cumulative number of PSGs identified in the published study when the top N genes, according to the test statistic, were considered. A dashed vertical line is drawn on each panel at the x-axis value corresponding to the number of $Emp_{0.01}$ PSGs in the Mammals species group.

detect PSGs with high overall dN/dS values as opposed to genes with smaller proportions of positively-selected sites.

The ROC curves for the two more recent studies showed noticeably greater, and roughly equivalent, performance. In both cases, all 4 curves showed nearly identical performance in the high-specificity region of the graph, identifying roughly 25% of the total number of PSGs were identified by all curves before the vertical dashed line was reached. At the same significance threshold, roughly 20% and 2% of PSGs were identified in the Nielsen et al. and Clark et al. graphs, respectively. The curves based on dN/dS values showed better performance than the sitewise methods at the higher end of the curve; this was somewhat

expected, as the $Emp_{0.01}$ and Hochberg methods both required reasonably strong sitewise evidence for positive selection to successfully distinguish between PSGs and non-PSGs. The observation that the sitewise methods were less able than dN/dS values to identify PSGs from the Rhesus consortium and Kosiol et al. in the low-specificity range was consistent with a slight lack of power to detect weak distributed positive selection resulting from the use of sitewise estimates to identify PSGs.

5.6 Gene families with many PSGs

Table 5.2 contained a relatively large number of PSGs from the same gene family (e.g., solute carrier family genes *SLC26A8*, *SLC16A7*, *SLC4A1*, *SLC13A2*, *SLC9A10*; collagen genes *COL1A2*, *COL4A3*, *COL16A1*; and toll-like receptor genes *TLR1* and *TLR4*). The clustering of PSGs within large gene families was not unexpected, as different members of a gene family may be more likely to have similar cellular functions; thus, a family of immune-related genes such as the TLR genes would be expected to be enriched for PSGs. The prevalence of positively-selected gene family members was concerning, however, as many gene families arise through segmental duplications [Ohno, 1970], and duplicate genes residing nearby on a chromosome are likely targets of ectopic gene conversion events [Ezawa *et al.*, 2006; Benovoy & Drouin, 2009]. Gene conversion is a non-reciprocal recombination process which is initiated by a double-stranded break in the DNA helix that is subsequently repaired through strand invasion by a homologous sequence; ectopic gene conversion events are defined as those that occur between homologous sequences not at the same genetic locus [Benovoy & Drouin, 2009]. The problem with gene conversion in comparative studies is that it breaks the assumption that the relationships of a set of genes can be well described by one bifurcating phylogenetic tree. Thus, when sequences with gene conversion are analyzed using the species tree, an incorrect sequence of substitution events is required to explain the observed sequences with respect to the phylogenetic tree, potentially leading to excessive estimates of substitution rates. In the case of detecting positive selection, gene conversion among paralogs has been observed to result in moderately elevated rates of false positives [Casola & Hahn, 2009].

I assessed the potential impact of gene conversion on the current dataset by identifying nearby paralog pairs (NPPs) and comparing those to the list of $Emp_{0.05}$ PSGs from the relaxed sitewise filter and the Mammals species group. I defined NPPs as pairs of genes which are members of the same Ensembl gene family and which reside on the same chro-

Gene Family	Genes	NPPs	PSGs	NPP–PSGs		Top 4 NPP–PSGs
Ensembl Families with > 4 PSGs						
ENSFM00600000921151	6	6	6	6		COL4A6, COL4A2, COL4A4, COL4A5
ENSFM00250000001219	5	5	5	5		CD1D, CD1C, CD1A, CD1E
ENSFM00250000000804	4	4	4	4		C6, C7, C8A, C8B
ENSFM00250000000852	4	4	4	4		GBP6, GBP4, GBP5, GBP3
ENSFM00500000269596	11	11	4	4		SERPINB3, SERPINB12, SERPINB13, SERPINB9
ENSFM00500000269665	4	4	4	4		CTSG, GZMB, GZMH, CMA1
ENSFM00250000000002	89	68	4	3		ZFP37, ZNF473, ZNF677
ENSFM00250000000948	4	3	4	3		ACOT6, ACOT4, ACOT2
ENSFM00350000105388	5	5	3	3		CES5A, CES2, CES1
ENSFM00400000131714	7	6	3	3		SLC22A25, SLC22A14, SLC22A8
ENSFM00400000131728	7	7	3	3		MMP3, MMP8, MMP1
ENSFM00470000251442	4	3	4	3		EMR2, EMR3, CD97
ENSFM00500000269709	5	4	4	3		ITGAL, ITGAX, ITGAM
ENSFM00500000269927	4	4	3	3		SLC17A3, SLC17A1, SLC17A4
ENSFM00570000851010	5	4	4	3		NLRP9, NLRP4, NLRP5
Manually Curated Families					FET p-value	
Toll-like Receptors	8	4	5	3	0.50	TLR8, TLR6, TLR1
Collagen	30	7	22	7	0.08	COL4A6, COL4A2, COL4A4, COL4A5
ADAM Family	42	11	7	4	0.06	ADAM32, ADAM2, ADAM28, ADAM7
Solute Carrier Family	338	51	37	10	0.03	SLC26A3, SLC17A3, SLC17A1, SLC17A4
All Genes	15946	1150	1898	200	0.00	AC090098.1, CDKN2A, FAM26F, ACOT6

Table 5.3

mosome within 2 Mb of each other; in total, 1,150 genes from 361 Ensembl families were identified as members of a NPP. The top section of Table 5.3 summarizes the coincidence of NPPs and PSGs within Ensembl gene families containing at least 3 PSGs, sorted by the number of genes which were both PSGs and part of a NPP. The list was topped by the collagen type IV family, with all 6 family members showing evidence of positive selection in mammals and residing within 2Mb of another family member. Other families containing many NPP–PSGs were the CD1 family of transmembrane glycoproteins [Joyce, 2001], several members of the complement immune system [Nonaka & Kimura, 2006], a family of guanylate-binding proteins located in a cluster on chromosome 1 [Olszewski *et al.*, 2006], and two families containing granzyme peptidases and serine peptidase inhibitors.

Every PSG from the aforementioned families was also a member of a NPP, suggesting that gene conversion may have led to the false detection of positive selection in these families. However, many of the same families contained genes involved in core immune system processes which have been consistently shown to harbor the highest fraction of PSGs. Thus, although these gene families exhibited a striking coincidence of NPPs and PSGs,

Human Gene			Evidence for Positive Selection				
Name	Chr.	Loc.	Conservative	Relaxed	Lit.	NPP	$Emp_{0.01}$ p-value
COL4A6	chrX	107.4	PgLMmHDi	PgLMmHDi	K	+	2.8e-04
COL4A2	chr13	111.0	P LMmH	P LMmH		+	2.8e-04
COL4A4	chr2	227.9	PgLMmHDi	PgLMmHDi		+	2.8e-04
COL4A5	chrX	107.7	PgLMmH i	PgLMmH i		+	2.8e-04
COL4A1	chr13	110.8	PgLMmH i	PgLMmH i		+	2.8e-04
COL4A3	chr2	228.0	PgLMmHDi	PgLMmHDi	K	+	2.8e-04
MMP3	chr11	102.7		Mm		+	8.3e-03
MMP8	chr11	102.6		MmH		+	8.3e-03
MMP1	chr11	102.7		m		+	3.2e-01

Table 5.4

the impact of such co-occurrence on the false detection of positive selection within any one family was highly dependent on the function of genes within that family and the associated prevalence of true PSGs. Regardless of this complication, it could be asserted that gene families from the top section of Table 5.3 represented those with the highest likelihood of false positives resulting from gene conversion. A more in-depth study of the evolution of each family would be necessary to confidently assess whether individual families or genes contained evidence of false positives resulting from gene conversion events. Of particular interest were the families without obvious involvement in well-known systems of genetic conflict and positive selection, such as the collagen (e.g., *COL4A6*), carboxylesterase (e.g., *CES5A*), solute carrier family (e.g., *SLC22A25* and *SLC17A3*), and matrix metallopeptidase (e.g., *MMP3*) families.

