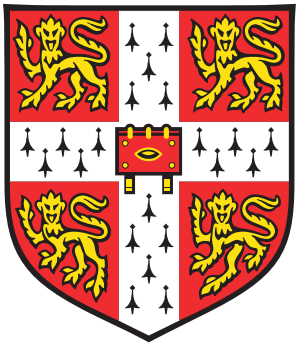
Analysis of alignment error and sitewise constraint in mammalian comparative genomics



Gregory Jordan

European Bioinformatics Institute

University of Cambridge

A dissertation submitted for the degree of

Doctor of Philosophy

November 30, 2011

To my parents, who kept us thinking and playing

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text and acknowledgements.

This dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university, and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This dissertation does not exceed the specified length limit of 60,000 words as defined by the Biology Degree Committee.

November 30, 2011 Gregory Jordan

Analysis of alignment error and sitewise constraint in mammalian comparative genomics

Summary

Gregory Jordan
November 30, 2011 Darwin College

Insight into the evolution of protein-coding genes can be gained from the use of phylogenetic codon models. Recently sequenced mammalian genomes and powerful analysis methods developed over the past decade provide the potential to globally measure the impact of natural selection on protein sequences at a fine scale. The detection of positive selection in particular is of great interest, with relevance to the study of host-parasite conflicts, immune system evolution and adaptive differences between species. This thesis examines the performance of methods for detecting positive selection first with a series of simulation experiments, and then with two empirical studies in mammals and primates.

Our ability to make confident estimates of the prevalence of positive selection in proteins has been hampered to some extent by uncertainty regarding the level of false positives resulting from alignment error. To this end, I conduct a simulation study to estimate the rate of false positive results attributable to alignment error. A variety of aligners and alignment filtering methods are compared, showing a striking difference between aligners in their tendency to produce false positive results. Under most conditions, the best aligners tend to produce very few false positives due to misalignment.

The rest of this thesis focuses on two genome-wide studies: an analysis of sitewise selective pressures across 38 mammalian genomes, and a genome-wide scan for genes with evidence of accelerated evolution in gorilla and the African great apes. In the broader mammalian analysis, the global distribution of sitewise evolutionary constraint is characterized and strong evidence is presented for less positive selection in the protein-coding genes of rodents compared to primates and other mammalian orders. New methods are developed for combining sitewise estimates across genes and protein-coding domains, revealing widespread signals of positive selection in genes and domains related to host defense and, surprisingly, centromere binding. The African great ape analysis uses phylogenetic codon models to identify genes which have experienced elevated evolutionary rates in gorilla, human and chimpanzee. Similar numbers of accelerated genes are identified in each of these genomes, and several accelerated genes are identified in gorilla with plausible relationships to its unique phenotypic and behavioral characteristics. Finally, genome-wide coding alignments are used to infer genome-wide selective pressures along each branch of the great ape tree, providing corroborating evidence of a trend towards decreasing population sizes in the recent evolutionary history of African great apes.

Acknowledgements

First and foremost I would like to thank Nick Goldman, who supervised my Ph.D. research at the EBI. I arrived with little knowledge and even less experience; it is a testament to Nick's patience, encouragement, and unequivocal support that I arrived at the other end some sort of scientist. Thanks especially for giving me the freedom to make my own mistakes—and for having the kindness to help mend them.

For enlightening my mind and enriching my spirits at work and beyond, I have the past, former and current members of the Goldman Group to thank. One would have some difficulty casting a more insightful, helpful, and playfully skeptical ensemble. The characters, in order of appearance: Tim, Ari, Martin, Fabio, Jacky, Stefan, Emeric, Botond, and Hazel (plus the cameos and guest appearances).

My work and my scientific growth have benefitted greatly from fruitful discussions and productive collaborations with several others who got caught in my academic web. My thesis advisory committee (Ewan Birney, Nick Mundy, and Jan Korbel) provided useful feedback and well-tempered advice; members of the Ensembl Compara team (Albert Vilella, Javier Herrero) kindly introduced me to their systems and taught me invaluable "farming techniques"; and other collaborators (including Stephen Montgomery, Aylwyn Scally, Tim O'Connor, Amalio Telenti, David Aanansen and the members of the Mammalian Genome Project Consortium) have together taught me innumerable lessons about how to do good work and play nice with others.

To my family, I owe everything for their love and support over the past years. I couldn't have dreamt of feeling so close to home while living thousands of miles away. Sure, Skype helped, but it was a feeling much more magical than the Internet that kept me close.

Throughout the years I've been blessed with great friends, who together made sure life in Cambridge never lost its lustre. My adopted Gates "family" embraced me with entertaining dinners, exciting trips and a new bunch of lifelong friends. Juggling pals Ben and Guy allowed me a curious peek into the world of May ball entertainment. And my amazing housemates, the fearless tenants of the pink house on Coleridge

Road—Niko, Lesley, Loizos, Jan, Felix, Tom and Karolina through the years—have always been there when I most needed to laugh, commisserate, eat good food, or enjoy yet another alphabetical party.

Finally, the EBI Predocs & Friends have been there all along, a core to my life in so many ways. The gang of 2007 predocs—Michele, Markus, Adam, Diva, Julia and Judith—have been constant companions from Heidelberg to Cambridge, as well as reliable scientists, friends, and travelers on our various acronymic holidays. To all the others, I am thankful for the never-ending mailing list hijinks, the 24-hour support network (for computers and for life) and the endless good cheer. It's a rare bunch that can keep you looking forward to what's next even after four years. And I would like to extend a very special thank-you to Anna, who as a friend and so much more has filled this "thesis year" of our lives with a much-needed blend of caring, encouragement and fun.

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. medicine that was optimistically predicted during the unveiling of the draft human genome sequence is still far from being realized [Collins and 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 and 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, certain evolutionary patterns can highlight genes and pathways involved in ongoing, universal mammalian genetic conflicts—for instance, genes related to host immune defense and reproduction [Castillo-Davis *et al.*, 2004]. On the other hand, when strong selective pressures are observed in just one or a few lineages, they may indicate more specific adaptations related to those species' unique evolutionary history [Messier and Stewart, 1997; Sawyer *et al.*, 2005; Nielsen *et al.*, 2007].

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 may combine to make it difficult to assess accuracy or to distinguish anomalous results from interesting biological signals [Mallick *et al.*, 2009; Schneider *et al.*, 2009; Fletcher and Yang, 2010; Markova-Raina and Petrov, 2011]. 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 contribute to our understanding of the impact of some sources of error on large-scale comparative analyses and to further develop methods for appropriately predicting and handling such error.

Although each subsequent chapter contains its own short introduction, this chapter presents some of the key concepts and methods which are recurrent throughout the thesis or provide an appropriate historical background. Section 1.1 first introduces the biology and evolution of the mammals, highlighting features of their evolutionary history which are important to the study of their genomes. Section 1.2 then presents a brief account of the development of mathematical models of sequence evolution and their application to the comparative analysis of DNA and protein sequences. The development of the field is traced from the first comparisons of amino acid sequences in the early 1960s through to the introduction and popularization of codon-based models in the 1990s and 2000s; along the way, key concepts such as the distinction between purifying, neutral and positive selection are introduced and several methods and techniques used throughout the remaining chapters are described.

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 (and continues to evolve) [Waterston *et al.*, 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. A brief review of some aspects of the evolutionary history of mammals and their genomes will thus provide some 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 and Reeder, 2005]. According to a comprehensive supertree constructed by Bininda-Emonds *et al.* [2007] 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 and 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, when up to two-thirds of all mammalian species went extinct [Alroy, 1999]. Most mammalian lineages experienced decreased diversification levels (defined by Bininda-Emonds *et al.* [2007] as the difference between the per-lineage rates of speciation and extinction) for 10 Myr after the K/T extinction event, after which point they continued to diversify at a relatively constant rate up to modern times [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 and Kumar [2009], disagree with the above description based on Bininda-Emonds *et al.* [2007]. This reflects the large amount of uncertainty regarding the dates of the earliest 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 (but note that the phylogeny 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 five species spanning 45 Myr. These features of the mammalian phylogeny make it well-suited for large-scale comparative analysis, as long evolutionary branches separating sequences, which are a major source of alignment error and of uncertainty in evolutionary estimates, 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 [Martin *et al.*, 2007; 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

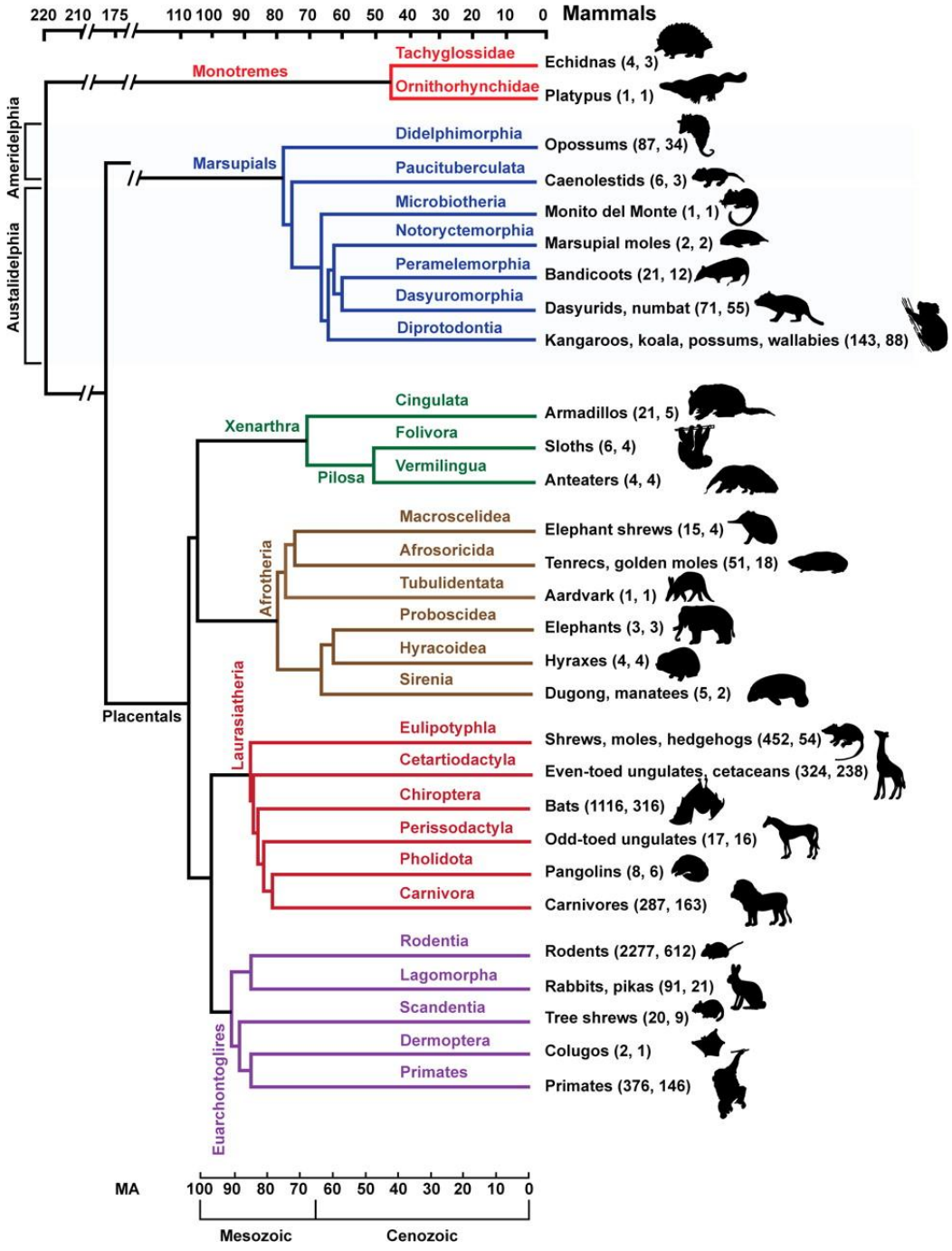


Figure 1.1: A time-resolved consensus phylogeny of the major mammalian lineages. Topologies and dates use data from Hedges and Kumar [2009]. Each terminal branch represents a mammalian family. The number of species contained in each family is included as the first number in parentheses after each family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

highly specialized morphological and behavioral adaptations and 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 and Flynn, 2006; Monroe and 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 [Waterston *et al.*, 2002; Hwang and 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 and 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 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 represent the neutral evolutionary rate (defined as the rate of evolution in genomic regions not subject to the pressure of natural selection; the distinction between neutrally and non-neutrally evolving sequence is covered in more detail in Section 1.2). Figure 1.2 shows a phylogenetic tree for 29 mammals where branch lengths correspond to the genome-wide mean expected number of substitutions per site within neutrally-evolving regions. This scaling emphasizes the high observed substitution rates of most rodents and the low rates of most hominids and some larger-bodied species from other mammalian orders. In comparative analyses, where a larger number of substitutions generally increases the power of a method to detect a genomic feature or estimate an evolutionary rate (as is the case for detecting conserved regulatory elements or positively-selected genes), the larger branch lengths of smaller-bodied 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. It should be noted, however, that in some cases increased divergence levels may not result in increased power. When sequences are extremely divergent, the accurate estimation of distances becomes difficult and some methods may suffer as a result; this effect is sometimes referred to as "saturation" of sites. Additionally, sequences separated by greater diver-

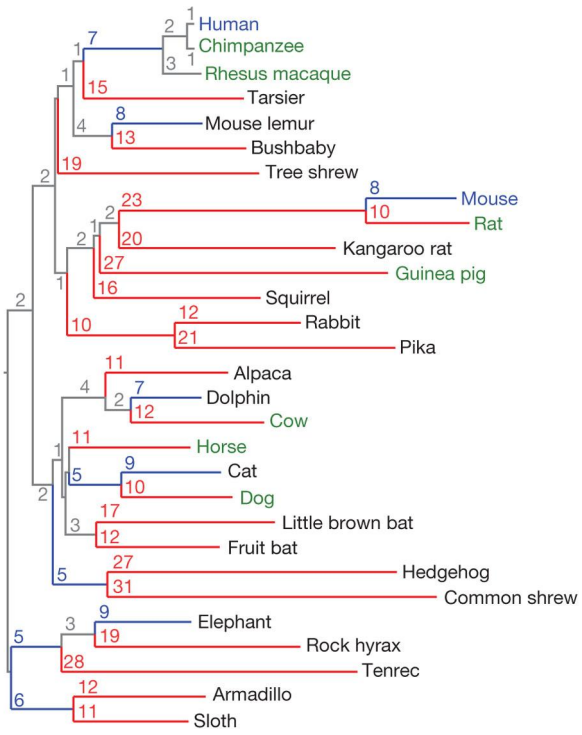


Figure 1.2: A phylogeny of 29 mammalian species, with branch lengths representing the neutral evolutionary rate estimated from genome-wide DNA alignments. Organisms with finished genome sequences are labeled in blue; those with high quality drafts are labeled in green; and those with low-coverage 2x assemblies are labeled in black. The number above each branch corresponds to the number of substitutions per 100bp along that branch. Branches with $\geq$10 substitutions are colored red, branches with 5–9 substitutions are colored blue, and branches with $<$5 substitutions are colored gray. 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].

gence levels may have experienced greater numbers of independent biological sequence insertions or deletions. These events can be difficult to accurately reconstruct when aligning sequences, and errors in the construction of alignments may also reduce power; this effect is explicitly tested in Chapter 2 for the sitewise detection of positive selection.

A second biological characteristic showing significant variation between mammals, the effective population size (N_e), has important consequences for the study of genomic regions 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]. The N_e is generally smaller than the census population size; the magnitude of this dif-

ference depends on how strongly the assumptions of the Fisher-Wright model of an idealized population—which include random mating, an equal sex ratio, non-overlapping generations and a Poisson-distributed number of offspring—are violated. As a result, many aspects of a natural 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 vary significantly 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. As it relates to this thesis, N_e can be viewed as a measurement of the influence of genetic drift, defined as the tendency for the frequencies of alleles within a population to change over time due to random sampling, within a population.

The main predicted impact of N_e on the study of fixed substitutions between species is that some slightly deleterious mutations are more likely to become fixed within a population having a small N_e versus a population with a large N_e. This results from the differential influence of genetic drift: in a population with large N_e, natural selection acts efficiently to weed out alleles containing slightly deleterious mutations, whereas a population with small N_e is more subject to the sampling effects of genetic drift and is more likely to see slightly deleterious mutations reach 100% frequency within the population (i.e. become fixed). 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 and Ohta, 1974; 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 and Rocha, 2011], birds [Axelsson and Ellegren, 2009] and mammals [Kosiol *et al.*, 2008; Ellegren, 2009] (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 dN/dS ratio estimates (a fundamental measurement in the study of protein evolution which will be introduced in Section 1.3), to compare the relative efficacy of natural selection (and through the prediction of the nearly neutral theory, the ancestral N_e) between our closest primate relatives and their ancestral lineages.

Some key features of the mammalian genome itself are also worth highlighting. Mammals contain relatively large genomes (containing 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 and Sankoff,

2003; Pevzner and Tesler, 2003]. Some regions termed "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 which are found in mammals [Eichler and Sankoff, 2003]. Breakpoints of mammalian chromosomal rearrangements tend to occur near transposable elements [Zhao and Bourque, 2009], which are small DNA sequences capable of replicating throughout the genome [Lander *et al.*, 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 significantly the ancient and ongoing evolution of mammalian genomes [Lander *et al.*, 2001; Cordaux and 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 gene [Lander *et al.*, 2001; Waterston *et al.*, 2002] in the human and mouse genomes have been lowered based on accumulating functional and phylogenetic evidence to roughly 21,000 genes [Gibbs *et al.*, 2007]; the most recent gene annotations from Ensembl [Flicek *et al.*, 2011] contain 20,599 human and 21,873 mouse "known" protein-coding genes. A majority of these genes are shared between all mammals: human and mouse share an estimated 80% of genes in a "one-to-one" fashion, meaning no apparent gene duplications or deletions occurred since the common ancestor [Waterston *et al.*, 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 *et al.*, 2001].

In any study of vertebrate protein-coding genes, the two rounds of genome duplication (2R) hypothesis looms large. Originally proposed by Ohno [1970], the 2R hypothesis suggests that two polyploidization 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 and 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 and 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 previous section described the major genomic and evolutionary features of mammalian species. It is important to note that the majority of those well-established observations were made by fitting mathematical models of sequence evolution to comparative genomics data, which is now a standard analytical approach. This section briefly introduces the methods and models of 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, special segments of DNA are replicated and passed on to all its descendants; importantly, the processes of DNA replication and repair are imperfect [Arnheim and Calabrese, 2009] and the resulting errors, called mutations, can be passed on to successive generations if they occur in germline cells. 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. Mutations arise in individuals, are passed on to descendants through DNA replication, and subsequently increase or decrease in their frequency within a population over time due to the survival or death of individuals containing that mutation. Sometimes a mutation reaches 100% frequency within the population, at which point it has become "fixed", or shared between all individuals of a population.

The fixation of mutations within independently evolving populations produces observed differences in the DNA sequences of different species at homologous locations. The most commonly observed type of fixed DNA mutation is where one nucleotide base is substituted for another; these differences, called point mutations or substitutions, can be reasonably modeled using phylogenetic trees and Markov models of sequence evolution [Yang]. Other common mutation types, however, are less amenable to modeling. Short sequence insertions or deletions, which occur roughly 10% as frequently as base substitutions, result in either the loss of genetic information from a lineage (in the case of a deletion) or the incorporation of genetic information which does not share a simple common ancestor with other lineages (in the case of an insertion). As a result, sequence insertions and deletions are not as easily modeled as base substitutions. The presence of many insertions and deletions within related sequences can make the assignment of homology between sequence positions—a process referred to as alignment—difficult. This thesis explores some aspects of alignment error on downstream evolutionary analyses, especially in Chapter 2. Other possible types of mutations, including chromosomal rearrangements, large-scale duplications and deletions of genomic regions, and horizontal gene transfer, are also important in genome evolution but will not be studied extensively in this thesis.

The rest of this section will introduce the many mathematical models constructed to describe the accumulation of point mutations over time. The earliest observations that biological sequences

tend to change randomly over time were made from sequences of proteins, the main molecules of cellular machinery comprised of amino acid units whose arrangement is encoded in the DNA sequences of exons within genes. In the early 1960s, Zuckerkandl and Pauling were analyzing the amino acid sequences of hemoglobin genes from various species. They noted 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 and Pauling, 1962; Morgan, 1998]. Zuckerkandl and Pauling continued to explore the implications and applications of this "molecular evolutionary clock" hypothesis, using hemoglobin and cytochrome C sequences to estimate the date of human-gorilla divergence (at 11 million years) and to infer the protein sequences of mammalian ancestors [Zuckerkandl and Pauling, 1965]. A wide variety of evolutionary models has subsequently been developed to describe observed patterns of amino acid and DNA substitutions. 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 provides a comprehensive mathematical treatment of the main models used in practice.

The simplest Markov model for DNA substitution, proposed by Jukes and Cantor [1969], assumes that every nucleotide has the same rate of changing into any other nucleotide. Although the assumption of equal rates is a reasonable starting point for modeling a random process, the 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), making the existence of biases towards or against certain types of mutations highly plausible. This was quickly discovered to be the case: analysis of the ever-increasing number of available biological DNA sequences showed that in most datasets *transitions*, defined as substitutions between two pyrimidine nucleotides (i.e., T→C or C→T) or between two purine nucleotides (i.e., 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, K80 extends JC69 by incorporating an additional parameter, κ, referred to as the transition/transversion ratio, representing the ratio of the rate of transition substitutions to 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 [Brown *et al.*, 1982]. The κ parameter of K80 a prototypical example of the parametric approach to building evolutionary models, whereby a parameter is introduced into the model which allows for a commonly-violated assumption of the simpler model to be relaxed. Note that the value of the parameter is not specified in the model; rather, it must be provided or estimated from the data on a case-by-case basis, usually by maximum likelihood (ML) estimation [Whelan *et al.*, 2001].

Several nucleotide models were subsequently described which relax various further assump-

tions of the JC69 and K80 models [Whelan *et al.*, 2001; Yang]. 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 mutational biases), making it inappropriate to assume an equal base composition [Yang]. Thus, models such as FEL [Felsenstein, 1981], HKY [Hasegawa *et al.*, 1985] and TN93 [Tamura and Nei, 1993] were developed to allow for various combinations of unequal base frequencies and unequal transition/transversion ratios. The most general reversible nucleotide model, REV, includes three parameters to describe the equilibrium nucleotide frequencies (where the frequency of the fourth nucleotide is determined by the requirement that all frequencies must sum to 1) and six rate parameters, one for each possible pair of substitutions [Tavaré, 1986].

In contrast to the primarily parametric DNA models, evolutionary models for amino acids have generally been estimated empirically [Whelan *et al.*, 2001]. The JTT [Jones *et al.*, 1992] and Dayhoff [Dayhoff and Schwartz, 1978] amino acid models were estimated using parsimony-based counting methods, using substitutions inferred from closely-related sequences to fill the entries of the 20x20 amino acid substitution matrix; these counts were then used to estimate reversible Markov substitution models. More recently, empirical amino acid models were estimated using ML methods, which improved upon a number of methodological deficiencies of the parsimony approach [Adachi and Hasegawa, 1996; Whelan and Goldman, 2001; Le *et al.*, 2008].

At this point it should be pointed out that 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, and (c) related by the same underlying phylogenetic tree. These assumptions are clearly violated in many real datasets, so I will briefly review the development of models which relax them to various degrees.

Independence between sites is generally a difficult assumption to relax for computational reasons [Kosiol *et al.*, 2006], 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 and Green, 2004; Siepel and 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 (whether methylated or not),

cytosine deamination produces uracil, which is removed from DNA strands by the enzyme uracil glycosylase, allowing for DNA repair mechanisms to replace the original cytosine. On the other hand, the deamination of 5-methylcytosine produces thymidine, which is not efficiently repaired and results in frequent C→T transitions [Ehrlich *et al.*, 1982; Hwang and 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 and Arndt, 2008].

The assumption of a homogeneous evolutionary process acting across all alignment sites is often violated in real datasets of all sequence types [Yang; Whelan, 2008], and the development of methods allowing for this assumption to be relaxed in the analysis of various types of sequences has long been an area of productive research. Most studies have focused on across-site variation of a single parameter such as the evolutionary rate [Uzzell and Corbin, 1971; Yang, 1994, 1996; Nielsen and Yang, 1998], but more complex types of heterogeneity, such as heterogeneity in the entire rate matrix which describes the evolutionary process, have also been explored [Lartillot and Philippe, 2004] . The approach generally taken for variation of a single parameter is to describe the rate (or whichever parameter is being modeled as varying across sites) for each site as a random draw from a statistical distribution. In this way, the mathematically convenient assumption of independence between sites is upheld while sites are allowed to vary according to the parameter of interest. The gamma distribution is commonly used for this purpose, as it contains only one parameter if the mean rate is normalized to 1. This parameter, α, is typically estimated from the data and has a very clear interpretation: low values of α reflect an L-shaped gamma distribution with a large amount of rate variation, while high values of α reflect a bell-shaped distribution where most sites evolve near to the average rate.

The final assumption of a single phylogenetic tree relating all sites within a sequence has been less studied. Simulations have been used to evaluate the potential impact of recombination [Anisimova *et al.*, 2003; Shriner *et al.*, 2003] and gene conversion [Casola and Hahn, 2009] on the use of evolutionary codon models (introduced in the next section) to detect positive selection, finding that moderate levels of false positives occur when the assumption of a single tree is violated. Simulations have also been used to estimate the impact of recombination on reconstruction of ancestral sequences [Busto and Posada, 2010] and phylogeny inference [Schierup and Hein, 2000]. In virus genetics where recombination is commonly encountered, methods have been developed to automate the identification and handling of recombination events in evolutionary analyses [Grassly and Holmes, 1997; Kosakovsky Pond *et al.*, 2006]. Additionally, in comparative genomic analyses of closely-related species, the tree relating the species' sequences may vary across the genome due to incomplete lineage sorting (ILS). This is encountered in the analysis of great ape genes presented in Chapter 6, where a simple filtering scheme was used to remove sites potentially subject to ILS.

1.3 Detecting purifying and positive selection in proteins

Proteins are functionally active as folded, structured amino acid molecules, but the genes which encode them replicate and mutate and evolve as DNA molecules. The connection between DNA and protein is mediated by the genetic code, which describes how non-overlapping codons, or triplets of nucleotides within the coding sequence of a protein-coding gene, are translated by the ribosome and tRNA molecules into polymers of the 20 common amino acids. Since there are $4^3 = 64$ possible codons and only 20 amino acids, the genetic code is degenerate (i.e., multiple codons are translated into the same amino acid). This degeneracy is concentrated in the third codon position, with many codons differing by only their third nucleotide coding for the same amino acid, but some degeneracy also exists in the first position. No codons differing by their second nucleotide encode the same amino acid. Only two amino acids, methionine and tryptophan, are encoded by just one codon, and three codons are stop codons used uniquely to signal the end of the peptide chain.

The degeneracy of the genetic code causes some changes on the DNA level to be synonymous, meaning they result in no change to the encoded amino acid sequence, while others are nonsynonymous, meaning they result in an altered protein sequence. The decoupled nature of the process of DNA mutation, which is presumably "unaware" of the genetic code and affects all nucleotides equally, and the process of natural selection, which acts on phenotypes typically (but not exclusively) affected by protein structure, suggests that a comparison of rates of nonsynonymous and synonymous substitution would allow the influence of natural selection acting on a protein to be detected while inherently correcting for the neutral mutation rate. Indeed, the potential of this approach was noted by various researchers as soon as large-scale DNA sequencing became practical [Kimura, 1977; Jukes and King, 1979], and the comparison of the rate of nonsynonymous substitution (dN) and the rate of synonymous substitution (dS) has been a cornerstone of the evolutionary analysis of proteins since then [Yang]. Various techniques were historically used to estimate dN and dS [Yang and Nielsen, 2000], but most modern software is based on one of the two parametric Markov models for coding sequence evolution independently proposed by Goldman and Yang [1994] and Muse and Gaut [1994] or, when an empirical codon model is desirable, a version of the model estimated by Kosiol *et al.* [2007].

Although the two parametric codon models differ in the details of how nucleotide and codon frequencies are handled [Yang and Nielsen, 2000; Bierne and Eyre-Walker, 2003], they are similar in that each incorporates a selection parameter ω, representing the ratio of dN and dS, into a Markov model of coding sequence evolution. The ω parameter has a simple interpretation, namely that it measures the "the net effect of selection at the protein level" [Yang *et al.*, 2000]. When

$\omega = 1$, natural selection acts neither for nor against protein change, and the sequence is said to be evolving neutrally; when $\omega < 1$, selection acts to conserve the protein sequence, exerting a so-called purifying selective pressure; when $\omega > 1$, selection acts to change the protein sequence, exerting a so-called positive selective pressure. Within the context of population genetics, the ω parameter can be linked to s, the selective coefficient of an allele segregating in the population [Nielsen and Yang, 2003; Nielsen, 2005; Kryazhimskiy and Plotkin, 2008], with $\omega > 1$ corresponding to $s > 0$ and $\omega < 1$ corresponding to $s < 0$. This interpretation involves many assumptions, however, and has found little practical use [Nielsen and Yang, 2003; Nielsen, 2005].

For either of these Markov codon models (as well as for the simpler models of DNA and protein evolution described earlier), parameters (here including ω, κ, and branch lengths for the model of Goldman and Yang [1994]) can be numerically optimized for a given alignment and phylogeny by maximum likelihood using Felsenstein's pruning algorithm [Felsenstein, 1981; Goldman and Yang, 1994; Yang and Nielsen, 2000]. In their initial description of the model, Goldman and Yang [1994] presented an example modification to the model which allowed for heterogeneous substitution rates across sites using an approach similar to that described above for nucleotide models. Much subsequent work has been focused on developing models allowing for variation of ω across sites in the alignment [Nielsen and Yang, 1998; Yang *et al.*, 2000; Yang and Swanson, 2002; Wong *et al.*, 2004; Yang *et al.*, 2005; Massingham and Goldman, 2005], between branches in the tree [Yang and Nielsen, 1998], or both [Yang and Nielsen, 2002; Zhang *et al.*, 2005]. A detailed account of these developments will not be presented here (Anisimova and Kosiol [2009] provide a comprehensive review of the state-of-the-art in probabilistic codon models), but three codon-based models in particular are used extensively throughout this thesis to examine patterns of natural selection within mammalian proteins: the branch model and sites models implemented in the Phylogenetic Analysis by Maximum Likelihood (PAML) program by Ziheng Yang [Yang, 2007] and the Sitewise Likelihood Ratio (SLR) method implemented in the SLR program by Tim Massingham [Massingham and Goldman, 2005]. Each of these models is introduced in more detail below.

Branch and sites models in PAML

For a long time, the ω ratio was almost always calculated as an average value across an entire protein [Sharp, 1997]. Functional and structural constraints within a protein sequence might dominate the gene-averaged signal of evolutionary constraint, however, even if positive selection has acted on some portion of the protein. In some cases, such as when the amount of computational power or data available is limited, averaging ω across sites is a reasonable compromise, but as computer and sequencing technologies rapidly developed in the late 1990s, that compromise was becoming less necessary and whole-gene estimates less justifiable. Whole-gene ω estimates

were shown to lack power [Endo *et al.*, 1996], and although ad-hoc analyses of ω within specific regions of proteins were more sensitive [Hughes and Nei, 1988], they were limited to cases where prior knowledge of the tertiary or domain structure of a protein could be used to identify subsets of the protein for analysis.

As a statistically rigorous alternative, Ziheng Yang and collaborators adopted a so-called "random sites" approach to modeling variation of the ω ratio across sites for the PAML software. This implementation typically favored modeling ω variation with a gamma or beta distribution, discretized into a predefined number of site classes for computational efficiency. To identify positive selection with statistical justification, a variety of likelihood ratio tests (LRTs) were developed.

The LRT is a widely-used statistical technique which tests the goodness of fit between two models. Before returning to a description of the LRTs designed for detecting positive selection with variation of ω across sites, the likelihood function and the LRT will be briefly introduced. Given a statistical model H and a vector of unknown parameters Θ, the likelihood function L describes the probability of observing a given dataset X as a function of the parameters, $L(\Theta; X) = Prob(X|\Theta)$. The maximum likelihood estimate (MLE) of the parameter vector $\hat{\Theta}$ is the set of parameters which maximizes the likelihood of the observed data X. Since likelihood values can be extremely small, the log-likelihood function $\ell = ln[L(\Theta; X)]$ is often preferred. The LRT statistic 2Δ compares the goodness of fit of two different models, H_0 and H_1, with parameter vectors Θ_0 and Θ_1 each containing n_0 and n_1 parameters. With parameters set to their MLEs, the maximum log-likelihood values for H_0 and H_1 are defined as $\ell_0 = \ell(\hat{\Theta_0})$ and $\ell_1 = \ell(\hat{\Theta_1})$, respectively, and the LRT statistic 2Δ for comparing goodness of fit is simply twice the difference of maximum log-likelihood values, $2\Delta = 2[\ell_1 - \ell_0]$. When H_0 is nested within H_1 (e.g., when H_0 is equivalent to H_1 with a single parameter held at a constant value, leading to $n_1 - n_0 = 1$) then 2Δ is asymptotically distributed as $\chi^2_{n_1-n_0}$ when the null model H_0 is true. By comparing the LRT statistic to the critical value of the appropriate χ^2 distribution, the hypothesis that a more complex model better describes the observed data can be tested [Wilks, 1938; Yang]. When comparing non-nested models or nested models where a parameter is fixed at the boundary of possible values, the χ^2 approximation does not hold; in some cases a mixture of distributions may be appropriate [Whelan and Goldman, 1999], but in other cases parametric simulations are needed to estimate the null distribution of the LRT statistic [Goldman, 1993].

In the context of detecting positive selection, the most effective LRTs typically compare a model allowing for some variation of ω, but not allowing $\omega > 1$, to a more complex model which additionally allows for $\omega > 1$ (the former model being nested within the latter). Nielsen and Yang [1998] initially proposed a few simple LRTs for detecting positive selection; these were further expanded and evaluated by Yang *et al.* [2000], and Anisimova *et al.* [2001] performed extensive simulations assessing the power and accuracy of these tests. The current release of

PAML recommends two LRTs for detecting positive selection acting at a subset of sites within genes: M2a–M1a and M8–M7 (see Table 1 in Wong *et al.* [2004] for a complete description of each model). PAML also implements a method for identifying which individual sites show strong evidence for positive selection. This method, called the Bayes Empirical Bayes method and described in Yang *et al.* [2005], calculates for each site an approximate posterior probability that it has been subject to positive selection. I evaluate the power of this method for detecting sitewise positive selection in Chapter 2.

Codon models relaxing the assumption of a constant ω throughout the branches of the tree, but not across sites, were also developed and implemented in PAML [Yang, 1998; Yang and Nielsen, 1998]. Although these models have received less attention and use than the branch-site models, which allow ω to vary across both sites and branches [Zhang *et al.*, 2005], they may be useful when branch lengths are small and parameter estimation under the complex branch-site models is difficult, or when variation of ω across sites is not of interest. These models are used in Chapter 6 to construct a series of LRTs for detecting accelerated evolution along specific branches of the great ape tree.

The Sitewise Likelihood Ratio method

In contrast to PAML's approach to site-specific evolutionary analysis, where a LRT for positive selection within a gene is first performed and, subject to a significant LRT result, sitewise posterior probabilities for positive selection are then inferred, the SLR method [Massingham and Goldman, 2005] was specifically designed for the sitewise estimation of purifying and positive selection. SLR is based on a Markov model of codon evolution similar to that of Goldman and Yang [1994]. No assumptions are made regarding the distribution of ω ratios within the alignment; instead, the value of ω is considered to be an independent parameter at each site, ω_i for site i. The idealized sitewise LRT for positive selection then compares the log-likelihood value of a null model where $\omega_i = 1$ to a model where ω_i is optimized to its MLE. As the estimation of model parameters and calculation of likelihood values to perform an exact LRT at each site would involve an expensive high-dimensional optimization, SLR uses two approximations which greatly reduce the computational complexity: first, parameters common to all sites (including branch lengths, κ and the equilibrium codon frequencies) are estimated under the M0 model [Yang *et al.*, 2000] with one ω for all sites instead of the more parameter-rich true null model, and second, the sitewise ω parameter is estimated independently at each site under the simplifying assumption that each site's contribution to the common parameters is minimal [Massingham and Goldman, 2005]. In practical terms, SLR uses the common parameters and the alignment 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 corresponds to neutral evolution by holding the ω parameter fixed

at 1 and the alternative model uses the MLE of ω. The raw SLR statistic measures the strength of evidence for non-neutral evolution at each site, and the observed statistic can be compared to its theoretical distribution under the null model, χ^2_1, to identify sites with evidence for purifying or positive selection at a desired significance threshold. Simulations performed by Massingham and Goldman [2005] showed SLR to perform as well as or better than PAML's random sites models at detecting positive selection within genes under some conditions; Chapter 2 provides further assessment of the power of the SLR method to detect sitewise positive selection under a wide range of conditions, and in Chapters 4 and 5 I apply SLR on a large scale to analyze genome-wide patterns of sitewise positive selection in mammals.

1.4 Outline of the thesis

This thesis describes three largely independent studies centered around the theme of using codon models to analyze patterns of selective constraint in mammalian genomes.

Chapter 2 describes a simulation study I conducted to evaluate the impact of alignment error on detecting sitewise positive selection. The widespread adoption of powerful methods for sequence analysis has been somewhat hampered by a lack of understanding of the limitations of, and sources of error inherent within, those methods. Without such knowledge, researchers could either avoid using such methods for fear of producing misleading results, or blindly apply such methods without regard for possible errors. Both paths have been taken by others with respect to the issue of alignment error in detecting positive selection [Thompson *et al.*, 1994a; Bakewell *et al.*, 2007; Studer *et al.*, 2008; Markova-Raina and Petrov, 2011]. As a first step towards an improved understanding of alignment error and its impact on downstream evolutionary analyses, I performed a thorough examination of the impact of alignment error by comparing different aligners, trees, and methods for detecting positive selection. This work has been recently published [Jordan and Goldman, 2011] and is presented in Chapter 2 largely unmodified from its published form.

With whole-genome sequences quickly accruing in the databases at an ever-increasing pace, I devoted a large part of my research effort towards applying evolutionary codon models on a large scale to mammalian and primate genomes. The rest of the thesis reflects this focus, presenting two major empirical analyses. Chapters 3, 4 and 5 describe in three sections a genome-wide analysis of sitewise selective pressures in mammals that was performed in collaboration with the Mammalian Genome Project (MGP); a more detailed description of the MGPs and my involvement in the analysis is provided in the introduction to Chapter 3. A highly summarized and edited version of the results of this analyses was recently published [Lindblad-Toh *et al.*, 2011], but the version presented here differs in that a more recent dataset was used and the methods and results are

described in considerably more detail. Particular attention was paid to identifying (and in some cases ameliorating) possible sources of error and to comparing the current results with those from previous similar studies in the literature.

Chapter 6 presents a more detailed survey of genome-wide evolutionary patterns within a much more closely-related group of mammals, the African great apes. This study was the result of a collaboration with Stephen Montgomery and Nick Mundy (Department of Zoology, University of Cambridge) as part of the gorilla genome analysis group led by the Wellcome Trust Sanger Institute, and a manuscript including the major results from our analysis is currently undergoing peer review. As a few aspects of the study design and interpretation of results were performed by collaborators as well as myself, those items which do not represent entirely my own work are clearly indicated. As in the mammalian analysis, attention was paid to assessing the potential impact of known and unknown sources of error on the conclusions drawn.

Finally, Chapter 7 briefly ties together the themes developed within each of the prior chapters, summarizing the main contributions of the work in a broader context.

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, 2007; Birney *et al.*, 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.*, 2001; Massingham and Goldman, 2005]. 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.*, 2006].

A major concern in the detection of positive selection in proteins 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. This chapter focuses on methods for the sitewise detection of positive selection, as they will be be widely used throughout this thesis.

In addition to the choice of aligner, the protein structure and evolutionary divergence of a dataset may 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 biological indels—and will therefore be less susceptible to alignment error—than less structured regions.

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 fixed substitutions between species, they show little power at low divergence and their highest power at intermediate to high divergence levels [Anisimova *et al.*, 2001]. 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).

Fortunately, 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 phy-

logeny inference [Wong *et al.*, 2008] and estimates of positive selection [Schneider *et al.*, 2009; Markova-Raina and Petrov, 2011]. Most recently, Markova-Raina and Petrov [2011] 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 and 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 [2011] 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 and Gadagkar, 2009; Ogden and Rosenberg, 2006; Löytynoja and Goldman, 2008], but those results cannot be easily extrapolated to the analysis of sitewise selective pressures. More recently, Fletcher and Yang [2010] 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.*, 2000] and PRANK [Löytynoja and 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 and 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 and 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.*, 2008; Aguileta *et al.*, 2009], continued investigation of potential benefits of alignment filtering to the detection of positive selection seems well-warranted.

This chapter 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. This approach builds on those of Anisimova *et al.* [2002], Fletcher and Yang [2010], and Privman *et al.* [2011], using simulated protein alignments including insertions and deletions to evaluate methods for detecting sitewise positive selection. Furthermore, a diverse sample of aligners and alignment filters were incorporated into an experimental design that differs from previous ones in a number of important ways.

I focused on sitewise detection of positive selection occurring throughout a phylogeny and evaluated the impact of a number of alignment filtering methods on the sitewise analysis. Thus, the biological hypothesis being investigated was different from that studied by Fletcher and Yang [2010], who focused on genewise selection acting at specific branches. Privman *et al.* [2011] recently published a related paper considering the evolutionary characteristics 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.*, 2008] and which comprise a larger diversity of protein structures and a wider range of species divergence levels.

I hypothesized 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, these simulations encompassed 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 2.1. Three rooted trees were used to guide the simulation of protein-coding DNA alignments: the artificial 6-taxon tree used by Anisimova *et al.* [2001] and Massingham and Goldman [2005] 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 [Birney *et al.*, 2007; Nikolaev *et al.*, 2007]. 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 2.2).

The INDELible program [Fletcher and 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 dataset of sitewise selective pressures estimated from mammalian gene trees (Lindblad-Toh *et al.* [2011]; log-normal parameters shown in Table 2.1). 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 and Ohta, 1974]. 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 $\omega > 3$.

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		

Table 2.1: Parameter Values Used in Simulations. MPL is the mean path length of the tree in units of substitutions per synonymous site (dS). Indel lengths are measured in units of codons, and the indel rate is defined as the number of insertion & deletion events per substitution.