The presence of metallopeptidase and collagen gene families in Table 5.3 was especially intriguing, as members of the metallopeptidase class of enzymes are responsible for breaking down collagen fibers in the extracellular matrix with various specificities [Sluijter *et al.*, 2006]. Table 5.4 summarizes the collagen and metallopeptidase genes which showed evidence of positive selection; the signal of positive selection was much stronger in the type IV collagen genes than in the metallopeptidases, and all genes with evidence for positive selection were members of a NPP. If gene conversion among these NPPs can be ruled out, then the presence of positive selection within these gene families may be suggestive of either an undescribed relationship between type IV collagen fibers and the immune system "arms race", or a novel type of genetic conflict underlying the presence of positive selection in these related gene families.

Within larger gene groups and across the genome-wide dataset, Fisher's exact test could be used to test the hypothesis of independence between NPPs and PSGs, providing some

quantitative evidence for or against the hypothesis that NPPs were involved in the false positive detection of PSGs. The bottom section of Table 5.3 shows the results of this test for four manually-curated gene superfamilies and the entire set of 15,946 genes from the relaxed dataset in the Mammals species group. While there was little evidence for non-independence between these two factors for the group of 8 toll-like receptors, FET yielded low p-values for the 30 collagen genes and 42 ADAM family genes, a significant p-value for the 338 solute carrier family genes, and a highly significant result for non-independence across the entire genome. This result provided strong evidence that the distribution of NPPs and PSGs was highly non-uniform; evaluated in the context of previous results showing that gene conversion can lead to false positives in detecting positive selection, this suggested that the evidence for positive selection within NPPs–PSGs, some of which have been identified in previous studies (e.g., *COL4A6* and *COL4A3* from Table 5.4 which were identified by Kosiol et al. [**?**]) should be treated with caution.

5.7 Identifying positive selection within protein-coding domains

Using the same methodology developed for genes, the sets of sitewise estimates could be grouped by other entities of interest to assess levels of purifying and positive selection within those groups. An interesting application of this approach was the use of sitewise data to identify protein-coding domains showing the strongest genome-wide evidence for positive selection. Although the gene-wise results could be used to some extent for this by identifying protein domains which are commonly seen in PSGs, the method would be noisy, as positive selection occurring within each PSGs may not be localized to the same shared domains. Instead, directly aggregating sitewise estimates from within the region covered by each domain had the potential to more sensitively and accurately detect positive selection within protein-coding domains.

To identify protein domains with significant evidence for positive selection, I mapped Pfam domain annotations from human genes onto the genome-wide set of mammalian alignments, yielding 2.5 million aligned sites with sitewise estimates and Pfam domain annotations, 5,805 of which contained evidence for positive selection at a nominal $p < 0.01$ threshold. For each Pfam domain, all sites were combined to produce domain-wise p-values using the previously described Hochberg, $Emp_{0.05}$ and $Emp_{0.01}$ methods. Domain-wise p-values were separately estimated for all species groups using the relaxed and conser-

vative sitewise filtered datasets, and multiple testing was controlled at FDR<0.1 using the Benjamini-Hochberg method. After correcting for multiple tests, Pfam domains of the "family" and "repeat" types were excluded from the analysis, so only entries of the "domain" type remained.

Table 5.5 summarizes the results of the Pfam domain analysis. Similar to Table 5.2, the presence or absence of significant evidence for positive selection is indicated by a string of characters; uppercase letters indicate a significant $Emp_{0.01}$ result (e.g., `P`, `L`, `M` for Primates, Laurasiatheria, and Mammals), lowercase letters indicate a significant $Emp_{0.05}$ result (e.g., `g`, `m` for Glires and Mammals), and `H` indicates a significant Hochberg result. Domains were categorized as primarily immune-related, protease, or protease inhibitor domains based on information from the Pfam database; domains with miscellaneous or uncharacterized functions are included in the "Other Domains" section of Table 5.5.

An initial observation was that the conservatively-filtered dataset did not yield as many significant results for most domains, indicating that a large portion of the signal for positive selection within these domains may reside in frequently duplicated genes or alignment regions with clusters of nonsynonymous substitutions (which constituted the major differences between the relaxed and conservative sitewise datasets). I chose to focus on the results from the relaxed dataset for the domain analysis, as the results were biologically interesting and the conservative filter appeared to remove a large number of sites within evolutionarily conserved domains.

As expected, immune-related functions dominated the list of significantly positively-selected Pfam domains. Together, the immunoglobulin and immunoglobulin V-set domains accounted for over 205 $p < 0.01$ PSCs spread across 94 proteins, and both domains were significant at FDR<0.1 for multiple species groups and methods using the relaxed sitewise filter. Interestingly, only a fraction of immunoglobulin-containing genes contributed to the significant evidence for positive selection, with only 51 out of 240 total immunoglobulin-annotated genes containing $p < 0.01$ sites within the domain. This was not unexpected for immunoglobulins, which are known as a highly diverse protein domains in mammals, with representation in several hundred human genes comprising many immune and non-immune functions [Lander & International Human Genome Sequencing Consortium, 2001]. However, the case of immunoglobulin suggests that for the study of less well-known domains or organisms, evidence for positive selection in a subset of domain instances could be taken as evidence of potential adaption of a domain for immune purposes.

Other immune domains with significant positive selection included lectin, a carbohy-

drate binding domain involved in cell adhesion and apoptosis [Cambi & Figdor, 2009], the IL-8 like cytokine domain involved in inflammation and chemotaxis in the immune response [Stein & Nombela-Arrieta, 2005], and the membrane attack complex (MAC) domain involved in the creation of membrane pores causing lysis of bacterial and virus-infected host cells [Lovelace *et al.*, 2011]. The presence of a number of protease and protease inhibitor domains was also consistent with the bulk of positive selection in mammals having resulted from the evolutionary "arms race" with invading pathogens; the buildup and continued evolution of proteases and their inhibitors may represent a significant evolutionary medium through which this conflict is expressed. Domains in the