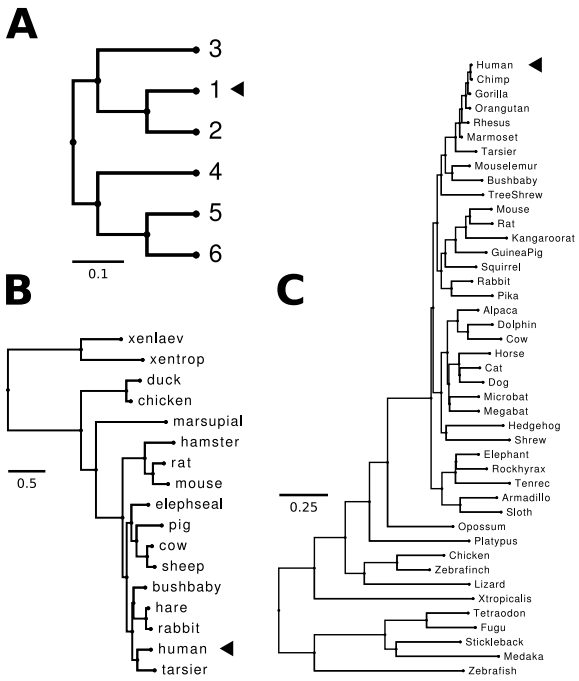


Figure 2.1: Phylogenetic trees used for simulation and analysis. The original scale for each tree is indicated by a scale bar, but trees were scaled to equal mean path length (MPL) divergence levels for simulation. (A) A 6-taxon artificial tree used in previous simulations [Anisimova *et al.*, 2001;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

Branch lengths for each of the simulation trees were scaled before simulation to correct for the difference between the current 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 and Yang, 2010]. 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 and Nielsen, 1998] and the mean ω ratio for the distribution used in this chapter 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.*, 1994b], MAFFT [Katoh *et al.*, 2005], Prob-

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

Table 2.2: Genome-wide Divergence Estimates for Commonly Analyzed Eukaryotes. The root-to-tip dS is equivalent to the MPL (mean path length) used in these 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.

Cons [Do *et al.*, 2005], T-Coffee [Notredame *et al.*, 2000] and two variants of PRANK [Löytynoja and 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.*, 2000; Notredame and 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 entire 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 and 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 an initial score cutoff 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 2007) and Sitewise Likelihood Ratio (SLR; Massingham and Goldman 2005) software packages. The major models implemented by these two programs were introduced in Chapter 1. To estimate sitewise selective pressures with Phylogenetic Analysis by Maximum Likelihood (PAML) I used the two models for which the recommended Bayes Empirical Bayes method are implemented, M2a and M8. Sitewise Likelihood Ratio (SLR) was run using default parameters. Following Massingham and Goldman

[2005], 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 analysis, 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 and Goldman, 2005]. 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.

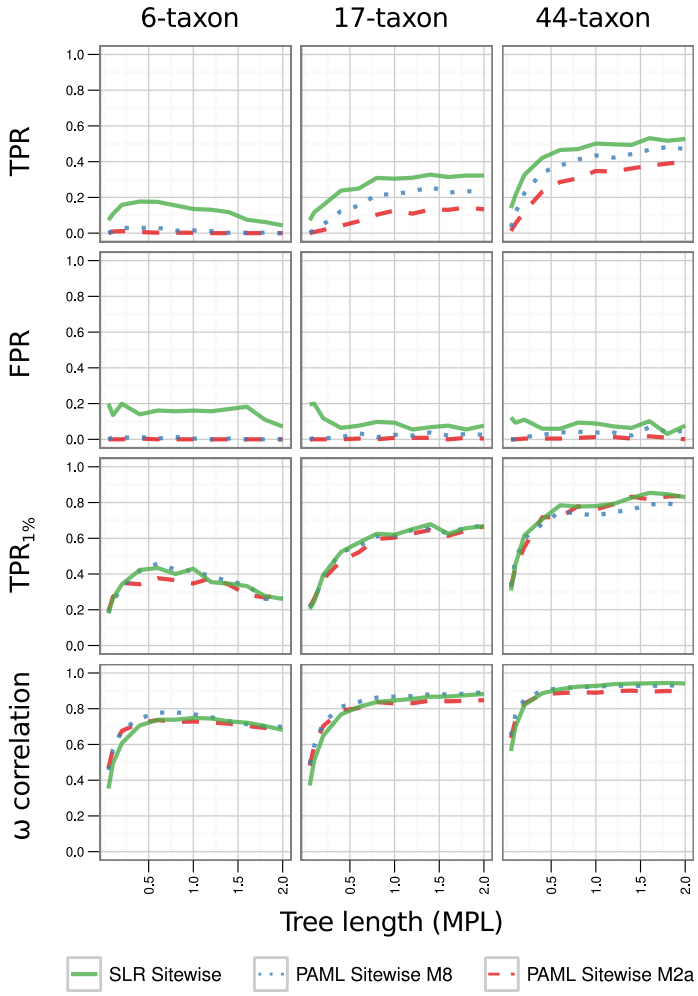


Figure 2.2: Alignments were simulated without indels for three tree shapes and analyzed with SLR, PAML M8, or PAML M2a. Fifty replicate alignments were simulated for each data point. The performance of each analysis method, as measured by four summary statistics, is plotted as a function of the mean 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.

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, $TPR_{1\%}$ 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.*, 2002]. 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 2.2).

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 the chosen simulation conditions.

The conservative nature of the default thresholds for PAML and SLR has been previously noted [Anisimova *et al.*, 2002; Yang *et al.*, 2005; Massingham and Goldman, 2005], but the extremely low false positive rates in these 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 the 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 these 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 2.5.) 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 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,

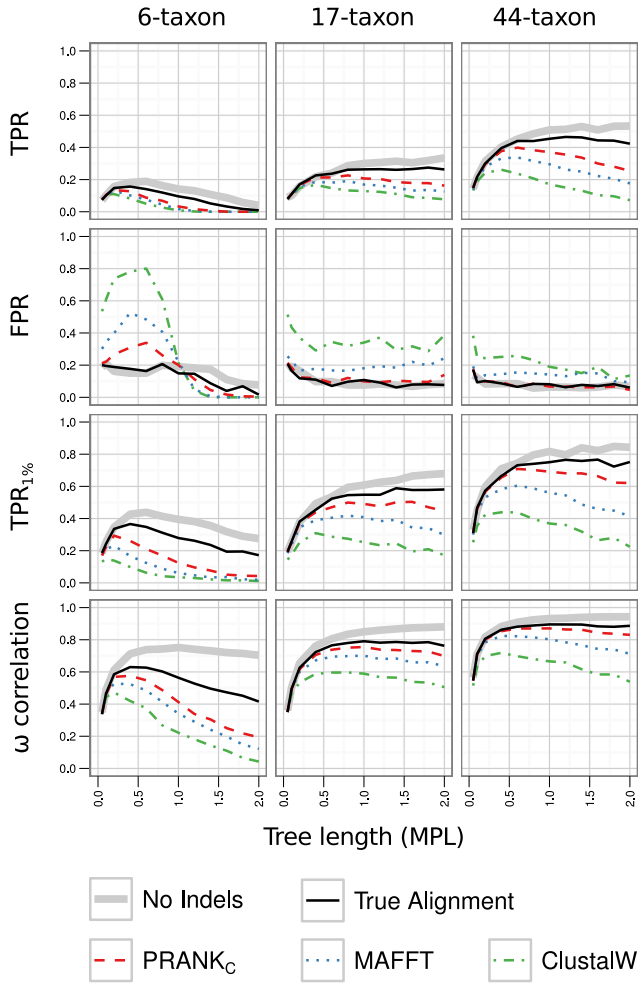


Figure 2.3: Performance of sitewise detection of positive selection with alignment error. Sequences were simulated without indels (solid gray lines) or with indels (solid black and textured lines) using one of three tree shapes, aligned with one of three aligners, and analyzed with SLR; true alignments 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.

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 perfor-

mance 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 and Yang, 2010; Markova-Raina and Petrov, 2011; Privman *et al.*, 2011]; my 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 inferred 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.

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

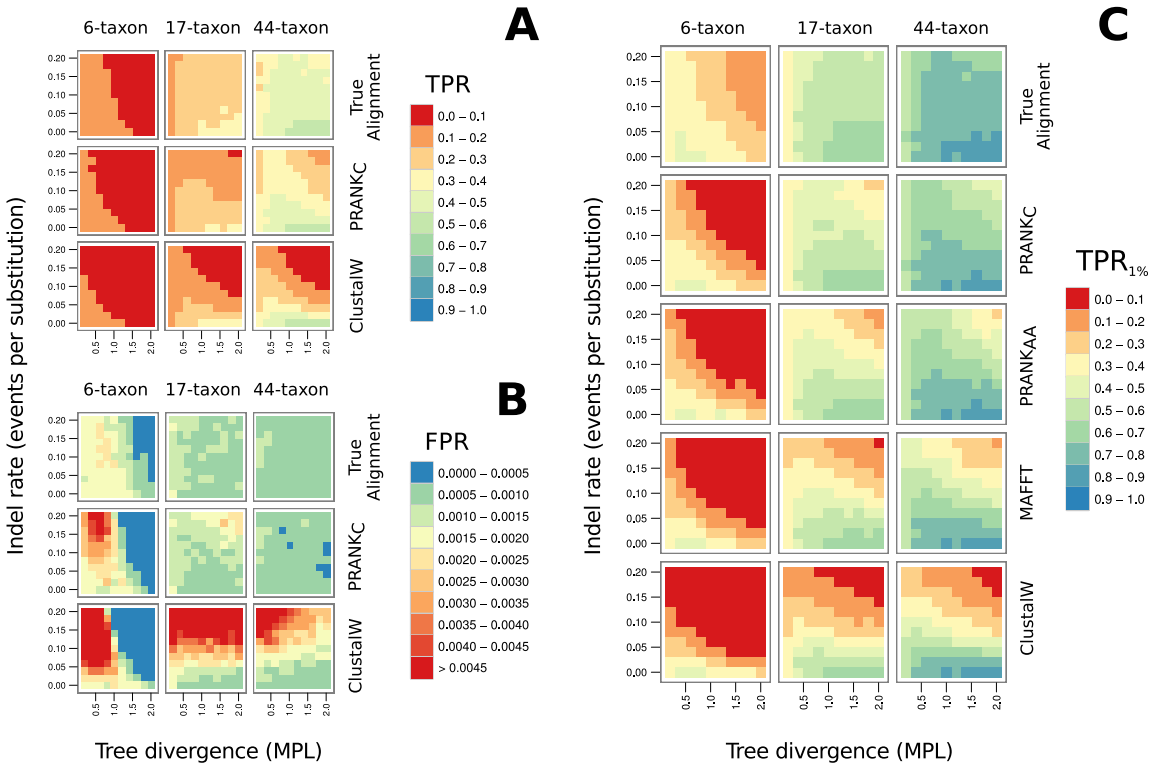


Figure 2.4: The TPR, FPR, and $TPR_{1\%}$ for sitewise detection of positive selection. Sequences were simulated with indels using one of three tree shapes (6-, 17- or 44-taxon) and a range of indel rates and mean path length (MPL) divergence levels. Alignments are inferred with one of four aligners (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 $PRANK_C$.

the error-controlled $TPR_{1\%}$ 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 for clarity. The performance of all these aligners fell between that of ClustalW and $PRANK_C$ for all measurements. $PRANK_{AA}$ slightly outperformed MAFFT, and ProbCons and T-Coffee showed similar performance to MAFFT. A comprehensive set of TPR, FPR, and $TPR_{1\%}$ results can be found in Figure 2.5.) 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 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 as expected, 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 $\text{TPR}_{1\%}$ 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 $\text{TPR}_{1\%}$ 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 $\text{TPR}_{1\%}$ in the 6-taxon tree, but in the larger two trees ClustalW and MAFFT alignments were unable to achieve high $\text{TPR}_{1\%}$ values, presumably due largely to their elevated FPR in those trees.

It is worth noting the 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$_{\mathrm{C}}$'s reduction in sitewise performance—a novel observation which provides insight into the nature of PRANK$_{\mathrm{C}}$ alignments and their application to sitewise evolutionary analysis.

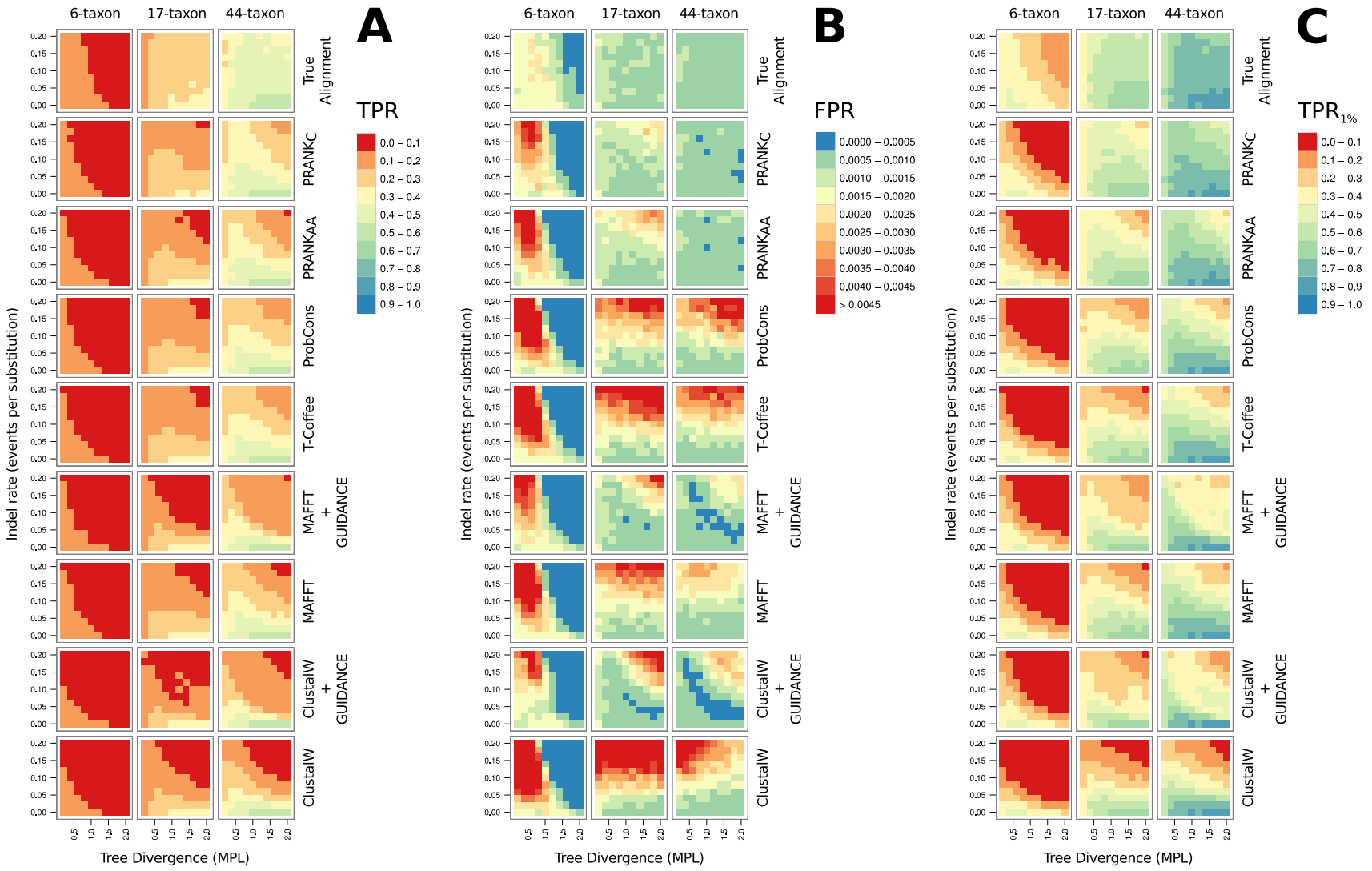


Figure 2.5: This figure depicts the same simulations and uses the same formatting as Figure 4, except that results for MAFFT and PRANK_{AA} have been added to sections (A) and (B) and results using ProbCons, T-Coffee, ClustalW + GUIDANCE, and MAFFT + GUIDANCE have been added to all sections.

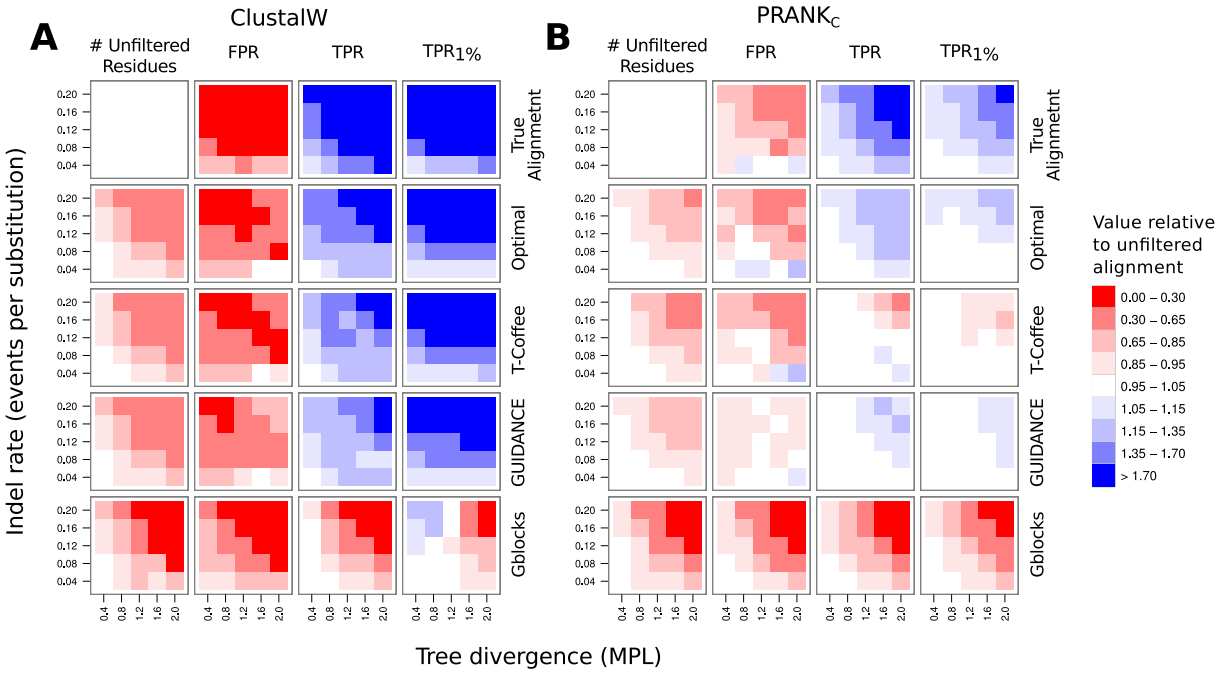


Figure 2.6: The effect of alignment filtering on sitewise detection of positive selection. Sequences were simulated using the 17-taxon tree and a range of indel rates and mean path length (MPL) divergence levels. Alignments were inferred using (A) ClustalW or (B) PRANK$_C$, either left unfiltered or 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 2.6. (Filtering results for PRANK_{AA}, ProbCons and T-Coffee were not calculated, and results for MAFFT are omitted from Figure 2.6 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 2.6). 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.

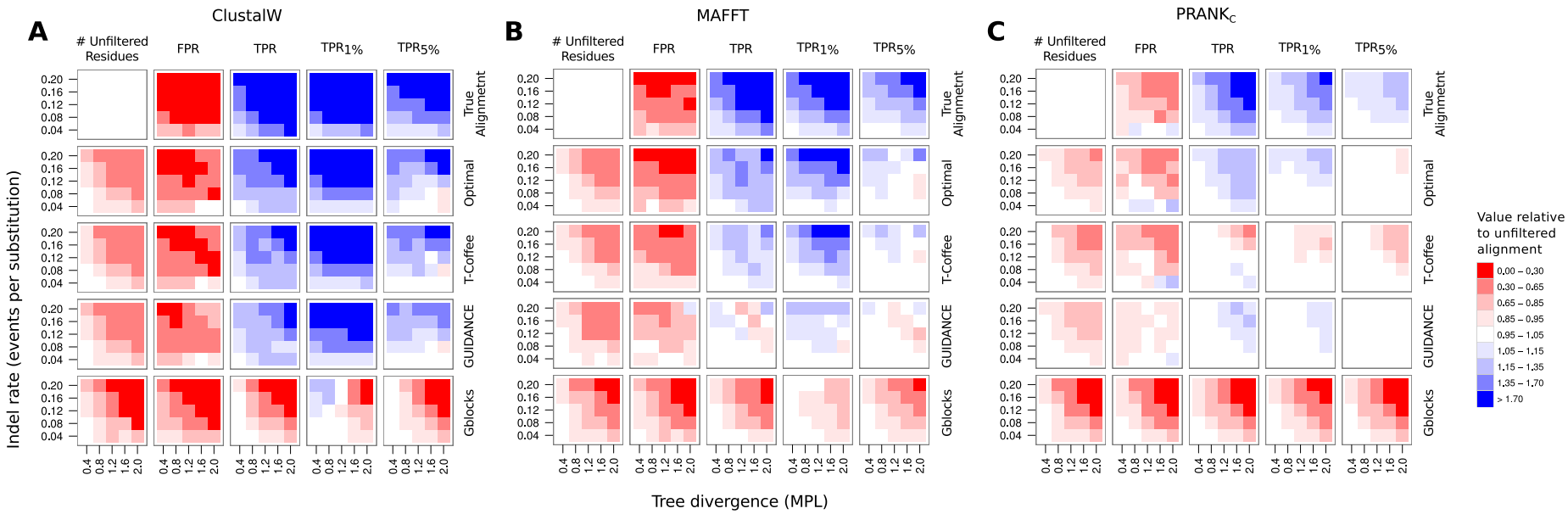


Figure 2.7: This figure depicts largely the same simulations and uses the same formatting as Figure 5 except for two changes. First, results for MAFFT have been added to section (B) and results for PRANK$_C$ have been moved to section (C). Second, the rightmost column has been added, showing the true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

The expectation 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 $TPR_{1\%}$ change slightly lower than for true alignment. For ClustalW alignments the amount of improvement was quite large, with >70% increase in $TPR_{1\%}$ 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% $TPR_{1\%}$ 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 effect on the accuracy of phylogenetic inference; the current 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 here. 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, the 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 was 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 2.6 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 $\text{TPR}_{1\%}$ 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 2.5. Comparing absolute performance, I found that GUIDANCE filtering generally improved the $\text{TPR}_{1\%}$ for ClustalW and MAFFT alignments, largely due to strong FPR reductions in regions of high indel rates and low divergence levels. The resulting $\text{TPR}_{1\%}$ performance for filtered ClustalW alignments was comparable to (but not better than) unfiltered MAFFT alignments, and the $\text{TPR}_{1\%}$ 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 $\text{TPR}_{1\%}$ 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 was 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, it should be noted that the performance of the 'optimal' filter on PRANK_C alignments suggests that mild further improvements to filtering strategies may be possible; however, the potential for improvement is small and may be of little practical value.

2.4 Conclusions

In this chapter, 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. I 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 $\text{TPR}_{1\%}$ of the sitewise detection of positive selection as 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 [2010], Markova-Raina and Petrov [2011], 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, these 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.*, 2001, 2002]. 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 conflicted somewhat with the results of Fletcher and Yang [2010], 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 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 control of false positives.

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, this chapter showed 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, the use of alignment filtering in the detection of sitewise positive selection cannot be uniquivocally recommended, except perhaps for the most divergent and indel-prone sequences. Importantly, sequences with these levels of divergence 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 the current 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 the range tested here (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 the 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 the simulations or filtering procedures, but in the end the results are largely complementary: Privman *et al.* [2011] 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.

It is important to note that the current 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 [2008]). Such heterogeneity should not affect the 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 the framework introduced here.

However, such heterogeneity is clearly important to the evolution of mammalian proteins [Fay and 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 here, making false positive results either less or more likely, respectively. As such, the appropriateness of any simulation scheme should be critically considered when evaluating specific power and error rate estimates in the context of real-world data analysis.

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 and 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.

Chapter 3

Curating a set of orthologous mammalian gene trees

3.1 The Mammalian Genome Project

This chapter, Chapter 4 and Chapter 5 describe different aspects of the research which the Goldman group and I performed in collaboration with the Mammalian Genome Project (MGP). As a preface to these chapters, this section introduces the MGP and the main goals of the analysis which we contributed to the consortium.

A major goal of mammalian comparative genomics has been to identify and understand regions of the human genome that are evolving under evolutionary constraint, as those regions are likely to be functionally important [Waterston *et al.*, 2002]. Protein-coding genes clearly fall into this category, but additional non-coding functional elements such as RNA genes, transcription factor binding sites, and enhancer elements are known to also be functionally important and subject to detectable purifying selection [Birney *et al.*, 2007]. Comparisons of the first non-human mammalian genomes to the human genome indicated that around 5% of the human genome has been evolving under purifying selection throughout the last 100 million years of evolution [Waterston *et al.*, 2002; Cooper *et al.*, 2004; Gibbs *et al.*, 2004; Lindblad-Toh *et al.*, 2005], but the small number of genomes available at the time limited the accuracy and resolution with which specific regions subject to purifying constraint could be identified and classified [Ponting and Hardison, 2011].

The signal used to detect constrained regions is most commonly a locally decreased rate of nucleotide substitutions [Cooper *et al.*, 2004]. The strength of this signal increases along with the expected number of substitutions per neutrally-evolving site [Siepel *et al.*, 2005], so the incorporation of additional genomes into the analysis was expected to be an effective way to

improve power. Following this line of reasoning, the MGP was proposed with the primary goal of increasing the accuracy and confidence with which evolutionarily constrained regions of the human genome could be identified [Margulies *et al.*, 2005, 2007].

The first phase of the MGP took the form of a coordinated series of genome sequencing projects initiated in 2005 and organised by the Broad Institute of MIT and Harvard. In total, 20 new mammalian genomes were sequenced, with species chosen in order to maximize the amount of evolutionary divergence available for comparative analysis when combined with the 9 mammalian genomes already available [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 the result of a deliberate compromise between the number of species sequenced and the quality of each genome. The 2x level of coverage was estimated to maximize the amount of additional branch length made available with a reasonably high level of sequence quality given a limited budget [Margulies *et al.*, 2007].

As the project proceeded from its sequencing to analysis phase in late 2008, it became apparent that the additional branch length afforded by the 29-species phylogeny would yield improved power for 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 in non-coding regions, and detection of purifying and positive selection in protein-coding genes. Groups working on topics in these areas used the data generated by the MGP to conduct more focused analyses incorporating the newly sequenced genomes. Given the prior involvement of the Goldman group in analysing the comparative sequencing data from the Encyclopedia of DNA Elements (ENCODE) project [Margulies *et al.*, 2007; Birney *et al.*, 2007] and Tim Massingham's work on the Sitewise Likelihood Ratio (SLR) method and software program [Massingham and Goldman, 2005], the group was recruited in late 2008 to perform the protein-coding evolutionary analysis for the MGP in close collaboration with members of the Ensembl Compara team led by Javier Herrero. The main goal was to use data from the newly sequenced mammalian genomes to analyze selective pressures in proteins at the level of individual codons. All of the work described in this and the following two chapters was performed by me, though the work has benefitted greatly from the advice and guidance of members of the Goldman group (Nick Goldman, Tim Massingham and Ari Löytynoja), the EnsEMBL Compara team (Albert Vilella, Javier Herrero and Ewan Birney) and various organisers and members of the MGP (Manolis Kellis, Kerstin Lindblad-Toh, Mike Lin and Katie Pollard).

3.2 Introduction

The main goal of the analysis presented in this and the following two chapters was to characterize the amount and location of purifying and positive selection within mammalian protein-coding genes. The primary data were the genome sequences of the species being compared. Originally, this consisted of the 29 Eutherian mammals chosen for analysis by the MGP consortium, but the analysis described here used a larger group of 38 mammalian genomes included in release 63 of the Ensembl database [Flicek *et al.*, 2011].

In order to analyze proteins using phylogenetic codon models, a significant amount of processing must be applied to the raw genomic data. The annotation of protein-coding genes, identification of orthologous sequences, and inference of phylogenetic trees and coding alignments are all necessary steps in producing the inputs required for analysis by methods such as PAML or SLR. The Ensembl gene annotation pipeline and the Ensembl Compara comparative genomics pipeline already perform many of these steps on a comprehensive set of mammalian and vertebrate genomes, so I used the Ensembl databases as a source for gene annotations, gene trees, and protein annotations where appropriate. In some cases, the Ensembl data required modification or filtering: this chapter describes how taxonomic constraints were used to extract a more useful set of orthologous gene trees from the set of "root" phylogenetic trees inferred by the Compara pipeline, and Chapter 4 describes how sequences were extensively re-aligned and filtered before being analyzed with SLR.

An important feature of the Ensembl gene annotation pipeline is that low-coverage genomes are processed using a modified version of the pipeline designed to make the best possible use of the fragmented and gappy assemblies resulting from low-coverage shotgun genome sequences. This was a significant factor behind my use of Ensembl as a primary source for gene annotation and orthology information, as Ensembl is unique in incorporating low-coverage genomes into a full-fledged annotation database. Hubbard *et al.* [2007] describe the modifications made to the Ensembl pipeline to accommodate low-coverage genomes, and I present the approach and discuss some of its limitations in Section 3.4.

The main data used in the current chapter was the set of vertebrate gene trees inferred by the Ensembl Compara pipeline [Vilella *et al.*, 2009]. The Compara pipeline uses the set of annotated proteins within each vertebrate genome to cluster homologs, align sequences and infer gene trees with a Bayesian gene tree reconciliation method; Vilella *et al.* [2009] describe and justify the approach and methods in more detail. This chapter begins with an overview the methods used by the Compara and other orthology pipelines, identifying some aspects of the approach taken by Compara which may have consequences for my use of the gene trees to infer sitewise selective pressures.

The gene trees inferred by the Compara pipeline include a reconstruction of the deep evo-

lutionary history of genes, often containing paralogous genes related by several ancient duplication events within one phylogenetic tree. Many of the deep duplication events reflect the two rounds (2R) whole-genome duplication events that occurred in the vertebrate ancestor, while more recent duplication events can be viewed as part of an ongoing process of gene duplication and loss within each genome. However, the inclusion of paralogous genes within the current analysis was undesirable for two reasons. First, the goal of this project within the scope of the MGP was to use mammalian evolutionary history to better understand the human genome, so gene trees that represented largely orthologous evolutionary relationships within mammals would provide the most useful context for understanding our own genome. While ancient duplications are an interesting part of the evolutionary history of genes, the evolution of duplicated genes was beyond the scope of this analysis.

Secondly, a number of studies have suggested that duplicated genes may evolve under different evolutionary constraints compared to singleton genes in the genome. Dermitzakis and Clark [2001] showed that members of a duplicate gene pair experienced different patterns of amino acid changes linked to likely subfunctionalization [Massingham *et al.*, 2001], and some studies have described increased rates of positive selection within duplicate genes [Lynch and Conery, 2000; Zhang *et al.*, 2002; He and Zhang, 2005; Hahn, 2009]. Although the extent to which the evolution of paralogs differs from that of orthologs is a matter of open debate [Nembaware *et al.*, 2002; Jordan *et al.*, 2004; Studer and Robinson-Rechavi, 2009], the potential for the adaptive evolution of duplicate genes to produce misleading results in the current analysis made it important to avoid including paralogous relationships within the gene trees to be analyzed.

After introducing the methods behind the Compara orthology pipeline, the rest of this chapter describes the development of a method I used to extract a set of "core" orthologous gene trees for further analysis. Although in theory each Compara gene tree contains a fully resolved history of gene duplication and loss events, in practice I found it unsatisfactory to use the simplest imaginable approach—splitting each tree at all duplication nodes—to extract a set of appropriate orthologs. Instead, I adopted a method based on applying simple taxonomic constraints to identify sub-trees with sufficient representation in the mammalian phylogeny to warrant inclusion in the codon-based analyses described in the following two chapters.

3.3 Methods for ortholog identification

Central to most 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, putative 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, the idea of homology across an entire gene may be unsatisfactory and a more localized concept of homology (based on shared domains or functions) may be more appropriate [Koonin *et al.*, 2001; Sjölander *et al.*, 2011]. Within closely related groups of organisms such as mammals, however, the process of gene duplication and divergence dominates patterns of relatedness between protein sequences [Ohno, 1970], making gene-wide 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 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 and Conery, 2000], the correct identification of orthologous versus paralogous relationships in vertebrates is critical for any detailed molecular evolutionary analysis.

Many methods for distinguishing orthologous from paralogous relationships within vertebrate genomes have been developed over the past decade [Yuan *et al.*, 1998; Remm *et al.*, 2001], but the amount of overlap between vertebrate orthologous groups identified by different methods has historically been disappointingly small [Chen *et al.*, 2007; Jun *et al.*, 2009], 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. Complete genomes are important in that the availability of a complete set of genes allows for duplication and loss events to be more confidently inferred. 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. This approach is most powerful when the species tree

can be confidently estimated, though some methods allow for species tree uncertainty [Vilella *et al.*, 2009] and there is not too much uncertainty in the relationships between most mammals. 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].

The Compara pipeline uses TreeBeST, a phylogeny-based method which uses a known species tree to guide the resolution of orthologs and paralogs within a gene tree in a Bayesian framework, to infer its gene trees, so I will refer mainly to this method throughout the rest of this chapter. TreeBeST is one of the more popular phylogeny-based methods, having also been used to infer vertebrate and eukaryotic gene trees for the OPTIC and TreeFam databases [Heger and Ponting, 2008; Ruan *et al.*, 2008; Vilella *et al.*, 2009].

3.4 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 were sequenced to low coverage by the 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 significant advantage over otherwise similar orthology databases due to the greater amount of covered branch length.

The prevalence of missing sequence data and fragmented contigs in low-coverage genomes presents a unique set of problems for the generation of transcript annotations in Ensembl, so 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]. Although the "normal" annotation pipeline varies somewhat between different high-coverage genomes, the general approach taken is to align experimentally observed transcripts and protein sequences from a variety of sources against the new genome in order to infer gene and transcript structures. The main difference for the low-coverage genomes is that a whole-genome alignment is produced between the human genome and each low-coverage target genome, and gene models are subsequently projected from human to the target genome based on this genome-wide alignment. Small frame-disrupting insertions or deletions within the low-coverage sequence are corrected based on this alignment, 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 cannot map complete genes or transcripts from human to the target genome, causing difficulty in identifying duplications [Hubbard *et al.*, 2007]. On one hand, the lack of completely assembled chromosomes means that segmental duplications in low-coverage genomes are 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 causes 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 the resulting gene trees.

The Compara pipeline for inferring gene trees uses the set of annotated transcripts from each species as its primary 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 criticized 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 sequencing errors from low-coverage genomes such as those described by [Hubisz *et al.*, 2011] seem less necessary.

3.5 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 is beyond the scope of this chapter, in this section I will briefly outline the major aspects of the approach used by the Compara gene tree inference pipeline, focusing on key details which are relevant to the current analysis.

The Compara pipeline begins with a set of protein-coding transcripts collected from each individual species' annotation database. This first step is not straightforward, however: the Compara pipeline only allows for one transcript per gene, while the prevalence of alternative splicing in Eutherian mammals makes it common for a single gene to harbor many different transcript structures [Mironov *et al.*, 1999]. The use of one transcript per gene has two drawbacks: first, some evolutionarily conserved exons may not be included in the chosen transcript, reducing the amount of orthologous material included in the source dataset, and second, different alternatively-spliced

transcripts may be chosen for different species, leading to patterns of missing or extra exons in the downstream multiple alignments. Although in theory it should be possible to incorporate multiple alternatively-spliced transcripts into a "meta-transcript" for each gene, the Compara pipeline currently only uses one "canonical" transcript per gene in the clustering tree-building process. At least one other pipeline for ortholog identification has dealt with this problem in similar ways: 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 and Ponting, 2008].

Once the set of transcripts is chosen for each species, the Compara pipeline then performs an all-against-all protein BLAST search is performed using the Washington University variant of BLAST [Chao *et al.*, 1992] and genes are clustered into homologous groups using *hcluster_sg* [Ruan *et al.*, 2008], an implementation of a hierarchical clustering algorithm for sparse graphs which is distributed with the widely-used TreeBeST package. Sequences are aligned using MCoffee, a meta-aligner algorithm which combines the results from different aligners into one alignment using a maximum-consistency criterion [Wallace *et al.*, 2006]. The aligners used for the M-Coffee alignment include MAFFT [Katoh *et al.*, 2005], MUSCLE [Edgar, 2004], KAlign [Lassmann *et al.*, 2009], and T-Coffee [Notredame *et al.*, 2000]. Finally, the aligned sequences are input to TreeBeST, which infers a gene tree (including gene duplication and loss events) given a set of aligned sequences and a known species tree [Ruan *et al.*, 2008]. The type of the homology relationship between each pair of genes (e.g., one-to-one ortholog, one-to-many ortholog, within-species paralog) is determined using a simple set of rules based on the structure of the inferred gene tree and the annotation of ancestral nodes where a duplication event has likely occurred.

3.6 Quantifying paralogous relationships within Ensembl gene trees

As stated in the introduction, the main goal of this chapter was to identify and extract a set of gene trees or gene subtrees from the Compara database comprising largely of orthologous relationships within mammals, avoiding as much as possible the inclusion of paralogous relationships. I will refer to trees with these characteristics as largely orthologous trees (LOTs).

It was necessary to use some sort of criterion to extract LOTs from the set of Compara gene trees, because many of the Compara trees contained multiple sets of complete mammalian orthologous trees linked by ancient gene duplication events. In other words, the Compara gene trees could be characterized 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

Tree Set	Tree Count	Sequence Counts			Human Content			Human Total	Med. MPL	Med. Species
		Med (Min / Max)	Total	N50	0	1	2+			
Ensembl Roots	18607	15 (2 / 400)	9.0e+05	139	0.50	0.30	0.20	19995	0.55	8
(<=15)	9378	3 (2 / 15)	3.6e+04	5	0.92	0.08	0.00	809	0.04	2
(>15)	9229	54 (16 / 400)	8.6e+05	146	0.07	0.53	0.40	19186	1.04	47

Table 3.1: 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.

data. This section is concerned with quantifying the amount of this 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 3.1 and Figure 3.1; a complete description of these data will follow, but the aspects relevant to characterizing the amount of paralogy contained within these trees will be discussed here.

The first row of Table 3.1 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 is further supported by Figure 3.1, 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 large number of Compara gene trees containing multiple mammalian LOTs.

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

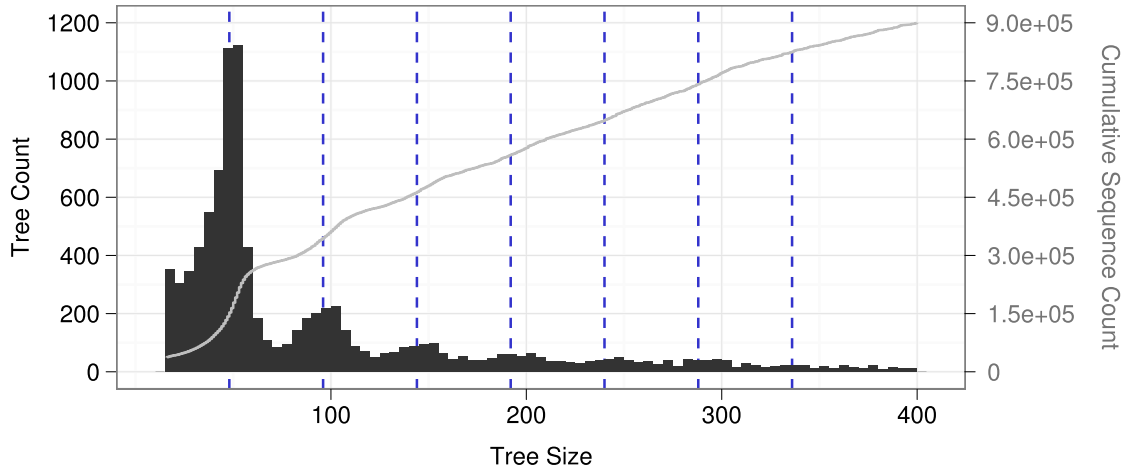


Figure 3.1: Tree sizes for the set of "root" Compara trees. Black bars show a histogram of tree sizes in bins of width 5, and a gray line shows the cumulative number of sequences contained within trees of that size or smaller. For clarity, 9,378 trees with 15 or fewer sequences are not shown (but they have been included in the calculation of the cumulative sequence count). Dashed blue lines are drawn at integral multiples of 48, the number of vertebrate species within Ensembl.

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 3.2. 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]. With Botond Sipos, I performed the evolutionary analysis of *NBEAL2* for the paper 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 (*NBEAL1*, *LRBA*, and *NBEA*) into one tree and similarly clustered four more distant *NBEAL2* paralogs (*WDFY4*, *WDFY3*, *LYST* and *NSMAF*) 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 3.2), 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 3.2), the connection between these eight genes would have been even more immediately apparent.

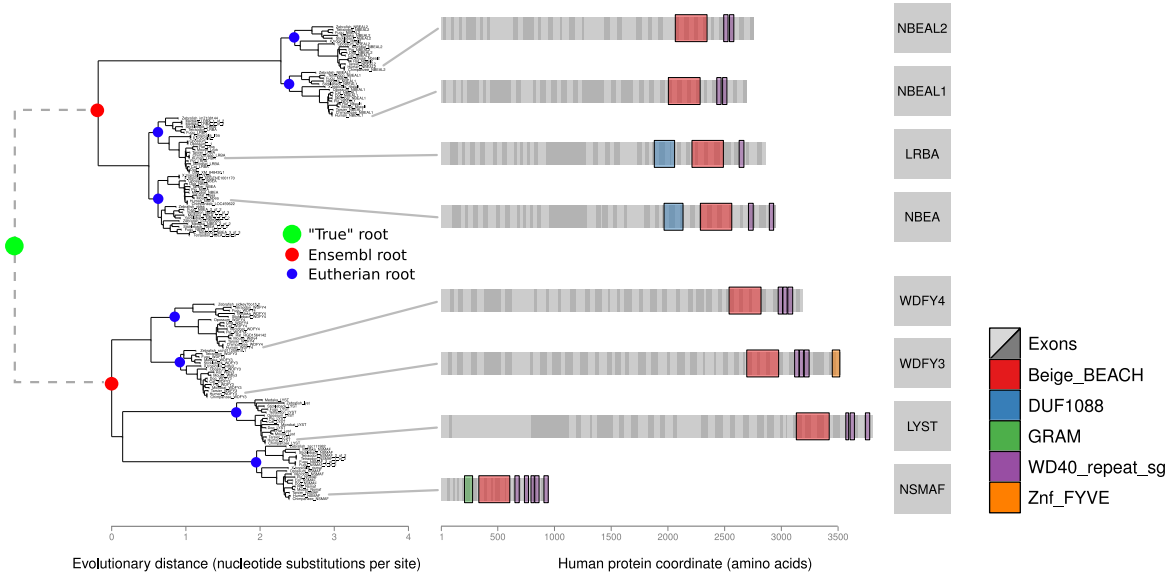


Figure 3.2: The evolutionary history of the human *neurobeachin-like 2* gene (*NBEAL2*) and its paralogs. Left, two phylogenetic trees from Ensembl Compara (release 60) are shown, summarizing the evolution of *NBEAL2* and its three paralogs (top), and *LYST*, a presumed distant paralog of *NBEAL2*, and 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]

3.7 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 taxonomic coverage (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 and 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; in that 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 ongoing gene duplication and loss in mammals [Demuth *et al.*, 2006] 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 relatively star-like shape of the mammalian tree [Bininda-Emonds *et al.*, 2007]: a star-like tree 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.

To test the validity of this hypothesis, I applied a variety of alternative TC-based methods for extracting LOT from the set of Compara gene trees. I compared these trees to the set of "root" Compara trees and to trees built using homology annotations from the Ensembl homology pipeline, with the goal of identifying the most appropriate set of LOTs to select for further analysis.

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 clades examined, and their associated taxonomic coverage constraints (TCCs) used for the tree splitting process, are shown in Table 3.2. 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 evaluated at each node from root to tip: if a given set of TCCs were satisfied by both subtrees below node N, then the tree was split into two subtrees at node N, with the new trees having root nodes placed at the two child nodes, N_a and N_b. The process of evaluating and splitting nodes continued recursively until every node was tested against the TCCs. The smallest subtrees which satisfied the TCCs were thus 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 clades and TCCs to evaluate against the set of Compara trees, all of which were run against the 18,607 "root" trees within the Compara database to generate several genome-wide subtree sets. Table 3.2 shows the details of the chosen clades and the TCCs used for each clade. 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 3.3, including labels on the internal nodes or on the right hand side corresponding to the clade names given in Table 3.2.

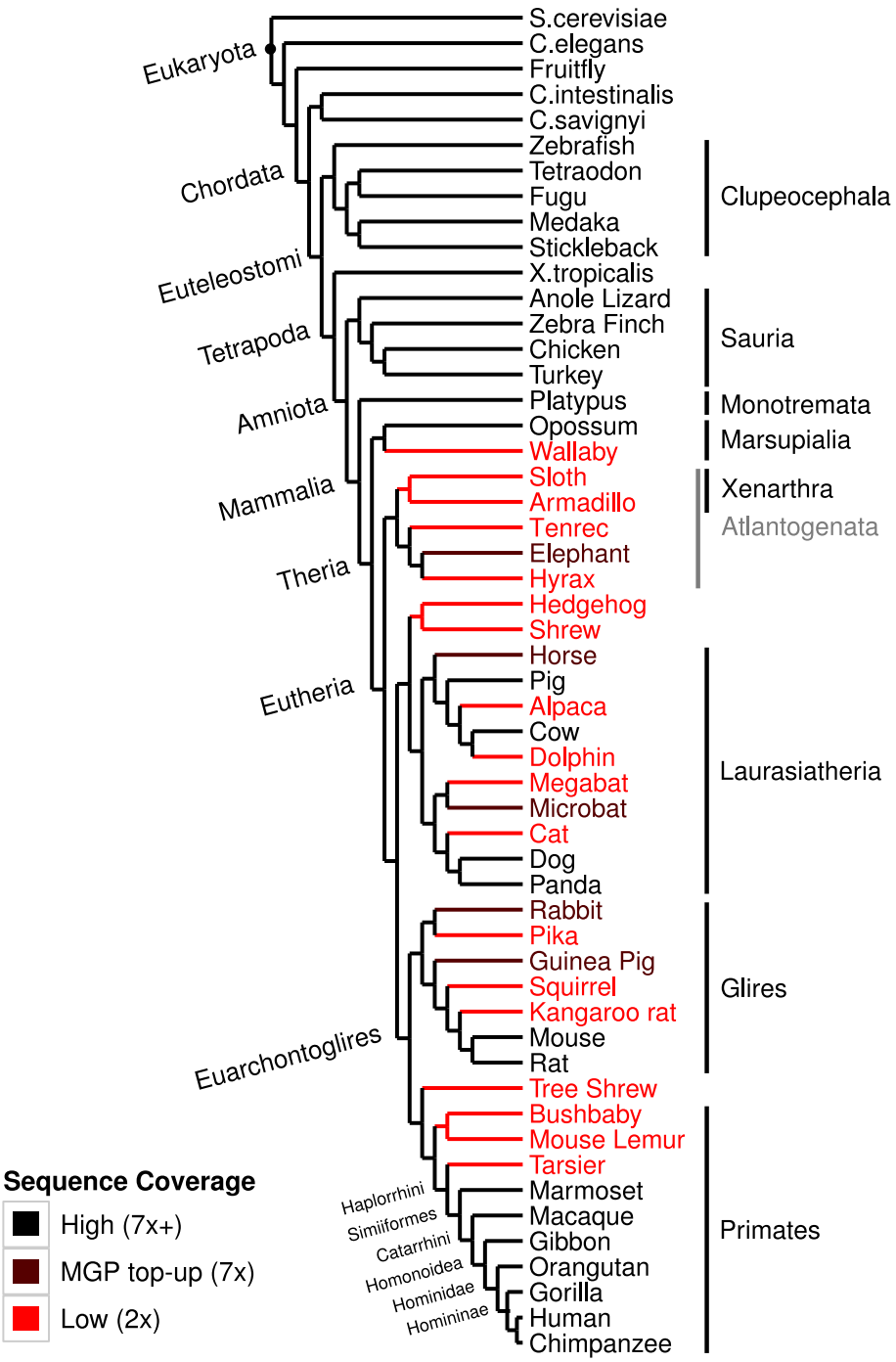


Figure 3.3: The NCBI taxonomy of species within the Ensembl Compara database. Note that branch lengths are not drawn to scale. Species with low-coverage genomes (e.g., those with approximately 2x sequence coverage) are labeled in red and those with high-coverage genomes are in black. Some species were originally sequenced at low coverage for the MGP but were subsequently "topped up" to higher coverage with additional sequencing; those species are labeled in dark red. Clade names are included on the left and on the right side of the tree.

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_{min} constraint required that the TC for all of the included clades was above a given minimum value. To complement the TC_{min} constraint, the TC_{max} constraint required that at least one of the included clades had a TC 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 the Ensembl homology database as orthologs to each gene from a given source species. Thus, for each gene from the source species, the Compara subtree containing all Ensembl-annotated orthologs was extracted and stored; this was guaranteed to yield exactly one subtree for every gene in the source species. This method was tested using human, mouse, zebrafish, and drosophila as source species.

In contrast to the tree-splitting strategy, the ortholog-based approach made use of the orthology annotations resulting from Ensembl's orthology pipeline. This pipeline uses the Compara gene trees as its source, applying a set of rules to each tree to classify the relationships between pairs of genes (e.g., one-to-one orthology, one-to-many orthology, paralogy, etc.). This pairwise approach implicitly uses taxonomic information when classifying gene relationships, allowing one to avoid the inclusion of paralogous genes. The main drawback of this approach was that it was based on annotations of orthology with respect to a single reference species. Unless only strict one-to-one orthology was required—which would have been overly conservative, given the large number of species included here—the use of a reference species resulted in each subtree not necessarily containing a completely unique set of genes. For example, a gene which was recently

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	$TC_{min}(Laur., Glires, Primates) > 0.1$
	MammalSubgroupsPlusOutgroup	$TC_{min}(Laur., Glires, Primates) > 0.1$ AND $TC_{max}(Sauria, Clupeo., Ciona, Marsup.) > 0)$
Orthologs	Human Orthologs	
	Mouse Orthologs	
	Zebrafish Orthologs	
	Drosophila Orthologs	
Root Nodes	Ensembl Roots	

Table 3.2: 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.

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. If those non-human genes evolved with a large amount of positive selection, then this signal would over-represented in the downstream analysis due to the inclusion of multiple copies of those sequences. The tree-based subtree splitting approach was preferable in this regard, guaranteeing that each LOT would only be included once in the analysis. 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 3.2. 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.

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 and 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 3.1). These methods will thus be ignored in the discussion through the remainder of this chapter.

The sets of trees resulting from the different Ingroups TCC methods might be expected to show different characteristics if different groups of species (e.g., fish versus birds or mammals) experienced different large-scale patterns of gene duplication or gene loss subsequent to the divergence of their lineages. Given the well-accepted evidence for a teleost-specific whole genome duplication [Jaillon *et al.*, 2004], I expected the Fish TCC to behave differently, but less certain was whether there would be significant differences between the trees resulting from Sauria or the three mammalian TCC.

The three 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 be due to an excess of primate-specific gene trees that were not captured by non-primate TCCs.

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 (Figure 3.3); 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).

Subtree Method		Tree	Sequence Counts			Human Content			Human	Med.	Med.
Category	Name	Count	Med (Min / Max)	Total	N50	0	1	2+	Total	MPL	Species
Ingroups	Primates	17673	46 (6 / 388)	8.0e+05	48	0.02	0.93	0.05	19024	0.68	42
	Glires	15786	48 (8 / 391)	7.8e+05	49	0.02	0.90	0.08	17904	0.73	44
	Laurasiatheria	15764	48 (8 / 391)	7.8e+05	49	0.01	0.90	0.09	17952	0.73	44
	Sauria	13046	49 (3 / 391)	6.9e+05	51	0.06	0.80	0.14	14988	0.78	45
	Fish	18291	40 (3 / 391)	5.9e+05	49	0.43	0.52	0.06	12183	0.58	38
Outgroups	Eutheria	16477	47 (21 / 391)	7.9e+05	49	0.01	0.92	0.07	18343	0.71	43
	Amniotes	15899	48 (26 / 391)	7.9e+05	49	0.01	0.91	0.08	18094	0.73	44
	Vertebrata	15634	48 (29 / 391)	7.9e+05	49	0.01	0.91	0.08	17938	0.74	44
	Fungi/Metazoa group	14957	48 (32 / 391)	7.8e+05	50	0.01	0.90	0.09	17623	0.76	44
Subgroups	MammalSubgroups	21179	40 (4 / 159)	7.7e+05	46	0.18	0.79	0.03	18595	0.54	37
	MammalSubgroupsPlusOutgroup	17155	46 (5 / 159)	7.8e+05	48	0.05	0.90	0.05	17640	0.71	43
Orthologs	Human Orthologs	19991	49 (2 / 367)	1.0e+06	52	0.00	1.00	0.00	19991	1.07	44
	Mouse Orthologs	21873	50 (2 / 352)	1.2e+06	54	0.10	0.81	0.09	28256	1.01	43
	Zebrafish Orthologs	24540	51 (2 / 392)	1.5e+06	56	0.11	0.76	0.13	30063	1.14	46
	Drosophila Orthologs	9210	60 (2 / 399)	8.6e+05	125	0.08	0.49	0.43	17625	1.22	50
Default Trees	Ensembl Trees	18607	15 (2 / 400)	9.0e+05	139	0.50	0.30	0.20	19995	0.55	8
	Optic Trees	17372	9 (2 / 789)	1.5e+05	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 3.1 and the set of trees from the OPTIC database [Heger and 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.

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 mammalian 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 3.4, shows that several thousand trees fit the expected model. If we assume that all trees from the Fish subset that contain no 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 15,000–17,000, 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 to produce a TC value of 0.6. The minimum subtree size ranged from

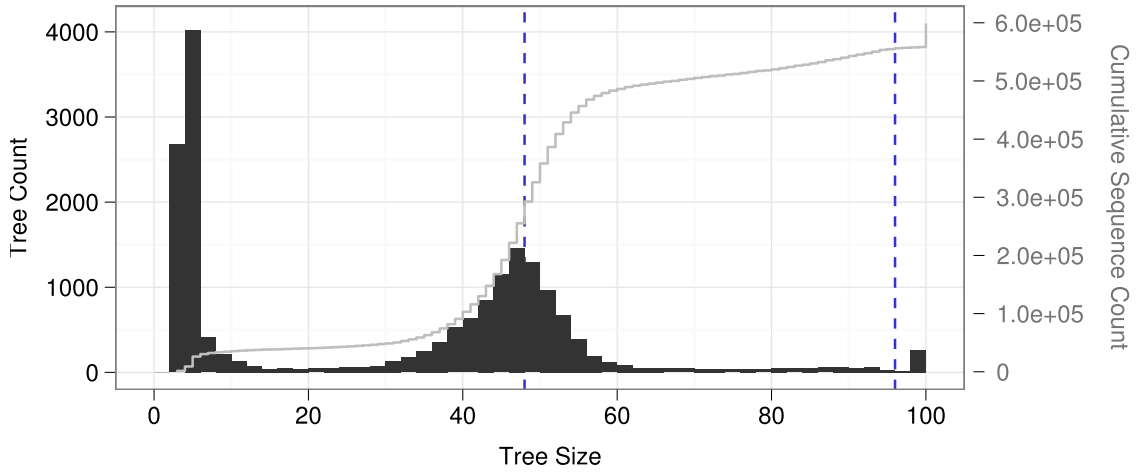


Figure 3.4: Tree sizes for the set of subtrees identified using the Fish clade taxonomic coverage constraint, showing an excess of small subtrees resulting from the teleost genome duplication. Black bars show a histogram of tree sizes in bins of width 2, a gray line shows the cumulative number of sequences 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.

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 very different from 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

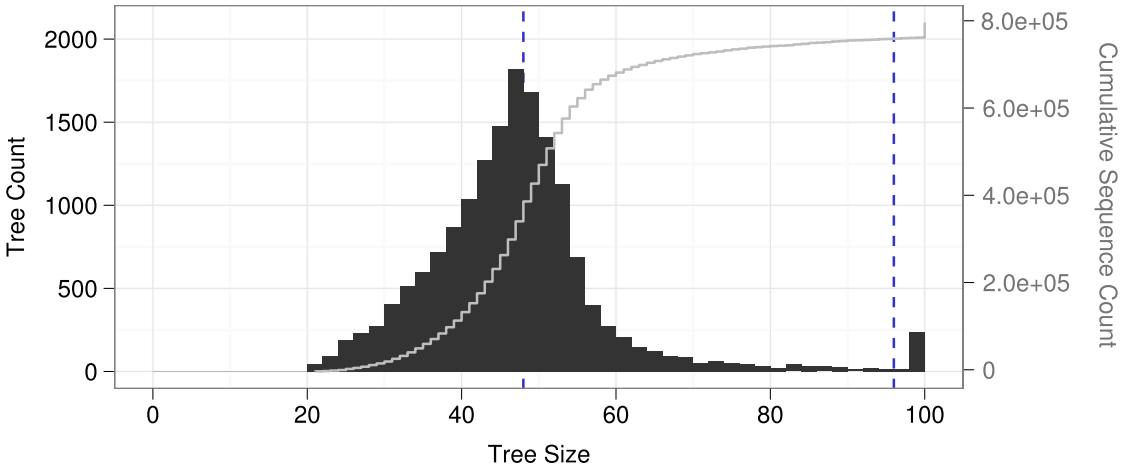


Figure 3.5: Tree sizes for the set of subtrees identified using the Eutheria clade taxonomic coverage constraint. Black bars show a histogram of tree sizes in bins of width 2, a gray line shows the cumulative number of sequences contained within trees of that size or smaller, and dashed blue lines are 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.

contained an apparent orthologous relationship with one human gene, which was only slightly lower than the 92% of Eutheria subtrees containing one human sequence.

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 3.5. 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.

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

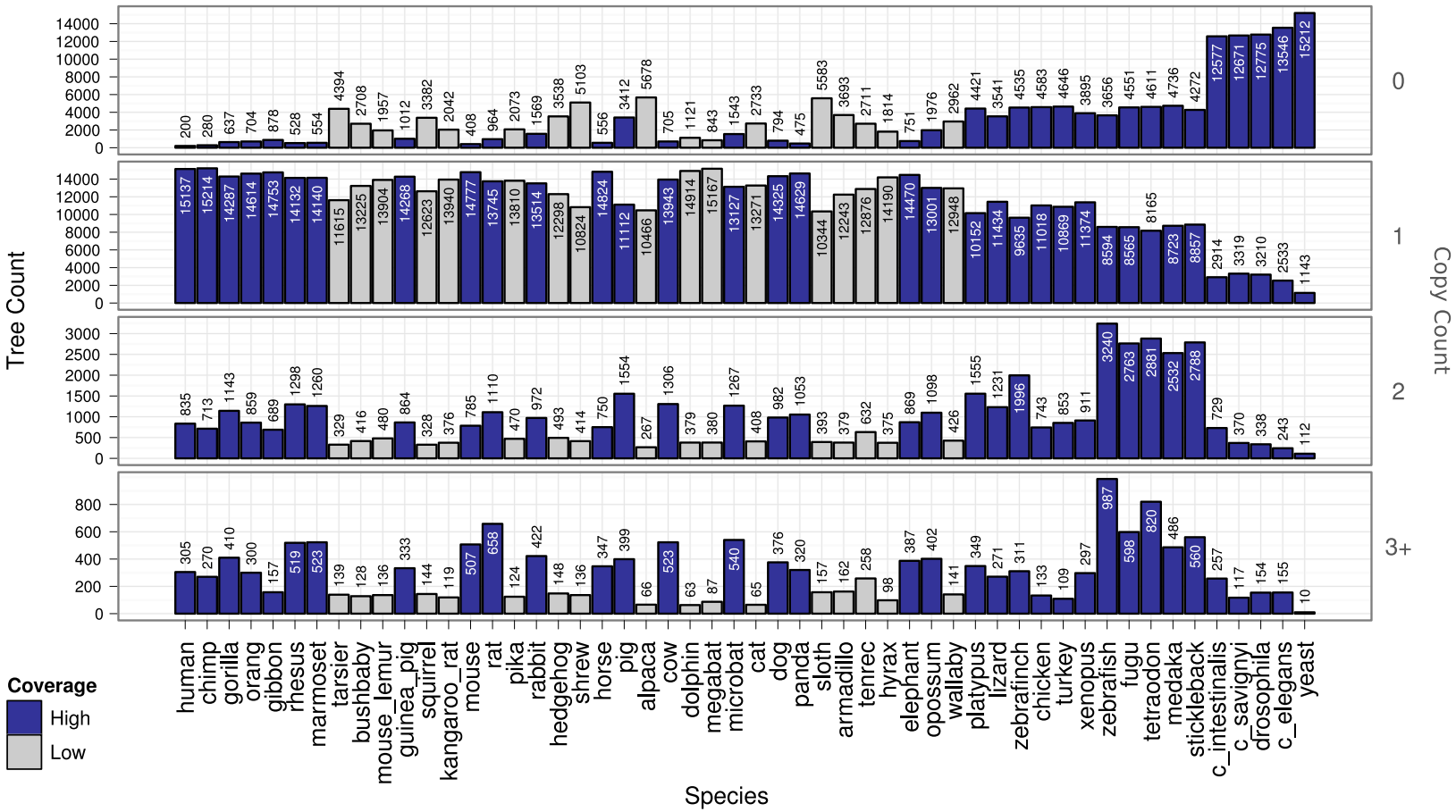


Figure 3.6: Taxonomic distribution of gene copy counts for the Eutheria subtrees defined by taxonomic coverage constraints. Results for the Primates, Glires, and Laurasiatheria are omitted for clarity; they showed similar characteristics to the Eutheria, Amniotes, and Vertebrata methods 3.3). Each panel 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.

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 3.6 shows the gene copy counts for the set of Eutherian subtrees across all Ensembl species. The excess of zero-copy trees and deficit of duplications in low-coverage genomes is immediately apparent from Figure 3.6, confirming the trend observed in the set of root Compara trees.

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: 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 trees; at the low end were chicken and turkey with 876 and 972 multi-copy trees, respectively, and at the high end (excluding the fish species) were zebrafinch and platypus with 1904 and 2307 multi-copy trees, respectively.

In the analysis of vertebrate genomes, one must be 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 comparison 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

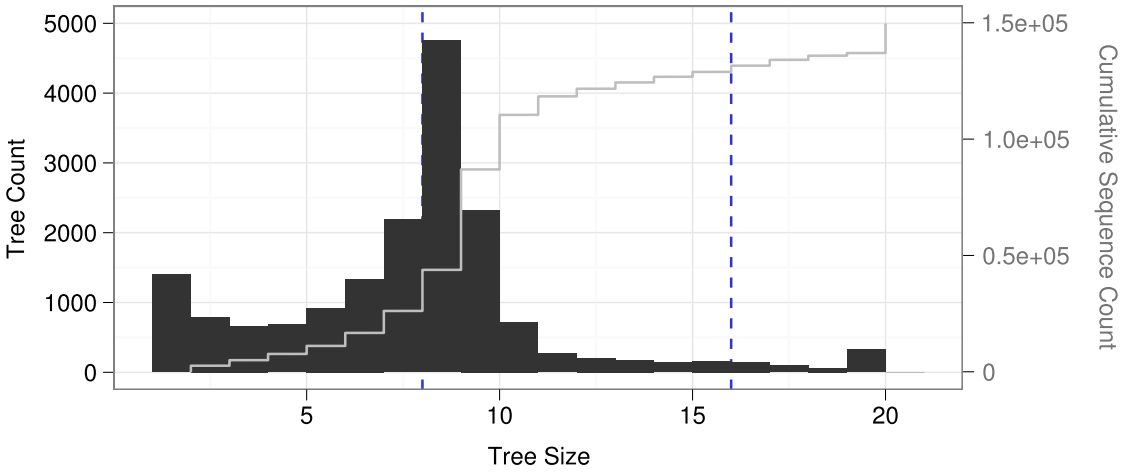


Figure 3.7: Tree sizes for gene trees in eight amniote species from the OPTIC orthology database [Heger and Ponting, 2008]. Black bars show a histogram of tree sizes, a gray line shows the cumulative number of sequences contained within trees of that size or smaller, and dashed blue lines are drawn at integral multiples of 8, the number of amniote species analyzed by OPTIC. Trees with more than 20 sequences are included in the topmost bin.

expect the number of zero-copy trees in these species to decrease as finished-quality genome sequences are 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 evaluate whether a different phylogenetic tree-based orthology pipeline produced a similar distribution of gene tree sizes. The OPTIC database, a project from Christ Ponting's 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 and 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 (included in the bottom row of Table 3.3)

and plotted the distribution of gene tree sizes in Figure 3.7.

After accounting for differences in the number of sampled species, the set of OPTIC gene trees more closely resembled the set of Eutherian subtrees than the root Compara gene trees (Table 3.3). 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 3.5 and 3.7 clearly show that the OPTIC and Eutherian trees contain similarly-shaped distributions of sequence counts; 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 3.1.

3.11 Conclusions

This chapter was concerned with identifying a set of largely orthologous trees (LOTs) for use in a downstream analysis of sitewise selective pressures in mammalian orthologs. 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. The Ensembl Compara database as a source of inferred gene trees due to its well-established methodology [Heger and Ponting, 2008; Vilella *et al.*, 2009] and because genes from low-coverage mammalian genomes were included in the inference pipeline, resultin gin 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 variety of ways, including using 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 taxonomic coverage constraints (TCCs) and applied the method 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.

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. The increased number of zero-copy trees reflected an expected amount of loss resulting from missing sequence data in low-coverage genomes and was not too worrying with respect to the intended downstream analysis. The decreased number of multi-copy genes was concerning, however, as it suggested that the assembly of duplicated genomic regions may frequently be "collapsed" in low-coverage assemblies, possibly resulting in genes containing chimeric sequence derived from different duplicated segments.

Other, more subtle trends in the patterns of gene copy counts 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.

A major conclusion of the analysis presented in this chapter was that the use of taxonomic criteria is important in identifying mammalian LOTs. Other gene tree databases, such as TreeFam [Ruan *et al.*, 2008] and OPIC [Heger and Ponting, 2008] explicitly use taxonomic information or outgroups to ensure that orthologs are consistently clustered, but Compara currently does not use such information in building its gene trees. As a result, I developed a simple method to extract LOT from the set of "root" Compara trees and evaluated the method by analyzing the distribution of tree sizes, the number of human gene copies per tree, and the taxonomic distribution of gene copy counts for the TC-defined LOT compared to the "root" Compara trees. Further comparison of these statistics to a separate dataset of amniote orthologs provided confirmation that the resulting trees more closely resembled previously-published sets of orthologous groups.

Chapter 4

Patterns of sitewise selection in mammalian genomes

4.1 Introduction

This chapter describes the use of sitewise evolutionary estimates to characterize the genome-wide distribution of selective constraint in mammals and within the major mammalian superorders. I apply the Sitewise Likelihood Ratio (SLR) test, which was introduced in Chapter 1 and evaluated in Chapter 2, to the set of orthologous gene trees from Chapter 3 to estimate statistics measuring sitewise selective constraint in several groups of mammalian species. Both this chapter and Chapter 5 are concerned with the analysis of these sitewise data: here I will analyze the overall distribution of constraint observed in several groups of mammalian genomes, and Chapter 5 will apply these sitewise data to identify genes, biological processes and protein domains with the strongest genome-wide enrichment for signals of positive selection.

The first section of this chapter describes the filtering and alignment of mammalian orthologs and introduces a protocol for filtering sitewise estimates of selective pressures. Although the simulations from Chapter 2 showed that sequences with mammalian-like divergence levels showing biological patterns of insertion and deletion can be aligned without introducing many false positive positively selected codons (PSCs), the analysis of real sequence data involves many potential non-biological 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, each one of which involves potential errors and biases. Errors may arise during the sequencing of DNA bases, assembly of genomic fragments, and annotation of gene-coding regions; each of these steps has been previously highlighted as an important source of error in the large-scale analysis of genomic alignments [Schneider *et al.*, 2009; Mallick *et al.*, 2009; Milinkovitch *et al.*, 2010; Hubisz *et al.*, 2011]. As such, care was taken in this

study to design and evaluate a variety of filters to reduce the probability of yielding misleading results.

The second portion of this chapter presents an analysis of the global distribution of mammalian selective constraint, using sitewise estimates to identify sites evolving under purifying and positive selection in different groups of species. In parallel with the major goal of the Mammalian Genome Project (MGP) to better identify and understand the nature of evolutionary constraint across mammalian genomes, the purpose of this analysis was to better characterize the distribution of evolutionary constraint within mammalian protein-coding regions. Thus, a major question was what proportion of protein-coding material has been evolving under purifying, neutral, or positive selection in mammals. Proteins are well understood to evolve under strong purifying constraint due to their functional importance [Fay and Wu, 2003], but some regions of proteins, such as disordered regions between two well-folded domains, may evolve under relaxed constraints. Furthermore, positive selection of beneficial substitutions can also play a role in shaping the evolutionary history of proteins [Pál *et al.*, 2006]. There has thus 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 [Ellegren, 2008; Marques-Bonet *et al.*, 2009].

While most genome-wide analyses of selective constraint have focused on the gene as the unit of analysis [Nielsen *et al.*, 2005; Mikkelsen *et al.*, 2005; Kosiol *et al.*, 2008], this chapter adopts a primarily sitewise approach, presenting distributions of sitewise estimates aggregated across all sites from all genes included in the analysis. The use of explicitly sitewise estimates allowed for various types of filtering to be applied, removing sites sites from the dataset according to different filtering criteria. Different sitewise filters could be applied without the computationally expensive step of re-estimating evolutionary parameters, allowing for the impact of various filters on the amount of inferred positive and purifying selection to be quickly and flexibly estimated.

4.2 Data quality concerns: sequencing, assembly and annotation error

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 genomes sequenced by the MGP. A number of issues relating to the impact of low-coverage genomes on the detection of orthologs were discussed in Chapter 3; here, the focus was on how, after orthology has been inferred, sequences from low-coverage genomes may contribute to the false detection of positive selection. Although the SLR test and other sitewise maximum likelihood methods have been shown to be conservative in their identification of positively selected

sites under most conditions, even when the amount of data is low or the null model is violated [Anisimova *et al.*, 2002, 2003; Massingham and Goldman, 2005], most evolutionary analyses are based on the assumption that all sites within an alignment column are truly homologous. This assumption can be violated in a number of ways, resulting in misalignment.

In Chapter 2 I explored misalignment resulting from errors in reconstructing the evolutionary history of sequences evolving with insertions and deletions. Simulations showed that PRANK_C alignments contained little misalignment and caused few false positives in the sitewise detection of positive selection. However, biological insertions and deletions are not the only potential source of misalignment. An additional concern was the potential for errors resulting from the inclusion of erroneous or non-homologous sequence regions within the mammalian orthologs. As the low-coverage genomes were not assembled into chromosomes and contained large amounts of missing sequence, the likelihood of miscalled bases, spurious insertions or deletions, or shuffled regions due to mis-assembly was relatively high [Green, 2007]. Most aligners were not designed to deal with these types of errors, so they may be expected to result in excess misaligned regions.

The impact of these types of errors on detecting positive selection at any given codon should depend on the model used to infer selection, the number and identify of the non-homologous nucleotides placed in the same aligned codon, and the branch length leading to the sequence containing the misaligned bases. Misalignment of multiple nucleotides in the same codon would tend to produce more false positives than single-nucleotide errors, and misalignment in sequences or trees with shorter branch length may have an overall greater impact on the estimated nonsynonymous substitution rates.

Empirical evidence for a strong impact of sequencing, assembly and alignment error

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 be difficult to predict without a reliable model for their generation. This is especially true for larger-scale errors from mis-assembly or mis-annotation, which are less easily modeled than some other types of error, e.g., base calling, 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. Two recent empirical studies in mammals provided convincing evidence that sequence, alignment and annotation errors can increase the number of false positive PSGs in the branch-site test for positive selection (Zhang *et al.* [2005]; introduced in Chapter 1).

Schneider *et al.* [2009] performed a genome-wide scan for positive selection in the terminal branches of 7 mammalian genomes using the branch-site test and analysed the fraction of PSGs

within subsets of high- or low-quality genes according to three sequence and alignment quality metrics. They found that the fraction of PSGs was significnatly higher for genes exhibiting lower quality sequence, annotation and alignments, with genes in the highest-quality and lowest-quality categories showing a 7.2-fold difference in the inferred fraction of PSGs [Schneider *et al.*, 2009]. This observation provided evidence of a correlation between the chosen quality metrics and the tendency of an alignment to exhibit positive selection. It did not necessarily imply causation, however, as the same result might have been observed—even in the absence of sequence error—if some biological properties of the true PSGs caused them to yield lower quality metrics than non-PSGs. Looking at the three metrics used in their study (sequencing coverage, gene annotation status, and alignment quality according to the heads-or-tails method), it is plausible that properties associated with elevated ω ratios and positive selection, such as recent gene duplication [Beisswanger and Stephan, 2008; Studer *et al.*, 2008; Casola and Hahn, 2009], high GC content [Ratnakumar *et al.*, 2010] or functional shifts [Storz *et al.*, 2008; Wang and 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 and Yang, 2010], so the authors' results based on this measurement were questionable. Despite these criticisms, overall Schneider *et al.* [2009] provided good evidence that some measurable 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 lineage in 59 genes which had been identified as chimpanzee PSGs by Bakewell *et al.* [2007]. Mallick *et al.* [2009] were concerned that the finding by Bakewell *et al.* [2007] of a larger proportion of PSGs in chimpanzee than in human was the result of the lower-quality chimpanzee genome rather than a biologically significant difference in levels of adaptation. Mallick *et al.* [2009] 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. This suggested that the original 4x coverage chimpanzee genome contained a number of sequencing and assembly errors leading to false inferences of positive selection. The authors' detailed analysis of 302 codons with multiple spurious nonsynonymous substitutions in the original chimpanzee assembly showed roughly comparable contributions from sequence error (explaining 23% of codons), assembly error (14% of codons) and local alignment error (30% of codons).

Taken together, the results of Schneider *et al.* [2009] and Mallick *et al.* [2009] provide strong evidence in support of the hypothesis that errors in sequencing, assembly, annotation and alignment can result in strongly elevated inferred ω values when using sensitive tests for detecting

positive selection. Furthermore, the detailed identification and quantification of error sources performed by Mallick *et al.* [2009] shows 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 while the current analysis focused on detecting sitewise positive selection throughout the tree, their results could be expected to generalize well enough to guide the design of filtering methods for the present sitewise analysis.

Three filtering steps were implemented 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

I first applied a conservative filter to the set of input sequences based on sequence quality scores associated with each genome for which such scores were available. Most automated genome assembly pipelines 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. These scores represent the probability, calculated by the sequencing and/or assembly program, that a given base call is incorrect. This probability is concisely expressed as the negative base-10 logarithm of the probability of an error multiplied by 10, 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].

Ensembl does not store quality scores from its source genome assemblies, so Phred quality scores were downloaded for all low-coverage genomes where Phred-like quality scores were 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.

A suitable score threshold for filtering coding regions was chosen based on a study by Hubisz *et al.* [2011], who performed a detailed analysis of Phred quality scores and 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 [Birney *et al.*, 2007]. They also 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.* [2011] noted that 85% of the 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 score threshold of 25 was chosen as a reasonable trade-off

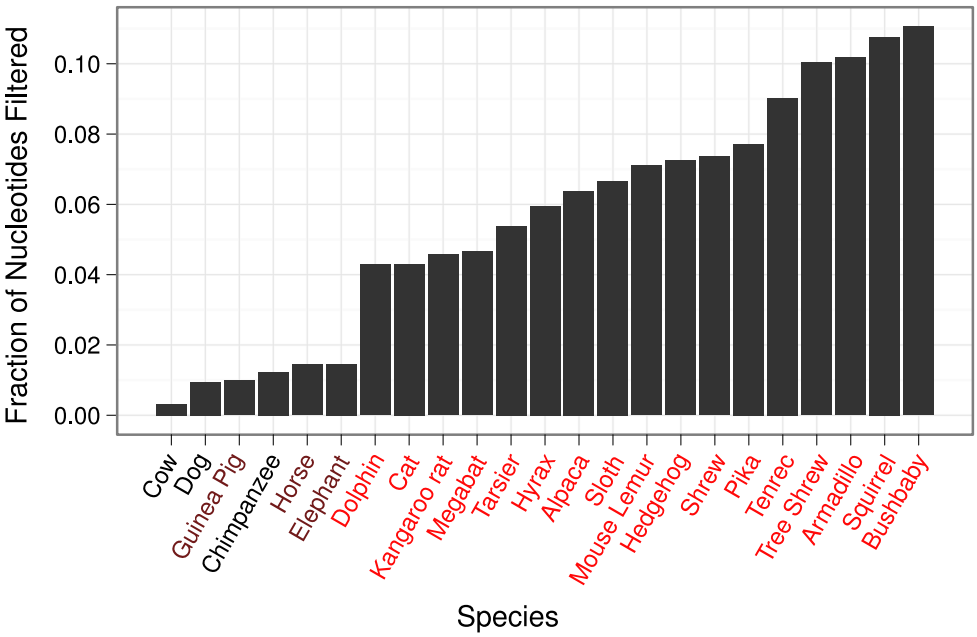


Figure 4.1: The fraction of nucleotides masked from protein-coding regions of mammalian genomes based on low predicted sequence quality. For genomes with Phred sequence quality scores available, all codons containing any nucleotides with a Phred score < 25 were masked with 'N's. Each bar shows the fraction of nucleotide sequences (compared to the total number of coding nucleotides from each genome) which were masked. As in Figure 3.3, high-coverage genomes are labeled in black, low-coverage genomes are labeled in red, and guinea pig and elephant—whose genomes were originally sequenced at 2x coverage but were later "topped up" to 7x coverage–are labeled in dark red. Phred scores were unavailable for most high-coverage genomes, and Phred scores for wallaby were unable to be used due to mismatched assembly versions (see text).

between the potential benefit of removing 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 above filtering scheme was applied to all coding sequences from each genome for which quality scores were available, which included all of the low-coverage genomes (except for wallaby) and six high-coverage genomes. Unfortunately, the low-coverage wallaby genome was not filtered based on sequence quality due to a mismatch in the sequence identifiers used by Ensembl and those found in the quality score files made available for download; wallaby was not one of the 29 species included in the original MGP analysis, but this error has been marked for correction in a future version of the 38-species analysis. Note also that the guinea pig, rabbit, microbat, horse and elephant genomes were originally sequenced at low 2x coverage for the MGP, but they since underwent additional sequencing to produce high-coverage 7x assemblies (these species are

labeled in dark red in Figure 3.3 from Chapter 3). These higher-coverage assemblies were used for release 63 of Ensembl and for the present analysis. Phred scores for the high-coverage guinea pig, horse, and elephant assemblies were used for filtering here, but Phred scores for the high-coverage rabbit and microbat assemblies were not provided by the sequencing institutions.

The fraction of nucleotides filtered from each genome is shown in Figure 4.1. Genomes with high-coverage assemblies contained fewer bases with low Phred scores, resulting in 1-2% of nucleotides being filtered, while the bulk of low-coverage genomes resulted in 4-8% of nucleotides being filtered. 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 based on Hubisz *et al.* [2011] that 5-10% of nucleotides would be filtered using a Phred score threshold of 25, and the variation in filtered nucleotide proportions between different low-coverage 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.

The lack of publicly available sequence quality scores for most high-coverage genomes was lamentable, especially for the closely-related primates. Taking chimpanzee as an example high-coverage genome with quality scores available (although it may have an exceptionally error-prone genome sequence [Mallick *et al.*, 2009]), the procedure described above resulted in 331,737 nucleotides, or roughly 1% of all protein-coding nucleotides, being masked. Clearly some of those masked nucleotides were of lower quality and some were of higher quality (i.e., high-quality nucleotides removed on account of being within a masked codon). But if one assumes a constant error probability among masked nucleotides of $3.16e-3$ (corresponding to the $Q = 25$ threshold), then an expected 1,049 of the masked chimpanzee nucleotides contained sequencing errors. An equivalent calculation for bushbaby, a low-coverage primate genome, yielded 2,694,054 filtered nucleotides and 8,519 expected sequencing errors.

These numbers may be evaluated in terms of signal and noise, where true substitution events are the desired evolutionary signal and errors are noise. To estimate the amount of "signal" present in each species' terminal lineage, the genome-wide set of filtered alignments was used to infer substitution events along each branch (the methods used for this are described in Section 4.2), yielding 85,368 substitutions along the chimpanzee branch and 892,890 substitutions along the bushbaby branch. Comparing the inferred signal to the estimated noise in each of these example genomes, one finds that for each true lineage-specific substitution there were approximately 0.0095 expected errors in the bushbaby genome and 0.0123 expected errors in the chimpanzee genome. Thus, even though the chimpanzee genome was sequenced to 6x coverage while bushbaby was only sequenced to 2x coverage, the unmasked chimpanzee genome contained perhaps a lower "signal-to-noise" ratio within coding sequences than the unmasked bushbaby genome.

These calculations were in general agreement with a study by Taudien *et al.* [2006] which

appraised the extent to which errors in an earlier working draft (WD) version of the chimpanzee genome affected the comparative analysis of coding regions. Although the overall error rate of the WD sequence was low, they found that up to 20% of the exonic sequence differences between human and chimpanzee were false positives resulting from errors in the chimpanzee sequence. When a quality score threshold of $Q > 20$ was used, however, the proportion of false positives decreased markedly to 8% [Taudien *et al.*, 2006]. The issue of sequence quality will be considered again in Chapter 6, where I examined the impact of sequence filtering on the estimation of genome-wide dN/dS values within primates.

Removing recent paralogs

Chapter 3 discussed a number of issues relating to the identification of homologous protein-coding sequences, and a set of largely orthologous gene trees was collected. The relatively low amount of sequence divergence within mammals suggested that the probability of erroneous inclusion of altogether non-homologous sequences was low, but the existence of paralogous relationships within the alignments used for sitewise analysis was still a concern.

The inclusion of paralogous gene relationships in a large-scale analysis of orthologous gene evolution may produce unwanted signals of adaptive evolution following gene duplication [Lynch and Conery, 2000], artifacts resulting from gene conversion [Casola and Hahn, 2009] or biases due to lineage-specific family expansion, a process which is relatively common in mammalian gene families [Gu *et al.*, 2002]. As a result, it has traditionally been considered important to filter out recently-duplicated genes (e.g., genes duplicated after the whole-genome duplication event in the vertebrate ancestor) in large-scale evolutionary analyses. Previous genome-wide scans for positive selection involving six or fewer mammalian genomes have either required strict one-to-one orthology [Clark *et al.*, 2003; Nielsen *et al.*, 2005] or allowed very limited numbers of recent duplications in specific lineages [Kosiol *et al.*, 2008]. With larger numbers of species included in a phylogenetic analysis, 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 included, 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 or shorter paralogous copy from the gene tree.

An additional complicating factor in the current analysis was the concern that many of the apparent gene duplications might actually be 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 intervening 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. This result might not be detrimental to sitewise analysis in itself, as each shortened gene may end up being correctly aligned and provide useful information to the alignment. However, 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 the shortest of these gene fragments.

Sequence divergence was the other criterion used to select which paralogous copy of recently-duplicated genes to retain. Models of evolution after gene duplication have tended to predict 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 analysis. Gene length was used primarily to discriminate spuriously shortened genes from true genes, and sequence divergence was used to distinguish between more- and less-diverged paralogs. First, the mean pairwise sequence distance was calculated between each putative paralog and all other sequences in the gene tree, resulting in one mean pairwise distance estimate per putative paralog (hereafter referred to as the mean distance). For these distance calculations, the coding sequence alignments provided by Ensembl Compara and the JC69 nucleotide model were used 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 twp or more genes from the same species 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 one sequence yielded a mean

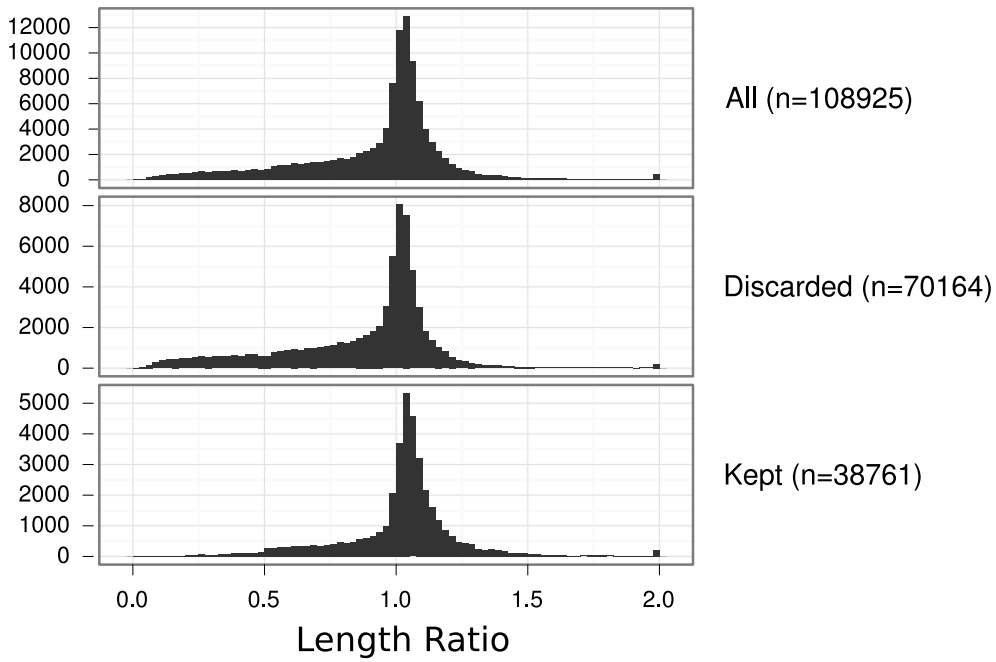


Figure 4.2: Length ratios of putative paralogs. The length ratio was calculated as the length of a putative paralogous copy divided by the mean length all sequences its corresponding gene tree. Putatively paralogous genes (top panel) were either discarded (middle panel) or kept (bottom panel) according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

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 108,925 putative paralogs within the 9,604 gene trees containing at least one set of putative paralogs. Figure 4.2 shows the distributions of length ratios for the set of all putative paralogs, those discarded from the alignments, and those kept for subsequent analysis. The overall distribution of length ratios showed that most putative paralogs had lengths similar to the mean length across the gene tree (with a peak at or slightly above 1), but the shape of the distribution was asymmetric, with a bias towards shorter lengths. The filtering protocol effectively removed these shortened genes, as evidenced by the enrichment of lower length ratios in the distribution of discarded genes and the less skewed distribution of length ratios in the set of 38,761 retained putative paralogs.