Pfam Domain		FDR < 0.1		All Sites		$p < 0.01$ Sites		
Accession	Description	Cons.	Relaxed	Genes	Sites	Genes	Sites	Top 5 Genes with $p < 0.01$ in Mammals
Immune Related Domains								
PF07686	Immunoglobulin V-set domain	MmH	PgLMmH	220	10851	58	210	TMIGD1, TREM1, CD2, PILRA
PF00047	Immunoglobulin domain	H	P MmH	240	30486	51	180	Pecam1, CD4, PIGR, FCRL4
PF00084	Sushi domain (SCR repeat)		P LMmH	37	7957	17	125	C1S, CD46, C4BPA, CR2
PF00059	Lectin C-type domain	g Mm	PgLMmH	51	4798	29	111	CD72, KLRB1, PRG3, MBL2
PF00048	Small cytokines (intecrine/chemokine), IL-8 like	LMmH	PgLMmH	26	1231	20	53	CXCL13, CXCL9, CCL23, CCL16
PF01823	MAC/Perforin domain	P MmH	P LMmH	9	1176	5	28	C9, C8A, C6, C7
PF00021	u-PAR/Ly-6 domain		P LMmH	16	775	7	27	CD59, TEX101, CD177, LYPD4
PF00340	Interleukin-1 / 18		PgLMm	8	523	6	27	IL1A, IL18, IL1F6, IL1F9
PF00969	Class II histocompatibility antigen, beta domain		P MmH	5	266	4	22	HLA-DMB, HLA-DRB1
PF02841	Guanylate-binding protein, C-terminal domain		PgLMm	4	970	4	21	GBP4, GBP5, GBP3, GBP6
PF00074	Pancreatic ribonuclease		Mm	5	338	5	17	RNASE7, RNASE12, RNASE4, RNASE10
PF00993	Class II histocompatibility antigen, alpha domain		gLMmH	5	364	4	16	HLA-DQA1, HLA-DMA
PF00354	Pentaxin family		Pg MmH	6	677	3	14	CRP, APCS, SVEP1
PF00062	C-type lysozyme/alpha-lactalbumin family		Mm	7	462	5	13	LALBA, LYZ, LYZL6, SPACA5B
Protease Domains								
PF00089	Trypsin	P L mH	P LMmH	82	10426	44	184	CFI, C1S, PRSS44, PRSS48
PF00656	Caspase domain		P LMmH	12	1474	10	40	CASP8, CASP10, CASP5, CFLAR
PF00031	Cystatin domain	P LMmH	P LMmH	11	780	7	33	HRG, KNG1, FETUB, CST9LP1
PF00246	Zinc carboxypeptidase	M H	MmH	22	4505	9	23	CPB1, CPB2, CPN1, CPM
PF07859	alpha/beta hydrolase fold	Mm	P LMm	6	804	5	16	AADAC, AADACL3, AADACL4, LIPE
Protease Inhibitor Domains								
PF00079	Serpin (serine protease inhibitor)		P Mm	33	7358	19	73	SERPINA3, SERPINB3, SERPINA4, SERPINB13
PF01835	MG2 domain		Mm	8	1615	7	21	C5, C4A, PZP, A2M
PF07678	A-macroglobulin complement component	m	Mm	8	1483	7	14	C5, C4A, A2M, A2ML1

Table 5.5 (*continued on next page*)

Pfam Domain		FDR < 0.1		All Sites		$p < 0.01$ Sites		
Accession	Description	Cons.	Relaxed	Genes	Sites	Genes	Sites	Top 5 Genes with $p < 0.01$ in Mammals
Other Domains								
PF00092	von Willebrand factor type A domain	H	LMmH	53	13254	25	108	ITIH4, COL6A5, CLCA1, CLCA4
PF00067	Cytochrome P450	mH	LMmH	39	11918	18	75	CYP7B1, CYP17A1, CYP2J2, CYP4A11
PF00530	Scavenger receptor cysteine-rich domain	H	P MmH	17	2865	8	42	CD5L, MARCO, MSR1, CD5
PF01284	Membrane-associating domain	M H	gLMmH	25	2499	10	29	SYPL1, CKLF, CMTM6, PLP2
PF01099	Uteroglobin family		P MmH	4	271	3	24	SCGB1D1, SCGB2A1, SCGB1A1
PF08840	BAAT / Acyl-CoA thioester hydrolase C terminal		P Mm	2	333	4	20	BAAT, ACOT4, ACOT2, ACOT6
PF01630	Hyaluronidase	m	gLMmH	5	1435	3	17	SPAM1, HYAL3, HYAL1
PF00049	Insulin/IGF/Relaxin family		Mm	4	232	3	15	RLN1, INSL6, INS-IGF2
PF04326	Divergent AAA domain		P Mm	4	225	3	15	SLFN12, SLFN11, SLFN14
PF00068	Phospholipase A2		P Mm	8	412	3	13	PLA2G2A, PLA2G2D, PLA2G10
PF01471	Putative peptidoglycan binding domain		LMm	16	796	5	11	Mmp12, MMP3, MMP7, MMP9

Table 5.5: Pfam domains with significant evidence for positive selection in mammals. Sitewise estimates from within each domain were combined to identify significant evidence for positive selection at FDR< 0.1 in 4 species groups (Primates, Glires, Laurasiatheria, and Mammals) using two sitewise filtered datasets (conservative and relaxed) and 3 methods to combine sitewise estimates ($Emp_{0.05}$, $Emp_{0.01}$, Hochberg). Characters used to indicate positive selection in the "Cons." and "Relaxed" columns are the same as those used in Tables 5.2 and 5.4. Columns under the "All Sites" heading contain the number of unique genes and sites annotated with each Pfam domain; columns under the "$p < 0.01$ Sites" heading contain the number of unique genes and sites with nominal $p < 0.01$ for positive selection. Only domains with 10 or more $p < 0.01$ sites, 3 or more genes, and FDR< 0.1 in Mammals using the $Emp_{0.05}$ and $Emp_{0.01}$ methods are shown.

5.8 Conclusions

Chapter 6

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

6.1 Introduction

The gorilla and other primate genome projects

Incomplete lineage sorting

Effective population sizes of extant and ancestral primate populations

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

Data quality concerns: sequencing, assembly and alignment error

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

Identification of genes with one-to-one homology

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

Filtering sequence regions with low sequence quality

Filtering sequence regions with high substitution counts

Filtering sequence regions with evidence of incomplete lineage sorting

%**********

Chapter 7

Conclusions

yadda yadda.

Bibliography

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073. 13

Adachi, J. & Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution*, **42**, 459–468.

Aguileta, G., Refrégier, G., Yockteng, R., Fournier, E. & Giraud, T. (2009). Rapidly evolving genes in pathogens: Methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infection, Genetics and Evolution*, **9**, 656–670. 17

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. 53, 54

Alexa, A., Rahnenführer, J. & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607. 136

Alroy, J. (1998). Cope's rule and the dynamics of body mass evolution in north american fossil mammals. *Science*, **280**, 731–734. 5

Alroy, J. (1999). The fossil record of north american mammals: Evidence for a paleocene evolutionary radiation. *Systematic Biology*, **48**, 107–118. 3

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402. 45

Anisimova, M. & Kosiol, C. (2009). Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol*, **26**, 255–71. 122

Anisimova, M., Bielawski, J. & Yang, Z. (2001a). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*, **18**, 1585–92. 122, 123

Anisimova, M., Bielawski, J.P. & Yang, Z. (2001b). Accuracy and power of the likelihood ratio test in detecting Adaptive molecular evolution. *Molecular Biology and Evolution*, **18**, 1585–1592. 13, 15, 18, 19, 37

Anisimova, M., Bielawski, J. & Yang, Z. (2002a). Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol*, **19**, 950–8. 82, 122

Anisimova, M., Bielawski, J.P. & Yang, Z. (2002b). Accuracy and power of bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution*, **19**, 950–958. 17, 19, 24, 25, 37

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. 82

Arbiza, L., Duchi, S., Montaner, D., Burguet, J., Uceda, D.P., Lucena, A.P., Dopazo, J. & Dopazo, H. (2006). Selective pressures at a codon-level predict deleterious mutations in human disease genes. *Journal of Molecular Biology*, **19**, 1390–1404. 13

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. 74

Archibald, J.D. & Deutschman, D.H. (2001). Quantitative analysis of the timing of the origin and diversification of extant placental orders. *Journal of Mammalian Evolution*, **8**, 107–124. 2