A more 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

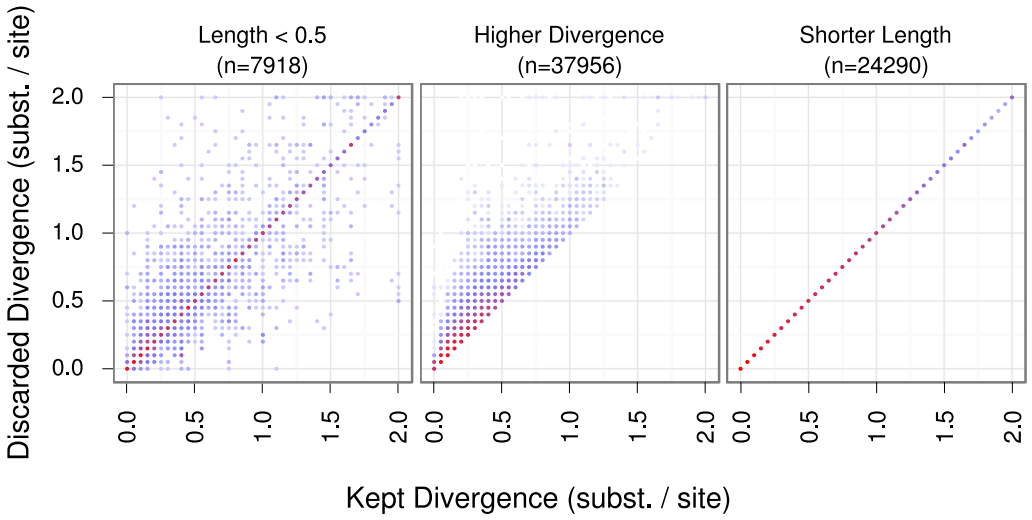


Figure 4.3: Sequence divergence of retained and discarded putative paralogs. Each data point was a gene which was discarded from the tree for one of three reasons: it had a length ratio of less than 0.5 while the retained copy had a length ratio greater than 0.5 (*Length* < 0.5; right panel), it had more sequence divergence than the retained gene (*Higher Divergence*; middle panel), or it had equal sequence divergence but shorter length than the retained gene (*Shorter Length*; right panel). Divergence was measured as the mean pairwise divergence between the gene and all other sequences in the tree. Each colored 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.

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 perhaps only a slight bias towards genes lying above the diagonal (i.e., in the direction of greater divergence in the discarded copy). This was in line with the expectation 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 by 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 all 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.

This was obviously not the most conservative approach to dealing with recent duplications. One could instead remove all putatively paralogous copies from the gene tree, creating an apparent gene deletion in that species, or simply ignore all gene families with any recent duplications (e.g., require one-to-one orthology allowing for gene deletions). The latter option would likely be overly restrictive for any sensible genome-wide analysis, 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 genes containing 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)

Identifying clusters of nonsynonymous substitutions

After filtering for sequence quality and removing paralogous genes and shortened gene fragments, PRANK_C was used to align the codon sequences of each of the 16,477 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. Examples of two such regions are shown in Figure 4.4. These obviously erroneous stretches were likely due to mis-assembly of a genomic region or misidentification of exon boundaries within the gene of one species. In the examples in Figure 4.4, most of the nonhomologous material was inferred to be a lineage-specific insertion in the mis-annotated species; these regions were not of concern, as SLR ignores single-species insertions because they contain no evolutionary information. More concerning, however, were the regions indicated with red brackets in Figure 4.4 where the apparently non-homologous material was aligned with other species. These regions 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 4.2, this type of error is particularly prone to cause false positives in the detection of positive selection.

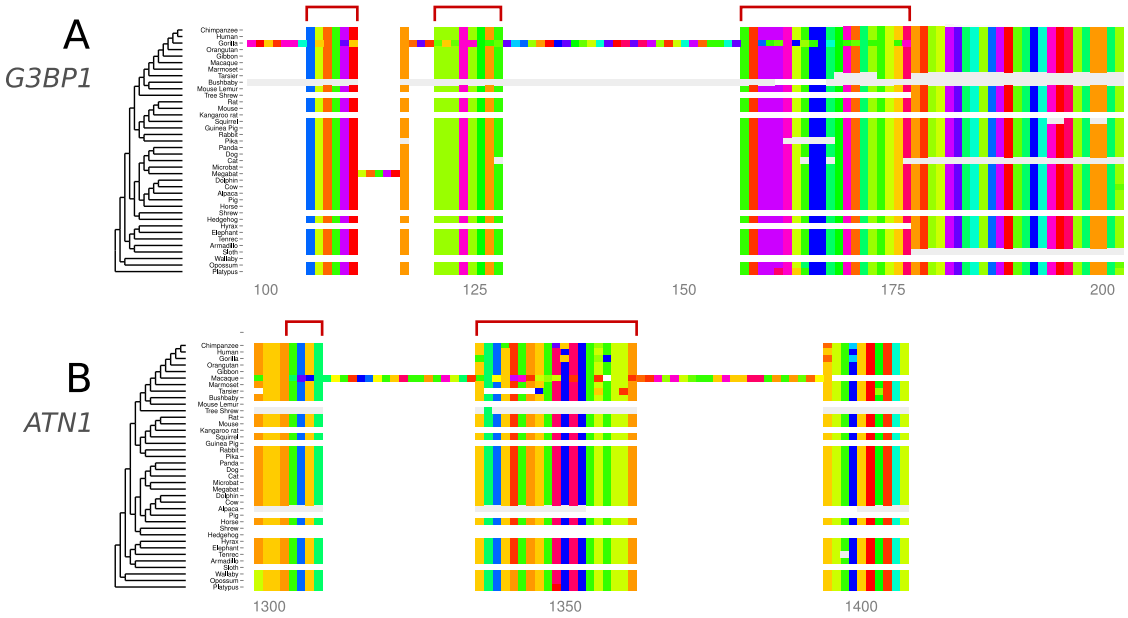


Figure 4.4: Two regions of protein-coding mammalian alignments with stretches of non-homologous sequence in one species. Alignments are shown with amino acids colored according to Taylor [1986]. (A) The *G3BP1* alignment, showing a mis-annotated exon in gorilla resulting in misalignment. (B) The *ATN1* alignment, showing a mis-annotated exon in macaque resulting in misalignment. The regions indicated by red brackets contained aligned non-homologous material that might cause false positives in detecting sitewise 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, while the platypus branch is very long. Thus, one would expect to observe many more lineage-specific substitutions in platypus than in human for a given alignment window. In contrast, a stretch of non-homologous aligned sequence should introduce, on average, a constant number of non-synonymous and synonymous substitutions into the branch ancestral to the sequence in which it

exists. For this reason it should be more difficult to distinguish homologous from non-homologous stretches in species with long terminal lineages. On the other hand, this trend should also serve to limit the 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; Wang *et al.*, 2007]. 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, 2007] under the M0 model (e.g., assuming one ω for all sites and all branches in the tree) to perform the marginal reconstruction of ancestral sequences at internal nodes [Yang *et al.*, 1995] and to identify the substitution events implied by the reconstructed ancestral sequences of each gene alignment. 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 confidently inferred substitution events along internal and terminal branches of the mammalian phylogenetic tree.

Counts of synonymous and nonsynonymous substitutions along each branch were separately collected for non-overlapping 15-codon alignment windows; the results for a selection of species and internal nodes are shown in Figure 4.5, 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. Each window is thus represented twice in Figure 4.5: once in the nonsynonymous histogram, and once in the synonymous histogram. Windows with no substitutions along the given branch are represented in the left-most synonymous and nonsynonymous bins. The mean length of the branch ancestral to the given node, calculated from the set of branch lengths estimated by *codeml*, is indicated in parentheses after each node name.

Figure 4.5 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 4.5, top panel),

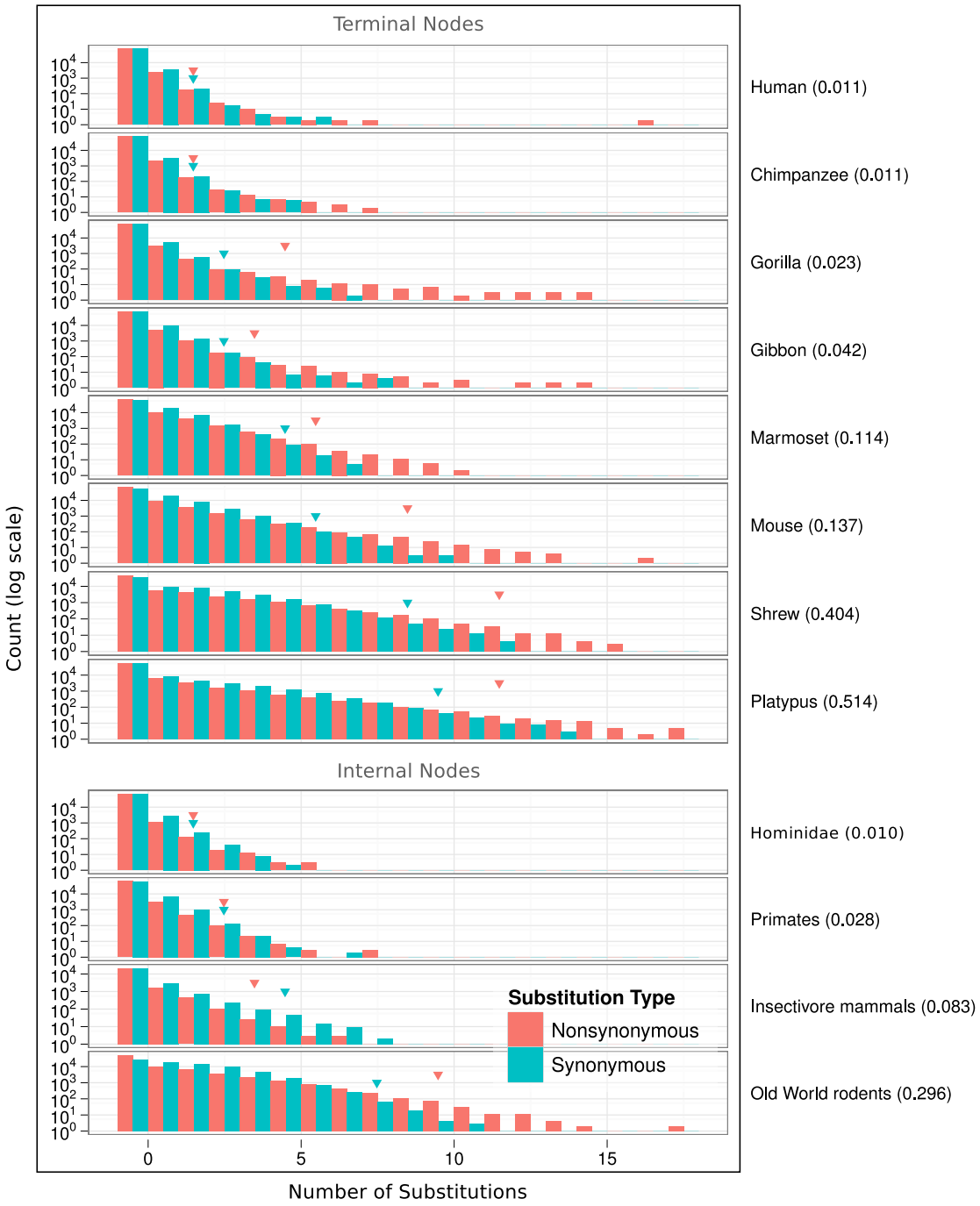


Figure 4.5: Counts of inferred nonsynonymous (red bars) and synonymous (blue bars) substitutions in 15-codon windows along terminal and internal branches of the mammalian tree. The leftmost two bars correspond to windows with 0 substitutions, the next two bars correspond to windows with 1 substitution, 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 mean length of the branch ancestral to each node is included in parentheses after the node label.

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 substitutions), 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 in most of the terminal branches 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 or lower quantile positions for nonsynonymous versus synonymous substitutions, but the rodent ancestral node showed a pattern more similar to the terminal nodes in Figure 4.5, with a higher 99% quantile for nonsynonymous substitutions. 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 raised the possibility that the entire rodent clade contains many misaligned non-homologous stretches due to shared differences in gene annotations between rodent and non-rodent species.

The end result of this analysis was the identification, for each terminal node of the mammalian tree, of windows with nonsynonymous substitution counts above the top 0.1% of 15-codon windows; 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. A qualitative analysis of regions containing windows at a variety of thresholds found the 0.1% threshold to strike a good balance between sensitivity and specificity.

In total, 37,824 windows containing potential stretches of non-homologous aligned sequence

	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 16,477 branch lengths estimated by SLR for each gene. MPL – mean path length.

were identified across 8,951 alignments, with 881 genes containing more than 10 windows each. The locations of these windows were stored for later use in defining the most conservatively-filtered sitewise dataset.

4.3 Species groups for sitewise analysis

For each alignment of mammalian 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 mean path length (MPL) and total branch length of their subtrees, estimated as the median value across all 16,477 genes' estimates of dS distances.

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

istics, 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.*, 2007; Smith and 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 (i.e., wallaby, platypus, armadillo) in favor of only the high-coverage Eutherian genomes (i.e., 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.

4.4 Evaluating and filtering sitewise results

Sitewise data were collected from SLR and stored in a database for storage and further analysis. The Mammals group, containing the greatest total branch length of all the datasets and repre-

senting 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.

Two genome-wide datasets were generated by processing sitewise data separately with two levels of filtering: 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 misalignment or large numbers of gene duplications.

I first examined the overall distributions of ω estimates and sitewise LRT statistics from SLR. Figures 4.6 and 4.7 show the distributions of six sitewise statistics for each group of species. Four of the statistics (Omega, LRT Statistic, Site Pattern and Random) were collected from the output of SLR and two of the statistics (Non-gap Codons and Non-gap Branch Length) were directly calculated from the codon alignments. The Omega statistic is SLR's maximum likelihood (ML) estimate of ω, hereafter referred to as ω_{ML}. The LRT Statistic is the raw statistic resulting from the sitewise likelihood ratio test (LRT) performed by SLR. Following Massingham [2005], a signed version of the LRT statistic, hereafter LRT_{SLR}, is used throughout this chapter. The LRT_{SLR} is formed by negating the raw LRT statistic for sites where $\omega_{ML} < 1$; the signed statistic is a useful measure by which to sort sites according to their evidence for purifying and positive selection. The Site Pattern is a categorical classification of the pattern of synonymous and nonsynonymous substitutions at each site: a site is "constant" if it contains no differences, "synonymous" if it only contains synonymous differences between sequences, and nonsynonymous if it contains at least one nonsynonymous difference. Sites designated as "Random" contain a pattern of codons not significantly different from random, as calculated by SLR based on the entropy and optimized likelihood at that site. Finally, the Non-gap Codons and Non-gap Branch Length statistics were calculated from the pattern of gaps and non-gaps at each alignment column: Non-gap Codons is a count of the number of sequences without gaps at that site, and Non-gap Branch Length measures the total branch length of the subtree connecting each of the non-gap sequences.

A prominent feature of the distribution of ω_{ML} values for the unfiltered Mammals data, shown in the top row of Figure 4.7, was the large number of sites with $\omega_{ML} = 0$. Further inspection of the data revealed that all $\omega_{ML} = 0$ sites contained either synonymous or constant site patterns. In fact, 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 and Goldman, 2005], the ω value which maximizes the likelihood of observing zero substitutions is zero, since that value minimizes both

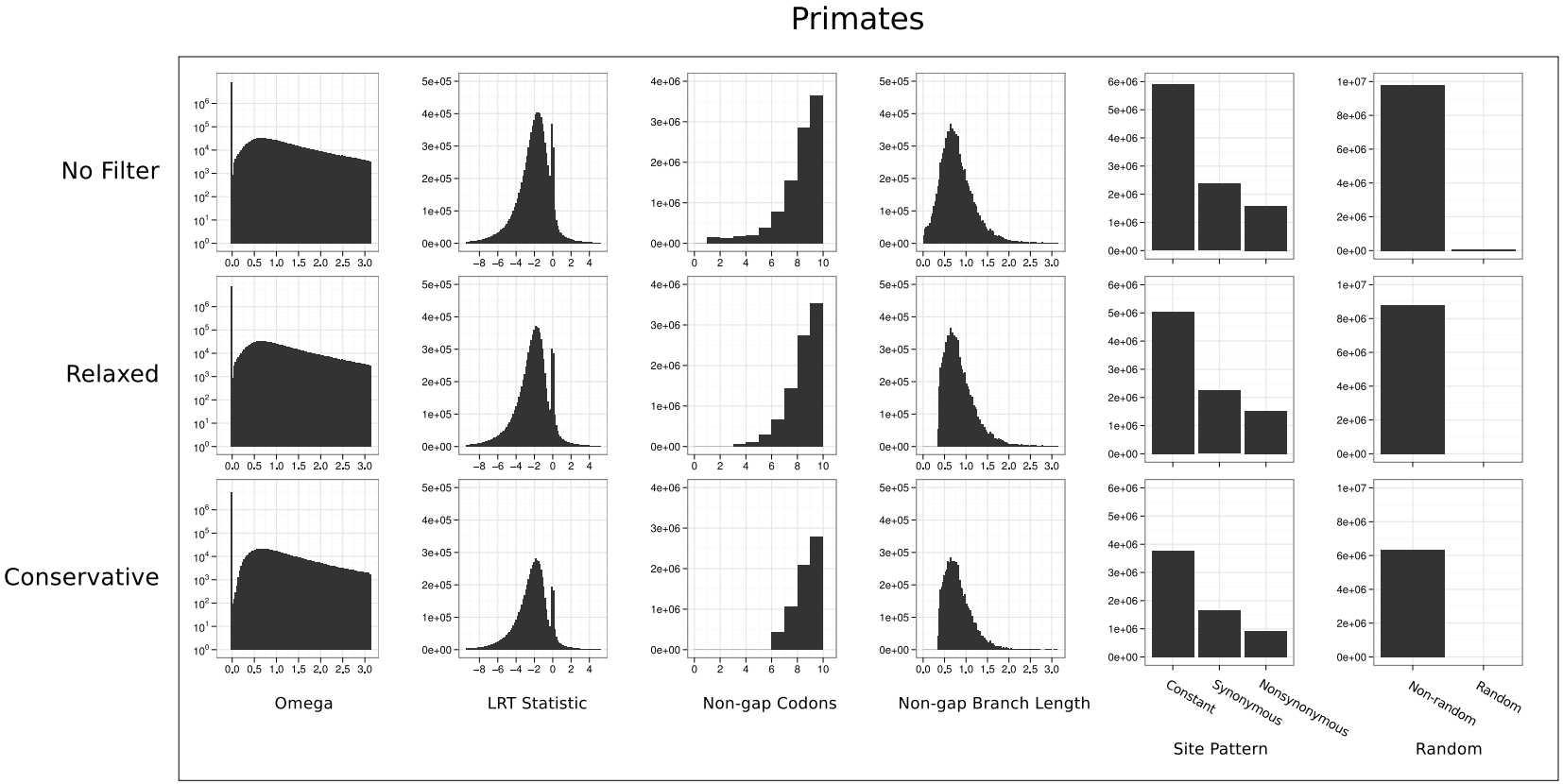


Figure 4.6: Distributions of sitewise values for the Primates species group, showing the raw data (top row) and the result of applying the relaxed (middle row) and conservative (bottom row) filters.

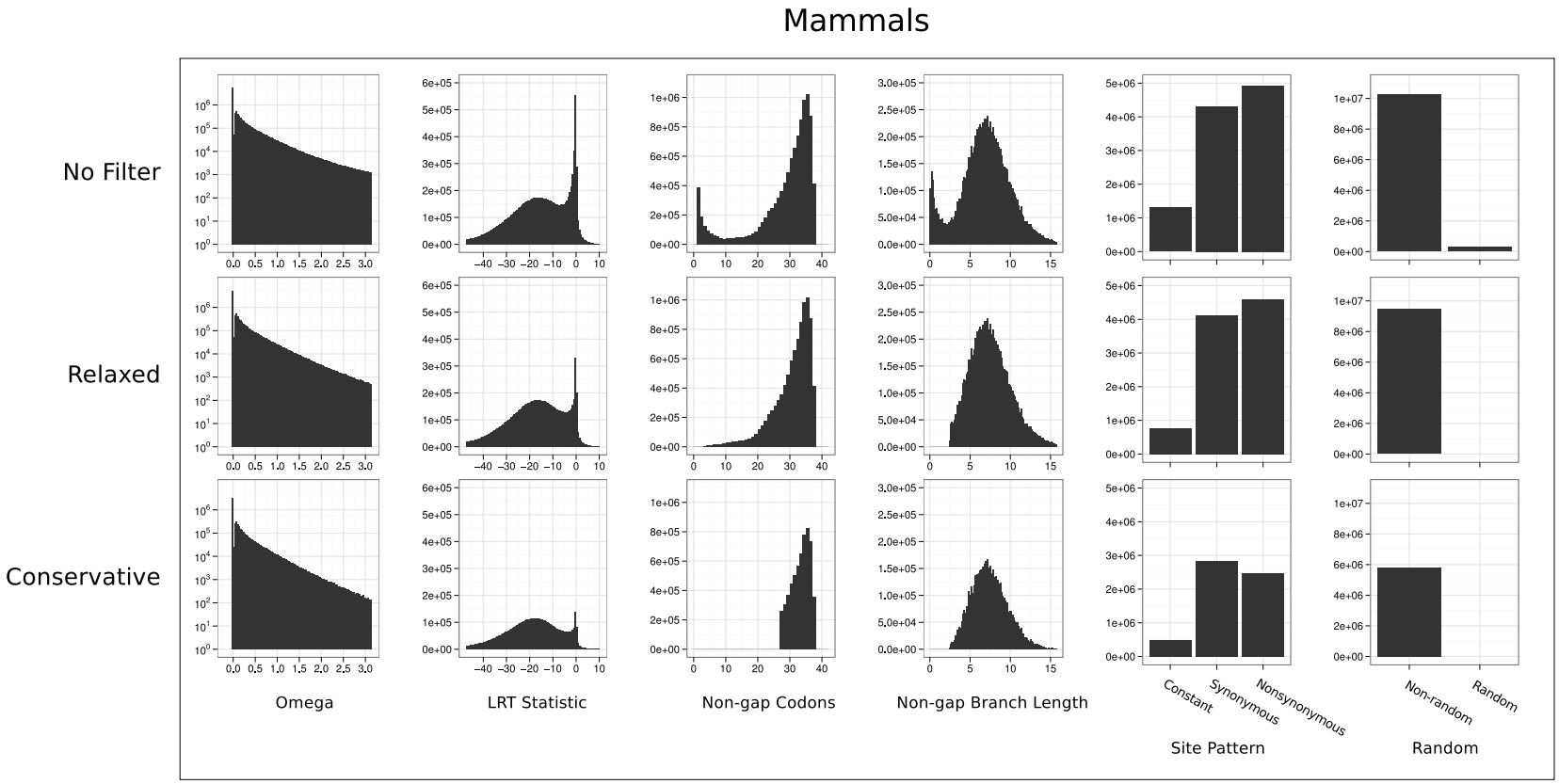


Figure 4.7: Distributions of sitewise values for the Mammals species group, showing the raw data (top row) and the result of applying the relaxed (middle row) and conservative (bottom row) filters.

the nonsynonymous and the total substitution rate.

It is not evident from either Figure 4.6 or 4.7, but a small proportion (ca. 0.2%) of sites containing synonymous site patterns resulted in ω_{ML} 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 and Goldman, 2004], and empirical results have supported the notion that codon models that allow for multiple simultaneous nucleotide changes better describe evolution than those that do not [Kosiol *et al.*, 2007]. However, the very low proportion of synonymous sites requiring non-zero nonsynonymous substitution rates suggested that the impact of these effects on the current dataset was minimal; this was likely due to the relatively short branch lengths separating the nodes of the mammalian tree, making it less probable that codons differing by 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 4.7 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 of sites contained only a few non-gap sequences. (For the unfiltered Primates histograms in Figure 4.6, there was a noticeable long tail of low Non-gap Codons values, but no excess of low Non-gap Branch Length value sas seen in Figure 4.7). 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 non-homologous 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

	BL	Non-gap BL			Non-gap Codons			ω_{ML}, %		
	Quantile	25%	50%	75%	25%	50%	75%	< 1	> 1	$\text{PSC}_{5\%}$, %
Mammals	0.20	3.31	3.80	4.18	19	30	35	95.59	4.41	0.43
	0.30	4.79	5.03	5.26	27	33	36	97.02	2.98	0.34
	0.40	5.69	5.89	6.07	28	33	36	97.00	3.00	0.35
	0.50	6.43	6.60	6.78	29	33	36	96.81	3.19	0.39
	0.60	7.12	7.29	7.47	29	33	36	96.46	3.54	0.43
	0.70	7.83	8.02	8.21	29	33	36	96.04	3.96	0.50
	0.80	8.64	8.87	9.12	29	33	35	95.48	4.52	0.60
	0.90	9.67	9.99	10.36	29	33	35	94.49	5.51	0.81
	1.00	11.33	12.14	13.42	29	33	35	93.02	6.98	1.12
Primates	0.20	0.39	0.42	0.45	8	9	10	94.69	5.31	0.27
	0.30	0.50	0.52	0.55	8	9	10	93.88	6.12	0.27
	0.40	0.59	0.61	0.63	8	9	10	93.19	6.81	0.32
	0.50	0.67	0.70	0.72	8	9	10	92.37	7.63	0.36
	0.60	0.76	0.78	0.80	8	9	10	91.35	8.65	0.45
	0.70	0.85	0.88	0.91	9	9	10	90.61	9.39	0.52
	0.80	0.97	1.01	1.05	8	9	10	89.01	10.99	0.66
	0.90	1.14	1.19	1.26	9	9	10	87.05	12.95	0.86
	1.00	1.45	1.63	2.01	8	9	10	84.81	15.19	1.21

Table 4.2: Proportions of sites with evidence for purifying and positive selection in the Mammals 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 PSC at a nominal 5% FPR. BL–branch length; PSC–positively selected codons.

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% FPR, hereafter referred to as positively selected codons (PSCs). The results of this analysis are presented in Table 4.2. The lowest bin was a clear outlier in the Mammals data, with approximately 18% of sites having $\omega_{ML} > 1$ and 2% of sites being PSCs. The other 9 bins with greater non-gap branch lengths showed fewer sites with $\omega > 1$ and less evidence for positive selection; within those 9 bins, a pattern of gradual increase in the proportion of sites with $\omega_{ML} > 1$ and PSCs was observed at progressively higher non-gap branch lengths. The increase in evidence for positive selection with increasing non-gap branch length could be explained by genes with higher overall $d\text{N}/d\text{S}$ ratios (and presumably more PSCs) having higher branch lengths due to the increased rate of nonsynonymous substitution; alternatively, longer branch lengths may lead to more statistical power to detect positive selection. 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. The reason for 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 shows 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 Figures 4.7 and 4.6, two other criteria were used to target sites for removal before analysis. First, the rightmost panels of Figures 4.7 and 4.6 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 and Goldman, 2005], and they were also targeted for removal before analysis of the global distribution. Second, all sites with fewer than four non-gap sequences were removed. This was done to avoid analyzing sites with very few sequences which were not within the bottom 10% of sites by non-gap branchlength.

At this point, all of the criteria used to define the relaxed filter have been described: non-gap branch lengths, the "random" flag, and the number of non-gap sequences at each site. The middle rows of Figures 4.6 and 4.7 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 4.2) 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 4.2) 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 two 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 4.6 and 4.7. Comparing between the distributions the three rows of Figure 4.7, 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, 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 overlapping with

Pfam domain annotations collected from Ensembl; as most Pfam domains represent well-folded protein modules [Finn *et al.*, 2010], the set of Pfam-annotated sites were expected to exhibit stronger purifying selection and be less prone to insertions or deletions and alignment error. The rows labeled in parentheses summarize the set of sites which were removed during the creation of the conservatively-filtered dataset, either due to overlap with a window of clustered substitutions (Clusters) or from being within a gene that contained more than two 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 made clear the impact of extensive filtering on the levels of positive and purifying selection observed in the data. The unfiltered data from the Primates group contained 9.03% 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 $\omega_{ML} > 1$ being reduced from 5.68% 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 anomalous or untrustworthy sites from the dataset. For sites removed due to 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 enrichment 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.

4.5 The global distribution of sitewise selective pressures in mammals

To produce high-confidence sitewise estimates across the 10 chosen species groups, sitewise data from each species group were processed with the conservative filter as described above. The re-

			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 method of Benjamini and 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.

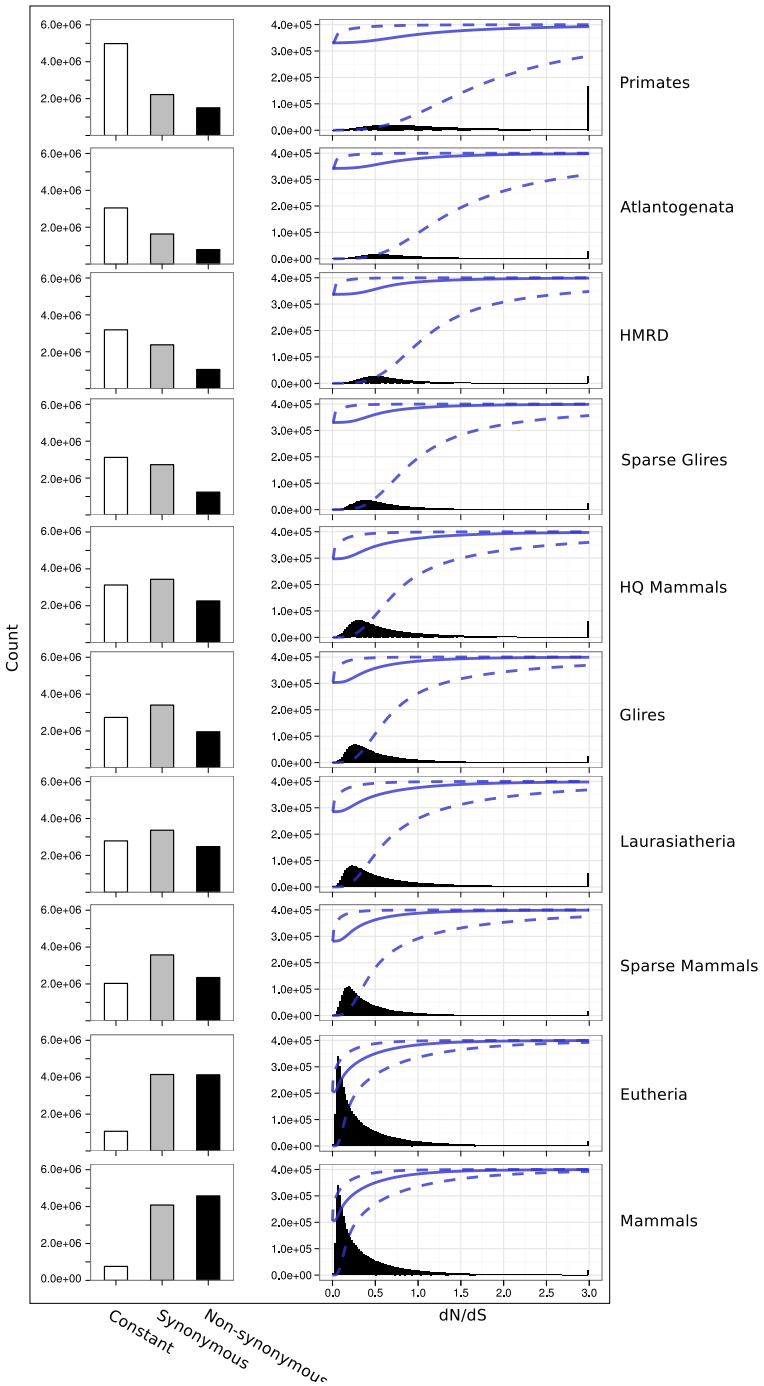


Figure 4.8: Global distributions of site patterns and ω estimates for 10 species groups. Left panels: bars represent the number of sites showing constant, synonymous, and nonsynonymous patterns. Note, the y-axis is held constant between rows. Right panels: bars represent histograms of ω_{ML} estimates only for sites where $\omega_{ML} > 0$. Sites with $\omega_{ML} > 0$ correspond to sites with nonsynonymous site patterns, and sites with $\omega_{ML} = 0$ correspond to constant or synonymous site patterns. 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.

sulting global distributions of site patterns, sitewise ω_{ML} estimates, and 95% confidence intervals are shown in Figure 4.8. The left panel in each row shows 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.

The site pattern counts in Figure 4.8 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.

The distributions of ω_{ML} estimates are shown in Figure 4.8 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 densities of the lower and upper limits of the 95% confidence interval resulting from each sitewise estimate. It was clear that the majority of protein-coding sites have evolved under purifying selection in mammals, a fact which was most easily seen in the species groups with large total branch length. The Mammals group 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 $\omega_{ML} < 1$.

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 Mammals 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 an apparent upwards shift in the distribution of the remaining nonzero ω_{ML} values.

4.6 Identifying sites with significant evidence for purifying and positive selection

An important component of SLR's output is sitewise information indicating 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. 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

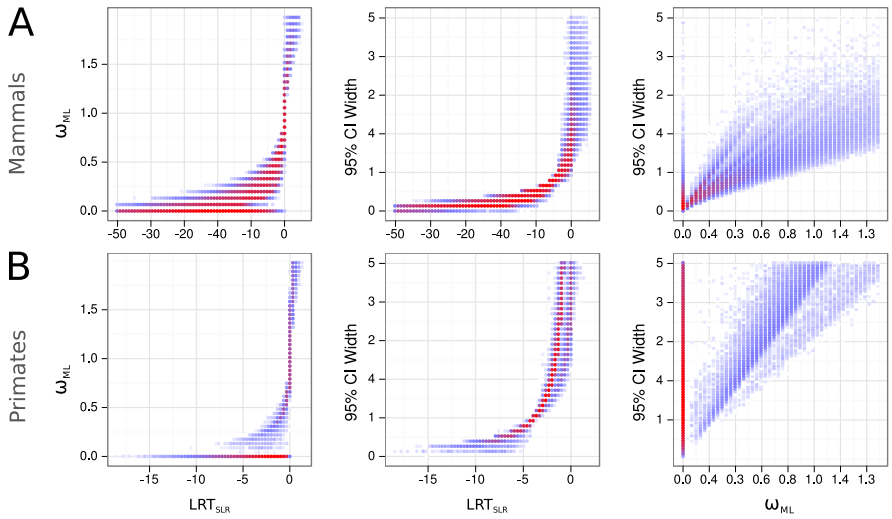


Figure 4.9: The relationship between LRT_{SLR}, ω_{ML}, and $\text{CI}_{95\%}$ width in (A) Mammals and (B) Primates datasets. Each point represents the binned density of sites; no points are drawn where no density exists, while blue and red points are drawn at areas of low and high density, respectively. The left panel shows sites where $\omega_{ML} > 0$, the middle panel shows all sites, and the right panel shows sites where $0 < \omega_{ML} < 1.3$. 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.

of the strength of evidence for purifying or positive selection, not of 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.

Figure 4.9A 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}. Correspondingly, the middle panel shows an even stronger relationship between the LRT_{SLR} magnitude and the $\text{CI}_{95\%}$ width, with the tightest confidence intervals at sites with very strong evidence for purifying selection. The rightmost panel compares the ω_{ML} of each site with the width of its $\text{CI}_{95\%}$, revealing a more linear and diffuse positive relationship between ω_{ML} and the size of the $\text{CI}_{95\%}$. The equivalent plots for Primates, shown in Figure 4.9B, reveal similar patterns, but with generally less-negative LRT_{SLR} values, higher ω_{ML}, and larger $\text{CI}_{95\%}$. These differences highlight the impact of branch length on the

amount of confidence with which ω can be estimated on a per-site basis. The low branch length of the Primates clade rarely yields ω_{ML} estimates with $\text{CI}_{95\%}$ intervals smaller than 1, while the bulk of sites from the Mammals dataset have relatively small $\text{CI}_{95\%}$. Thus, the distribution of ω_{ML} estimates from datasets with low branch lengths (e.g., the histogram densities seen in the top few panels of Figure 4.8) 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.

Instead, the confidence intervals and LRT statistics calculated by SLR for each site can 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 using conservative filtering, showing the same statistics 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 and in different species groups, demonstrating that anywhere between 0.01% to 0.73% of sites were identified as under positive selection, depending on the nominal FPR threshold and the species group used.

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 and Hochberg [1995] method, however, far fewer PSCs were identified. Only the Eutheria and Mammals 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 Mammals groups all produced

			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	6.35e+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
Atlantogenata		4.16e+06	57.25	30.05	12.70	5	0.94	1.01	0.37	0.13	0.46	90.01	95.38	4.62	2.29
HMRD		5.10e+06	49.75	36.40	13.85	4	0.96	1.01	0.43	0.12	0.41	90.05	96.36	3.64	1.73
Sparse Glires		5.43e+06	45.37	39.10	15.53	5	1.24	1.32	0.72	0.12	0.39	90.91	96.70	3.30	1.51
HQ Mammals		6.54e+06	37.08	40.65	22.27	8	1.46	1.55	0.66	0.17	0.47	88.29	94.96	5.04	2.51
Glires		5.93e+06	34.71	43.66	21.63	7	1.77	1.88	0.85	0.14	0.38	90.54	96.52	3.48	1.50
Laurasiatheria		5.41e+06	33.38	41.98	24.64	11	2.03	2.15	0.86	0.17	0.44	88.89	95.36	4.64	2.22
Sparse Mammals		5.74e+06	25.82	46.74	27.44	6	2.56	2.75	1.45	0.13	0.34	91.66	97.28	2.71	1.10
Eutheria		5.78e+06	11.95	49.78	38.27	32	5.80	6.01	1.96	0.15	0.34	90.17	96.76	3.24	1.19
Mammals		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

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; SD—standard deviation.

		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	46140	(0.73)	26261	(0.41)	7801	(0.12)	50	(0.001)	33.85	65.42	0.73	15.87	83.71	0.41
Atlantogenata		8301	(0.20)	3918	(0.09)	764	(0.02)	0	(0.000)	46.88	52.93	0.20	23.70	76.21	0.09
HMRD		6670	(0.13)	3105	(0.06)	540	(0.01)	0	(0.000)	63.68	36.19	0.13	37.40	62.54	0.06
Sparse Glires		7326	(0.13)	3362	(0.06)	650	(0.01)	0	(0.000)	70.32	29.54	0.13	49.05	50.89	0.06
HQ Mammals		30102	(0.46)	16795	(0.26)	4662	(0.07)	0	(0.000)	74.82	24.72	0.46	61.49	38.25	0.26
Glires		11416	(0.19)	5710	(0.10)	1252	(0.02)	0	(0.000)	78.91	20.90	0.19	67.89	32.02	0.10
Laurasiatheria		29464	(0.54)	17870	(0.33)	6016	(0.11)	43	(0.001)	78.30	21.16	0.54	68.70	30.97	0.33
Sparse Mammals		8065	(0.14)	3977	(0.07)	869	(0.02)	0	(0.000)	81.99	17.87	0.14	75.27	24.66	0.07
Eutheria		35650	(0.62)	24513	(0.42)	11134	(0.19)	1014	(0.018)	89.00	10.38	0.62	86.53	13.04	0.42
Mammals		29432	(0.51)	20150	(0.35)	9140	(0.16)	795	(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 method of Benjamini and 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.

lower 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 species groups were widely distributed 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 both the lowest (Primates) and the highest (Mammals) total branch lengths.

In Mammals, 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 4.8. A large amount of purifying constraint (83.87% of sites at 5% FPR), a smaller proportion of neutrally-evolving sites (15.57%), and a diminishing fraction of positively-selected sites (0.55%) were observed. 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 Mammals than in Primates to confidently detect purifying selection (83.87% vs. 15.97%).

The strong correlation between branch length and the amount of purifying selection detected was consistent with a relatively constant level of purifying selection in mammalian proteins and greater power to detect such selection in larger alignments. The overall proportion of protein-coding sites subject to purifying selection appeared to be between 90-95% based on the proportion of sites under purifying selection with $p < 0.1$, $\omega_{ML} < 0.5$ and $\omega_{ML} < 1$ (90.61%, 90.86% and 97.25%, respectively, in Mammals). The effect of branch length could clearly be seen by comparing the Mammals and Glires species groups with their "sparse" counterparts: in both comparisons, the proportion of sites with $\omega_{ML} < 0.5$ was nearly identical, but the proportion of purifying sites at $p < 0.1$ was much lower in the "sparse" group.

The detection of positively-selected sites within different species groups showed no such clear trend. As was perhaps expected, Mammals and Eutheria, the two species groups with the greatest branch length, contained the greatest proportion of PSCs at a $p < 0.01$ threshold: 0.16% and 0.19%, respectively. However, there appeared to be significant variation—independent of branch length—in the amount of positive selection detected within different species groups. Comparing the percentage of PSCs at $p < 0.01$ between Mammals, Sparse Mammals, and HQ Mammals, the "sparse" dataset contained by far the fewest PSCs, and HQ Mammals showed roughly half as many $p < 0.01$ PSCs as Mammals. At a more relaxed $p < 0.1$ threshold, HQ Mammals and Mammals contained roughly the same proportion of PSCs, while Sparse Mammals still had far fewer. The difference between Mammals and HQ Mammals at the two thresholds suggested that the larger branch length of the Mammals species group allowed PSCs to be more readily detected at higher significance levels. In contrast, the consistently lower proportion of PSCs in Sparse

Mammals could only be readily explained by a species sampling effect. The Glires and Sparse Glires species groups showed similarly low levels of positive selection, suggesting that species sampling within this clade did not strongly affect the prevalence of PSCs.

Despite the large number of species groups considered in this analysis, it was difficult to assess whether the observed variation in levels of positive selection was associated with true biological variation or with some artifact of genome quality or branch length. With respect to genome quality, the strong presence of PSCs within the HQ Mammals group provided some evidence that low-coverage genomes were not solely responsible for the increased level of positive selection in certain species groups. Due to the overlap between many of the species groups, the effects of branch length and species sampling were difficult to disentangle. One hypothesis, motivated by the few PSCs in Sparse Mammals compared to HQ Mammals, is that a mammalian phylogeny with a greater proportion of its branch length in shorter or more recent branches may either contain more positive selection or have more power to detect it. A thorough exploration of this hypothesis is beyond the scope of this thesis, but a simple simulation study could be performed to test whether the LRT performed by SLR is more sensitive to episodic positive selection occurring within short branches than within longer branches of the phylogeny.