Arnheim, N. & Calabrese, P. (2009). Understanding what determines the frequency and pattern of human germline mutations. *Nature Reviews Genetics*, **10**, 478–488. 9

Averof, M., Rokas, A., Wolfe, K. & Sharp, P. (2000). Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science*, **287**, 1283–6. 103

Axelsson, E. & Ellegren, H. (2009). Quantification of adaptive evolution of genes expressed in avian brain and the population size effect on the efficacy of selection. *Mol Biol Evol*, **26**, 1073–9. 7

Bachmann, K. (1972). Genome size in mammals. *Chromosoma*, **37**, 85–93. 7

Bachtrog, D. (2008). Similar rates of protein adaptation in drosophila miranda and d. melanogaster, two species with different current effective population sizes. *BMC Evolutionary Biology*, **8**, 334. 7

Baer, C.F., Miyamoto, M.M. & Denver, D.R. (2007). Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics*, **8**, 619–631. 5

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. 87

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. 97

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. 87

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. 110, 115, 118, 130

Benner, S.A., Cohen, M.A. & Gonnet, G.H. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *Journal of Molecular Biology*, **229**, 1065–1082. 18

Benovoy, D. & Drouin, G. (2009). Ectopic gene conversions in the human genome. *Genomics*, **93**, 27–32. 146

Bininda-Emonds, O. (2007). Fast genes and slow clades: comparative rates of molecular evolution in mammals. *Evol Bioinform Online*, **3**, 59–85. 5

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. 1, 2, 3, 55

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. 50

Blake, R.D., Hess, S.T. & Nicholson-Tuell, J. (1992). The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *Journal of Molecular Evolution*, **34**, 189–200.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L. & Rubin, E.M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394. 1

Boyko, A., Williamson, S., Indap, A., Degenhardt, J., Hernandez, R., Lohmueller, K., Adams, M., Schmidt, S., Sninsky, J., Sunyaev, S., White, T., Nielsen, R., Clark, A. & Bustamante, C. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, **4**, e1000083. 81

Bromham, L. (2011). The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philos Trans R Soc Lond B Biol Sci*, **366**, 2503–13. 5

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. 66, 76

Caballero, A. (1994). Developments in the prediction of effective population size. *Heredity*, **73 (Pt 6)**, 657–79. 6

Callahan, B., Neher, R., Bachtrog, D., Andolfatto, P. & Shraiman, B. (2011). Correlated evolution of nearby residues in drosophilid proteins. *PLoS Genet*, **7**, e1001315. 97

Cambi, A. & Figdor, C. (2009). Necrosis: C-type lectins sense cell death. *Curr Biol*, **19**, R375–8. 151

Cartwright, R.A. (2009). Problems and solutions for estimating indel rates and length distributions. *Molecular Biology and Evolution*, **26**, 473–480. 18, 40

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. 87, 90, 146

Castillo-Davis, C.I., Kondrashov, F.A., Hartl, D.L. & Kulathinal, R.J. (2004). The functional genomic distribution of protein divergence in two animal phyla: Coevolution, genomic conflict, and constraint. *Genome Research*, **14**, 802–811. 1

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540–552. 16, 20

Cepas, H., Bueno, A., Dopazo, J. & Gabaldón, T. (2007). PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucl. Acids Res.*, **36**, gkm899. 46

Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*, **10**, 195–205. 6

Chen, F., Mackey, A., Vermunt, J. & Roos, D. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*, **2**, e383. 45

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. 101

Clark, A., Glanowski, S., Nielsen, R., Thomas, P., Kejariwal, A., Todd, M., Tanenbaum, D., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T., Sninsky, J., Adams, M. & Cargill, M. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–3. 91, 121, 137, 142, 143, 144

Clark, A.G. & Civetta, A. (2000). Evolutionary biology: Protamine wars. *Nature a - z index*, **403**, 261–263. 122

Clark, N. & Swanson, W. (2005). Pervasive adaptive evolution in primate seminal proteins. *PLoS Genet*, **1**, e35. 140

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. 88

Collins, F.S. & McKusick, V.A. (2001). Implications of the human genome project for medical science. *JAMA: The Journal of the American Medical Association*, **285**, 540–544. 1

TOCITE (????). Citation will be inserted at a later point in time. 11, 79

Cordaux, R. & Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, **10**, 691–703. 7

Cousins, R. (2007). Annotated bibliography of some papers on combining significances or p-values. 125

Crespi, B.J. & Summers, K. (2006). Positive selection in the evolution of cancer. *Biological Reviews*, **81**, 407–424. 140

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. 48

Darlington, R.B. & Hayes, A.F. (2000). Combining independent p values: Extensions of the stouffer and binomial methods. *Psychological Methods*, **5**, 496. 125, 126

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London. 9

Datta, R., Meacham, C., Samad, B., Neyer, C. & Sjölander, K. (2009). Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Res*, **37**, W84–9. 46

Dayhoff, M.O. & Schwartz, R.M. (1978). A model of evolutionary change in proteins. *in Atlas of Protein Sequence and Structure*.

de la Chaux, N., Messer, P.W. & Arndt, P.F. (2007). DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evolutionary Biology*, **7**, 191. 14

Dehal, P. & Boore, J. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, **3**, e314. 8, 54

Demuth, J., De Bie, T., Stajich, J., Cristianini, N. & Hahn, M. (2006). The evolution of mammalian gene families. *PLoS One*, **1**, e85. 55

Dessimoz, C. & Gil, M. (2010). Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biology*, **11**, R37. 16, 35

Do, C.B., Mahabhashyam, M.S., Brudno, M. & Batzoglou, S. (2005). ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Research*, **15**, 330–340. 20

Duret, L. & Arndt, P. (2008). The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*, **4**, e1000071.

Duret, L., Eyre-Walker, A. & Galtier, N. (2006). A new perspective on isochore evolution. *Gene*, **385**, 71–4.

Dwivedi, B. & Gadagkar, S.R. (2009). Phylogenetic inference under varying proportions of indel-induced alignment gaps. *BMC Evolutionary Biology*, **9**, 211. 16

Eddy, S. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform*, **23**, 205–11. 45

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–7. 50

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., McCune, R.A. & Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research*, **10**, 2709–2721.

Eichler, E.E. & Sankoff, D. (2003). Structural dynamics of eukaryotic chromosome evolution. *Science*, **301**, 793–797. 7

ELLEGREN, H. (2008). Comparative genomics and the study of evolution by natural selection. *Molecular Ecology*, **17**, 4586–4596. 122

Ellegren, H. (2009a). A selection model of molecular evolution incorporating the effective population size. *Evolution*, **63**, 301–305. 40

Ellegren, H. (2009b). A selection model of molecular evolution incorporating the effective population size. *Evolution*, **63**, 301–5. 7, 82, 117

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. 80, 88

Endo, T., Ikeo, K. & Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. *Molecular Biology and Evolution*, **13**, 685–690. 121

Eyre-Walker, A. & Keightley, P. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet*, **8**, 610–8. 81

Eyre-Walker, A., Keightley, P.D., Smith, N.G.C. & Gaffney, D. (2002). Quantifying the slightly deleterious mutation model of molecular evolution. *Molecular Biology and Evolution*, **19**, 2142–2149. 6

Eyre-Walker, A., Woolfit, M. & Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, **173**, 891–900. 81