4.7 The impact of effective population size on protein-coding constraint in mammals

Another possible source of variation in levels of positive selection between different species groups was true biological differences in the proportion of purifying and positively-selected sites. 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. However, this number varied strongly between different species groups. Comparing between the four phylogenetically independent mammalian superorders (Primates, Glires, Laurasiatheria, and Atlantogenata), 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. The Glires group showed strikingly lower levels of positive selection compared to the other mammalian superorders. Despite the relatively large amount of branch length covered by the Glires group (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 (N_e) on the efficacy of natural selection in mammals [Popadin *et al.*, 2007; Nikolaev *et al.*, 2007; Ellegren, 2009]. Rodents are known to have N_e well above that of primates [Kosiol *et al.*, 2008], and given the strong correlation between body size, generation time and N_e [Nikolaev *et al.*, 2007] one

can infer that species within the Laurasiatheria group, with generally longer generation times and larger body sizes than rodents [Hou *et al.*, 2009], have N_e 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 N_e. Nevertheless, Ohta's nearly neutral theory [Ohta, 1992] predicts that species with lower N_e 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, differences in the proportion of sites under purifying selection were well explained by the difference in N_e between primates and rodent-like mammals.

Theory also predicts that positive selection should be more efficient in populations with high N_e [Ellegren, 2009; Halligan *et al.*, 2010], and a number of empirical studies have supported this prediction. The prevalence of adaptive evolution has been extensively studied in different species using a combination of within-species diversity and between-species divergence data, finding high proportions of amino acid substitutions driven by positive selection in species with high N_e such as *Drosophila* [Bierne and Eyre-Walker, 2004] and *E. coli* [Charlesworth and Eyre-Walker, 2006], and low proportions in species with low N_e such as human [Zhang and Li, 2005; Mikkelsen *et al.*, 2005; Boyko *et al.*, 2008] and chicken [Axelsson and Ellegren, 2009]. Fewer studies have assessed differences in levels of positive selection between species using only divergence data and dN/dS-based tests for selection, but some such analyses have been published. Clark *et al.* [2007] used PAML [Yang, 2007] to estimate the proportion of positively selected genes (PSGs) and PSCs in 12 *Drosophila* genomes, finding evidence for positive selection within 33% of single-copy orthologs, affecting 2% of codons within those PSGs. Kosiol *et al.* [2008] found a lower proportion of PSGs in their analysis of 6 mammalian genomes, with only 544 candidate PSGs out of 16,529 orthologs tested (3.3%). The authors then compared the levels of apparent positive selection in different lineages, finding that 72% of the 544 candidate PSGs showed evidence of positive selection in rodents; they concluded that "whether because of power or a genuine increase in selection, the rodent branch appears to play a major role in the identification of PSGs" [Kosiol *et al.*, 2008]. Together, these two studies based on dN/dS estimates largely support the theoretical prediction of increased levels of positive selection in populations with large N_e.

The current results appeared to contradict the theoretical prediction and empirical evidence for a positive correlation between N_e and the prevalence of positive selection in mammalian species. The Primates, Laurasiatheria and HQ Mammals species groups showed greater levels of positive selection than the Glires group, as measured by both the proportion of sites with $\omega_{ML} > 1$ and the proportion of PSCs identified at all FPR and false discovery rate (FDR) thresholds shown in Table 4.6. These species groups were the same groups which showed evidence for lower long-term N_e due to their elevated mean ω_{ML} values and decreased proportion of sites with $\omega_{ML} < 1$ (Table 4.5). Although estimates of ω_{ML} should be treated with caution

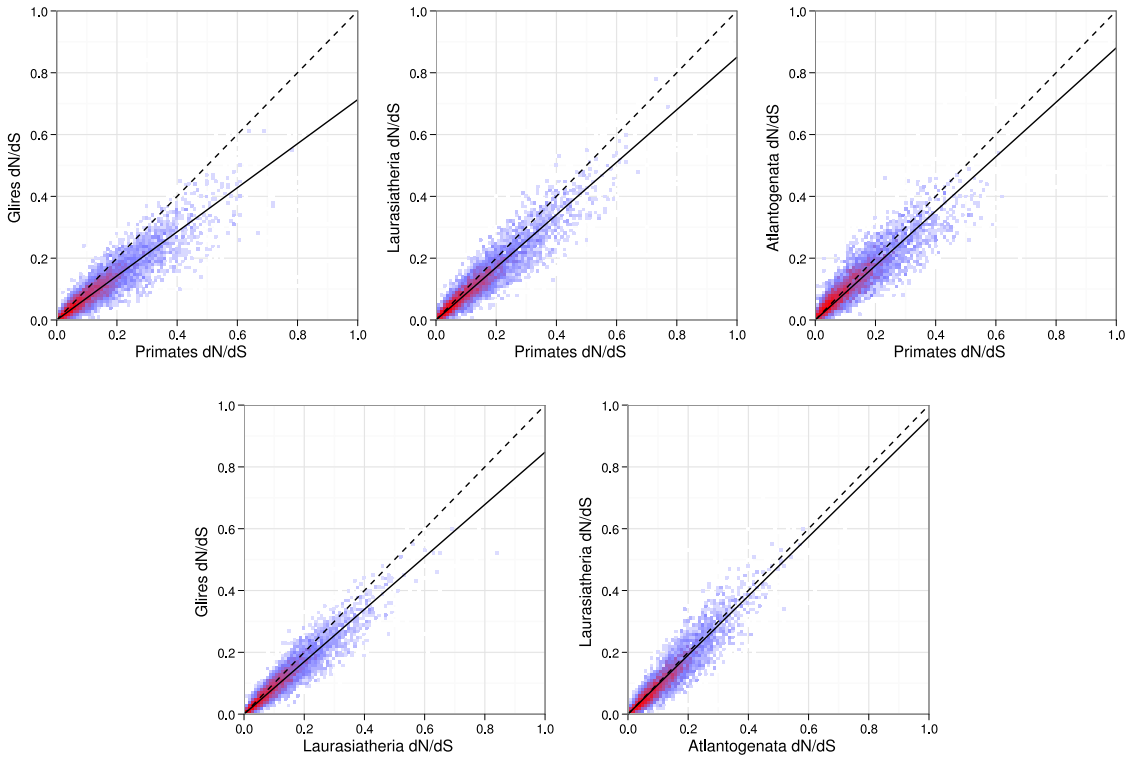


Figure 4.10: Correlations between gene-wide dN/dS ratios estimated for 16,477 orthologous genes in four different species groups. dN/dS ratios were estimated for each alignment by SLR using the M0 (one ratio for all sites) codon model of evolution. Genes were plotted according to their dN/dS ratio in each species group, and each panel shows the density of genes as a colored heat map with red areas corresponding to the highest density of points. A dashed line with slope of 1 is drawn for reference, and a solid line is drawn in each panel along the first principal component axis of those data points.

due to stochasticity in the ML estimate at a single alignment site, further evidence for lower N_e in these species groups came from the estimates of gene-wide dN/dS values calculated by SLR based on the M0 codon model of evolution. Figure 4.10 shows a comparison of gene-wise dN/dS in these species groups, providing strong evidence for the following ordering of N_e: Glires $>$ Laurasiatheria $\approx$ Atlantogenata $>$ Primates. Thus, rather than a positive correlation between N_e and the prevalence of positive selection, these sitewise data seemed to exhibit a negative correlation between these two factors.

This suggested 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 N_e. A difference in N_e should impact slightly deleterious and slightly advantageous mutations equally, with a greater proportion of both types of mutations being effectively neutral in the population with lower N_e. In comparing the Primates and Glires groups, the expected

result was that a subset of mutations which were under purifying selection in Glires would be effectively neutral in Primates, bringing the expected ω for those sites from < 1 to 1. The same should occur for slightly advantageous sites, but if the proportion of slightly deleterious sites is much greater than the proportion of slightly advantageous sites, it may have a much more measurable impact. Furthermore, since sites with $d\mathrm{N}/d\mathrm{S} = 1$ are more likely to produce false positive PSCs, it is plausible that the greater proportion of effectively neutral sites in a species with low N_e would lead to an increased proportion of PSCs detected by methods based on $d\mathrm{N}/d\mathrm{S}$ ratios.

This relaxed constraint argument tempers the interesting observation of strong differences in the numbers of PSCs between different species groups. A lower historical N_e 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 and Hochberg [1995] method for FDR control used to identify PSCs at an expected FDR< 0.05 in Table 4.6, 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. This apparent discrepancy with previously-published results should be an interesting area for continued investigation.

4.8 Conclusions

This chapter described the filtering, alignment, and analysis of sitewise selective pressures in a comprehensive set of orthologs across 38 mammalian 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 gene 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. 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.

Sitewise selection pressures were then calculated for several groups of mammalian species. The impact of the total branch length of a species group on the estimation of sitewise selective pressures could be clearly seen from these results, with the Mammals group containing many more non-constant alignment sites and a more accurate distribution of ω_{ML} estimates than groups with little total branch length, due to the greater amount of evolutionary information.

The MGP analysis consortium 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. Comparing sitewise results to the same reference species groups, 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 in levels of purifying and positive selection in different groups, likely resulting from differences in N_e. Although the impact of N_e is well known and has been previously studied in mammalian species, 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.

More work needs to be done to evaluate what might be causing these strongly different estimates in different mammalian superorders and to correctly control for the possible effect of relaxed constraint on the identification of positive selection in primate genomes. Differences between the current results and those obtained from diversity-based estimates in contemporary populations could be attributed to differences in the timescale under examination [Ellegren, 2009], as divergence-based estimates of selective constraint are obviously sensitive to the long-term N_e within each clade, while diversity-based estimates measure the selective constraints based on more recent N_e. However, the contrast with the results of Kosiol *et al.* [2008], who used similar dN/dS based methods yet found higher levels of positive selection in the rodent lineage than the primate lineage, is less easily explained.

An interesting question for future work is whether the observed species group differences in

positive selection result more from differences in ancient or more recent branches in the phylogeny. In other words, at what point in evolutionary time did the major mammalian orders begin to experience different levels of positive selection (or relaxed constraint)? Although these groups of species show different contemporary N_e, they have evolved independently for several dozen million years, during which time several N_e expansions or contractions could have occurred. With more dense species sampling and branch-specific dN/dS estimates, a clearer picture of the connection between the evolution of N_e, relaxed constraint and positive selection may begin to emerge.

Chapter 5

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

5.1 Introduction

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 positively selected genes (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 [Gibbs *et al.*, 2007]; analyses in other mammalian genomes have revealed largely similar patterns [Kosiol *et al.*, 2008; 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 genes involved in host–pathogen interactions [Yang, 2005; Meyerson and Sawyer, 2011], sexual selection or genetic conflict between the sexes [Wyckoff *et al.*, 2000; Clark and Civetta, 2000], and functional adaptation following gene duplication or environmental changes [Zhang *et al.*, 2002].

As the power of phylogenetic analysis using codon models depends strongly on the amount of branch length encompassed by the species being compared [Anisimova *et al.*, 2001, 2002], there was some reason to believe *a priori* that the detection of PSGs using mammalian alignments

incorporating low-coverage genomes would be more powerful than in previous whole-genome analyses, which typically included 12 or fewer species across mammals and lower total branch length [Ellegren, 2008]. However, differences in the alignments and methods used to detect positive selection may act to limit the amount of overlap in PSGs between different 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 Sitewise Likelihood Ratio (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 explicitly compared the differences in PSGs identified by site-specific and branch-site methods—or the differences in PSGs identified by the same method in different studies—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.

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 first develop a number of methods for using sitewise estimates to identify signals of positive selection within genes and apply these methods to the sitewise data generated in Chapter 4. Next, to provide a higher-level interpretation of these results I use Gene Ontology (GO) annotations [Ashburner *et al.*, 2000] to identify categories enriched for genes with evidence of positive selection in different species groups. A quantitative comparison to results from the literature is provided through a direct comparison of the PSGs and GO term results to previously-published studies. Finally, I apply the same methods for combining sitewise estimates to identify protein domains with the strongest enrichment for positive selection throughout mammalian evolution.

5.2 Combining sitewise estimates to identify positive selection

In Chapter 4 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 assessed.

The more traditional approach in comparative genomics has been to model the protein-coding gene as the unit of analysis. For detecting positive selection, the grouping of alignment sites into

genes—which results in identification of PSGs instead of positively selected codons (PSCs)—has three main advantages. First, the combined analysis of many alignment sites improves the accuracy of estimated evolutionary parameters and boosts the power of likelihood ratio tests (LRTs) for detecting positive selection. This is seen in the simulations of Anisimova *et al.* [2001, 2002], 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 within genes [Sawyer *et al.*, 2005; Kosiol *et al.*, 2008], suggesting that the evolutionary dynamics causing detectable signals of positive selection tend to affect many functionally or structurally related amino acid sites within genes as opposed to acting on randomly distributed sites. The third argument in support a gene-centric analysis of positive selection is that in the absence of a protein structure for every gene, 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 the function of 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 and Goldman, 2005]. When combining these p-values to decide whether a gene contains significant evidence for positive selection, the number of tests performed must be taken into account. 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 false positive rate (FPR) of 0.05; correspondingly, the chance that at least one site within the gene would have $p < 0.05$ is 99.4%. 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 three 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; and third, estimating empirical gene-wise p-values based on the genome-wide distribution of sitewise estimates. Each approach makes different use of the sitewise data from each gene to identify a set of significant PSGs and thus had the potential to identify slightly different sets of PSGs. The remainder of this section

provides a description of how each of the three approaches was applied to the sitewise data.

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 popular Bonferroni correction is a more easily computed approximation) identifies the p-value threshold x which is necessary to control the FWER at the desired level α [Sidak, 1967]. 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 and Goldman, 2005].

To identify PSGs by controlling the FWER, 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 expected weakness of this approach was that the evidence for a PSG comes only from the site in each gene 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 tend to contain multiple sites subject to similarly strong amounts of positive selection [Sawyer *et al.*, 2005; Kosiol *et al.*, 2008], the gene-wise p-values resulting from the FWER-controlling approach were expected to lack some power. The next two methods described are both sensitive to more than just the most significant site, making them potentially more powerful for identifying PSGs.

Combining p-values

The second approach to multiple testing addresses the potential weakness of the FWER-controlling approach by combining p-values from all sitewise tests performed, producing an overall p-value for the null hypothesis given the overall set of tests. 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 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). Both methods first combine the set of 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. The combined statistic is then compared to the expected null distribution given the same number of input p-values. A comparison of both tests by Darlington and Hayes [2000] suggests that they provide similar power overall, but that 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.

When the distribution of input p-values is nonuniform or the number of tests is large, however, performance of Fisher's and Stouffer's methods can be reduced. Zaykin *et al.* [2002] noted that a relatively small number of large p-values can limit the power of Fisher's test, and the Stouffer method can be expected to be equally sensitive to a bias towards large p-values. Since the majority of the sitewise estimates in mammals showed moderately strong signals of purifying selectio, the distribution of one-sided p-values for positive selection was heavily weighted towards 1. This can be seen clearly in Figure 5.1, which shows a histogram of genome-wide one-tailed p-values based on the Mammals species group. The standard versions of both Fisher's and Stouffer's methods were expected to lack power to identify PSGs given the strong impact of purifying selection on the sitewise data.

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 and 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 of which five have $p < 0.05$, the test statistic would be the product of those five 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

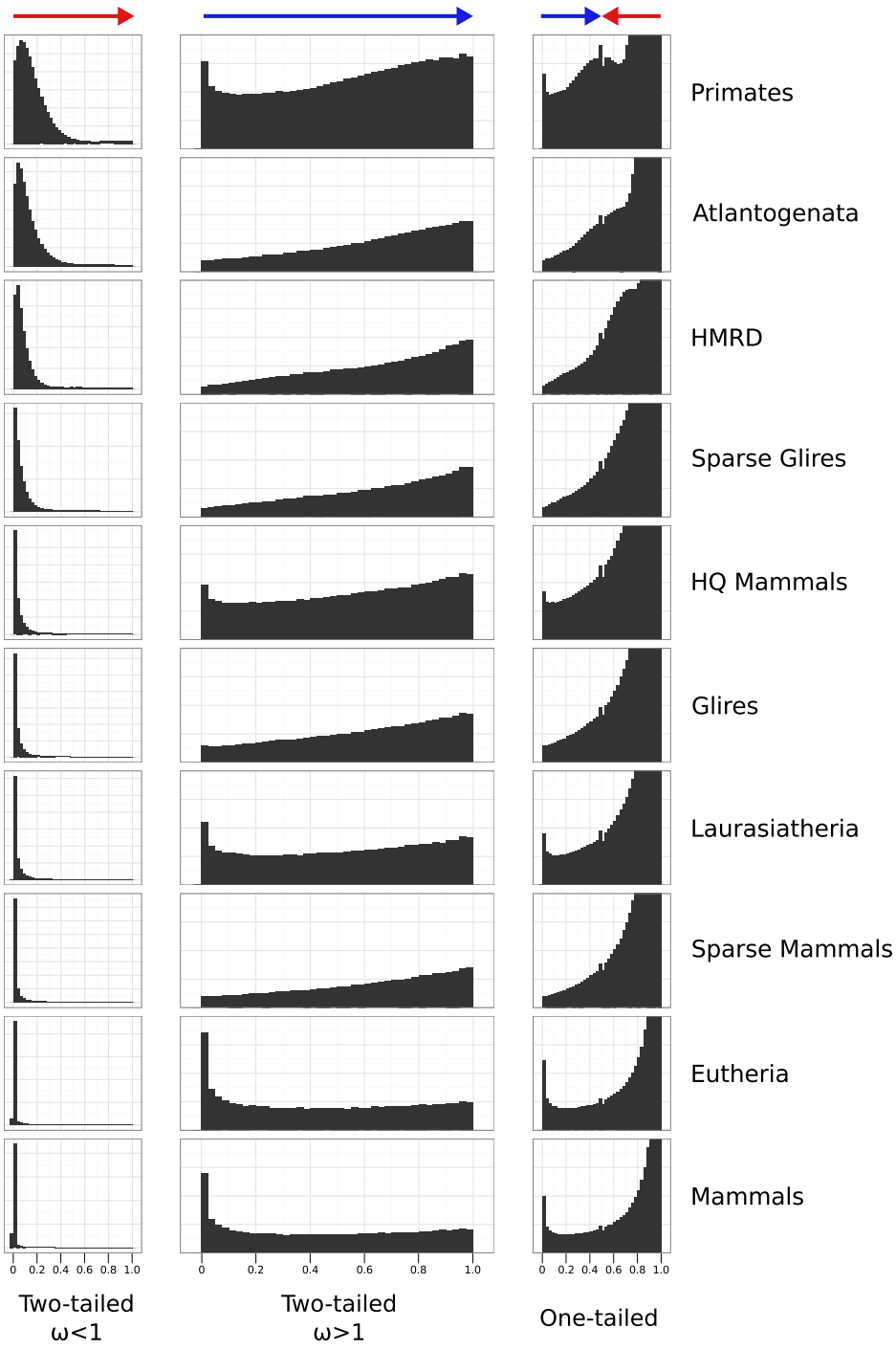


Figure 5.1: Histograms of sitewise p-values in 10 species groups using the conservative filter. Two-tailed p-values for sites with $\omega_{ML} < 1$ (left panels) and for sites with $\omega_{ML} > 1$ (middle panels) are shown. One-tailed p-values for positive selection (right panels) were constructed by halving the two-tailed p-values for sites with $\omega_{ML} < 1$ and $\omega_{ML} > 1$ and negating the $\omega_{ML} < 1$ p-values (shown visually by the blue and red arrows). Note: the y-axis is held fixed in the middle and right panels, but not in the left panels. Some histogram bars in the right panels are cut off at high p-values.

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 PSG. 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 are excluded from the calculation of the combined statistic, thus avoiding undue influence from non-significant p-values; however, the TPM method still assessed the significance of the combined statistic against a uniform distribution of p-values. Thus, 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, allowing for strong statements to be made about genes showing significance with these methods. However, for a global analysis of functional trends in genes subject to positive selection, a less conservative approach would provide more signal and may be preferable. 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 as opposed to the product of p-values. The decision to use the count of significant sites as the test statistic was made primarily due to its simplicity and ease of implementation; further testing of the empirical approach described here could evaluate methods using the product of p-values or other test statistics. 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 an appropriate 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 10,000 replicates; as with the TPM method, the effect of different truncation thresholds was assessed by separately calculating empirical p-values for each gene using nominal 0.05%, 1%, 5%, and 10% FPR thresholds.

In calculating gene-wise p-values using the empirical method described above, the empirical distribution of sitewise estimates (from which the pseudo-replicate genes were sampled) was chosen to match the species group and sitewise filter used to generate the observed test statistic. For example: to test the significance of a 500-codon gene with 10 $p < 0.01$ sites in the Glires species group using the conservative sitewise filter, replicates were randomly sampled from the genome-wide distribution of sitewise p-values using the same species group and filter. The resulting gene-wise p-value measured the significance of the test statistic for that gene relative to the statistic expected from a gene with sitewise selective pressures randomly drawn from the genome-wide distribution. This interpretation was slightly different from the Hochberg and TPM methods described above, as the significance of a given test statistic for those methods would not vary depending on the species group or filtering protocol used. Instead, for the empirical method the significance was tested relative to the sitewise observations across all genes with a matched species group and filter. For example, the observation of a given number of $p < 0.01$ sites in the Glires species group would yield a more significant gene-wise empirical p-value than in the Primates species group, as the genome-wide "null" distribution of sitewise p-values in Primates contained many more sites with $p < 0.01$ (as seen in Figure 5.1).

5.3 Analysis of PSGs identified using sitewise selective pressures

The methods described above were applied to the one-tailed sitewise p-values for positive selection, calculated by SLR from each of the 10 species groups and three levels of sitewise filtering described in 4. To assess the overall behavior of each method for combining sitewise p-values from each gene into a gene-wise p-value, I first looked at the distribution of gene-wise p-values for different species sets using the conservatively-filtered sitewise data. Figure 5.2 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

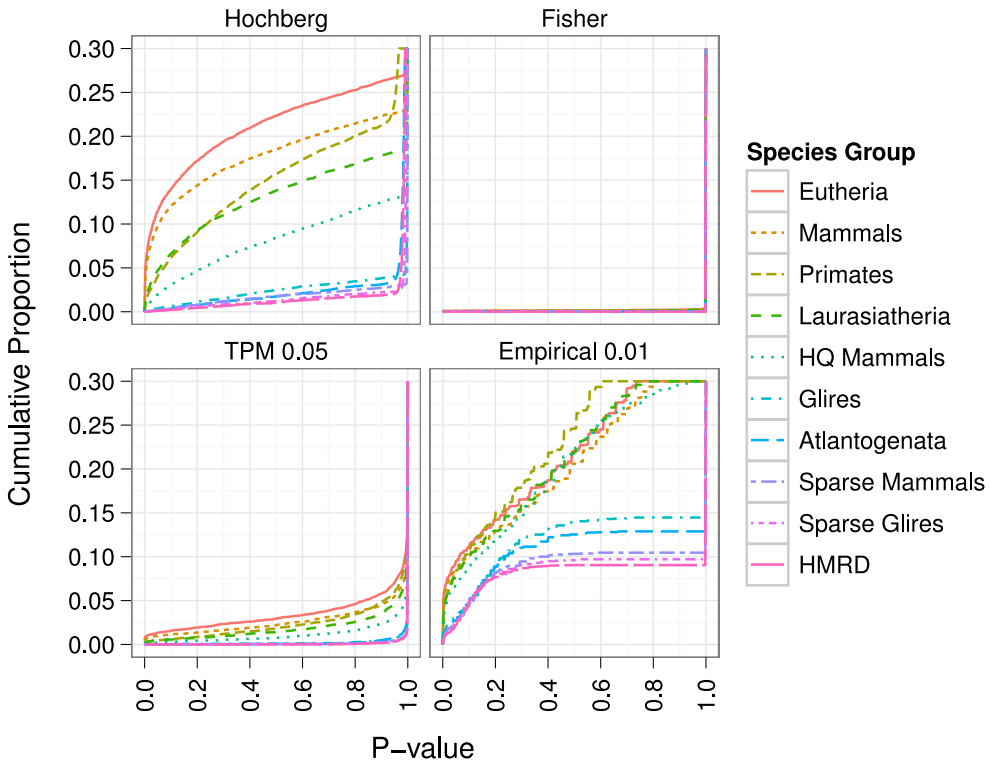


Figure 5.2: Cumulative distributions of gene-wise p-values for positive selection resulting from 4 different methods for combining sitewise estimates within genes. Note that the species groups are listed in order of their cumulative proportions at a p-value of 0.5 for the Hochberg method. To more clearly show the separation between species groups at lower y-values, cumulative proportions above 0.3 are not shown.

similar.)

As expected, Fisher's product method produced very few p-values below 1, showing little or 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 five 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 clustering of the empirical p-values for the five species groups with greater evidence for positive selection and the greater proportion of low empirical p-values for the five species groups with less evidence for positive selection. Both differences could be explained by the fact that the Hochberg method assesses significance based on the absolute magnitude of the LRT statistic for positive selection, while the empirical method assesses 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 and 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 and Hochberg [1995] method.

Table 5.1 provides a summary of PSGs identified by each method for the 10 species groups

Filter	Species Group	Gene Count	$\bar{\omega}_A$	$\bar{\omega}_G$	Hochberg	Fisher	TPM $p < 0.5$	TPM $p < 0.2$	TPM $p < 0.1$	TPM $p < 0.05$	Empirical $p < 0.005$	Empirical $p < 0.01$	Empirical $p < 0.05$	Empirical $p < 0.1$
Relaxed	Primates	15142	0.20	0.12	89	16	17	35	67	95	1500	884	1856	2457
	Atlantogenata	11155	0.17	0.11	0	0	0	0	1	1	167	59	426	777
	HMRD	13137	0.14	0.09	0	0	0	0	0	0	95	53	368	803
	Sparse Glires	13588	0.14	0.08	0	0	0	0	0	0	106	39	360	809
	HQ Mammals	15296	0.16	0.10	0	1	2	9	17	31	1090	737	1561	2061
	Glires	14717	0.14	0.08	0	0	0	0	0	0	262	122	627	1155
	Laurasiatheria	15091	0.17	0.10	74	6	13	32	50	80	1215	856	1631	2125
	Sparse Mammals	14943	0.14	0.09	0	0	1	0	1	1	227	107	557	994
	Eutheria	15800	0.17	0.11	1434	55	59	134	227	364	1687	1351	1976	2332
	Mammals	15890	0.16	0.11	1243	50	52	107	194	312	1594	1251	1901	2306
Conservative	Primates	10676	0.17	0.12	36	1	2	2	5	11	805	460	1142	1520
	Atlantogenata	7715	0.15	0.11	0	0	0	0	0	1	90	39	267	472
	HMRD	9242	0.12	0.09	0	0	0	0	0	0	46	28	201	499
	Sparse Glires	9587	0.12	0.08	0	0	0	0	0	0	68	19	193	527
	HQ Mammals	10898	0.14	0.10	0	0	0	0	5	7	608	342	929	1319
	Glires	10123	0.12	0.08	0	0	0	0	0	0	151	51	409	728
	Laurasiatheria	9660	0.15	0.10	46	2	2	6	9	15	618	400	904	1166
	Sparse Mammals	10254	0.12	0.09	0	0	0	0	0	0	98	54	302	557
	Eutheria	10184	0.14	0.11	611	8	10	14	34	73	804	629	1033	1254
	Mammals	10241	0.14	0.11	476	8	9	13	23	49	725	586	920	1134
Pfam	Primates	11867	0.18	0.12	22	6	9	13	24	34	399	279	681	987
	Atlantogenata	8521	0.15	0.11	0	0	0	0	0	0	28	18	98	243
	HMRD	10199	0.12	0.09	0	0	0	0	0	0	24	8	91	229
	Sparse Glires	10557	0.12	0.08	0	0	0	0	0	0	18	7	73	189
	HQ Mammals	12020	0.14	0.10	0	1	2	3	10	15	321	230	598	828
	Glires	11489	0.12	0.08	0	0	0	0	0	0	42	13	144	345
	Laurasiatheria	11805	0.15	0.10	25	4	5	15	21	34	393	284	636	854
	Sparse Mammals	11684	0.12	0.09	0	0	0	0	0	0	40	26	131	266
	Eutheria	12465	0.15	0.11	468	32	33	60	99	156	568	503	806	964
	Mammals	12554	0.14	0.11	418	29	33	51	84	130	544	464	757	905

Table 5.1: Counts of PSGs identified using sitewise data with three sitewise filters, 10 species groups and different methods to combine p-values across sites. The Benjamini and Hochberg [1995] method was used to control for multiple tests; counts of PSGs significant at FDR< 0.1 are shown. The columns $\bar{\omega}_A$ and $\bar{\omega}_G$ represent 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. TPM—truncated product method.

and for the relaxed, conservative, and Pfam sitewise filters. These three filters were chosen to represent a range of filtering protocols; results for the unfiltered sitewise data were not shown due to the elevated levels of spurious positive selection identified within unfiltere data in Chapter 4. The Pfam filtered dataset was filtered using the same rules as the relaxed filter, but only sites within annotated Pfam domains were retained for analysis. 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 analyzed 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 PSG 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 5.2. 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 proved 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 5.2; 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 four 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 two clusters, each with a very different proportion of identified PSGs. This was consistent with the results from Chapter 4, where the species groups also clustered into two groups based on the prevalence of positively-selected codons within the genome-wide distribution.

Overlaps between positively-selected genes in different species groups

Using the sets of significant PSGs summarized in 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 1,035 out of 11,520 genes, or 8.9% of those

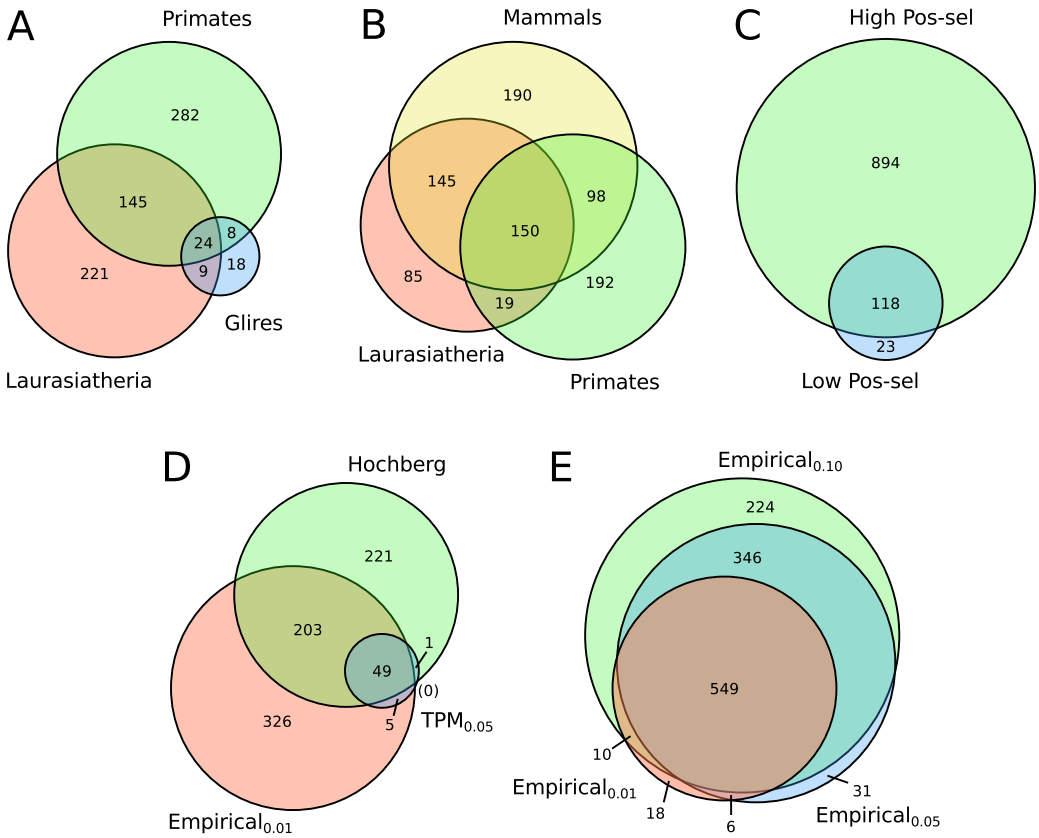


Figure 5.3: Venn diagrams of PSGs identified in different species groups and using different methods. (A) PSGs identified using empirical p-values in Primates, Glires and Laurasiatheria. (B) PSGs identified using empirical p-values in Mammals. Laurasiatheria and Primates. (C) PSGs identified using empirical 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).

investigated, were identified as a PSG in at least one of the species groups. Figure 5.3 shows a more detailed breakdown of how many PSGs were shared between various species groups. Figure 5.3A 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, Laurasiatheria, or both. Figure 5.3B 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 5.3C 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 1,300, or 11.3% of all genes tested. Compared to the 1,035 from the empirical method alone, the additional 265 genes came from the Hochberg method in the Eutheria and Mammals groups. Figure 5.3D 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 5.3E 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

The Gene Ontology (GO) [Ashburner *et al.*, 2000] is a structured ontology for describing the biological functions performed by the proteins encoded in genes. A major focus of gene annotation projects has been to accurately assign GO terms to genes, and genome-wide GO annotations have often been used to summarize the types of biological functions associated with different groups of genes [Camon *et al.*, 2004]. I used 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 [Gibbs *et al.*, 2007]. One assumption of this approach was that gene function is conserved between a gene and all of its mammalian orthologs. This assumption is true for any evolutionary analysis using experimental or functional information derived from a single extant species. At least within mammals, all of whom share similar developmental regimes and core biochemical capabilities, this assumption seems unlikely to be violated by too many genes with largely 1-to-1 orthology throughout the phylogenetic tree.

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 $\text{Emp}_{0.01}$ PSGs (`M`), the 934 $\text{Emp}_{0.05}$ PSGs (`m`), and the 202 genes in the top 2% genome-wide by overall $d\text{N}/d\text{S}$ value calculated by SLR using the M0 codon model (`D`). The latter group was defined using gene-wide $d\text{N}/d\text{S}$ 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 $\text{Emp}_{0.01}$ PSGs from Primates (`P`), the 409 $\text{Emp}_{0.05}$ PSGs from Glires (`g`), and the 400 $\text{Emp}_{0.01}$ 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 $\text{Emp}_{0.05}$ 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 $\text{Emp}_{0.05}$ 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`), Gibbs *et al.* [2007] (`R`), and Kosiol *et al.* [2008] (`K`).

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 $\text{Emp}_{0.05}$ 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 evaluated 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 $\text{Emp}_{0.05}$ 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 two of the three previously published studies. The first five 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.* [2008] 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-

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 two of three 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 six 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; Ann.—the number of genes annotated with the term; Sig.—the observed number of significant genes annotated with the term; Exp.—the number of significant genes expected to be annotated with the term given random assortment.

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 three 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*, *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 and 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 and 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 evolution of increased cancer risk. Although similar trends were observed by Nielsen *et al.* [2005], the current study provides 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 and 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.*, 2008]. 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], Gibbs *et al.* [2007], and Kosiol *et al.* [2008]. 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.

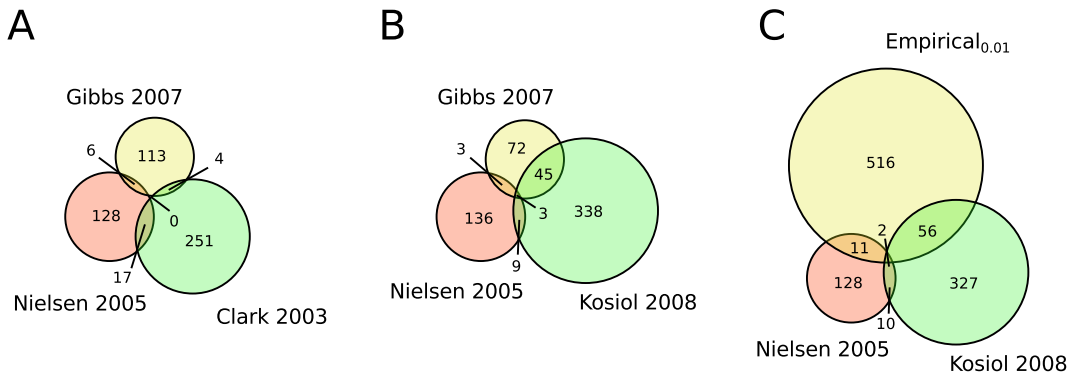


Figure 5.4: Venn diagrams of PSGs identified in different studies. (A) PSGs identified in primates by Clark *et al.* [2003], Nielsen *et al.* [2005] and Gibbs *et al.* [2007]. (B) PSGs identified in primates and mammals by Nielsen *et al.* [2005], Gibbs *et al.* [2007] and Kosiol *et al.* [2008]. (C) PSGs identified in primates and mammals by Nielsen *et al.* [2005], Kosiol *et al.* [2008] and this study using the Mammals species group, conservative filter, and the $\text{Emp}_{0.01}$ method.

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 PSGs 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 151 successfully mapped genes with LRT values greater than that value. Gibbs *et al.* [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.* [2008] 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 5.4 shows the overlap between PSGs identified in this and previously-published studies. (Note that all of these comparisons were made using results from the conservative filter; comparisons using the relaxed filter were qualitatively similar to those in Figure 5.4.) Overall, the lack of overlap in identified PSGs was striking: Figure 5.4A shows the overlap between the

three past studies in primates, with no genes shared by all three studies and from 4 to 17 genes shared between any pair. Although each analysis 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.* [2005] also noted that only 1 of their top 50 genes was also identified by Clark *et al.* [2003] as evolving under positive selection. Figure 5.4B shows slightly more overlap between the two most recent studies, with 45 PSGs shared between Kosiol *et al.* [2008] and Gibbs *et al.* [2007]. The comparison between PSGs from Nielsen *et al.* [2005], Kosiol *et al.* [2008], and the set of $\text{Emp}_{0.01}$ PSGs from the Mammals species group shown in Figure 5.4C revealed a similar number of overlapping genes, even though a slightly higher number of overlapping PSGs may have been expected due to 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 5.4C 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 four gene-wide p-values or $d\text{N}/d\text{S}$ 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 $d\text{N}/d\text{S}$ estimates for Primates and Mammals (where the test statistic was the negative $d\text{N}/d\text{S}$ value, so the genes with highest $d\text{N}/d\text{S}$ were sorted first), and to test whether the method used to combine sitewise estimates within genes had an effect, I included $\text{Emp}_{0.01}$ and Hochberg p-values as test statistics.

Figure 5.5 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 $\text{Emp}_{0.01}$ p-values in Mammals, making the intersection of the $\text{Emp}_{0.01}$ ROC curve at the vertical lines equivalent to the numbers of overlapping PSGs seen in Figure 5.4C for the Nielsen et al. [2005] and Kosiol et al. [2008] sets of PSGs.

The Clark *et al.* [2003] curves hardly strayed from the diagonal line, showing little ability beyond random chance to identify the PSGs from that study. This was not necessarily unexpected,

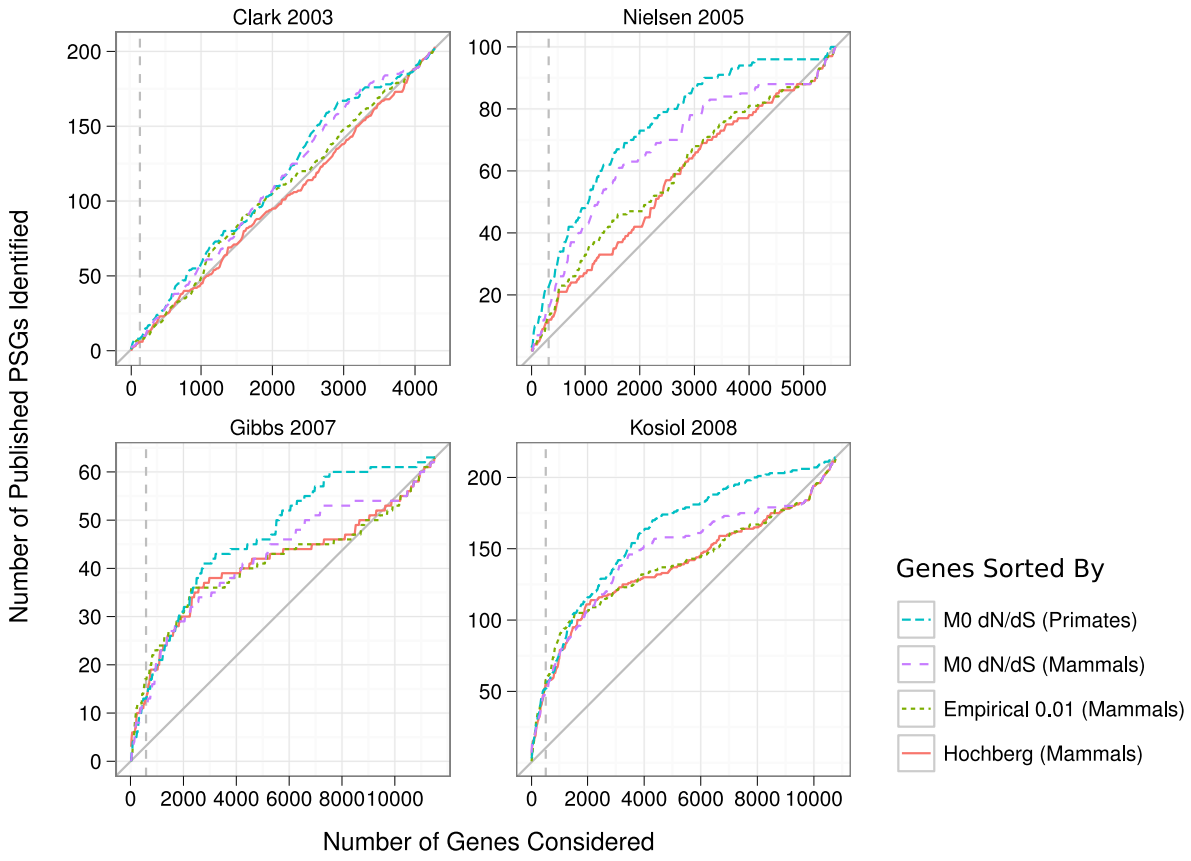


Figure 5.5: ROC curves for using dN/dS estimates and different PSG identification methods to identify PSGs in previously-published studies. The conservatively-filtered sitewise data were used for all comparisons. Within each panel, the x-axis represents all genes successfully matched between the published 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 $\text{Emp}_{0.01}$ PSGs in the Mammals species group.

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 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 four 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 $\text{Emp}_{0.01}$ 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 on the part of the sitewise methods to detect weak positive selection distributed throughout a protein.

5.6 Gene families with many PSGs

Table 5.2 contained a relatively large number of PSGs from certain gene families (e.g., solute carrier family genes *SLC26A8*, *SLC16A7*, *SLC4A1*, *SLC13A2*, *SLC9A10*; collagen genes *COL1A2*, *COL4A3*, *COL16A1*; and Toll-like receptor (TLR) 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 and be subject to similar selective pressures; thus, a family of largely immune genes such as the TLR family may be expected to be enriched for PSGs. The prevalence of positively-selected gene family members was at the same time 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 and 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 and 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

Gene Family	Genes	NPPs	PSGs	NPP–PSGs		Top 4 NPP–PSGs
Ensembl Families with $\geq$ 3PSGs						
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: Coincidence of PSGs and NPP within Ensembl gene families. Top: Counts of NPPs and PSGs in Ensembl gene families with at least three PSGs, sorted by the number of genes within that family which are both PSGs and NPPs. Bottom: the same summaries for manually-curated families and the set of all genes. The FET p-value for independent assortment of NPPs and PSGs within each group of genes is shown. NPP—nearby paralogous pair; PSG—positively-selected gene; FET—Fisher's exact test.

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 and Hahn, 2009].

I performed an assessment of the potential impact of gene conversion on the current dataset by identifying NPPs and comparing those to the list of $\text{Emp}_{0.05}$ 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 chromosome within 2Mb 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 three 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 six 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 and 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 moderately increased rates of 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, the impact of such co-occurrence on the false detection of positive selection within any one family was highly correlated with 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, or a novel type of genetic conflict underlying the presence of positive selection in these related gene families. Interestingly, *Staphylococcus aureus* bacteria express collagen-binding factors on their cell wall surface that appear to be vital to pathogenicity [Foster and Höök, 1998; Ohbayashi *et al.*, 2011], suggesting a possible host-pathogen genetic conflict underlying the signal of positive selection within collagen and metallopeptidase family proteins.

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 quantitative

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: Collagen and matrix metallopeptidase genes with evidence for positive selection. Strings of characters under the "Evidence for Positive Selection" heading indicate which methods from the current study ("Conservative" and "Relaxed") and which previously-published studies ("Lit.") identified each gene as a PSG (see text for definitions). Genes which are members of a NPP are marked with a "+" in the NPP column. Note that all of the collagen genes shared the same $Emp_{0.01}$ p-value of 2.8e-04; this was the minimum possible empirical p-value based on the 10,000 pseudo-replicates tested by the empirical method, indicating that the number of observed $p < 0.01$ sites within each gene was greater than in any of the pseudo-replicates. Chr. Loc.—the chromosomal location of the human gene in Mb; NPP—nearby paralogous pair.

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.* [2008]) 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 in other biologically interesting ways to assess levels of purifying and positive selection within

those groups. One obvious 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, that approach would lack power, as positive selection occurring within each PSGs may not necessarily be occurring in 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 these functionally related sequence regions.

To link sitewise estimates to the protein domains in which they reside, I mapped Pfam domain annotations from human genes in Ensembl onto the genome-wide set of mammalian alignments, yielding 3.73 million alignment columns with both sitewise estimates and Pfam domain annotations. Of these sites, 6,159 contained evidence for positive selection at a nominal $p < 0.01$ threshold in the Mammals group. (These numbers can be found in the summary statistics for Pfam-annotated sites contained in Table 4.4 of Chapter 4.) All the sites annotated with each Pfam domain were combined to produce domain-wise p-values using the same Hochberg, $\text{Emp}_{0.05}$ and $\text{Emp}_{0.01}$ methods described in Section 5.2. Domain-wise p-values were separately estimated for all species groups using the relaxed and conservative sitewise filtered datasets, and multiple testing was controlled at FDR< 0.1 using the Benjamini-Hochberg method in the same way as was already described for genes. 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, showing all domains with significant evidence for enrichment of positive selection at FDR< 0.1 in the Mammals species group using the $\text{Emp}_{0.01}$ and $\text{Emp}_{0.05}$ methods. As in 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 $\text{Emp}_{0.01}$ result (e.g., `P`, `L`, `M` for Primates, Laurasiatheria, and Mammals), lowercase letters indicate a significant $\text{Emp}_{0.05}$ 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.

Pfam Domain		FDR < 0.1		All Sites		$p < 0.01$ Sites		
Accession	Description	Cons.	Relaxed	Genes	Sites	Genes	Sites	Top 5 Genes with $p < 0.01$ PSCs
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
PF00530	Scavenger receptor cysteine-rich domain	H	P MmH	17	2865	8	42	CD5L, MARCO, MSR1, CD5
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		MmH	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
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
PF00031	Cystatin domain	P LMmH	P LMmH	11	780	7	33	HRG, KNG1, FETUB, CST9LP1
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 $p < 0.01$ PSCs
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
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 four species groups (Primates, Glires, Laurasiatheria, and Mammals) using two sitewise filtered datasets (conservative and relaxed) and three methods to combine sitewise estimates ($\text{Emp}_{0.05}$, $\text{Emp}_{0.01}$, 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, three or more genes, and FDR< 0.1 in Mammals using the $\text{Emp}_{0.05}$ and $\text{Emp}_{0.01}$ methods are shown.

An initial observation is 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. 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 have removed a large number of sites within evolutionarily conserved domains. As noted in Chapter 4, the conservative filter removed over 3,000 genes with more than two sets of putative mammalian paralogs; the results in this chapter suggested that filtering so many genes based on apparent paralogy is perhaps overly stringent. Further refinement of the criteria used to define the different levels of sitewise filters may be an interesting direction for future research.

As expected, immune-related functions dominate the list of significantly positively-selected Pfam domains. Together, the immunoglobulin and immunoglobulin V-set domains accounted for over 390 $p < 0.01$ PSCs spread across 109 proteins, and both domains were significant at FDR< 0.1 for multiple species groups and methods using the relaxed sitewise filter. Interestingly, only a relatively small fraction of immunoglobulin-containing genes contributed to this significant evidence for positive selection, with only 51 out of 240 total immunoglobulin-annotated genes containing $p < 0.01$ sites within the domain. A similar fraction of genes containing $p < 0.01$ sites was seen for the immunoglobulin V-set domain. This was not unexpected for immunoglobulins, which are known to be highly diverse protein domains in mammals, with representation in several hundred human genes comprising many immune and non-immune functions [Lander *et al.*, 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 the lectin C-type domain, a carbohydrate binding domain involved in cell adhesion and apoptosis [Cambi and Figdor, 2009], the IL-8 like cytokine domain involved in inflammation and chemotaxis in the immune response [Stein and Nombela-Arrieta, 2005], and the membrane attack complex (MAC) domain, also known as perforin, 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 in Table 5.5 was consistent with a large amount of positive selection in mammals having resulted from an evolutionary "arms race" between invading microbes and the host mammalian immune system, as the buildup and continued evolution of proteases and their inhibitors is known to be a significant medium through which this conflict is expressed. Invading pathogens express a diverse array of proteases designed to facilitate physical dispersal within the host, subvert the blood clotting system which would otherwise immobilize the pathogen, or directly attack proteins of the host immune system through proteolysis [Armstrong, 2006]. For example, *Plasmodium falicparum*, one of the

agents of malaria, expresses a variety of secreted and surface-bound proteases which aid infection [McKerrow *et al.*, 1993] and *Yersinia pestis*, the causative agent of plague, contains a surface protease which activates plasminogen, thus reducing the blood clotting reaction and significantly increasing the pathogen's virulence [Sodeinde *et al.*, 1992].

Host immune systems have accordingly evolved proteases and protease inhibitors which can act to inhibit or destroy various pathogenic factors: this analysis found that three classes of protease inhibitors, the serine protease inhibitors (or *serpins*), cystatin cysteine protease inhibitors, and the α_2-macroglobulin domain, showed evidence for positive selection in mammals. Meanwhile, a number of mammalian proteases have themselves been subject to significant positive selection, including trypsin, caspase, zinc carboxypeptidase, and the alpha/beta hydrolase fold. Interestingly, trypsin—one of the protease domains under the strongest positive selection in Table 5.5, containing 184 $p < 0.01$ sites—is a typical serine protease domain, which can be inhibited by members of the *serpin* family noted above. Caspases, which are centrally involved in apoptosis and also help trigger the inflammatory response to viral infection [Nicholson and Thornberry, 1997], are inhibited by two known viral proteins, CrmA from the cowpox virus and p35 from baculovirus [Villa *et al.*, 1997]. The evidence for positive selection on genes containing this domain may reflect the importance of viral infection throughout mammalian history. Other explanations are possible, of course; caspase proteins were highlighted by da Fonseca *et al.* [2010] in their review of PSGs involved in apoptosis, showing how they interact with a number of homeostatic, immune, and reproductive pathways.

Most of the domains contained within the "Other Domains" section of Table 5.5 are less well-characterized than the categorized domains or exhibit a less-specific range of functions within the context of mammalian biology. Some of these domains may ultimately belong within the immune category, or some of the genes containing PSCs within those domain regions may have obvious immune functions or be subject to other forms of ongoing genetic conflict. It is worth noting that a number of the domains and genes within this section have been implicated in spermatogenesis, the main non-immune function often associated with positive selection [Wyckoff *et al.*, 2000; Torgerson *et al.*, 2002; Nielsen *et al.*, 2005]: defects in phospholipases are associated with sperm immobility and infertility in mice [Murakami *et al.*, 2010; Sato *et al.*, 2010] and the *SPAM1* gene containing PSCs within its hyaluronidase domain is a major sperm cell antigen [Martin-DeLeon, 2006]. Other domains are known to have important cellular roles, but evoke no immediate hypothesis for why they contain an excess of PSCs. For example, cytochrome P450 proteins are involved in largely metabolic and biosynthetic processes including hormone biosynthesis and drug metabolism [Werck-Reichhart and Feyereisen, 2000], and the *Schlafen* genes containing the AAA domain have been characterized as being involved in the immune response and cell proliferation [Bustos *et al.*, 2009].

5.8 Case study: the sitewise evolutionary history of mannose-binding lectin 2 (*MBL2*)

I chose to highlight the human gene *MBL2*, which codes for the MBL protein, as an example of how sitewise patterns of purifying and positive selection can provide a new perspective the evolutionary history of immune proteins throughout the history of mammals. MBL is a member of the collectin protein family which is secreted by the liver and binds carbohydrate molecules of invading pathogens, facilitating phagocytosis and activating the complement system of the innate immune response through its association with MBL-associated serine proteases (MASPs) [Seyfarth *et al.*, 2005]. There has been some controversy surrounding the exact immune function of MBL, as alleles of the *MBL2* gene containing SNPs in the N-terminal collagen-like domain persist at high frequencies in a number of human populations (up to 30%). These alleles disrupt the MBL protein's collagenous structure and are associated with decreased levels of the protein in blood serum. Thus, the high prevalence of MBL-disrupting alleles seemed at odds with its supposed important role in the innate immune response [Seyfarth *et al.*, 2005] and some studies postulated a selective advantage for low MBL levels, which would help explain the high frequencies of MBL-disrupting alleles [Garred *et al.*, 2006]. One study of *MBL2* alleles in humans found high levels of heterozygosity supporting a selective interpretation [Bernig *et al.*, 2004], but a second study [Verdu *et al.*, 2006] found no evidence of strong purifying or positive selection, leading the authors to endorse a purely neutral interpretation, writing that "The evolutionary neutrality of *MBL2* strongly supports [...] that this lectin is largely redundant in host human defences." Thus, the evidence from within human populations was somewhat conflicting.

Although data from recent human evolution may be more directly applicable than between-species comparative analyses to our understanding of *MBL2* in human health and disease, the contradictory nature of the results from studies using *MBL2* allele suggested that a better understanding of the evolutionary history of *MBL2* in non-human primates and mammals could provide useful additional information. A comparative study in 12 non-human primates was indeed conducted by Verga Falzacappa *et al.* [2004]. Their analysis revealed a strong signal of purifying constraint within the collagen helix region, but they found no evidence for positive selection, concluding that "*MBL2* is well conserved in agreement with its important role in the immune system."

The sitewise view of selective pressures in *MBL2*, shown in Figure 5.6, paints a detailed picture of the evolutionary forces shaping this gene within mammals. The protein alignment is shown alongside the exon and Pfam domain organization, and sitewise selective pressures in six species groups are displayed below the protein alignment. (A detailed description of the colors and conventions used is included in the caption to Figure 5.6.) In this view, the effect of branch

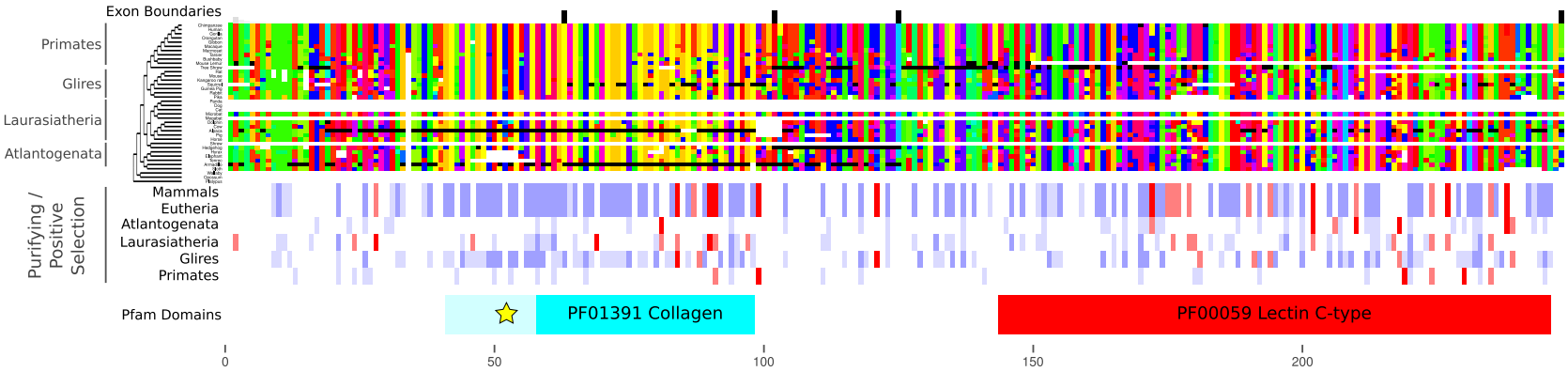


Figure 5.6: Alignment of mammalian orthologs, domain structure and sitewise selective pressures in 6 species groups for the *MBL2* gene. Protein sequences are displayed in rows, colored according to Taylor [1986], with the tree relating the sequences shown to the left of the alignment (branch lengths are not shown to scale). Blank rows correspond to species for which no orthologous sequences were found (panda, dog, cat, microbat, shrew, wallaby, and opposum). Black residues represent codons in low-coverage genomes which were either missing or filtered for low sequence quality . Alignment columns corresponding to a gap in the human sequence have been removed (i.e., the alignment is human-flattened). Although few indels were present in the full alignment, long unaligned regions in the mouse sequence, likely resulting from an altered transcript structure, necessitated flattening the alignment for clarity. Above the alignment, gray peaks show the location of boundaries between exons. Below the alignment, rows of blocks indicate the sitewise selective pressures estimated by SLR for each of the 6 labeled species groups. Light gray and gray blocks show sites with nominal $p < 0.05$ and $p < 0.01$ evidence for purifying selection, respectively; light red and red blocks show sites with nominal $p < 0.05$ and $p < 0.01$ evidence for positive selection, respectively. The bottom row shows the domain structure of *MBL2*: the blue bar represents the collagen helix domain (the light blue region shows a portion of the collagen helix not annotated by Ensembl but which contains the characteristic Gly-X-Y collagen motif), and the red bar represents the lectin C-type carbohydrate recognition domain. A star is drawn near the location of the nonsynonymous human SNPs in the collagen helix region (codons 52, 54, and 57).

length on each species group can be clearly seen by the different numbers of sites with $\omega < 1$ at $p < 0.05$ and $p < 0.01$ significance, shown as gray boxes below the alignment: Mammals and Eutheria detect significant purifying selection in a majority of sites, with slightly fewer in Glires and far fewer in Atlantogenata, Laurasiatheria and Primates. On the other hand, for sites with $\omega > 1$ at $p < 0.05$ and $p < 0.01$, shown as red boxes, no clear branch length-related pattern exists. Instead, Laurasiatheria exhibits the greatest number of PSCs, Mammals and Eutheria have slightly fewer, and the other mammalian orders show some, but significantly less, positive selection. Two regions of *MBL2* in particular show strong evidence for positive selection in multiple independent species groups: first, the last 20 codons of the collagen helix domain contain at least two $p < 0.01$ or $p < 0.05$ PSCs in each of Primates, Glires, and Laurasiatheria and one $p < 0.01$ PSCs in Atlantogenata, and second, the last 30 codons of the lectin domain contain multiple PSCs in all four species groups at $p < 0.05$ significance. A total of 9 sites elsewhere in the lectin domain are also significant in Mammals, but the evidence for those PSCs seems weak and is not corroborated by individual mammalian orders.

So what conclusions can be made regarding the evolution of *MBL2* and its potential immune function from these sitewise data? First, it was clear from the high number of purifying sites in Mammals that *MBL2* has been well-conserved across mammals, confirming the findings of Verga Falzacappa *et al.* [2004]. There appeared to be especially strong conservation within the collagen helix domain. Second, in contrast to Verga Falzacappa *et al.* [2004], the distribution of PSCs showed that certain regions of *MBL2* do show strong evidence for positive selection, but these regions do not overlap with the location of the human variants in question. Finally, the existence of PSCs across a number of different mammalian species groups suggested that the evolutionary forces causing the observed positive selection in *MBL2* have been phylogenetically widespread and ongoing in the evolution of different mammalian lineages. These observations are consistent with *MBL2* having performed an important immune function throughout the evolution of mammals. Furthermore, some regions of this protein appear to be more prone to experiencing positive selective pressure than others. Numerous previous studies have found PSCs clustered together near the site of protein-protein or protein-ligand interactions involved in antagonistic host-parasite evolution [Sawyer *et al.*, 2005; Kosiol *et al.*, 2008; Guirao-Rico and Aguadé, 2009; Huang *et al.*, 2011], and the evidence presented here suggests that the two identified regions of *MBL2* may be subject to similar evolutionary pressures.

5.9 Conclusions

In this chapter I developed and evaluated some methods for using genome-wide estimates of sitewise selective pressures to identify positively-selected genes and domains. The family-wise

error rate (FWER)-controlling approach (which is already implemented internally within SLR) performed well but lacked power to discriminate between genes with one PSC and genes with many PSCs of equal strength. Although generic statistical methods have been described for combining p-values from independent tests into an overall p-value, they mostly lacked power due to the overwhelming influence of purifying selection on the distribution of sitewise p-values within each gene. The third approach I tested was to calculate empirical p-values based on the number of $p < 0.01$ or $p < 0.05$ sites within each gene, the gene's length, and the genome-wide set of sitewise estimates.

The results from applying these methods were compared to each other and to previously-published sets of mammalian or primate positively selected genes (PSGs). Different species groups and different methods for combining sitewise estimates showed only moderate overlap, and previously-published sets of PSGs showed little to no overlap with each other or with the sitewise results. ROC curves were used to compare this overlap across the entire range of significance thresholds, showing similar patterns for sitewise p-values and whole-gene $d\mathrm{N}/d\mathrm{S}$ estimates.

The low number of shared PSGs was somewhat disheartning, as one might have hoped for greater overlap between studies. Some, but certainly not all, of these differences could be explained by differences in the methods used to detect selection. While Clark *et al.* [2003] used a branch-specific test to infer positive selection in human and chimpanzee, the PSGs that I compared from Nielsen *et al.* [2005], Gibbs *et al.* [2007] and Kosiol *et al.* [2008] all tested for positive selection across the entire tree under consideration. Perhaps a larger component of the variation in results is an unavoidable result of different species being analyzed: even with a consistent alignment and filtering strategy, I found that the Primates and Laurasiatheria species groups shared only roughly 20% of their PSGs (Figure 5.3A) and Glires contained far fewer PSGs.

There was more agreement between enriched functional categories: the sitewise-based PSGs were enriched in most of the well-established functions such as immunity and inflammation, but some commonly-identified terms such as sensory perception and olfaction were not. The somewhat unexpected enrichment for PSGs involved in mitosis and centrosome organization was interesting and evocative of a form of hypothesized chromosomal genetic conflict [Malik and Henikoff, 2002]. A number of gene families contained several PSGs, which provided some cause for concern due to the potential for gene conversion between family members to produce false positive results. On a genome-wide scale, PSGs and nearby paralogous pairs co-occurred more often than expected by chance, but the identification of false positives due to gene conversion was deferred for future research.

The last analysis performed in this chapter was the application of the same sitewise methods to identifying positive selection within Pfam domains. These results largely recapitulated the trends seen in the GO term enrichments, showing in another way that the evolution of the mammalian immune system has been strongly driven by positive selection due to interactions

between pathogens and their hosts. An interesting pattern seen in the domains with significant evidence for positive selection was that the PSCs were often located in a minority of the domain "instances" throughout the genome. For example, only 58 out of the 220 genes annotated with an immunoglobulin v-set domain contained any $p < 0.01$ PSCs within that domain. Other domains contained slightly higher proportions of positively-selected "instances", with 44 of 82 trypsin-containing genes, 18 of 39 cytochrome P450-containing genes, and 19 of 33 serpin-containing genes having $p < 0.01$ sites within the annotated domain region. These patterns reflected the diverse immune and non-immune roles acquired by various domain types in the mammalian repertoire of protein-coding genes.

Chapter 6

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

6.1 Introduction

of molecules to compare humans with chimpanzee and gorilla, our closest living relatives. In the 1960s Zuckerkandl and Pauling applied their molecular clock hypothesis to estimate the gorilla-human divergence time, and King and Wilson's (in)famous 1975 paper, "Evolution at Two Levels in Humans and Chimpanzees," proposed that changes to the expression of genes, rather than changes within protein-coding regions themselves, may be primarily responsible for the observed anatomical and behavioral differences between humans and chimpanzees [King and Wilson, 1975]. Figure 6.1 shows the schematic used by King and Wilson to illustrate the proposed disconnect between organismal and molecular rates of evolution: whereas human shows signs of having changed more from the human-chimpanzee ancestor in terms of behavior and anatomy, more equal rates of molecular evolution were observed.

Continued genetic, anatomical and behavioral study of primates has somewhat tempered the imbalance in the "organismal" rates of human and chimpanzee evolution as drawn by King and Wilson: for example, the evolution of human form does not appear to be exceptional within the context of other mammalian and primate species [Carroll, 2003], and humans variously share and differ in many behavioral and sexual characteristics with chimpanzee and gorilla [Harcourt *et al.*, 1980; Goodall, 1986; Vigilant and Bradley, 2004]. On the molecular level, however, few plausible connections between human-specific phenotypes and molecular changes have been found, lending support to King and Wilson's 30-year old prediction [Clark *et al.*, 2003; Mikkelsen *et al.*, 2005; Bradley, 2008]. In the first comparison of the complete human and chimpanzee genomes, Mikkelsen *et al.* [2005] wrote that "We thus find minimal evidence of acceleration unique to either

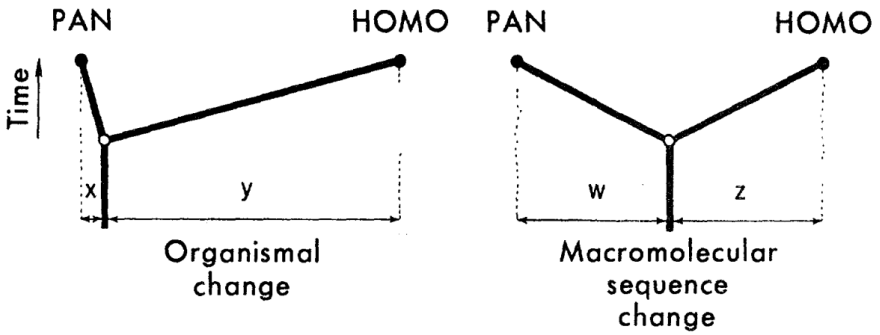


Figure 6.1: Schematic diagram of the rates of organismal and molecular change in human (Homo) and chimpanzee (Pan). Figure taken from King and Wilson [1975].

the human or chimpanzee lineage across broad functional categories." Still, these general findings do not preclude the possibility of *some* amount of protein-coding change playing a role in the differences between humans and our close relatives. Studies of protein-coding evolution in humans and other great apes have provided insight into the evolution of primates as a whole, identifying patterns of adaptive evolution or relaxed constraint in genes related to sound perception [Clark *et al.*, 2003], coloration [Mundy, 2007], language [Enard *et al.*, 2002] and brain size [Montgomery *et al.*, 2011]. Adaptive evolution in genes related to sperm production [Clark and Swanson, 2005] and immune defense [Sawyer *et al.*, 2005] has also repeatedly been found in primates, though these patterns of selective pressures appear to be shared throughout the mammalian clade, as evinced in Chapter 5.

Within this context, the recent sequencing of a western lowland gorilla genome provided an opportunity to further examine patterns of molecular evolution within human and our closest living relatives, even if the power to detect lineage-specific adaptive events was likely to be limited and the prevalence of such events is still debated. Previous phylogenetic estimates indicated that the human-chimpanzee-gorilla (HCG) common ancestor lived 6-10 million years (Myr) ago, with humans and chimpanzees diverging a few Myr after that [Bradley, 2008]. The inclusion of gorilla into a sequence analysis would thus allow methods to make a distinction between substitutions along the more distant HCG and the more recent human-chimpanzee (HC) ancestral branches of the phylogeny. The inclusion of gorilla within a comparative analysis would also provide 6-10 Myr of additional branch length to the primate tree, possibly providing additional information regarding the constancy or variability of evolutionary rates in the recent evolution of the African great apes (AGAs) (i.e., human, chimpanzee and gorilla). Previous analyses have estimated a greater effective population size (N_e) in chimpanzee compared to human [Mikkelsen *et al.*, 2005; Siepel, 2009], and somewhat surprisingly, one study identified more lineage-specific positively selected genes (PSGs) in chimpanzee than in human (Bakewell *et al.* [2007], but see Mallick *et al.*

[2009] for evidence of false positives in these results). Gorilla, representing a third independent AGA lineage with a slightly longer terminal branch length than human and chimpanzee, could provide an important additional data point in this regard.

In collaboration with Stephen Montgomery and Nick Mundy from the Zoology Department, University of Cambridge, I performed an analysis of the evolution of protein-coding genes in gorilla and the AGA. The analysis was jointly designed all three of us; all data collection, calculations and statistical analyses were performed by me; and results were interpreted by us and other members of the Gorilla Analysis Consortium. As with the analysis for the Mammalian Genome Project, a summary of our main findings was contributed to the manuscript describing the gorilla genome, which is currently undergoing peer review. The description of the methods and results presented here is similar to the text included in the submitted manuscript and supplementary documents.

The main focus of this analysis was to use codon models of evolution to identify genes with accelerated nonsynonymous substitution rates in the terminal and ancestral branches of the AGA. To do this, I collected a highly filtered set of coding alignments in six primate species (the AGAs plus orangutan, rhesus macaque, and marmoset) and used a series of likelihood ratio tests (LRTs) based on the branch models implemented in PAML to identify accelerated genes—defined here as genes with significant evidence of an increased $d\mathrm{N}/d\mathrm{S}$ ratio—along different branches of the primate phylogeny. (The same LRTs could also be used to identify decelerated genes, but the biological significance of a decreased $d\mathrm{N}/d\mathrm{S}$ ratio is less clear and so was not a focus of this study.) Within the wider context of the consortium's analysis of the gorilla genome, which included an investigation of incomplete lineage sorting (ILS) in coding and non-coding regions, a secondary goal of this study was to identify patterns of ILS within and near protein-coding genes. A third goal was to use the set of genome-wide coding alignments to estimate lineage-specific average $d\mathrm{N}/d\mathrm{S}$ ratios. Through the connection between N_e and the strength of purifying selection, these results could be used to place gorilla and the ancestral AGA branches within the wider context of changing primate population sizes through time [Mikkelsen *et al.*, 2005].

6.2 Data collection and quality control

Primate one-to-one orthologous genes

All orthologous gene sets, gene trees, and sequence alignments were collected from Ensembl Compara release 60 using the Ensembl Perl API [Vilella *et al.*, 2009; Flicek *et al.*, 2011]. I first identified the set of genes sharing one-to-one orthology among all six primates by collecting homology annotations from the Ensembl Compara database: for each human protein-coding gene, the orthology status for each non-human species was assigned to different categories based on

the number of homologous genes and the type of homology as annotated by Ensembl. Homologs were classified as either one-to-one (e.g., one homolog available and either an "ortholog_one2one" or "apparent_ortholog_one2one" homology type), deleted (e.g., no homolog available), duplicated (e.g., multiple homologs available), or human duplication (e.g., one homolog available but containing an "ortholog_one2many" annotation, indicating that there are multiple human homologs for that single non-human homolog). From an initial set of 20,746 human protein-coding genes, this procedure identified 12,652 genes with 6-way 1-to-1 orthology; 4,809 genes with primate deletions; 1,171 genes with primate duplications; 308 genes with human-specific duplications; and 1,806 genes with mixed patterns of duplication and deletion.

Collecting and filtering six-way codon alignments for one-to-one genes

Codon alignments of all one-to-one orthologs were extracted from the 6-way primate genomic alignments pre-calculated by the Ensembl pipeline using the Enredo-Pecan-Ortheus (EPO) pipeline [Paten *et al.*, 2008a,b]. I extracted and concatenated primate alignment blocks corresponding to the protein-coding portion of each exon from the consensus coding transcript of each human gene. These alignments were then flattened to the human reference by removing all columns with insertions in non-human primates or deletions in the human lineage. The flattening was done to ensure that the alignment remained "in-frame" with respect to the human transcripts; the number of sites removed by this flattening was expected to be low due to the low divergence within primates, and the lower sequence quality of the non-human primates led to a concern that stretches of sequence not aligning to human may be the result of sequencing or assembly error. Since the EPO alignments are generated on the DNA level and the current evolutionary analysis was to be performed on the codon level, I cleaned each alignment for codon analysis by masking out any triplets containing stop codons or out-of-frame gaps. Of the original 12,562 one-to-one genes identified above, 11,538 codon alignments were successfully collected. The reduction numbers came from 1,024 genes which were discarded because an entire species was missing from that region of the primate EPO alignment; the species most often missing in the alignments were orangutan (520 genes), marmoset (475 genes), and macaque (328 genes).

The low levels of divergence between the primate species being analyzed made it extremely important to avoid the inclusion of any incorrectly-aligned material. The expected number of lineage-specific substitutions per gene scales linearly with the length of that species' terminal branch, meaning a small number of sequencing, assembly or alignment errors causing apparent nonsynonymous substitutions along one of the short AGA terminal lineages could easily lead to a false positive inference of accelerated evolution. As such, an aggressive set of filters was applied to each alignment prior to the PAML analysis.

First, the chimpanzee, orangutan, macaque and marmoset sequences were filtered using Phred or Phred-like quality scores downloaded from the UCSC (chimpanzee and macaque) or WUSTL (orangutan and marmoset) websites. Any bases with a quality score lower than 30 (corresponding to an expected error rate of 1 in 1000 bases) were replaced with 'N's.

An initial analysis of the Phred filtered 6-way alignments revealed many stretches of alignment with obviously non-homologous sequence in one or more non-human genomes. These regions appeared similar to the clustered substitutions seen in Chapter 4, but the causative factor was not likely the same: since the EPO alignments were built on the DNA level without the use of any gene annotation information, the presence of alternative exons or mis-annotated genes in different species could not explain these misaligned regions. Instead, the most likely cause was some combination of mis-assembled genomic sequence and misalignment by the EPO pipeline. I found that genes containing these dubious aligned stretches were prominent among the initial list of top accelerated genes based on these alignments.

To exclude such regions from analysis, I applied a filter based on windows of inferred lineage-specific substitutions, using an approach similar to that described for the filtering of mammalian alignments in Chapter 4. First, the codon alignment was analysed with the codeml program of PAML v4.14 using a M0 model to infer substitutions in the terminal lineages of the 6-species tree. Using the branch lengths (expressed as the expected number of substitutions per codon) and substitution events inferred by PAML, I analyzed the density of codons containing lineage-specific non-synonymous substitutions within every 15-codon window along the alignment. Any window containing more than 10 nonsynonymous substitutions per codon per unit of branch length was masked with 'N's. After analyzing the preliminary results from this method, a number of additional heuristic corrections were made to avoid excess stringency or lenience: first, branch lengths below 0.05 substitutions per codon were set to 0.05 during the filtering step in order to avoid too small a denominator; second, the masking threshold was decreased from 10 to 5 for any windows overlapping alignment gaps or ambiguous nucleotides; third, codons containing two or three nucleotide substitutions along one branch were counted as two nonsynonymous substitutions. The third correction was included to provide additional weight to codons with multiple apparent substitutions, which were especially likely to be the result of misalignment given the short branch lengths in the primate phylogeny.

This procedure resulted in a total of 72,729 nucleotides being masked from 1,156 genes, from a total of 11,538 genes and 7.28 million alignment sites. The breakdown of the number of genes in which each species had at least one nucleotide masked was as follows: 12 human, 195 chimpanzee, 232 gorilla, 296 orangutan, 271 macaque and 324 marmoset. The low number of genes from which any human sequence was masked indicated that the filtering was not overly conservative, while the high numbers in non-human primates indicated that the current assemblies of those genomes are more likely to contain highly localized, apparently spurious runs of nonsynonymous substitutions

in the regions of EPO alignments corresponding to human transcripts.

A final filter was applied to avoid a potential bias from substitutions in regions of ILS between sequences in human, chimpanzee, and gorilla. ILS regions are genomic segments where either gorilla-chimpanzee or gorilla-human share a most recent common ancestor, deviating from the "canonical" relationship where human and chimpanzee share the most recent ancestor. Roughly 20-30% of the genome shows evidence of ILS within the African great apes [Hobolth *et al.*, 2007]. In cases where a synonymous or nonsynonymous substitution occurs along the ancestral branch of a genomic segment subject to ILS, the assumption of a single phylogenetic tree per gene is violated and PAML cannot correctly infer a single substitution event. Instead, two substitutions must be inferred in order to fit the observed site pattern to the canonical phylogenetic tree. The method can choose to infer either two identical substitutions (one along each of the terminal "ILS" branches sharing the non-canonical common ancestor) or one substitution along the HCG ancestral branch and a second reversion substitution along the non-ILS terminal branch. In normal usage of PAML, the maximum likelihood (ML) estimation of parameters integrates over all possible substitution histories, but certain paths may be more favored than others. To test whether PAML the former (two identical substitutions along terminal branches) or the latter (one ancestral substitution and one reversion) sequence of events was more probable in great ape alignments, I used PAML to reconstruct ancestral sequences by ML [Yang *et al.*, 1995] and inferred the most likely substitution history. I observed that PAML tended to infer the latter sequence of events as the most likely substitution history. The long length of the HCG ancestral branch may have contributed to the substitution-reversion path being the more likely one.

Given the non-negligible proportion of expected sites under ILS, I applied a simple filtering method to mask out codons that were likely the result of a single substitution in an ancestral ILS lineage. Any codon where either gorilla-human or gorilla-chimpanzee shared a codon sequence that was different from both orangutan and the non-ILS species (either human or chimpanzee) was considered likely to contain an ancestral ILS substitution. The human, chimpanzee and gorilla sequences at these sites were all masked with 'N's, causing PAML to treat those nucleotides as missing data. This resulted in 7,841 codons being masked from 4,340 genes prior to analysis with PAML. Although the ILS masking was relatively widespread across genes, its effect on the results was conservative with respect to the number of inferred substitutions, and likely had a minimal impact on the identification of accelerated genes: the majority of masked genes (2,605) contained only one masked codon, which would be unlikely to seriously attenuate an otherwise strong signal of lineage-specific acceleration.

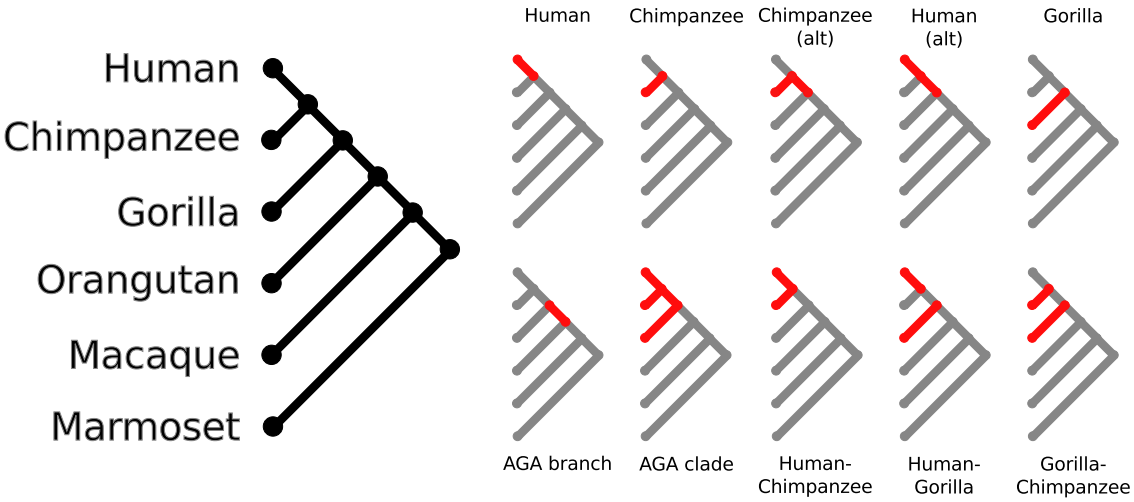


Figure 6.2: Branch models used to construct LRTs for detecting accelerated genes in various branches of the AGA phylogeny. The foreground branches for each model are highlighted in red. The label above or below each tree describes which species or group of species is under investigation with each model. AGA—African great apes; alt.—alternative model.

Manual identification of remaining alignment or assembly errors

A manual analysis of genes with suspiciously strong evidence for an accelerated nonsynonymous substitution rate identified some genes with apparent alignment or assembly error that escaped the various filtering steps described above. For each potentially erroneous gene, a manual analysis of the alignment was undertaken by visually inspecting the codon alignment, the locations of inferred substitutions, and the protein-based alignment of the same gene downloaded from v60 of the Ensembl Compara database. In total, four genes—*ITPK1*, *POLR2A*, *ATN1*, and *GAS6*—were found to show evidence of serious misalignment. Two other genes, *SUPT16H* and *POLR1A*, contained similarly strong signals of lineage-specific acceleration and elevated numbers of nonsynonymous and synonymous substitutions. These genes were also manually assessed, but no obvious signs of misalignment were found. The four genes with clear errors were removed from the set of genes analyzed for the remainder of the analysis, resulting in 11,534 total genes analyzed by the LRTs described below.

6.3 Codon model evolutionary analysis

The set of evolutionary models and LRTs used to identify accelerated genes in the 6-way primate alignments were designed by Stephen Montgomery, Nick Mundy and myself.

In contrast to the sitewise methods used in Chapters 4 and 5 to identify sites and genes subject to positive selection throughout an entire tree, the goal of this analysis was to identify

genes with evidence of increased or decreased rates of evolution along different branches of the primate tree. To detect signals of accelerated and decelerated evolution in gorilla and the AGA, a total of 10 so-called branch models of evolution were created. Each branch model separates the phylogenetic tree into two categories of branches, foreground and background branches, which are then modeled as evolving with separate dN/dS ratios by including distinct foreground and background ω parameters in the model [Yang, 1998; Yang and Nielsen, 1998]. Figure 6.2 shows the models used in this study, with background branches in gray and foreground branches in red. Each branch model was designed to allow detection of elevated (accelerated) or decreased (decelerated) dN/dS ratios in branches of particular interest to the study of AGA evolution. PAML was used to estimate parameters and calculate the likelihood value for each model in Figure 6.2 applied to each coding alignment. A LRT was performed for each model by comparing the likelihood of the alignment under that branch model to the likelihood of the alignment under the simpler M0 model, which uses a single ω parameter for the entire tree. This test will be referred to as a branch-LRT to distinguish it from the more commonly used branch-site LRT described below. The branch-LRT statistic represents the strength of evidence that a given branch model is a better fit than the simpler M0 model to a given alignment; in other words, a large statistic can be interpreted as an indication that the evolution of the gene is well-explained by different dN/dS ratios in the foreground and background branches of the tree. The ML estimate of the two ω parameters could be used to identify genes where the estimated foreground ω was higher or lower than the background. Genes where the foreground ω was higher than the background were categorized as accelerations, and genes where the foreground ω was lower than the background were categorized as decelerations. Using the LRT statistic and the distinction between accelerations and decelerations, a signed LRT statistic was constructed for each branch-LRT, where accelerated genes were assigned the branch-LRT statistic and decelerated genes were assigned the negative of the branch-LRT statistic. In this way, a single number was used to encapsulate the direction and strength of evidence in support of a shifted dN/dS ratio in the foreground branch of each model presented in Figure 6.2.

A highly positive signed branch-LRT score represented strong evidence for a lineage-specific elevated dN/dS ratio. Such an elevated ratio could be explained either by the effects of positive selection or relaxed constraint [Nielsen *et al.*, 2005; Mikkelsen *et al.*, 2005]. To attempt to distinguish the former from the latter, the branch-site LRT implemented in PAML [Zhang *et al.*, 2005] was also used to identify genes with significant evidence for positive selection acting along a branch or clade. Similar to the branch-LRT, the branch-site LRT requires a predefined separation of branches into foreground and background categories. However, the branch-site LRT is specifically tuned towards identifying temporally and spatially localized episodes of positive selection [Nielsen and Yang, 1998; Yang and Nielsen, 2002; Zhang *et al.*, 2005] by performing a LRT between a model allowing for positive selection and a model allowing only for purifying

and neutral selection. The branch-site LRT was run for the Human, Chimpanzee, Gorilla, AGA branch and AGA blade models shown in Figure 6.2.

In the analysis and interpretation of results, greater weight was placed on the branch-LRT results than on the branch-site LRT. Although a significant branch-site LRT result allows for a stronger interpretation than the branch-LRT (i.e. the former indicates significant evidence for positive selection acting along the foreground branch while the latter only indicates a significantly increased $d\mathrm{N}/d\mathrm{S}$ along the foreground branch), the models used for the branch-site LRT are more parameter-rich than those for the branch-LRT (containing four vs. two parameters, aside from the branch lengths and κ). When the sequences in an alignment are very similar, parameter estimates will be noisier due to the lack of data. The branch-site LRT was expected to suffer more from this problem due to its greater number of parameters, and the terminal branches of the AGA clade are very short (ca. 0.005 substitutions per synonymous site for human and chimpanzee), suggesting that parameter estimation in the branch-site LRT may be especially difficult. One recent study, in which empirical results from 175 transcription factor genes and a detailed simulation study showed that the branch-site LRT has little power to detect positive selection within closely-related primate genes [Nickel *et al.*, 2008], provided some empirical support for the expectation of limited power. Furthermore, the branch-LRT is sensitive to the same signal as the branch-site LRT—an increased number of substitutions in the foreground lineage—so it was expected that the genes with the strongest branch-LRT results genome-wide would be likely candidates for genes having undergone adaptive evolution in the foreground branch(es) of interest.