Ezawa, K., OOta, S. & Saitou, N. (2006). Genome-Wide search of gene conversions in duplicated genes of mouse and rat. *Molecular Biology and Evolution*, **23**, 927–940. 146

Fay, J. & Wu, C. (2003). Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet*, **4**, 213–35. 39, 80

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376. 10

Finarelli, J.A. & Flynn, J.J. (2006). Ancestral state reconstruction of body size in the caniformia (carnivora, mammalia): The effects of incorporating data from the fossil record. *Systematic Biology*, **55**, 301–313. 5

Finn, R., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J., Gavin, O., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E., Eddy, S. & Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Res*, **38**, D211–22. 109

Fisher, R. (1932). *Statistical methods for research workers*. Oliver and Boyd, London. 125

Fletcher, W. & Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, **26**, 1879–1888. 18, 40

Fletcher, W. & Yang, Z. (2010a). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol*, **27**, 2257–67. 2, 87

Fletcher, W. & Yang, Z. (2010b). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular Biology and Evolution*, **27**, 2257–2267. 16, 17, 19, 28, 37

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. 8, 48

Galtier, N., Blier, P. & Nabholz, B. (2009). Inverse relationship between longevity and evolutionary rate of mitochondrial proteins in mammals and birds. *Mitochondrion*, **9**, 51–7. 5

Green, P. (2007a). 2x genomes–does depth matter? *Genome Res*, **17**, 1547–9. 83

Green, P. (2007b). 2x genomes: does depth matter? *Genome Research*, **17**, 1547–1549. 13

Gu, X., Wang, Y. & Gu, J. (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet*, **31**, 205–9. 90

Hahn, M., Han, M. & Han, S. (2007). Gene family evolution across 12 drosophila genomes. *PLoS Genet*, **3**, e197. 46

Halligan, D., Oliver, F., Eyre-Walker, A., Harr, B. & Keightley, P. (2010). Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet*, **6**, e1000825. 6

Han, M., Demuth, J., McGrath, C., Casola, C. & Hahn, M. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome Res*, **19**, 859–67. 92

Hasegawa, M., Kishino, H. & Yano, T.a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174. 10

Haussler, D., O'Brien, S., Ryder, O., Barker, F., Clamp, M., Crawford, A., Hanner, R., Hanotte, O., Johnson, W., McGuire, J. *et al.* (2009). Genome 10k: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J Hered*, **100**, 659–674. 3, 4

He, L., Vasiliou, K. & Nebert, D. (2009). Analysis and update of the human solute carrier (slc) gene superfamily. *Hum Genomics*, **3**, 195–206. 140

Hedges, S. & Kumar, S. (2009). *The timetree of life*. The Timetree of Life, Oxford University Press. 3, 4

Heger, A. & Ponting, C. (2008). OPTIC: orthologous and paralogous transcripts in clades. *Nucleic Acids Res*, **36**, D267–70. 46, 50, 65, 69, 74, 75, 76, 77

Henikoff, S., Ahmad, K. & Malik, H.S. (2001). The centromere paradox: Stable inheritance with rapidly evolving DNA. *Science*, **293**, 1098–1102. 141

Hillier, L., W. Miller, E. Birney et al. (178 co-authors) (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716. 18

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple significance testing. *Biometrika*, **75**, 800–803. 124

Hoffmann, J.A., Kafatos, F.C., Janeway, C.A. & Ezekowitz, R.A.B. (1999). Phylogenetic perspectives in innate immunity. *Science*, **284**, 1313–1318. 8

Hokamp, K., McLysaght, A. & Wolfe, K.H. (2003). The 2R hypothesis and the human genome sequence. *Journal of Structural and Functional Genomics*, **3**, 95–110. 8

Holmes, I. (2005). Using evolutionary expectation maximization to estimate indel rates. *Bioinformatics*, **21**, 2294–300. 40

Hou, Z., Romero, R. & Wildman, D. (2009). Phylogeny of the ferungulata (mammalia: Laurasiatheria) as determined from phylogenomic data. *Molecular phylogenetics and evolution*, **52**, 660–664. 118

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. 47, 88

Hubisz, M., Lin, M., Kellis, M. & Siepel, A. (2011). Error and error mitigation in low-coverage genome assemblies. *PLoS One*, **6**, e17034. 48, 88, 89

Hughes, A. (1999). *Adaptive evolution of genes and genomes*. Oxford University Press, USA. 121

Hughes, A.L. & Yeager, M. (1997). Molecular evolution of the vertebrate immune system. *BioEssays*, **19**, 777–786. 8

Huttley, G.A., Easteal, S., Southey, M.C., Tesoriero, A., Giles, G.G., McCredie, M.R.E., Hopper, J.L. & Venter, D.J. (2000). Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. *Nature Genetics*, **25**, 410–413. 140

Hwang, D.G. & Green, P. (2004). Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 13994–14001. 5, 11

Jaffe, D., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J., Zody, M. & Lander, E. (2003). Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res*, **13**, 91–6. 84, 88

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. 63

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences : CABIOS*, **8**, 275–282.

Joyce, S. (2001). CD1d and natural T cells: how their properties jump-start the immune system. *Cellular and Molecular Life Sciences*, **58**, 442–469. 147

Jukes, T. & Cantor, C. (1969). Evolution of protein molecules, 21–132. *Mammalian Protein Metabolism. Academic Press, New York*. 9

Jun, J., Mandoiu, I. & Nelson, C. (2009). Identification of mammalian orthologs using local synteny. *BMC Genomics*, **10**, 630. 45

Kasahara, M. (2007). The 2R hypothesis: an update. *Curr Opin Immunol*, **19**, 547–52. 8

Katoh, K., Kuma, K., Toh, H. & Miyata, T. (2005a). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, **33**, 511–8. 50

Katoh, K., Kuma, K.i., Toh, H. & Miyata, T. (2005b). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**, 511–518. 20

Kellis, M., Birren, B. & Lander, E. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, **428**, 617–24. 50

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120. 10

Kimura, M. (1985). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. 7

Kimura, M. & Ohta, T. (1974a). On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences*, **71**, 2848–2852. 7

Kimura, M. & Ohta, T. (1974b). On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences*, **71**, 2848–2852. 19

Kircher, M., Stenzel, U. & Kelso, J. (2009). Improved base calling for the illumina genome analyzer using machine learning strategies. *Genome Biol*, **10**, R83. 84

Koonin, E. & Wolf, Y. (2006). Evolutionary systems biology: links between gene evolution and function. *Curr Opin Biotechnol*, **17**, 481–7. 49

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. 52

Koonin, E.V., Makarova, K.S. & Aravind, L. (2001). HORIZONTAL GENE TRANSFER IN PROKARYOTES: Quantification and classification1. *Annual Review of Microbiology*, **55**, 709–742. 45

Kosiol, C., Holmes, I. & Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Mol Biol Evol*, **24**, 1464–79. 106

Kosiol, C., Vinar, T., da Fonseca, R., Hubisz, M., Bustamante, C., Nielsen, R. & Siepel, A. (2008a). Patterns of positive selection in six mammalian genomes. *PLoS Genet*, **4**, e1000144. 7, 82, 91, 117, 122, 123, 125, 137, 142, 143, 144

Kosiol, C., Vina, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R. & Siepel, A. (2008b). Patterns of positive selection in six mammalian genomes. *PLoS Genet*, **4**, e1000144. 17

Lander, E. (2011). Initial impact of the sequencing of the human genome. *Nature*, **470**, 187–97. 1

Lander, E. & International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921. 7, 8, 150