For each model tested, the full length codon alignment was input to the `codeml` program from the PAML package along with a phylogenetic tree corresponding to the accepted species tree structure (the labeled tree in Figure 6.2). The 'cleandata' option was set to 0 (i.e., alignment columns containing gaps were not removed from the analysis but were treated as ambiguous data), and branch lengths inferred by `codeml` based on the initial M0 model analysis were used as the initial branch lengths for all other models tested.

When the null model of evolution in the branch-LRT is true, the LRT statistic—measured as twice the difference in log-likelihood values between the branch model and the null M0 model—should be distributed according to a χ^2_1 distribution with one degree of freedom. The same null distribution was assumed for the branch-site LRT. Strictly speaking, the branch-site LRT null distribution should be a 50:50 mixture of a point mass at 0 and χ^2_1 with one degree of freedom, but the more conservative χ^2_1 distribution is recommended to guard against violations of model assumptions [Yang, 2007]. p-values for each branch-LRT and branch-site LRT result were thus calculated by comparing the absolute value of the LRT statistic to a chi-squared distribution with 1 degree of freedom. The Benjamini-Hochberg method [Benjamini and Hochberg, 1995] was used to correct for multiple testing within each branch model by controlling the false discovery rate (FDR).

Model / Species	Acceleration $p < 0.05$	Acceleration FDR< 0.1	Deceleration $p < 0.05$	Deceleration FDR< 0.1	Branch-site LRT $p < 0.05$	Branch-site LRT FDR< 0.1
Human	663	10	151	2	142	8
Chimpanzee	562	18	157	1	192	40
Gorilla	535	14	226	1	183	27
AGA Branch	299	3	314	1	152	15
AGA Clade	869	52	289	9	341	21
HC Parallel	51	0	14	0	-	-
GH Parallel	40	0	10	1	-	-
GC Parallel	25	1	8	0	-	-

Table 6.1: A summary of the branch-LRT and branch-site LRT results. The first five rows correspond to the equivalently-named models in Figure 6.2; the HC Parallel, GH Parallel, and GC Parallel models correspond to the LRT_{min} and LRT_{max} between pairs of species, as described in Section 6.4. Each cell represents the number of genes for which the LRT was significant at the specified threshold for the given model. p-values were calculated by comparing the LRT statistic to a χ^2_1 distribution with one degree of freedom. The false discovery rate (FDR) was controlled within each model separately for the branch-LRTs and for the branch-site LRTs using the Benjamini and Hochberg [1995] method.

The "alternative" models shown in Figure 6.2, which were designed to detect accelerations along the human and chimpanzee lineages while correcting for the difference in branch length between the human-chimpanzee terminal branches and the gorilla terminal branch, did not yield qualitatively different results from the equivalent uncorrected models. Furthermore, the fact that the foreground branches in the two "alternative" models shared the same human-chimpanzee ancestral branch made the models non-independent and difficult to compare. To simplify the discussion, results from those tests were discarded from the rest of the analysis.

Branch-LRT and branch-site LRT results

Table 6.1 presents, for each model and at two significance thresholds ($p < 0.05$ and FDR< 0.1), the number of significantly accelerated and decelerated genes according to the branch-LRT and the number of positively-selected genes according to the branch-site LRT. The number of $p < 0.05$ accelerated genes for almost all of the branch models was greater than the expected number under the null model. Assuming equivalent amounts of acceleration and deceleration, the expected number of $p < 0.05$ accelerations and decelerations using the branch LRT would each be roughly $11,538 \times 0.05/2 = 288$ genes. All models showed an excess of accelerated genes, with between 300 and 873 genes accelerated at the nominal $p < 0.05$ threshold. Significant evidence for deceleration was found at levels equal to or slightly below the null expectation, with between 151 and 314 decelerations per model. The "AGA Branch" model showed a notable tendency towards lower dN/dS ratios, with the fewest accelerations (300) and most decelerations (314) out of all models tested.

Looking at genes with strong evidence for acceleration or deceleration, defined as those corresponding to FDR< 0.1, roughly equivalent numbers were found in the three terminal lineage models (human / chimpanzee / gorilla) with between 10–19 strong accelerations and between 1–2 strong decelerations for each model. The other models were much more variable, possibly due to differences in power resulting from different foreground branch lengths, with the AGA Clade model showing many strongly-shifted genes (56 strong accelerations and 9 strong decelerations) and the AGA Stem model showing very few (3 strong accelerations and 1 strong deceleration). The Human-Chimpanzee, Gorilla-Human, and Gorilla-Chimpanzee models, designed to detect evidence for parallel accelerations and decelerations, showed roughly twice as many strongly accelerated and decelerated genes as their terminal-branch counterparts (29–45 strong accelerations and 3–6 strong decelerations), as might be expected based on the doubled amount of branch length in the foreground portions of their models.

The overview of the LRT results presented in Table 6.1 showed human, chimpanzee and gorilla to contain largely similar numbers of accelerated and decelerated genes according to the branch-LRTs. At a nominal $p < 0.05$ threshold, human yielded slightly more accelerations than chimpanzee and gorilla, but at a FDR< 0.1 threshold it contained slightly fewer. The most striking difference between the three AGA species was perhaps the number of significant branch-site results in chimpanzee at FDR< 0.1, with 40 PSGs compared to 27 in gorilla and 8 in human. The branch-site LRT results were interpreted with caution, however, due to the problems with parameter estimation noted above and a number of previous studies which showed that the low quality of the chimpanzee genome assembly may contribute to inflated estimates of positive selection [Schneider *et al.*, 2009; Mallick *et al.*, 2009]. Finally, the "AGA Clade" model, which included all three AGA species as foreground branches, showed relatively high numbers of significant results for all tests performed, which likely reflected a power increase resulting from the greater amount of foreground branch length. The parallel accelerations and decelerations included in the bottom three rows of Table 6.1 will be discussed in Section 6.4, but the main impressions from the summary of non-parallel results were that accelerated genes are more prominent than decelerated genes in the AGA species relative to the primate background and that the AGA terminal lineages shared similar numbers of accelerated genes based on the branch-LRT results.

6.4 Parallel accelerations

I used the lineage-specific gene acceleration LRT results to evaluate the prevalence and strength of parallel gene accelerations between gorila and human (GH), gorila and chimpanzee (GC), and chimpanzee and human (CH) during the time period since the speciation of each pair. Parallel accelerations were analyzed on three levels: first, quantifying genome-wide signals for shared gene

acceleration; second, identifying Gene Ontology (GO) terms enriched for shared accelerations; and third, identifying genes with the strongest evidence of parallel accelerations for each species pair.

A suitable statistic by which to measure the amount of evidence for parallel accelerations was first developed. Three of the branch models shown in Figure 6.2 were designed to be sensitive to parallel accelerations in the species pairs of interest (i.e. the models labeled "Human-Chimpanzee", "Human-Gorilla", and "Gorilla-Chimpanzee"), but I found that many of the strongest branch-LRT results for these three models were driven by nonsynonymous substitutions in primarily one of the two species pairs. For example, *SUPT16H*, the gene with the highest branch-LRT under the "Human-Gorilla", yielded a LRT score of 51.14. However, it appears that most of this signal was due to substitutions in the gorilla lineage: looking at the lineage-specific human and gorilla LRT values for the same gene, I found a gorilla LRT of 61.1 and a human LRT of -1.34 (where a negative value indicated an estimated decrease in dN/dS relative to the background branches). Thus, human and gorilla clearly did not both experience independently accelerated dN/dS levels in *SUPT16H*, despite the strong LRT result from the gorilla-human branch model. For this reason, LRT results from these three "parallel" branch models were excluded from further analysis. To ensure that genes identified as undergoing parallel acceleration showed independent evidence in each lineage of having experienced dN/dS accelerations, I instead used the minimum of both lineages' independent branch model LRT value as the statistic for parallel acceleration in each pair of species. This statistic will be referred to as LRT_{min}.

The counts of parallel accelerations and decelerations shown in Table 6.1 were calculated using the LRT_{min} statistic for accelerations and an equivalent LRT_{max} statistic for decelerations (as the signed LRT statistic was negative for decelerated genes). Note that the LRT_{min} and LRT_{max} statistics do not share the same null distribution as each individual branch-LRT statistic, so the χ^2_1 thresholds used to identify genes significant at $p < 0.05$ and FDR< 0.1 for the branch-LRT were not strictly appropriate for assessing the significance of LRT_{min} and LRT_{max} values. For this reason, the analyses described below used a variety of LRT_{min} thresholds to investigate levels of parallel acceleration in the AGA species. For the sake of consistency, however, the same $p < 0.05$ χ^2_1 threshold and the Benjamini and Hochberg [1995] method for FDR control used for the individual branch-LRT results were applied to the LRT_{min} and LRT_{max} values for the three AGA species pairs (HC, GH, and GC); the number of significant genes for these species pairs are shown in Table 6.1.

Genome-wide rates of shared acceleration

A randomized resampling strategy was used to determine whether the number of parallel accelerations at a given branch-LRT cutoff threshold was significantly greater than that expected given independent distribution of each species' accelerated genes. For each iteration of the randomiza-

tion, a set of pseudo-"accelerated" genes for each paired species was chosen by randomly sampling N_{acc} genes (where N_{acc} is the number of observed lineage-specific accelerations for each species at the given LRT cutoff threshold) from among the 11,534 total genes. The number of overlapping accelerated genes was counted at each iteration, and the fraction of iterations which yielded a greater number of overlapping accelerations than the observed number of parallel accelerations was taken as the p-value for the significance of the observed number of overlapping accelerations at the given cutoff threshold. This was repeated for each species pair, and for cutoff thresholds ranging from 0 to 5.

The magnitude of over- or under-representation of parallel accelerations was also estimated by calculating the co-occurrence excess (defined as $N_{obs}/N_{exp} - 1$) of accelerated genes for each pair of species and the same range of branch-LRT cutoff thresholds. The expected number of parallel accelerations was calculated using the same null expectation as the randomization test: that each lineage has a proportion of accelerated genes, and that the accelerations for each lineage are independently distributed amongst the 11,534 total genes. N_{exp}, the expected number of overlapping accelerated genes, is thus proportional to the product of each lineage's proportion of accelerated genes, $N_{exp} = (N_{accA}/N) \times (N_{accB}/N) \times N$, where N is the total number of genes and N_{accA} and N_{accB} are the numbers of accelerated genes in each lineage. For each species pair and cutoff threshold, 100 bootstrap replicate datasets were sampled from the 11,534 genes. The co-occurrence excess was calculated for each replicate, and confidence intervals were calculated at the 50% level from the set of bootstrap values.

The results of the randomisation test and co-occurrence calculations are shown in Figure 6.3. For each species pair, the co-occurrence excess is plotted as a function of the LRT cutoff threshold with a solid line drawn at the co-occurrence value and a shaded area drawn around the 50% bootstrap confidence interval. The results of the randomization test are indicated by one or two stars drawn adjacent to the co-occurrence value for the given dataset. For reference, a dotted vertical line is drawn at the LRT cutoff corresponding to a nominal $p < 0.05$ χ^2_1 cutoff value. At lower (i.e. more lenient) LRT cutoff values a greater number of genes in each of the paired species were accelerated, and at higher (i.e. more stringent) LRT cutoff values fewer genes are accelerated. This sample size effect can be seen in the wider confidence intervals and larger amounts of apparent stochastic noise at higher LRT cutoffs.

A trend was clear when comparing the co-occurrence excess levels and randomisation test p-values between the GH, GC and HC species pairs: HC showed the largest excess of parallel accelerations, GH showed an intermediate amount of excess, and GC showed the least excess. This trend was consistent across a wide range of threshold cutoff values and was supported by both the co-occurrence excess values and the results of the randomisation tests. The HC species pair showed randomisation $p < 0.01$ for all but the two highest (i.e., most stringent) threshold cutoffs and a maximum co-occurrence excess of nearly 60%. The GH species pair showed significantly

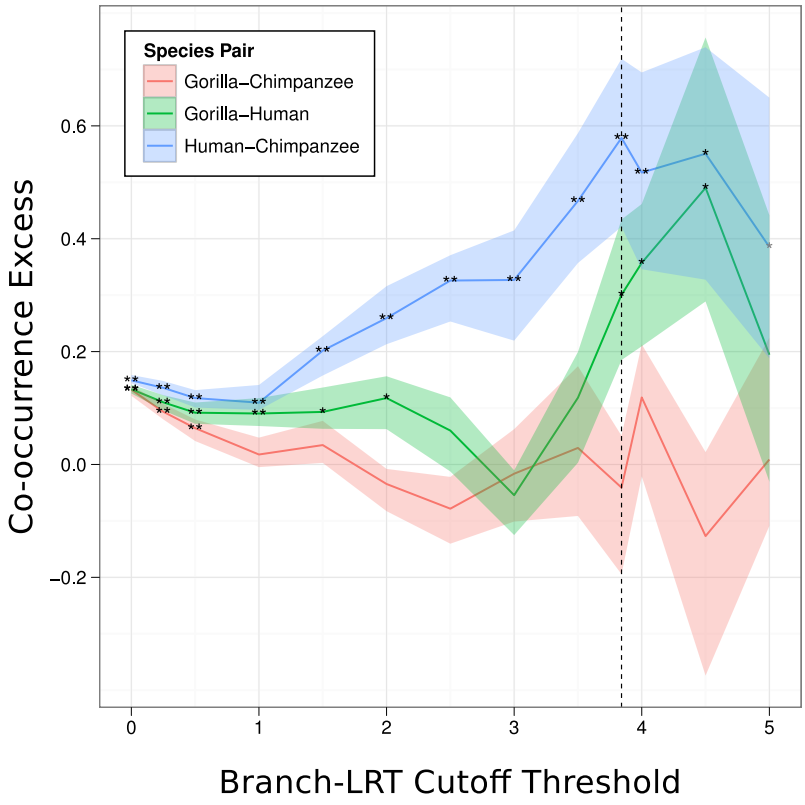


Figure 6.3: Genome-wide excess of parallel accelerations in pairs of AGA species at various LRT cutoff values. Parallel gene accelerations in three species pairs (Gorilla-Chimpanzee, red; Gorilla-Human, green; Human-Chimpanzee, blue) were identified using branch model LRT cutoffs from 0 to 5 (x-axis; see text for details on how parallel accelerations were identified), and the co-occurrence excess between lineage-specific and parallel accelerations at each LRT cutoff was calculated (y-axis; the 50% bootstrap confidence interval is shaded). A randomization procedure was used to identify significant enrichment for parallel accelerations: two black stars indicate $p < 0.01$, one black star indicates $p < 0.05$ for the given species pair and LRT cutoff. A vertical dotted line indicates the LRT cutoff corresponding to the 95% χ^2_1 significance value.

enriched acceleration overlap at $p < 0.05$ for threshold cutoffs below 2 and at 3.84, 4 and 4.5. The GH co-occurrence excess was noticeably lower than that of CH, but above zero for all but one threshold cutoff. The GC species pair showed little evidence of genome-wide enrichment for parallel accelerations, with a significant overlap only at weak threshold cutoffs of 0.5 or below and a co-occurrence excess hovering around zero across the range of cutoff thresholds.

That the HC pair showed the largest number of overlapping accelerations was not entirely surprising, as human and chimpanzee share the most recent speciation event among the three species pairs. Thus, they presumably share the greatest number of environmental and behavioral traits that might have caused a gene to experience increased nonsynonymous substitutions in both

lineages. More interesting was the difference in overlap levels between the GC and GH species pairs. Whereas the GC pair showed little genome-wide evidence for excess parallel acceleration, the GH pair showed a slight but consistent signal for more parallel accelerations than expected by chance. This could be due either to a greater degree of biological or environmental similarity in GH compared to GC, or it could be the result of some underlying bias in the data, such as differences between human and chimpanzee in population size or genome quality.

GO terms enriched in shared accelerated genes

The second approach used to characterize genes with parallel accelerations was to identify GO terms enriched for genes with evidence for parallel acceleration in each pair of AGA species. The methodology used to identify enriched GO terms will be described Section 6.5, and in that section the terms enriched for genes with lineage-specific accelerations will be discussed at more length; the parallel results are described here for continuity with the discussion of parallel accelerations.

For each species pair, genes with $\text{LRT}_{min} > 1.5$ were considered significant in the GO enrichment tests. This more lenient threshold was used since few genes were independently accelerated at the 95% chi-squared threshold of 3.84 in both lineages (25 genes for GC, 40 for GH, and 51 for HC; Table 6.1). At a LRT_{min} threshold of 1.5, the GC pair yielded 206 accelerated genes, GH yielded 238, and HC yielded 286. The sections of Table 6.2 labeled "Human-Chimpanzee Parallel", "Gorilla-Chimpanzee Parallel" and "Gorilla-Human Parallel" show the terms most enriched for parallel accelerations.

Although no species pair yielded GO terms significantly enriched after correction for multiple testing, a number of terms were enriched at a nominal $p < 0.05$ significance using Fisher's Exact Test (FET). The top three terms for HC parallel accelerations were "neuropeptide signaling pathway", "regulation of DNA-dependent transcription" and "microtubule-based movement"; for GC parallel accelerations, "protein autophosphorylation", "inner ear development and positive regulation of transcription factor activity"; and for GH parallel accelerations, "Wnt receptor signaling pathway", "sensory perception of sound" and "skeletal system morphogenesis" showed some enrichment for significant genes.

Ann.	Sig.	Exp.	ID	Definition	FET	topGO	goseq	Length	Top 5 Significant Genes
Human									
31	6	1.59	GO:0006368	RNA elongation from RNA polymerase II pr...	4.3e-03	4.3e-03	2.8e-03	421	ERCC3, POLR2E, GTF2E1, TAF12, TAF11
76	10	3.90	GO:0010552	positive regulation of gene-specific tra...	5.2e-03	5.2e-03	7.3e-03	620	NKX2-1, CELA1, TFEB, LEF1
46	7	2.36	GO:0006367	transcription initiation from RNA polyme...	8.4e-03	8.4e-03	7.3e-03	542	ERCC3, POLR2E, GTF2E1, TAF12, TAF11
112	11	5.75	GO:0009952	anterior/posterior pattern formation	2.9e-02	9.0e-03	4.0e-02	683	APC, LEF1, HOXD4, T, ALDH1A2
50	7	2.57	GO:0007588	excretion	1.3e-02	9.2e-03	1.5e-02	506	SLC22A18, OXT, TRPV1, NR3C2, AQP4
46	7	2.36	GO:0051028	mRNA transport	8.4e-03	9.3e-03	1.4e-02	766	NUP210, SMG5, SEH1L, NUP155, MYO1C
74	9	3.80	GO:0007605	sensory perception of sound	1.3e-02	1.9e-02	3.7e-02	1,091	EYA1, LOXHD1, OTOG, MYO7A, CHRNA10
38	5	1.95	GO:0001707	mesoderm formation	4.3e-02	5.3e-02	5.8e-02	616	EYA2, LEF1, SNAI1, SRF, TWSG1
182	15	9.34	GO:0007017	microtubule-based process	4.7e-02	5.9e-02	1.5e-01	1,057	APC, BRCA1, C14orf153, RFX3, TACC2
50	6	2.57	GO:0048704	embryonic skeletal system morphogenesis	4.2e-02	6.5e-02	5.0e-02	604	EYA1, HOXD4, HOXA3, DSCAML1, TGFBR1
189	20	9.70	GO:0048568	embryonic organ development	1.6e-03	8.7e-02	2.9e-03	652	EYA1, MYO7A, TFEB, LEF1, HOXD4
33	5	1.69	GO:0021983	pituitary gland development	2.5e-02	8.9e-02	2.6e-02	459	NKX2-1, EYA1, ALDH1A2, SALL1
855	56	43.89	GO:0007399	nervous system development	3.2e-02	9.8e-02	1.3e-01	747	NKX2-1, RET, APC, BRCA1
122	11	6.26	GO:0001822	kidney development	4.9e-02	1.1e-01	7.6e-02	700	RET, APC, EYA1, LEF1, EPCAM
82	10	4.21	GO:0050954	sensory perception of mechanical stimulu...	8.9e-03	1.4e-01	2.8e-02	1,069	EYA1, LOXHD1, OTOG, MYO7A, TRPV1
1322	84	67.87	GO:0030154	cell differentiation	1.9e-02	1.8e-01	8.9e-02	684	NKX2-1, RET, APC, POU2F3
68	8	3.49	GO:0035270	endocrine system development	2.2e-02	1.9e-01	2.6e-02	494	NKX2-1, EYA1, RFX3, ALDH1A2
Chimpanzee									
34	5	1.54	GO:0006809	nitric oxide biosynthetic process	1.8e-02	6.3e-03	9.6e-03	468	SLC7A2, DDAH2, EGFR, NOS1, GIMAP5
154	13	6.99	GO:0006260	DNA replication	2.3e-02	1.7e-02	2.7e-02	705	CCDC111, RFC3, TNFAIP1, ORC2L, NFIA
150	12	6.80	GO:0019318	hexose metabolic process	4.0e-02	5.9e-02	3.7e-02	569	B4GALT1, OGDHL, PDX1, MYC, TSTA3
29	5	1.32	GO:0006096	glycolysis	9.0e-03	6.7e-02	1.1e-02	625	OGDHL, GPI, PRKAG1, ALDOC, PRKAG2
1070	60	48.54	GO:0006355	regulation of transcription, DNA-depende...	4.6e-02	1.1e-01	5.7e-02	643	SIM1, PRDM15, TRERF1, FOXN1, DMRT2
73	9	3.31	GO:0006865	amino acid transport	5.5e-03	1.5e-01	8.3e-03	620	SLC1A1, SLC22A4, SLC7A2, SLC3A1, CACNA1A
41	5	1.86	GO:0045088	regulation of innate immune response	3.7e-02	1.7e-01	3.0e-02	511	TBK1, PELI1, GIMAP5, TLR3, TRAFD1
Gorilla									
48	8	2.05	GO:0042472	inner ear morphogenesis	8.5e-04	1.3e-03	1.3e-03	780	SIX4, ITGA8, FGF3, MYO15A, SOX2
40	6	1.70	GO:0001942	hair follicle development	6.5e-03	3.0e-03	7.5e-03	691	KRT27, EDAR, HOXC13, DSG4, SOX9
48	6	2.05	GO:0008584	male gonad development	1.5e-02	1.5e-02	1.3e-02	511	PDGFRA, SOX9, ERCC1, RARA, NR5A1
250	17	10.65	GO:0007420	brain development	3.8e-02	1.8e-02	6.8e-02	809	CNTN4, LMX1B, NKX2-6, SLIT1
74	8	3.15	GO:0007605	sensory perception of sound	1.3e-02	2.1e-02	3.1e-02	1,091	STRC, LOXHD1, OTOF, MYO15A, SOX2

Table 6.2 (*continued on next page*)

Ann.	Sig.	Exp.	ID	Definition	FET	topGO	goseq	Length	Top 5 Significant Genes
52	6	2.22	GO:0006364	rRNA processing	2.2e-02	2.2e-02	2.0e-02	563	ERI1, NOP2, RRP9, EMG1, EXOSC2
44	5	1.87	GO:0032526	response to retinoic acid	3.8e-02	4.5e-02	4.3e-02	698	WNT5B, DKK1, SYNJ1, RARA, SOX2
1579	85	67.28	GO:0045449	regulation of transcription	1.0e-02	6.2e-02	2.0e-02	675	SUPT16H, RBBP4, TBX4, ZNF165, HOXB5
1070	58	45.59	GO:0006355	regulation of transcription, DNA-depende...	3.0e-02	6.5e-02	3.7e-02	643	TBX4, ZNF165, HOXB5, SPEN, IKZF3
105	10	4.47	GO:0048705	skeletal system morphogenesis	1.4e-02	7.1e-02	1.4e-02	623	IMPAD1, PDGFRA, HOXB5, SIX4, SOX9
78	7	3.32	GO:0030308	negative regulation of cell growth	4.8e-02	7.2e-02	4.1e-02	543	RERG, PRDM4, SLIT1, ING1, GNG4
101	9	4.30	GO:0043161	proteasomal ubiquitin-dependent protein ...	2.8e-02	9.5e-02	2.1e-02	492	PSMD2, KCTD10, PLK1, FZR1, CRBN
58	9	2.47	GO:0042471	ear morphogenesis	7.0e-04	1.2e-01	1.0e-03	763	SIX4, ITGA8, FGF3, MYO15A, SOX2
124	10	5.28	GO:0045165	cell fate commitment	3.9e-02	1.9e-01	3.5e-02	581	TGFB1I1, SOX9, DKK1, DLX1, FGF3
164	12	6.99	GO:0046578	regulation of Ras protein signal transdu...	4.7e-02	1.9e-01	8.7e-02	891	FGD2, RABGAP1L, TBC1D12, IQSEC3, SH2B2
50	6	2.13	GO:0009116	nucleoside metabolic process	1.9e-02	2.3e-01	1.3e-02	413	MAT2B, PRPS2, PRTFDC1, DUT, PANK4
Human-Chimpanzee Parallel									
58	6	1.91	GO:0007218	neuropeptide signaling pathway	1.1e-02	1.1e-02	1.8e-02	758	GPR98, CELSR2, GPR125, TAC1, GPR56
1070	48	35.15	GO:0006355	regulation of transcription, DNA-depende...	1.4e-02	2.8e-02	3.7e-02	643	ELF5, FOXN1, NR2E3, SUPT6H, DMRT3
70	6	2.30	GO:0007018	microtubule-based movement	2.7e-02	3.6e-02	1.6e-01	1,347	KIF25, KIF2C, KIF1C, C14orf153, DNAH6
112	9	3.68	GO:0009952	anterior/posterior pattern formation	1.1e-02	4.0e-02	2.1e-02	683	SRF, ARC, HOXA3, CDX1, RIPPLY1
60	5	1.97	GO:0007156	homophilic cell adhesion	4.6e-02	4.6e-02	1.4e-01	1,188	FAT1, CELSR2, CLSTN2, CDH24, CDH23
47	5	1.54	GO:0043624	cellular protein complex disassembly	1.8e-02	4.8e-02	4.7e-02	1,049	MTERFD3, C12orf65, KIF2C, SHROOM2, SPTA1
74	6	2.43	GO:0007605	sensory perception of sound	3.4e-02	6.5e-02	1.2e-01	1,091	LOXHD1, GPR98, CDH23, DIAPH1, MYCBPAP
50	5	1.64	GO:0006413	translational initiation	2.3e-02	1.7e-01	1.8e-02	497	RPS6KB2, LGTN, EIF3B, IMPACT, EIF5B
Gorilla-Chimpanzee Parallel									
69	6	1.63	GO:0046777	protein amino acid autophosphorylation	5.7e-03	5.7e-03	1.5e-02	914	MEX3B, MAP3K13, NEK2, PDGFRA, VRK2
59	6	1.40	GO:0051091	positive regulation of transcription fac...	2.6e-03	1.4e-02	2.5e-03	550	HMGA2, AGER, MAP3K13, PLA2G1B, FZD4
70	7	1.66	GO:0048839	inner ear development	1.3e-03	3.0e-02	3.7e-03	820	FGF3, GPR98, CDH23, ATOH1, PDGFRA
70	5	1.66	GO:0007018	microtubule-based movement	2.5e-02	4.3e-02	1.2e-01	1,347	KIF27, KIF25, KIF20A, C14orf153, DNAH2
67	5	1.59	GO:0009791	post-embryonic development	2.1e-02	4.5e-02	4.1e-02	822	IREB2, CDH23, SCUBE1, INVS, SGPL1
74	5	1.75	GO:0007605	sensory perception of sound	3.0e-02	7.0e-02	9.9e-02	1,091	LOXHD1, OTOF, GPR98, CDH23, FZD4
69	5	1.63	GO:0006633	fatty acid biosynthetic process	2.3e-02	1.9e-01	1.9e-02	465	PRKAG2, PLA2G1B, PRG3, SCD, ACADVL
58	5	1.37	GO:0042471	ear morphogenesis	1.2e-02	2.1e-01	2.1e-02	763	FGF3, CDH23, ATOH1, SALL1, CEP290
Gorilla-Human Parallel									
133	8	3.64	GO:0016055	Wnt receptor signaling pathway	2.9e-02	6.6e-03	4.2e-02	630	SALL1, CELA1, DKK2, BCL9, RYK
74	7	2.02	GO:0007605	sensory perception of sound	4.0e-03	8.9e-03	2.9e-02	1,091	LOXHD1, CDH23, EYA1, GPR98, USH1C
105	8	2.87	GO:0048705	skeletal system morphogenesis	7.9e-03	3.5e-02	1.1e-02	623	ANKRD11, EYA1, RYK, DSCAML1, HOXB7

Table 6.2 (*continued on next page*)

Ann.	Sig.	Exp.	ID	Definition	FET	topGO	goseq	Length	Top 5 Significant Genes
150	9	4.10	GO:0050953	sensory perception of light stimulus	2.2e-02	5.3e-02	5.2e-02	744	CDH23, GPR98, MYO3A, HMCN1, C2orf71
137	8	3.74	GO:0007548	sex differentiation	3.4e-02	5.3e-02	4.3e-02	600	SALL1, GATA6, DMRT3, NR5A1, CDKL2
524	21	14.32	GO:0009790	embryonic development	5.0e-02	5.6e-02	1.6e-01	728	PTK7, ANKRD11, CDH23, SALL1, EYA1
82	6	2.24	GO:0032318	regulation of Ras GTPase activity	2.4e-02	7.9e-02	5.9e-02	826	TBC1D1, AGAP2, C6orf170, AGFG1, SFRP1
148	8	4.05	GO:0007601	visual perception	5.0e-02	1.6e-01	9.7e-02	724	GPR98, MYO3A, HMCN1, C2orf71, TULP1
227	11	6.20	GO:0007389	pattern specification process	4.7e-02	1.7e-01	8.9e-02	685	EYA1, HIPK1, DAND5, FOXF1, DSCAML1
57	5	1.56	GO:0006937	regulation of muscle contraction	1.9e-02	2.4e-01	2.3e-02	604	MYL5, DMPK, MYBPH, MYBPC3, TACR3
AGA Branch									
41	5	1.03	GO:0051591	response to cAMP	3.4e-03	3.4e-03	3.1e-03	594	CCND2, PAX4, CARM1, AQP9, ALDH3A1
70	6	1.75	GO:0007018	microtubule-based movement	8.0e-03	1.3e-02	1.4e-02	1,347	TUBD1, KIF5B, RSHL1, DNAH2, BBS2
140	8	3.51	GO:0007584	response to nutrient	2.4e-02	5.0e-02	2.5e-02	700	SLC27A4, AQP3, ZNF354A, OGT, EP300
68	5	1.70	GO:0010639	negative regulation of organelle organiz...	2.7e-02	6.5e-02	3.4e-02	1,028	TPM1, ESPL1, BCOR, CAPZA3, ACD
61	5	1.53	GO:0045444	fat cell differentiation	1.8e-02	7.2e-02	1.8e-02	660	NOC3L, BBS9, ALDH6A1, MB, BBS2
52	5	1.30	GO:0009108	coenzyme biosynthetic process	9.4e-03	9.0e-02	8.1e-03	457	MTHFD1, PANK1, GCLM, ASPDH, NAPRT1
AGA Clade									
30	5	1.96	GO:0048813	dendrite morphogenesis	4.3e-02	5.9e-03	9.2e-02	965	DCX, ROBO1, CELSR2, DCLK1, CACNA1A
28	6	1.83	GO:0048675	axon extension	8.3e-03	7.9e-03	2.1e-02	935	RYK, ULK1, DCX, SLIT3, DCLK1
74	10	4.84	GO:0007605	sensory perception of sound	2.1e-02	2.7e-02	1.1e-01	1,091	DFNB31, EYA1, TECTA, TMC1, MYO3A
28	5	1.83	GO:0007608	sensory perception of smell	3.3e-02	3.3e-02	2.5e-02	463	OR10G3, OR13G1, OR52H1, OR10G2, CNGA2
69	9	4.51	GO:0046777	protein amino acid autophosphorylation	3.5e-02	3.5e-02	9.1e-02	914	CRKRS, MEX3B, FLT3, MYO3A, EIF2AK3
23	5	1.50	GO:0021953	central nervous system neuron differenti...	1.5e-02	3.7e-02	2.0e-02	707	SMO, DCX, DCLK1, EPHB1, CACNA1A
48	8	3.14	GO:0042472	inner ear morphogenesis	1.2e-02	4.3e-02	2.3e-02	780	EYA1, ATOH1, FOXI1, FGF3, CEP290
40	8	2.61	GO:0021510	spinal cord development	3.8e-03	5.3e-02	5.4e-03	676	LHX3, SMO, LBX1, RFX4, SLIT3
250	24	16.34	GO:0007420	brain development	3.7e-02	1.0e-01	1.2e-01	809	CNTN4, EYA1, ATOH1, ECE2, NFIB

Table 6.2: GO terms enriched for lineage-specific or parallel accelerated genes. All terms with FET $p < 0.05$ are shown sorted by their topGO p-value. Non-significant p-values from the topGO [Alexa *et al.*, 2006] and goseq [Young *et al.*, 2010] methods, which account for the GO ontology structure and gene length bias, respectively, are shown in gray. The five genes within each category with the largest branch-LRT values are included for illustrative purposes. For full details of methodology, see Section 6.5. Ann.—the number of genes annotated with a given term; Sig.—the number of significant genes annotated with the term; Exp.—the number of significant genes expected given independent association between significant genes and GO terms.

Interestingly, the term "sensory perception of sound" was enriched at $p < 0.05$ in all three species pairs, but only three genes (*LOXHD1*, *CDH23* and *GPR98*) were significant at $\text{LRT}_{min} > 1.5$ in all three pairs; all other significant genes were unique to each species pair. The sound perception genes which were uniquely significant in the GH pair were *EYA1*, *USH1C*, *MYO3A* and *SLC1AC*; for the GC pair, *OTOF* and *FZD4*; and for the HC pair, *DIAPH1*, *MYCBPAP* and *DFNB31*. This suggested that the tendency of genes involved in sound perception to experience mildly to moderately elevated $d\text{N}/d\text{S}$ levels was relatively widespread in all three AGA genomes, with variation in the specific genes having undergone acceleration in each species or species pair. It is also worth noting that the enrichment for this term among parallel accelerated genes was strongest in the GH species pair, where the enrichment was significant at $p < 0.05$ in both the `topGO` and `goseq` tests. The term had $p > 0.05$ for those tests in the GC and HC pairs, indicating that human and gorilla share a slightly stronger signal for parallel accelerated evolution in hearing-related genes than do the other pairs of AGA species.

Figure 6.4 shows the patterns of nonsynonymous and synonymous substitutions throughout the 6-way primate phylogeny for two hearing-related genes, *GPR98* and *LOXHD1*, with strong signals of parallel elevated $d\text{N}/d\text{S}$. The much longer *GPR98* has accrued more substitutions in each branch of the primate tree than *LOXHD1*, but both genes showed large numbers of nonsynonymous substitutions in each of the AGA species relative to their short overall branch lengths. Within each branch of the phylogeny, substitutions are sorted first by their nonsynonymous or synonymous classification and then by their alignment position, with synonymous substitutions to the left of nonsynonymous substitutions. The relative number of nonsynonymous and synonymous substitutions in each branch is related to the estimated $d\text{N}/d\text{S}$ along that branch of the tree.

Top parallel accelerated genes

The third method of analysis was a survey of the top parallel accelerated genes for each species pair and for all three species using the same LRT_{min} statistic described above. Table B.1 (located in Appendix B at the end of this thesis) shows the top 10 accelerated and positively-selected genes for a variety of AGA lineages and tests; the top parallel accelerations according to the LRT_{min} statistic can be found towards the bottom of Table B.1.

Among the top genes accelerated across all three lineages are *LOXHD1*, a gene comprised of PLAT domains which was recently shown to be necessary for auditory hair cell function [Edvardson *et al.*, 2011]; *ITIH3*, a plasma serine protease inhibitor potentially involved in prevention of tumor metastasis and associated with risk of myocardial infarction [Ebana *et al.*, 2007]; and *PARP3*, a member of the ADP ribosyl transferase family which has recently been characterized as playing a role in telomeric stability, response to DNA damage, and neural crest development

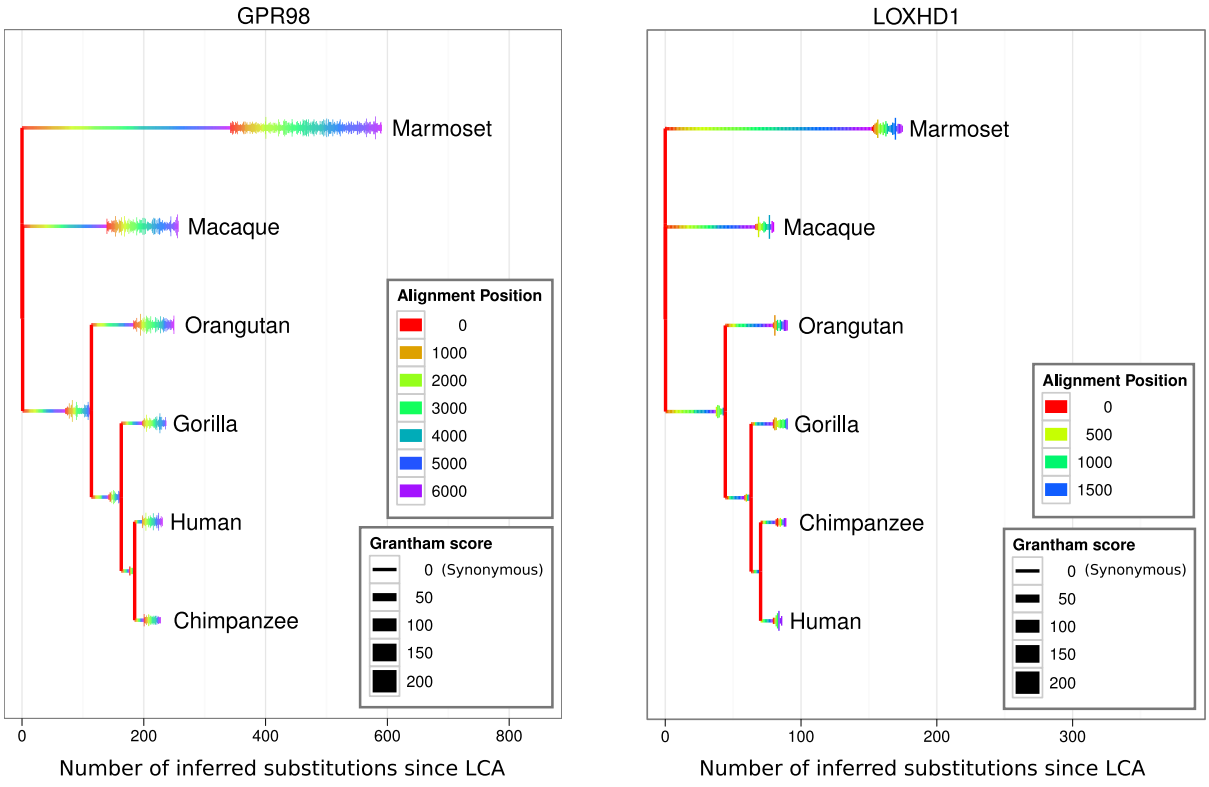


Figure 6.4: Inferred substitution events in the 6-way primate alignments of *GPR98* and *LOXHD1*, two hearing-related genes with evidence for independent elevated $d\mathrm{N}/d\mathrm{S}$ in AGA species. PAML was used to infer ML ancestral sequences [Yang *et al.*, 1995] from the 6-way alignment of each gene and substitution events were assigned to branches in the phylogeny. Only substitutions between nodes where the ancestral reconstruction had a posterior probability of > 0.9 are shown. Substitutions along the branch connecting marmoset to the remaining primates were arbitrarily assigned to the marmoset terminal branch in the rooted view of the phylogeny shown here. Substitutions are drawn as rectangles, arranged in a horizontal line for each branch in the tree. Within each branch, all synonymous substitutions are drawn to the left of all nonsynonymous substitutions; substitutions are then sorted by their position in the alignment. Each substitution is colored according to its alignment position and scaled vertically according to the Grantham score [Grantham, 1974] between the ancestral and derived amino acid.

[Rouleau *et al.*, 2011; Boehler *et al.*, 2011]. The molecular and medical evidence for the functional activity of these genes suggests that the elevated $d\mathrm{N}/d\mathrm{S}$ levels across all three African great apes are not well explained by a substantial loss of functional constraint; other possible causes might be relaxed evolutionary constraint due to a decreased N_e relative to the primate background, positive selection due to functional adaptation, or some combination of the two.

6.5 Gene Ontology (GO) term enrichments

Gene ontology (GO) term annotations for the "biological process" ontology tree were downloaded from release 60 of the Ensembl human database [Flicek *et al.*, 2011] and assigned to the alignment corresponding to each human gene. Three complementary methods were used to assign p-values for GO term enrichment among the most accelerated genes for each branch-LRT performed and genes with evidence of parallel acceleration in a pair of species. For lineage-specific accelerations, the 95% χ^2_1 cutoff value of the branch-LRT was used to identify accelerated genes. Parallel accelerations for each species pair were identified by genes with a LRT_{min} of 1.5 in both species of interest (more detail on the analysis of parallel accelerations is included in Section 6.4).

In general, the methods used to detect GO terms enriched for accelerated genes were similar to those used in Chapter 5 for detecting GO terms enriched in PSGs. The first test was a standard one-tailed FET applied to the 2x2 contingency table of significant / non-significant genes which were annotated / not-annotated with a given GO term. The second method, implemented in the `topGO` program [Alexa *et al.*, 2006], is also based on the FET statistic but additionally compensates for the structure of the GO hierarchy by iterating through the directed acyclic graph and removing nodes from consideration when certain descendant nodes have already shown significant enrichment (see Alexa *et al.* [2006] for complete details of the algorithm). The main effect of the `topGO` algorithm is to identify and remove semantically repetitive terms (e.g., terms that are nearby in the GO ontology and are annotated with similar sets of genes) from the set of most significantly enriched results by reducing the p-values of terms with more highly-enriched neighboring terms. The third method, implemented in the `goseq` program [Young *et al.*, 2010], accounts for a potential gene length bias in the propensity for a gene to yield a significant LRT results. As the sequence length can have a strong impact on the significance of LRT results [Anisimova *et al.*, 2001] and some GO terms tend to contain longer genes [Young *et al.*, 2010], I found it important to correct for this when identifying enriched terms. The `goseq` program first uses the set of gene-wise p-values and gene lengths to fit a smoothed probability weighting function (PWF) which predicts the expected proportion of significant accelerations given a gene's length. This PWF is then used to adjust the identification of significantly-enriched GO terms to correct for potential over-representation of terms with significantly longer or shorter mean gene lengths. Although the `goseq` program was designed primarily for the functional analysis of RNA-seq data (where gene length bias is a widely-acknowledged confounding factor) I found it to be effective in identifying potentially misleading GO enrichment results for the current analysis.