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. 50

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O.A., Leung, F.C.C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D.,

Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C.C., Lam, T.T.Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M.W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.W., Yiu, S.M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G.K.S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J. & Wang, J. (2009). The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317. 122

Lin, Y.S., Hsu, W.L., Hwang, J.K. & Li, W.H. (2007). Proportion of solvent-exposed amino acids in a protein and Rate of protein evolution. *Molecular Biology and Evolution*, **24**, 1005–1011. 13

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C., Mauceli, E., Xie, X., Breen, M., Wayne, R.K., Ostrander, E.A., Ponting, C.P., Galibert, F., Smith, D.R., deJong, P.J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.W., Cook, A., Cuff, J., Daly, M.J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.P., Parker, H.G., Pollinger, J.P., Searle, S.M.J., Sutter, N.B., Thomas, R., Webber, C., Baldwin, J., Broad Sequencing Platform Members & Lander, E.S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803–819. 79

Lindblad-Toh, K., Garber, M., Zuk, O. & ,et al. (64 co-authors) (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*. 3, 6, 80

Lindblad-Toh, K., M. Garber, O. Zuk et al. (64 co-authors) (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature (in press)*. 19

Loewe, L. & Charlesworth, B. (2006). Inferring the distribution of mutational effects on fitness in drosophila. *Biol Lett*, **2**, 426–30. 81

Lovelace, L.L., Cooper, C.L., Sodetz, J.M. & Lebioda, L. (2011). Structure of human C8 protein provides mechanistic insight into membrane pore formation by complement. *Journal of Biological Chemistry*, **286**, 17585–17592. 151

Löytynoja, A. & Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–1635. 15, 16, 20

Lynch, M. & Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–5. 45, 90

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. 50

Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. (2005). Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research*, **33**, D54–D58. 142

Malik, H. & Henikoff, S. (2002). Conflict begets complexity: the evolution of centromeres. *Curr Opin Genet Dev*, **12**, 711–8. 141

Malik, H. & Henikoff, S. (2009). Major evolutionary transitions in centromere complexity. *Cell*, **138**, 1067–82. 141

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. 2, 87, 102

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. 79

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. 79, 80

Markova-Raina, P. & Petrov, D. (2011a). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 drosophila genomes. *Genome Research*, **21**, 863–874. 2, 102

Markova-Raina, P. & Petrov, D. (2011b). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 drosophila genomes. *Genome Research*, **21**, 863–874. 15, 28, 37

Marques-Bonet, T., Ryder, O.A. & Eichler, E.E. (2009). Sequencing primate genomes: What have we learned? *Annual Review of Genomics and Human Genetics*, **10**, 355–386. 81

Martin, A.P. & Palumbi, S.R. (1993). Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences*, **90**, 4087–4091. 5

Martin, R., Soligo, C. & Tavaré, S. (2007). Primate origins: implications of a cretaceous ancestry. *Folia Primatol (Basel)*, **78**, 277–96. 3

Massingham, T. & Goldman, N. (2005a). Detecting amino acid sites under positive selection and purifying Selection. *Genetics*, **169**, 1753–1762. 13, 18, 19, 21, 22, 23, 25

Massingham, T. & Goldman, N. (2005b). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–62. 80, 82, 103, 108, 113, 123, 124

McLysaght, A., Hokamp, K. & Wolfe, K. (2002). Extensive genomic duplication during early chordate evolution. *Nat Genet*, **31**, 200–4. 8

Meyerson, N. & Sawyer, S. (2011). Two-stepping through time: mammals and viruses. *Trends Microbiol*, **19**, 286–94. 122

Milinkovitch, M., Helaers, R., Depiereux, E., Tzika, A. & Gabaldón, T. (2010). 2x genomes–depth does matter. *Genome Biol*, **11**, R16. 51, 63, 77

Miller, J., Koren, S. & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315–27. 61

Mironov, A., Fickett, J. & Gelfand, M. (1999). Frequent alternative splicing of human genes. *Genome Res*, **9**, 1288–93. 48

Monroe, M.J. & Bokma, F. (2010). SHORT COMMUNICATION: Little evidence for cope's rule from bayesian phylogenetic analysis of extant mammals. *Journal of Evolutionary Biology*, **23**, 2017–2021. 5

Moran, N.A., McCutcheon, J.P. & Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. *Annual Review of Genetics*, **42**, 165–190. 7

Morgan, G.J. (1998). Emile zuckerkandl, linus pauling, and the molecular evolutionary clock, 1959–1965. *Journal of the History of Biology*, **31**, 155–178. 9

Morrison, D.A. (2009). A framework for phylogenetic sequence alignment. *Plant Systematics and Evolution*, **282**, 127–149. 14

Mouse Genome Sequencing Consortium & Mouse Genome Analysis Group (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–62. 2, 5, 7, 8, 79

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., Powell, S., von Mering, C., Doerks, T., Jensen, L. & Bork, P. (2010). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res*, **38**, D190–5. 46

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. 101

Nabholz, B., Glémin, S. & Galtier, N. (2008). Strong variations of mitochondrial mutation rate across mammals–the longevity hypothesis. *Molecular Biology and Evolution*, **25**, 120–130. 5

Nei, M., Suzuki, Y. & Nozawa, M. (2010). The neutral theory of molecular evolution in the genomic era. *Annual Review of Genomics and Human Genetics*, **11**, 265–289. 18

Nielsen, R. (2005). MOLECULAR SIGNATURES OF NATURAL SELECTION. *Annual Review of Genetics*, **39**, 197–218. 139

Nielsen, R. & Yang, Z. (2003). Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol*, **20**, 1231–9. 82

Nielsen, R., Bustamante, C., Clark, A., Glanowski, S., Sackton, T., Hubisz, M., Fledel-Alon, A., Tanenbaum, D., Civello, D., White, T., J Sninsky, J., Adams, M. & Cargill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*, **3**, e170. 91, 121, 139, 140, 142, 143, 144

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. & Clark, A. (2007). Recent and ongoing selection in the human genome. *Nat Rev Genet*, **8**, 857–68. 1

Nikolaev, S., Montoya-Burgos, J., Popadin, K., Parand, L., Margulies, E., National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program & Antonarakis, S. (2007a). Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc Natl Acad Sci U S A*, **104**, 20443–8. 101, 117

Nikolaev, S., Montoya-Burgos, J.I., Margulies, E.H., Program, N.C., Rougemont, J., Nyffeler, B. & Antonarakis, S.E. (2007b). Early history of mammals is elucidated with the ENCODE multiple species sequencing data. *PLoS Genetics*, **3**, e2. 18, 19

Noble, W.S. (2009). How does multiple testing correction work? *Nature Biotechnology*, **27**, 1135–1137. 127

Nonaka, M. & Kimura, A. (2006). Genomic view of the evolution of the complement system. *Immunogenetics*, **58**, 701–713. 147

North, B., Curtis, D. & Sham, P. (2002). A note on the calculation of empirical P values from monte carlo procedures. *Am J Hum Genet*, **71**, 439–41. 127

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology*, **3**, e123. 14

Notredame, C. & Abergel, C. (2003). Using multiple alignment methods to assess the quality of genomic data analysis. *Bioinformatics and Genomes: Current Perspectives. Horizon Scientific Press, Wymondham, UK*, 30–55. 20, 21

Notredame, C., Higgins, D. & Heringa, J. (2000a). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**, 205–17. 50

Notredame, C., Higgins, D.G. & Heringa, J. (2000b). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**, 205–217. 16, 20

O'Brien, S., Menotti-Raymond, M., Murphy, W., Nash, W., Wienberg, J., Stanyon, R., Copeland, N., Jenkins, N., Womack, J. & Marshall Graves, J. (1999). The promise of comparative genomics in mammals. *Science*, **286**, 458–62, 479–81. 1

Ogden, T.H. & Rosenberg, M.S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology*, **55**, 314–328. 16

Ogurtsov, A.Y., Sunyaev, S. & Kondrashov, A.S. (2004). Indel-based evolutionary distance and mouse-human divergence. *Genome Research*, **14**, 1610–1616. 18

Ohno, S. (1970). *Evolution by gene duplication.*. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag. 8, 45, 146

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, **23**, 263–286. 7, 118

Olszewski, M.A., Gray, J. & Vestal, D.J. (2006). In silico genomic analysis of the human and murine Guanylate-Binding protein (gbp) gene clusters. *Journal of Interferon & Cytokine Research*, **26**, 328–352. 147

Pál, C., Papp, B. & Lercher, M. (2006). An integrated view of protein evolution. *Nat Rev Genet*, **7**, 337–48. 80

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. 48

Penn, O., Privman, E., Landan, G., Graur, D. & Pupko, T. (2010). An alignment confidence score capturing robustness to guide tree uncertainty. *Molecular Biology and Evolution*, **27**, 1759–1767. 16, 20

Pevzner, P. & Tesler, G. (2003). Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes. *Genome Research*, **13**, 37–45. 7

Pollard, K., Hubisz, M., Rosenbloom, K. & Siepel, A. (2010). Detection of non-neutral substitution rates on mammalian phylogenies. *Genome Res*, **20**, 110–21. 56

Ponting, C.P. & Hardison, R.C. (2011). What fraction of the human genome is functional? *Genome Research*, **21**, 1769–1776. 1

Popadin, K., Polishchuk, L., Mamirova, L., Knorre, D. & Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc Natl Acad Sci U S A*, **104**, 13390–5. 6, 117

Privman, E., Penn, O. & Pupko, T. (2011). Improving the performance of positive selection inference by filtering unreliable alignment regions. *Molecular Biology and Evolution*. 16, 17, 20, 28, 37, 39

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. 49

Putnam, N.H., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.K.,

Benito-Gutiérrez, E., Dubchak, I., Garcia-Fernàdez, J., Gibson-Brown, J.J., Grigoriev, I.V., Horton, A.C., de Jong, P.J., Jurka, J., Kapitonov, V.V., Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L.A., Salamov, A.A., Satou, Y., Sauka-Spengler, T., Schumtz, J., Shin-I, T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L.Z., Holland, P.W.H., Satoh, N. & Rokhsar, D.S. (2008). The amphioxus genome and the evolution of the chordate karyotype. **453**. 8

Ramsey, D.C., Scherrer, M.P., Zhou, T. & Wilke, C.O. (2011). The relationship between relative solvent accessibility and evolutionary Rate in protein evolution. *Genetics*, **188**, 479–488. 13

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. 50

Rat Genome Sequencing Project Consortium (2004). Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521. 79

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. 87

Remm, M., Storm, C. & Sonnhammer, E. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, **314**, 1041–52. 45, 46

Rhesus Macaque Genome Sequencing and Analysis Consortium (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–34. 8, 121, 136, 137, 142, 143

Rivals, I., Personnaz, L., Taing, L. & Potier, M.C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407. 136

Romiguier, J., Ranwez, V., Douzery, E. & Galtier, N. (2010). Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res*, **20**, 1001–9. 5

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. 46, 50, 51

Sawyer, S.L., Wu, L.I., Emerman, M. & Malik, H.S. (2005). Positive selection of primate trim5α identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 2832–2837. 1, 123, 125, 140

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. & Graur, D. (2009a). Estimates of positive darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol*, **1**, 114–8. 2, 86, 87, 102

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G.H. & Graur, D. (2009b). Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biology and Evolution*, **1**, 114–118. 15

Schueler, M.G., Swanson, W., Thomas, P.J., Program, N.C.S. & Green, E.D. (2010). Adaptive evolution of foundation kinetochore proteins in primates. *Molecular Biology and Evolution*, **27**, 1585–1597. 141

Siepel, A. & Haussler, D. (2004). Phylogenetic estimation of Context-Dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, **21**, 468–488.

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. 1, 50

Siepel, A., G. Bejerano, J. S. Pedersen et al. (16 co-authors) (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, **15**, 1034–1050. 18

Sipos, B., Massingham, T., Jordan, G. & Goldman, N. (2011). Phylosim–monte carlo simulation of sequence evolution in the r statistical computing environment. *BMC Bioinformatics*, **12**, 104. 41

Sjölander, K., Datta, R., Shen, Y. & Shoffner, G. (2011). Ortholog identification in the presence of domain architecture rearrangement. *Brief Bioinform*, **12**, 413–22. 45

Sluijter, J.P., deKleijn, D.P. & Pasterkamp, G. (2006). Vascular remodeling and protease inhibition—bench to bedside. *Cardiovascular Research*, **69**, 595–603. 148

Smith, F.A., Boyer, A.G., Brown, J.H., Costa, D.P., Dayan, T., Ernest, S.K.M., Evans, A.R., Fortelius, M., Gittleman, J.L., Hamilton, M.J., Harding, L.E., Lintulaakso, K., Lyons, S.K., McCain, C., Okie, J.G., Saarinen, J.J., Sibly, R.M., Stephens, P.R., Theodor, J. & Uhen, M.D. (2010). The evolution of maximum body size of terrestrial mammals. *Science*, **330**, 1216–1219. 3

Smith, J.M. (1970). Natural selection and the concept of a protein space. *Nature*, **225**, 563–564. 19

Smith, S. & Donoghue, M. (2008). Rates of molecular evolution are linked to life history in flowering plants. *Science*, **322**, 86–9. 101

Stein, J.V. & Nombela-Arrieta, C. (2005). Chemokine control of lymphocyte trafficking: a general overview. *Immunology*, **116**, 1–12. 151

Storz, J., Hoffmann, F., Opazo, J. & Moriyama, H. (2008). Adaptive functional divergence among triplicated alpha-globin genes in rodents. *Genetics*, **178**, 1623–38. 87

Stouffer, S., DeVinney, L. & Suchmen, E. (1949). *The American soldier: Adjustment during army life*, vol. 1. Princeton University Press, Princeton, NJ. 125

Studer, R., Duret, L., Penel, S. & Rechavi, M.R. (2008a). Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Research*, **18**, 1393–1402. 17

Studer, R., Penel, S., Duret, L. & Robinson-Rechavi, M. (2008b). Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res*, **18**, 1393–402. 87

Suraweera, A., Becherel, O.J., Chen, P., Rundle, N., Woods, R., Nakamura, J., Gatei, M., Criscuolo, C., Filla, A., Chessa, L., Fusser, M., Epe, B., Gueven, N. & Lavin, M.F. (2007). Senataxin, defective in ataxia oculomotor apraxia type 2, is involved in the defense against oxidative DNA damage. *The Journal of Cell Biology*, **177**, 969–979. 140

Swanson, W.J., Nielsen, R. & Yang, Q. (2003). Pervasive adaptive evolution in mammalian fertilization proteins. *Molecular Biology and Evolution*, **20**, 18–20. 140

Talavera, G. & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, **56**, 564–577. 16

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, **10**, 512–526. 10

Tavaré, S. (1986). *Some mathematical questions in biology: DNA sequence analysis*. Lectures on mathematics in the life sciences, American Mathematical Society. 10

Teytelman, L., Ozaydin, B., Zill, O., Lefrançois, P., Snyder, M., Rine, J. & Eisen, M. (2009). Impact of chromatin structures on DNA processing for genomic analyses. *PLoS One*, **4**, e6700. 89

The 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. 13, 18, 19

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680. 20

Torgerson, D.G., Kulathinal, R.J. & Singh, R.S. (2002). Mammalian sperm proteins are rapidly evolving: Evidence of positive selection in functionally diverse genes. *Molecular Biology and Evolution*, **19**, 1973–1980. 140

Van de Peer, Y., Maere, S. & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, **10**, 725–732. 8

Varmus, H. (2010). Ten years on—the human genome and medicine. *New England Journal of Medicine*, **362**, 2028–2029. 1

Venditti, C., Meade, A. & Pagel, M. (2011). Multiple routes to mammalian diversity. *Nature*. 1

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. 46, 47, 48, 50, 51, 76

Vilella, A., Birney, E., Flicek, P. & Herrero, J. (2011). Considerations for the inclusion of 2x mammalian genomes in phylogenetic analyses. *Genome Biol*, **12**, 401. 51

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. 50

Wang, Y. & Gu, X. (2001). Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics*, **158**, 1311–20. 87

Warnecke, T. & Rocha, E. (2011). Function-specific accelerations in rates of sequence evolution suggest predictable epistatic responses to reduced effective population size. *Mol Biol Evol*, **28**, 2339–49. 7

Warren, W.C., Hillier, L.W., Marshall Graves, J.A., Birney, E., Ponting, C.P., Gr—[uuml]—tzner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A.T., Yang, S.P., Heger, A., Locke, D.P., Miethke, P., Waters, P.D., Veyrunes, F., Fulton, L., Fulton, B., Graves, T., Wallis, J., Puente, X.S., L—[oacute]—pez-Ot—[iacute]—n, C., Ord—[oacute]——[ntilde]—ez, G.R., Eichler, E.E., Chen, L., Cheng, Z., Deakin, J.E., Alsop, A., Thompson, K., Kirby, P., Papenfuss, A.T., Wakefield, M.J., Olender, T., Lancet, D., Huttley, G.A., Smit, A.F.A., Pask, A., Temple-Smith, P., Batzer, M.A., Walker, J.A., Konkel, M.K., Harris, R.S., Whittington, C.M., Wong, E.S.W., Gemmell, N.J., Buschiazzo, E., Vargas Jentzsch, I.M., Merkel, A., Schmitz, J., Zemann, A., Churakov, G., Kriegs, J.O., Brosius, J., Murchison, E.P., Sachidanandam, R., Smith, C., Hannon, G.J., Tsend-Ayush, E., McMillan, D., Attenborough, R., Rens, W., Ferguson-Smith, M., Lef—[egrave]—vre, C.M., Sharp, J.A., Nicholas, K.R., Ray, D.A., Kube, M., Reinhardt, R., Pringle, T.H., Taylor, J., Jones, R.C., Nixon, B., Dacheux, J.L., Niwa, H., Sekita, Y., Huang, X., Stark, A., Kheradpour, P., Kellis, M., Flicek, P., Chen, Y., Webber, C., Hardison, R., Nelson, J., Hallsworth-Pepin, K., Delehaunty, K., Markovic, C., Minx, P., Feng, Y., Kremitzki, C., Mitreva, M., Glasscock, J., Wylie, T.,

Wohldmann, P., Thiru, P., Nhan, M.N., Pohl, C.S., Smith, S.M., Hou, S., Renfree, M.B., Mardis, E.R. & Wilson, R.K. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, **453**, 175–183. 8

Welch, J., Bininda-Emonds, O. & Bromham, L. (2008). Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evol Biol*, **8**, 53. 5

Whelan, S. (2008a). Spatial and temporal heterogeneity in nucleotide sequence evolution. *Molecular Biology and Evolution*, **25**, 1683–1694. 39

Whelan, S. (2008b). Spatial and temporal heterogeneity in nucleotide sequence evolution. *Mol Biol Evol*, **25**, 1683–94. 62

Whelan, S. & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a Maximum-Likelihood approach. *Molecular Biology and Evolution*, **18**, 691–699.

Whelan, S. & Goldman, N. (2004). Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*, **167**, 2027–43. 103

Whelan, S., Liò, P. & Goldman, N. (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet*, **17**, 262–72. 10

Whitlock, M. (2005). Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *J Evol Biol*, **18**, 1368–73. 125

Wilson, D. & Reeder, D. (2005). *Mammal Species of the World: A Taxonomic and Geographic Reference*. No. v. 1 in Mammal Species of the World: A Taxonomic and Geographic Reference, Johns Hopkins University Press. 2

Wolf, J.B., Künstner, A., Nam, K., Jakobsson, M. & Ellegren, H. (2009). Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. *Genome Biology and Evolution*, **1**, 308–319. 13

Wong, K.M., Suchard, M.A. & Huelsenbeck, J.P. (2008). Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476. 15

Woolfit, M. (2009). Effective population size and the rate and pattern of nucleotide substitutions. *Biol Lett*, **5**, 417–20. 6

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, **16**, 97–159. 6

Wyckoff, G.J., Wang, W. & Wu, C.I. (2000). Rapid evolution of male reproductive genes in the descent of man. *Nature*, **403**, 304–309. 122

Yang, W., Bielawski, J. & Yang, Z. (2003). Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *Journal of Molecular Evolution*, **57**, 212–221, 10.1007/s00239-003-2467-9. 17

Yang, Z. (2005). The power of phylogenetic comparison in revealing protein function. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 3179–3180. 122

Yang, Z. (2006). *Computational Molecular Evolution (Oxford Series in Ecology and Evolution)*. Oxford University Press, USA. 9, 10

Yang, Z. (2007a). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591. 21

Yang, Z. (2007b). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**, 1586–91. 97

Yang, Z. & Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, **46**, 409–418, 10.1007/PL00006320. 20

Yang, Z., Kumar, S. & Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**, 1641–50. 97

Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A.M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449. 18, 22

Yang, Z., Wong, W.S.W. & Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, **22**, 1107–1118. 22, 23, 25

Yuan, Y., Eulenstein, O., Vingron, M. & Bork, P. (1998). Towards detection of orthologues in sequence databases. *Bioinformatics*, **14**, 285–9. 45

Zaykin, D., Zhivotovsky, L., Westfall, P. & Weir, B. (2002). Truncated product method for combining p-values. *Genet Epidemiol*, **22**, 170–85. 125, 126

Zaykin, D., Zhivotovsky, L., Czika, W., Shao, S. & Wolfinger, R. (2007). Combining p-values in large-scale genomics experiments. *Pharm Stat*, **6**, 217–26. 126

Zhang, J., Zhang, Y. & Rosenberg, H. (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet*, **30**, 411–5. 122

Zhang, J., Nielsen, R. & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, **22**, 2472–9. 122

Zhao, H. & Bourque, G. (2009). Recovering genome rearrangements in the mammalian phylogeny. *Genome Research*, **19**, 934–942. 7

Zhu, J., He, F., Hu, S. & Yu, J. (2008). On the nature of human housekeeping genes. *Trends Genet*, **24**, 481–4. 49

Zuckerkandl, E. & Pauling, L. (1962). Molecular disease, evolution, and genie diversity. *Horizons in biochemistry*, 189–225. 9

Zuckerkandl, E. & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. 9