The results of the GO enrichment analysis are summarized in Table 6.2; the results for parallel accelerations were already discussed in Section 6.4. For each branch model (or pair of species for parallel accelerations), all terms with a FET over-representation $p < 0.05$ and five or more significant genes are shown, sorted by their `topGO` p-value. Any `topGO` or `goseq` p-values above

$p = 0.05$ are colored gray instead of black. Terms with non-significant `topGO` p-values are likely to have a closely-related term with stronger enrichment higher in the list, while terms with non-significant `goseq` p-values should be treated with caution due to a detected length bias in the detection of accelerated genes.

None of the GO term enrichments for any of the branch-LRTs shown in Table 6.2 remained significant at FDR< 0.1 after applying the Benjamini and Hochberg [1995] correction (results not shown). The lack of significance after correcting for multiple tests could be taken as evidence that no strong associations between accelerated genes and GO terms existed in these results. However, it may also be due to a variety of other factors, including limited power of branch models to detect dN/dS shifts, noise in the GO annotation of genes, or the specific choice of LRT cutoffs used to identify significant accelerations. Other studies have avoided the use of cutoff thresholds by using different tests such as the Mann-Whitney U test to identify terms with a significantly lower distribution of p-values than expected [Clark *et al.*, 2003; Kosiol *et al.*, 2008], but this was not done here. Despite the lack of strongly-controlled statistical significance, the use of a nominal $p < 0.05$ cutoff to identify enriched GO terms for display in Table 6.2 yielded a limited set of enriched terms for each branch-LRT that summarized the strongest functional associations with moderately to strongly accelerated genes.

In general, the GO terms enriched for accelerations along the terminal AGA branches were very different from the largely immune-related functions found in Chapter 5 for genes subject to positive selection throughout mammals. Terms such as "anterior/posterior pattern formation" (human FET p =2.9e-02), "mRNA transport" (human FET p =8.4e-03), "DNA replication" (chimpanzee FET p =2.3e-02) and "inner ear morphogenesis" (gorilla FET p =8.5e-04) indicated that the branch-LRTs were sensitive to dN/dS changes in genes with core metabolic and developmental functions, and the presence of several sensory and brain-related terms enriched in the AGA clade accelerated genes ("dendrite morphogenesis", "axon extension", "sensory perception of sound", "brain development") provided some evidence that the AGA clade as a whole has experienced increased levels of nonsynonymous substitutions in genes with roles in neural functioning and development. A likely explanation for the stark contrast with the largely immune-related terms from Chatper 5 was that the strength of the branch-LRT statistic for a given foreground branch was likely to be strongest for genes with a low background dN/dS. For example: a gene which has experienced positive selection throughout primates would have a high dN/dS ratio throughout the tree, while a gene which has only experienced positive selection along one branch could have a very low dN/dS in the rest of the tree. The branch-LRT measures the significance of evidence for a *change* in dN/dS along the foreground branch, so genes with a large contrast between the true foreground and background dN/dS ratios should yield the strongest branch-LRT results. For this reason, the genes producing significant results for tree-wide versus branch-specific tests were expected to be quite different.

6.6 Comparison with previous genome-wide scans for accelerated or positively-selected genes

A number of studies have previously investigated the prevalence and functional associations of PSGs and genes with elevated dN/dS in primate genomes, often using branch-LRTs or branch-site LRTs similar to those used here. Although these studies have varied widely in the exact datasets and analytical methods employed, a qualitative and quantitative comparison of their main results helps us appreciate the variability of previously published genome-wide results in primates. Table 6.3 presents a summary of results from the current analysis plus 8 previous genome-wide scans for accelerated or positively-selected genes. This table was originally compiled by Stephen Montgomery for the Gorilla Consortium, but it has been heavily modified and condensed into the current form.

The proportion of accelerated genes detected using branch-LRT methods (or close equivalents) ranged from 7.07% to 20.24%; results from the current study, which ranged from 4.64% in chimpanzee to 5.75% in human, were only slightly lower than the typical range of previously-published values. For the proportion of genes experiencing positive selection under the branch-site LRT (or similar), the current results results (ranging from 1.23% in human to 1.66% in gorilla) again fell towards the lower end of the the published range of 0.43% to 8.72%. Most published studies did not show a large difference in the proportion of accelerated or positively-selected genes between chimpanzees and humans; our results further confirm this trend and extend to gorilla the observed consistency in numbers of lineage-specific accelerated and positively-selected genes between different AGA species. It should be emphasized that this consistency only pertains to the number of accelerated or positively-selected genes along different AGA lineages within a single study; given the surprisingly low amount of overlap between PSGs identified from different studies found in Chapter 5, it was considered unlikely that the specific genes identified as accelerated in any of the studies surveyed here shared much overlap with the results from other studies.

Table 6.3 highlights the wide range of biological functions and processes that have commonly been found enriched for genes subject to accelerated evolution or positive selection. Terms involving immune functions, olfaction, and amino acid metabolism have most commonly been identified. The GO term enrichments based on the current branch-LRT results did not recover many terms in common previous studies. This may be the result of a different type of sensitivity in the specific branch-LRTs used here, where gene accelerations in AGA lineages *relative to the primate background rate* were detected, as opposed to high rates of evolution on their own. For example, immune genes with high dN/dS ratios across all primates were not likely to be identified as accelerated in this study, since immune genes tend to be subject to positive selection throughout the mammalian phylogeny as opposed to one particular lineage or another (see Chapter 5). This

could explain the lack of any apparent immune enrichment in the current results.

Interestingly, the GO term "sensory perception of sound" was among the top enriched terms in both the gorilla lineage and gorilla-human parallel accelerated genes. Although previous studies have detected enrichment for olfaction and visual perception [Clark *et al.*, 2003; Nielsen *et al.*, 2005; Gibbs *et al.*, 2007], this appears to be the first genome-wide analysis to provide strong evidence for an abundance of genes involving sound perception to have been subject to elevated $d\mathrm{N}/d\mathrm{S}$ ratios in the AGA. Some evidence was also found for enrichment of brain-related terms in human and gorilla, including "brain development" (gorilla $p = 0.038$, Table 6.2) and "nervous system development" (human $p = 0.032$, Table 6.2), although both terms may be subject to a gene length bias and both fail to reach $p < 0.05$ using the `goseq` method for detecting enriched terms.

Study	FG Species	Gene Count	Test	Ontology	Accel.	PSGs	Top 5 enriched terms
Clark *et al.* [2003]	H	7,645	B/BS	PAN	20.24%	8.72%	Olfaction Sensory perception Cell surface receptor-med. signal transd. Chemosensory perception Nuclear transport
Ibid.	C	7,645	B/BS	PAN	20.07%	-	Signal transduction Amino acid metabolism Amino acid transport Cell proliferation and differentiation Cell structure
Mikkelsen *et al.* [2005]	H/C	7,043	B	GEN	8.31%	-	Sensory perception of chemical stimulus Perception of smell Xenobiotic metabolism Complement activation Regulation of cytokine biosynthesis
Nielsen *et al.* [2005]	H/C	8,079	B ($\omega > 1$)	PAN	-	0.43%	Immunity and defense T-cell mediated immunity Chemosensory perception Unclassified Olfaction
Gibbs *et al.* [2007]	H/C	10,376	BS	PAN	-	0.65%	Physiological response to stimulus Immune response Immune system process Physiological response to wounding Sensory perception of chemical stimulus Olfactory receptor activity
Ibid.	M	10,376	BS	PAN	-	1.26%	Extracellular region Physiological response to wounding Physiological response to stimulus Response to wounding Defense response
Bakewell *et al.* [2007]	H	13,888	BS	PAN	-	1.11%	Anion/Ion transport Phosphate transport Ectoderm development Fatty acid metabolism G-protein mediated signaling
Ibid.	C	13,888	BS	PAN	-	0.96%	Protein metabolism and modififcation mRNA transcription Proteolysis mRNA transcription regulation Stress response

Table 6.3 (*continued on next page*)

Study	FG Species	Gene Count	Test	Ontology	Accel.	PSGs	Top 5 enriched terms
Kosiol *et al.* [2008]	H/C/M	12,823	BS	PAN	-	4.24%	Olfactory receptor activity Sensory perception of chemical stimulus Sensory perception of smell G-protein coupled receptor activity Sensory perception
Uddin *et al.* [2008]	H/C/M	23,945	FR	DAVID	7.07%	-	Glycoprotein Olfaction Transmembrane Mitochondrion Oxidative phosphorylation Cytokine activity Immune response
Locke *et al.* [2011]	H/C/M	13,882	BS	GO/PAN	-	-	Immunity and defense Visual perception Glycolipid metabolic processes
This Study	H	11,534	B	GO	5.75%	1.23%	RNA elongation from PolII Positive regulation of gene-specific transcr. Transcription initiation from RNA polymerase Anterior/posterior pattern formation Excretion
Ibid.	G	11,534	B	GO	4.87%	1.66%	Inner ear morphogen. Hair follicle development Male gonad development Brain development Sensory perception of sound
Ibid.	C	11,534	B	GO	4.64%	1.59%	Nitric oxide biosynthesis DNA replication Hexose metabolic process Glycolysis Regulation of DNA-dependent transcription
Ibid.	AGA Branch	11,534	B	GO	2.59%	1.32%	Response to cAMP Microtubule-based movement Response to nutrient Negative regulation of organelle organization Fat cell differentiation
Ibid.	H-C	11,534	B	GO	2.48%	-	Neuropeptide signaling pathway Regulation of DNA-dependent transcription Microtubule-based movement Anterior/posterior pattern formation Homophilic cell adhesion

Table 6.3 (*continued on next page*)

Study	FG Species	Gene Count	Test	Ontology	Accel.	PSGs	Top 5 enriched terms
Ibid.	G-C	11,534	B	GO	1.79%	-	Protein amino acid autophosphorylation Positive regulation of TF activity Inner ear development Microtubule-based movement Post-embryonic development
Ibid.	G-H	11,534	B	GO	2.06%	-	Wnt receptor signaling pathway Sensory perception of sound Skeletal system morphogenesis Sensory perception of light stimulus Sex differentiation

Table 6.3: A comparison of genome-wide studies of primate gene evolution. Nine studies of positive selection or accelerated dN/dS in primates, including the current study, are summarized by various factors including the number of genes analyzed, the tests of selection performed, the accelerated or positively-selected gene count, and the most strongly enriched functional terms. The current results are largely consistent with previous results in the percentage of accelerated and positively-selected genes identified. In contrast, the functional terms detected as enriched for accelerated genes varied widely, both within the set of previously-published studies and with the current results. FG—the foreground species tested for dN/dS acceleration or positive selection; Accel.—the percentage of genes identified as showing significant evidence for acceleration; B/BS/FR—the branch test for accelerated dN/dS, the branch-site test for positive selection, and the free-ratios model (i.e., one dN/dS ratio per branch) for lineage-specific dN/dS estimation; PAN/GO/DAVID—the Panther gene classification [Thomas *et al.*, 2003], Gene Ontology [Ashburner *et al.*, 2000], and DAVID [Huang *et al.*, 2008] gene classifications; H/C/G/M—human, chimpanzee, gorilla, macaque.

6.7 Genome-wide dN/dS ratios in the African great ape phylogeny

The gorilla genome also provided an opportunity to examine global trends in the evolutionary dynamics of the AGAs and their ancestral populations. I used the genome-wide set of coding alignments to examine lineage-specific dN/dS estimates across the six-primate phylogenetic tree.

Two broad categories of methods have commonly been employed to estimate ancestral N_e from comparative genomics data: methods based on estimating the variance of species divergence times or tree topologies in samples of coding or noncoding DNA, and methods based on estimating lineage-specific dN/dS ratios in protein-coding genes.

Over the past two decades a number of groups have applied methods based on the first approach (using variance in divergence times / tree topologies) to the estimation of primate ancestral populations (Takahata and Satta [1997]; Chen and Li [2001]; Hobolth *et al.* [2007]; Burgess and Yang [2008]; reviewed by Siepel [2009]). Despite significant differences in the inference methods and sizes of datasets used, most analyses were in agreement that the N_e of ancestral hominoids was considerably larger than that found in most present-day populations. However, no consensus appears to have emerged regarding the absolute N_e values of primate ancestral populations, and the precision of estimates has been lacking [Siepel, 2009].

The second general approach is based on comparing lineage-specific dN/dS values estimated from a large amount of aligned protein-coding sequence. Theory predicts that larger populations should exhibit, on average, lower dN/dS values due to the increased efficacy of purifying selection [Ellegren, 2008]. Although under certain assumptions a direct mathematical relationship between dN/dS and N_e can be derived [Nielsen and Yang, 2003; Kryazhimskiy and Plotkin, 2008], the assumptions involved are relatively unrealistic and few empirical studies have attempted to directly estimate the absolute N_e of a species or clade directly from divergence data. Instead, studies have tended to estimate lineage-specific mean dN/dS ratios across many aligned genes, using the relative estimates of dN/dS as a proxy for relative N_e values. Using this approach, results from several genome-wide analyses have consistently confirmed the correlation between N_e and mean dN/dS [Lindblad-Toh *et al.*, 2005; Gibbs *et al.*, 2007; Mikkelsen *et al.*, 2005; Warren *et al.*, 2008; Kosiol *et al.*, 2008]. All of these studies were based on alignments of orthologous genes in at least human and mouse (in addition to other species), and in every case human had a higher mean dN/dS value than mouse. This was consistent with the theory, given the expectation based on ecological studies of a much large N_e in mouse compared to human [Ellegren, 2008]. Furthermore, Ellegren [2008] identified a strong negative correlation between the lineage-specific dN/dS estimates from Kosiol *et al.* [2008] and the log-N_e as estimated from polymorphism data. Within primates, all of the above studies which included macaque showed it to have a lower dN/dS than

human and chimpanzee. There does exist some disagreement regarding the relative dN/dS of human and chimpanzee, however: Gibbs *et al.* [2007] found a mean dN/dS of 0.175 for chimpanzee and 0.169 for humans, while Kosiol *et al.* [2008] found a mean dN/dS of 0.245 for chimpanzee and 0.249 for humans. While the two values are very similar in both cases, this discrepancy indicates some remaining uncertainty regarding the N_e of humans and chimpanzees since their divergence 4–6 Myr ago. Furthermore, the association between sequencing and assembly errors and inflated estimates of positive selection [Schneider *et al.*, 2009] suggests that lower-quality genomes may be prone to increased dN/dS estimates as a result of such errors, especially in closely-related primate genomes.

I used PAML to estimate genome-wide dN/dS levels for each branch in the six-species primate phylogeny from the 11,534 aligned 1-to-1 orthologous genes assembled here. Two alignments were created: an unfiltered alignment, created by concatenating the alignment of each gene after sequences were filtered for sequence quality but before the window-based filter for clustered nonsynonymous substitutions and the filter for ILS-patterned substitutions were applied; and a filtered alignment, created by concatenating the final alignment of each gene that was used for the branch-LRT analysis. Both alignments were 7.263 million codons in length. Before being input to PAML, all columns containing a gap character or 'N' in any species were removed (this was done to avoid a human-specific bias, as the human-flattened alignments contained some gaps in non-human species but never in human). As expected, more columns were removed from the filtered alignment due to additional 'N's from the window-based masking procedure: the final unfiltered alignment contained 5.91 million codons and the final filtered alignment contained 5.89 million codons. A single dN/dS ratio was first estimated for each alignment using the M0 codon model, yielding 0.220 for the unfiltered and 0.218 for the filtered alignment. Separate dN/dS values for each branch were then estimated from each alignment using the "free-ratios" model implemented in PAML (parameter `model=1`). Because PAML estimates parameters based on an unrooted tree using reversible models of evolution, any estimates of dN/dS on the outermost branch (i.e., the branch connecting the H/C/G/O/M ancestor to marmoset) were ambiguous as to whether they occurred on the branch leading to marmoset or the branch leading to the other primates; thus, marmoset was considered an outgroup in this analysis.

The resulting genome-wide estimates of dS and dN/dS for each branch are given in Table 6.4 and plotted in Figure 6.5. I found that human had a slightly but significantly higher overall dN/dS than both chimpanzee and gorilla (dN/dS =0.256, 0.249, 0.239, respectively) and that orangutan, macaque, and the ancestral lineages all had lower overall dN/dS values than the terminal AGA branches, ranging from 0.195 for macaque to 0.211 for the human-chimpanzee ancestor. These results were in close agreement with equivalent estimates from Kosiol *et al.* [2008], who found overall dN/dS values of 0.249, 0.245, and 0.191 for human, chimpanzee, and macaque, respectively.

A comparison of the global dN/dS values from the unfiltered versus the filtered alignments in

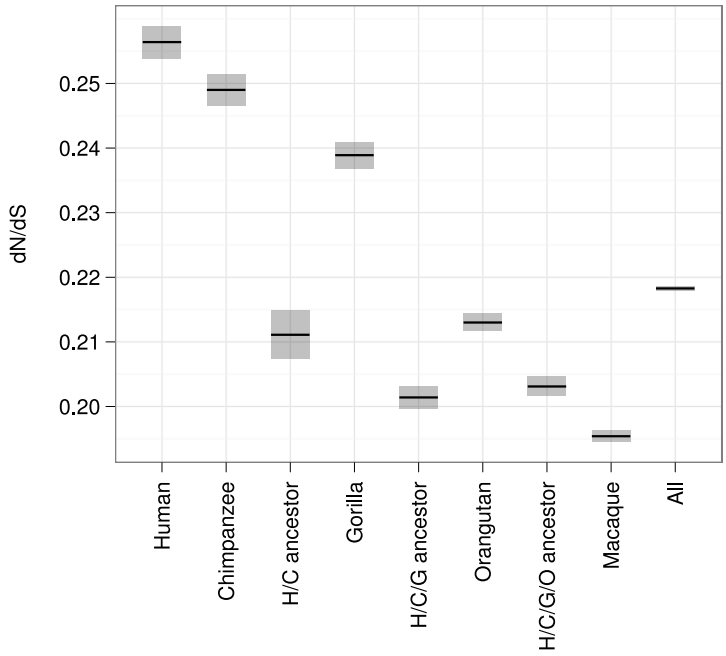


Figure 6.5: Genome-wide dN/dS values in 5 primate species and their ancestral lineages. dN/dS ratios were estimated from concatenated alignments of 1-to-1 orthologs from six primate species using the free-ratios model in PAML. The most distant aligned species, marmoset, was used as an outgroup. Each estimated dN/dS ratio is plotted as a horizontal line surrounded by a gray box corresponding to the standard error estimated by PAML. Ancestral lineages are labeled with the first characters of their descendant species.

Table 6.4 revealed that the unfiltered alignments generally yielded slightly higher dN/dS values, but the magnitude of change varied noticeably between branches. The human terminal branch and all of the ancestral branches showed less than a 1% decrease in dN/dS as a result of the alignment filtering, while chimpanzee, gorilla, orangutan, and macaque all showed greater than 1% decrease in dN/dS. Since the window-based masking procedure was only applied to clusters of substitutions inf the terminal branches, the lack of change in dN/dS ratios along the ancestral branches was expected. The smaller magnitude of change in human (0.39%) compared to the other extant primate genomes (ranging from 1.03% to 4.18%) indicated that the filtering procedure resulted in far more nonsynonymous substitutions being removed from non-human sequences than from human. Interestingly, the magnitude of the dN/dS change for non-human terminal branches appeared to correlate negatively with the dS values of those branches. Gorilla and chimpanzee, with dS = 0.0077 and dS = 0.0055 respectively, both had a ∼4% shift; orangutan, with dS =0.016, had a ∼2% shift; and macaque, with dS = 0.032, had a ∼1% shift. These observations

	Filtered			Unfiltered		% dN/dS change
Branch	dS	dN/dS	S.E.	dN/dS	S.E.	(unfiltered vs. filtered)
Human	0.0057	0.2564	2.5e-03	0.2567	2.5e-03	0.39
Chimpanzee	0.0055	0.2490	2.5e-03	0.2588	2.4e-03	4.02
H/C ancestor	0.0019	0.2111	3.7e-03	0.2117	3.1e-03	0.47
Gorilla	0.0077	0.2389	2.0e-03	0.2489	2.1e-03	4.18
H/C/G ancestor	0.0087	0.2014	1.8e-03	0.2015	1.8e-03	0.50
Orangutan	0.016	0.2130	1.3e-03	0.2184	1.3e-03	2.35
H/C/G/O ancestor	0.0131	0.2031	1.5e-03	0.2034	1.5e-03	0.00
Macaque	0.0324	0.1954	8.9e-04	0.1967	8.9e-04	1.03
All		0.2183	3.8e-04	0.2200	3.8e-04	0.92

Table 6.4: Genome-wide dS and dN/dS values in five primate species and their ancestral lineages using the "Filtered" genome-wide alignments. dN/dS ratios were estimated from concatenated alignments of 1-to-1 orthologs from six primate species using the free-ratios model in PAML. The unfiltered alignments were collected before clustered nonsynonymous substitutions and sites with ILS patterns were removed. The most distant aligned species, marmoset, was used as an outgroup. Ancestral lineages are labeled with the first characters of their descendant species. The M0 model was used to estimate values for the row labeled "All". S.E.—standard error of the dN/dS ratio estimate calculated by PAML.

were consistent with the trend expected if each genome contained a similar number of erroneous or misaligned bases, as the larger number of true nonsynonymous and synonymous substitutions species with longer terminal branch lengths would tend to "dilute out" the signal of elevated dN/dS resulting from misaligned bases. In sum, the comparison between filtered and unfiltered dN/dS levels provided further validation of the use of the window-based substitution filter. The application of the filter hardly affected the estimated dN/dS ratio of the finished-quality human genome, but it resulted in branch-length dependent decreases in dN/dS in the lower-quality nonhuman primate genome assemblies. Notably, chimpanzee showed a marginally higher dN/dS than human in the unfiltered alignment and a significantly lower dN/dS than human in the filtered alignment.

An additional consideration in the interpretation of global dN/dS values was that genes are evolutionarily heterogeneous entities, with each gene composed of sites evolving under different amounts of purifying and positive selection due to varying functional and biological constraints [Whelan, 2008]. The genome-wide dN/dS estimates in Table 6.4 were obtained using the free-ratios codon model, with one dN/dS ratio per branch but a constant dN/dS across all alignment sites. Each dN/dS ratio estimated under this model could thus be considered an "average" value, resulting from the combination of sites under strongly purifying, slightly purifying, neutral and positive selection into a single alignment.

This averaging across genes and sites causes two problems for comparative studies: first, the comparison of genome-wide dN/dS values from different studies is difficult, as the specific genes chosen for analysis can have a significant impact on the overall results [Ellegren, 2008]. Second,

the inclusion of sites with very different selective pressures might decrease the resolution with which lineage-specific differences in dN/dS (and thus N_e) could be detected. This is because some fraction of protein-coding sites may evolve neutrally or under positive selection. Neutrally-evolving sites in genes should show no relationship with N_e, and sites under constant positive selection should show the opposite effect, with a positive correlation between dN/dS and N_e due to the increased efficacy of natural selection. Although the proportion of such sites is likely small, they may obscure the connection between dN/dS and N_e.

One way to better understand the effect of heterogeneously evolving sites on genome-wide dN/dS analyses was to separate out sites with different selective pressures and analyze each group independently. I took this approach by separating protein-coding sites into bins based on their estimated sitewise selective pressure across mammals and separately analyzing the set of sites within each bin. Sitewise selection pressures were estimated for all 11,534 genes by applying Sitewise Likelihood Ratio (SLR) [Massingham and Goldman, 2005] to coding alignments of all Eutherian orthologs downloaded from the Ensembl Compara database. (Note that for this analysis, the protein-based MCoffee alignments calculated by the Ensembl pipeline were used directly, as the filtering methods performed on the mammalian alignments in Chapter 4 were not yet developed. As these alignments were only being used to sort sites into large bins based on the evidence for purifying or positive selection, misalignment was not expected to significantly bias the results.) The LRT_{SLR} statistic calculated by SLR was used to sort all sites by their strength of evidence for non-neutral selection. Sites were then split into five bins corresponding to the following cumulative percentiles of the LRT statistic: 0–0.05, 0.05–0.33, 0.33–0.67, 0.67–0.98, and 0.98–1.0. These ranges were chosen so that three bins of roughly equal size covered the bulk of sites, while two bins focused on the 5% of sites with the strongest evidence for purifying selection and the 2% of sites with strongest evidence for positive selection. All sites from each bin were concatenated into one alignment and analyzed with the PAML free-ratios model as described above.

The lineage-specific dN/dS ratios estimated from alignments binned by sitewise mammalian selective pressure are shown in Figure 6.6. Results from each bin are spread across the x-axis, and the dN/dS ratio for each branch is plotted on the y-axis as a fraction relative to the human dN/dS in that bin.

The four lowest dN/dS bins showed the same general trends in dN/dS levels as the combined analysis: with human, chimpanzee, and gorilla yielded the highest dN/dS values, followed by orangutan, and finally a cluster of macaque and the ancestral lineages contained the lowest dN/dS ratios. The H/C/G/O ancestral lineage appeared to have evolved with a slightly lower dN/dS than the other ancestral lineages, though this difference was only apparent in the second and third bins.

Moving from bins with lower human dN/dS to bins with higher human dN/dS, there was a

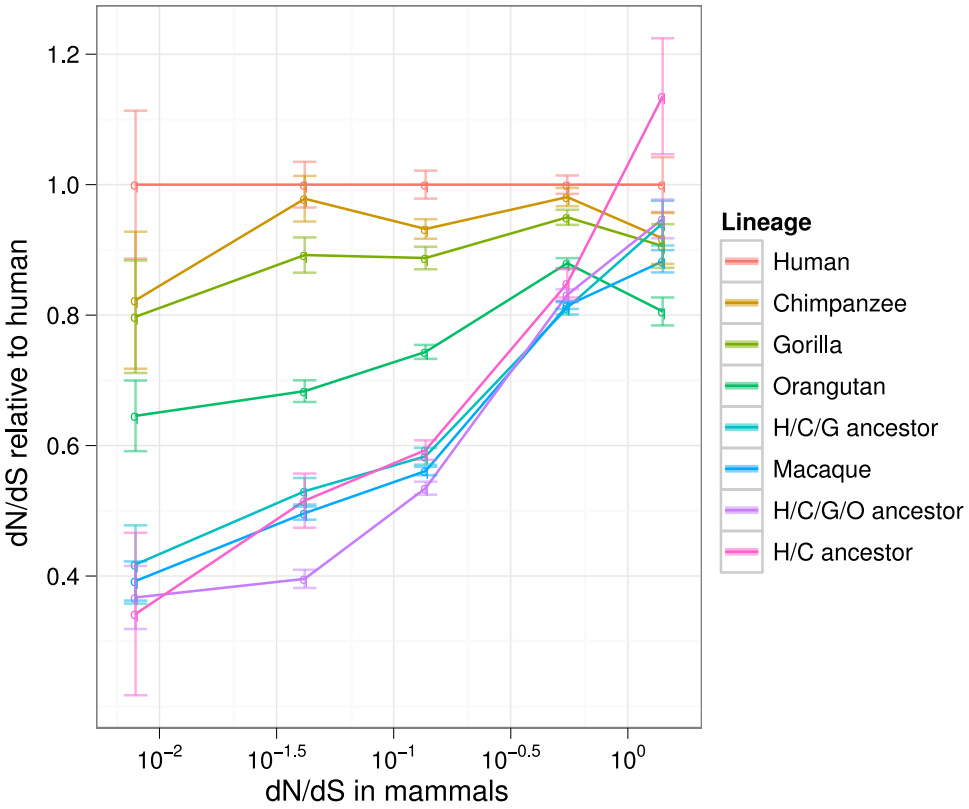


Figure 6.6: Genome-wide branch-specific dN/dS ratios in 6-way primate alignments binned by sitewise selection pressure. Alignment sites were sorted by the sitewise selection pressure estimated by SLR in mammals, assigned to one of five bins (more details in text), concatenated within each bin, and analyzed with PAML using the free-ratios codon model. Estimated dN/dS ratios are plotted with the human dN/dS for each bin on the x-axis and the dN/dS for each branch (expressed as a fraction relative to the human dN/dS in that bin) on the y-axis. Lines are drawn connecting estimates from the same branch in different bins, and standard errors reported by PAML are shown with error bars. Note the log scale on the x-axis.

distinct trend across all lineages towards increased dN/dS values relative to human. In the second-lowest bin, where human had a dN/dS ratio of ∼0.05, the dN/dS difference between lineages was much stronger than in the second-highest bin, where human had a dN/dS ratio of ∼0.55. This trend was consistent with the expected decrease in the differential effects of N_e in sites subject to weaker purifying (e.g., more nearly neutral) selection.

The pattern of genome-wide dN/dS levels in the highest bin—representing the top 2% of alignment sites ordered by the LRT statistic, and thus the 2% of sites with the greatest site-specific evidence for positive selection across Eutherian mammals—was distinct from the other four bins and warrants further mention. The human dN/dS in this bin was ∼1.4, confirming that this subset of alignment sites did indeed contain a number of sites subject to positive selection. Most of the terminal and ancestral branches showed a continuation within this bin of the general

trend of increasingly similar dN/dS estimates between lineages, with values for most branches clustered to around ∼88%–95% of the human dN/dS. However, the HC ancestral lineage showed strikingly increased dN/dS values in the highest bin, and the orangutan branch showed strikingly decreased values. Whereas the HC branch was at ∼85% of the human value in the 4th bin, its value was ∼110% that of human in the highest bin. In contrast, orangutan went from ∼90% in the 4th bin to ∼80% in the highest bin. Such a strong deviation from the consistent trends observed in the other lineages and other bins suggested that some effect other than a difference in N_e may have caused an increase or decrease in the prevalence of nonsynonymous substitutions at sites with evidence for positive selection across Eutherian mammals.

6.8 Conclusions

The gorilla genome sequence provided an opportunity to investigate patterns of recent molecular evolution through comparison to the genomes of other AGAs and more distantly-related primates. Relative to human, gorilla was neither the closest [Mikkelsen *et al.*, 2005] nor the most distantly-related primate to have its genome sequenced [Gibbs *et al.*, 2007; Locke *et al.*, 2011], limiting the expectation that its sequence might help identify entirely novel human-specific or primate-specific evolutionary trends. Instead, the aim of the work presented in this chapter was to leverage gorilla's intermediate phylogenetic position to help assess how variable or consistent patterns of evolutionary constraint have been throughout the evolution of the AGA clade.

This was done through a number of distinct but complementary analyses. First, a series of branch-LRTs were used to identify genes with evidence of elevated dN/dS along different branches of the primate phylogeny. The three AGAs showed similar overall counts of accelerated genes, and a number of distinct functional categories were enriched for accelerated genes; sound perception and hearing genes showed especially strong evidence for independent acceleration in AGA species. These LRT results were also used to quantify levels of parallel acceleration between pairs of AGA species, showing a greater amount of parallel acceleration between gorilla and human than between gorilla and chimpanzee. Finally, the set of genome-wide primate alignments was used to estimate lineage-specific dN/dS ratios. Since the historical effective population size (N_e) is a major factor behind variation in mean dN/dS ratios between closely-related species [Kosiol *et al.*, 2008; Ellegren, 2009], these precise dN/dS ratio estimates provided information regarding the relative historical N_e along each branch of the AGA phylogeny.

In conclusion, this study used codon models of evolution to place the gorilla genome within the context of the other AGAs and primates. The wealth of data afforded by genome-wide datasets allowed conclusions to be made regarding the variability of molecular evolutionary patterns among the AGAs species with more confidence than previously possible, revealing a largely uniform

landscape of recent AGA gene evolution which was generally in accordance with previously-published studies. On the other hand, the extent to which strong statements could be made about individual genes appeared somewhat limited by the small amount of divergence within the AGA clade. Furthermore, the evidence for lower N_e in the recent history of AGA species suggests that patterns of recent AGA and human evolution over the past 6–10 Myr have been dominated by genetic drift and relaxed constraint.

The use of sitewise estimates of selective pressures across mammals to guide the estimation of $d\mathrm{N}/d\mathrm{S}$ ratios within primates (for the creation of Figure 6.6) was a novel, if somewhat *ad hoc* approach to connecting the deeper evolutionary history of protein-coding sites with more recent patterns within closely-related primates. Future work stemming from this study could focus on formalizing the relationships between functional constraints acting on protein-coding material, the impact of population genetics on the efficacy of such selection, and the presumably infrequent—but highly interesting—occurrence of positive selection within recent primate lineages. A desirable result of this effort would be the development of methods which can efficiently use the wider evolutionary context of a gene or amino acid site to identify those substitution events most likely to be responsible for phenotypic variation between closely-related species such as the AGA.

A second area of important future work would be to extend the comparison between previously-published studies of gene acceleration and positive selection within primates into a more comprehensive analysis of the repeatability of results from these types of genome-wide scans. The quantitative comparison between published results in Chapter 5 was a first step in this direction, but a more thorough analysis would be helpful in understanding why different studies tend to show such poor overlap in identified accelerated and positively-selected genes. A key question is whether the discrepancies are due more to differences in the data (i.e., aligned sequences) or to differences in the methods. In the longer term, as more and more mammalian genomes are sequenced and small-scale analyses become less and less worthwhile, a reliable and fully-automated system for performing such analyses would seem like a worthy goal.

Chapter 7

Conclusions

In this thesis, I proposed to explore the application of mathematical models of evolution to better understand the patterns of natural selection acting on mammalian protein-coding genes. Throughout the analyses and discussions presented in the five preceding chapters, two recurrent dichotomies underscored significant remaining challenges and opportunities in contemporary comparative genomics: the distinction between truth and error in identifying orthologs and aligning protein-coding sequences, and the distinction between neutral evolution and natural selection in explaining their evolution.

The theme of error was at the forefront of Chapter 2, where simulated protein-coding sequence evolution was used to investigate the impact of alignment error on the detection of sitewise positive selection. The best aligners showed a good ability to accurately identify homologous codons, even in very divergent sequences prone to large amounts of biological insertion and deletion. On the other hand, *post-hoc* methods for alignment filtering seemed unable to improve on the best aligners in distinguishing true from erroneous homology. The parameters used for simulation were chosen to approximate the evolution of mammalian or vertebrate genes, and a wide range of divergence levels (from primate-like divergences to yeast-like divergences) was tested. Likely owing to differences in the prevalence of positive selection and in the distribution of selective pressures, some discrepancy was observed between the results of the current simulation and those of a similar study focused on the application of alignment filters to the study of HIV-1 evolution.

Even with powerful aligners available, errors were abundant in the alignments of mammalian and primate genes. Difficulties in identifying orthologs (Chapter 3), sequencing and assembling DNA (Chapter 4), gene conversion events (Chapter 5) and incomplete lineage sorting (Chapter 6) were all identified as plausible, and in some cases unavoidable, sources of error in the studies presented here. In Chapters 4 and 6 I described a heuristic approach to masking sequences or alignment regions with suspiciously dense clusters of nonsynonymous substitutions; further development of this approach, including quantification of its ability to reduce false positives in

downstream analyses, may be a fruitful area for future research.

The ability to distinguish between neutral evolution and natural selection is a major advantage of codon-based models of evolution in comparison to their nucleotide or amino acid counterparts. The application of codon models to the analysis of a large number of mammalian genomes showed how they can be used to explore the patterns of selective constraint experienced by protein-coding genes. It was clear that the additional mammalian genomes made available by the Mammalian Genome Project increased the power to detect purifying and positive selection, expanding the catalogue of genes with statistically significant evidence for positive selection and showing that positively selected genes (PSGs) often contain interwoven patterns of purifying and positive selection. In comparing the evolution of different mammalian groups, however, the distinction between drift and constraint was less certain. Chapters 4 and 5 found lower numbers of positively selected codons (PSCs) and PSGs, and lower average $d\mathrm{N}/d\mathrm{S}$ ratios within genes, in glires compared to primates and laurasiatheria. Given the well-established differences in effective population size (N_e) between glires and primate species, the nearly neutral theory provided a good explanation for the different $d\mathrm{N}/d\mathrm{S}$ ratios. The difference in levels of positive selection was harder to explain with confidence, but a number of factors may have contributed: widespread fixation of deleterious mutations in primates and laurasiatheria as a result of lower long-term N_e, higher error rates for detecting PSCs in primates due to the shorter total branch length, or a historically greater prevalence of positive selection in primates and laurasiatheria could all plausibly be responsible for the observed species-dependent differences in patterns of positive selection.

Future work could be directed towards an improved understanding of these differences. Results from the study of genetic variation in present-day populations may help shed light on levels of purifying and positive selection in the recent history of diverse mammals, and reasonable extrapolations deeper into history may provide new insight into the patterns observed here. Alternatively, the development of evolutionary models that explicitly account for changing N_e may help us better understand the impact of N_e on the evolution of mammalian genomes. The results from Chapter 6, which estimated a lower historical N_e for human than for all other African great ape (AGA) lineages examined, provided additional support for the development of advanced evolutionary models incorporating the effects of N_e within the framework of the nearly neutral theory.

The cost of sequencing a human-sized genome has dropped nearly 700-fold during the four years of my Ph.D. research (from \$7M to \$10k per genome, [Wetterstrand, 2011]), and ambitious yet realistic plans have been drawn to sequence several thousand vertebrate genomes in the near future [Haussler *et al.*, 2009]. With respect to the rapidly-developing technology of genome sequencing, two concluding points seem especially pertinent. First, the increasing amount of available genomic data will be matched perhaps only by the increasing number of potential false discoveries made possible by the error-prone nature of such high-throughput data collection and

analysis. Whereas researchers used to manually fix alignment or sequencing errors "by eye", this approach clearly does not work at a genomic scale, and well-designed automated methods should almost always outperform manual assessment. As a result, a rigorous and comprehensive understanding of sources of error in comparative genomics, combined with widespread adoption of best practices for reducing their impact on all types of downstream evolutionary analyses, will become increasingly important. Second, given the small number of observed fixed differences in comparative studies of humans and closely-related primates, it seems likely that the continued development of more complex evolutionary models and inference methods, rather than the sequencing of more primate genomes, has the most potential to significantly improve our power to identify and understand the molecular signatures of adaptive changes in our recent evolutionary past.

Appendix A

Publications

During the course of my Ph.D. research I contributed to the following published articles:

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. *et al.* (2011). Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. *Nat Genet*, **43**, 735–7.

Bishop C., Aanensen D., **Jordan, GE**, Kilian M., Hanage W. and Spratt B. (2009). Assigning strains to bacterial species via the internet. *BMC Biol*, **7**, 3.

Lindblad-Toh K., Garber M., Zuk O., Lin M., Parker B., Washietl S., Kheradpour P., Ernst J., **Jordan, G**, Mauceli E., Ward L., Lowe C., Holloway A., Clamp M., Gnerre S. *et al.* (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476–82.

Sipos B., Massingham T., **Jordan, G.E.** and Goldman N. (2011). PhyloSim–Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC Bioinformatics*, **12**, 104.

Jordan, G and Goldman N. (2011). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol*, **in press**.

Jordan, GE and Piel W.H. (2008). PhyloWidget: web-based visualizations for the tree of life. *Bioinformatics*, **24**, 1641–1642.

Appendix B

Top accelerated and positively-selected genes in the African great apes

			M0 dN/dS		Subst.		Branch-LRT			Branch-site LRT		
Test	Gene	Len.	Mam.	Pri.	N	S	LRT	p	FDR	LRT	p	FDR
Human	*IQSEC1*	1160	0.04	0.03	2	5	31.92	1.6e-08	1.8e-04	0.5	4.8e-01	1.0e+00
	FAM120B	911	0.25	0.31	24	5	22.14	2.5e-06	7.3e-03	15.46	8.4e-05	1.1e-01
	GRIP1	1129	0.08	0.06	4	0	17.26	3.3e-05	7.5e-02	2.74	9.8e-02	1.0e+00
	SLC23A1	603	0.12	0.1	7	1	16.8	4.1e-05	7.6e-02	1.22	2.7e-01	1.0e+00
	GPR98	6307	0.22	0.31	34	12	16.6	4.6e-05	7.6e-02	0.6	4.4e-01	1.0e+00
	MYO6	1296	0.05	0.09	5	0	16.32	5.3e-05	7.7e-02	2.98	8.4e-02	1.0e+00
	AC137834.1	4545	0.03	0.02	7	15	15.6	7.8e-05	9.0e-02	3	8.3e-02	1.0e+00
	SYNE1	8798	0.14	0.15	36	40	15.6	7.8e-05	9.0e-02	0	1.0e+00	1.0e+00
	DHX35	704	0.05	0.11	7	2	15.32	9.1e-05	9.5e-02	0.34	5.6e-01	1.0e+00
	CYB5R3	302	0.09	0.05	4	0	15.1	1.0e-04	9.8e-02	1.72	1.9e-01	1.0e+00
Chimpanzee	*HPS3*	1005	0.21	0.23	3	0	32.96	9.4e-09	1.1e-04	2.1	1.5e-01	1.0e+00
	UBR5	2800	0.03	0.03	7	4	30.6	3.2e-08	1.8e-04	10.4	1.3e-03	2.6e-01
	TRPC1	794	0.04	0.07	8	5	25.98	3.4e-07	1.0e-03	11.5	7.0e-04	1.7e-01
	ATP11A	1192	0.08	0.09	2	6	23.58	1.2e-06	2.8e-03	0	1.0e+00	1.0e+00
	POLR1A	1721	0.12	0.21	17	4	22.96	1.7e-06	3.2e-03	27.08	2.0e-07	4.5e-04
	DPP9	972	0.05	0.05	9	5	21.88	2.9e-06	4.5e-03	4.28	3.9e-02	1.0e+00
	KIAA0284	1591	0.15	0.13	25	14	21.76	3.1e-06	4.5e-03	2.46	1.2e-01	1.0e+00
	OTOF	1998	0.06	0.07	16	9	20.2	7.0e-06	8.9e-03	5	2.5e-02	1.0e+00
	FAM57B	275	0.07	0.03	6	3	19.5	1.0e-05	1.2e-02	0	1.0e+00	1.0e+00
	PAX8	451	0.08	0.05	6	4	19.24	1.2e-05	1.2e-02	3.98	4.6e-02	1.0e+00
Gorilla	*SUPT16H*	1048	0.02	0.04	14	7	61.1	5.0e-15	2.1e-11	0	1.0e+00	1.0e+00

Table B.1 (*continued on next page*)

Test	Gene	Len.	M0 dN/dS		Subst.		Branch-LRT			Branch-site LRT		
			Mam.	Pri.	N	S	LRT	p	FDR	LRT	p	FDR
	RBBP4	426	0.03	0.06	11	10	32.68	1.1e-08	3.1e-05	0	1.0e+00	1.0e+00
	MOBKL3	226	0.04	0.18	11	3	24.7	6.7e-07	1.3e-03	4.7	3.0e-02	1.0e+00
	ATP11A	1192	0.08	0.09	5	13	23.78	1.1e-06	1.8e-03	0	1.0e+00	1.0e+00
	ARHGAP5	1503	0.06	0.09	17	13	21.28	4.0e-06	5.7e-03	0	1.0e+00	1.0e+00
	PPIG	755	0.11	0.16	18	7	18.66	1.6e-05	1.8e-02	1.92	1.7e-01	1.0e+00
	EVPL	2034	0.08	0.1	36	24	18	2.2e-05	2.0e-02	0	1.0e+00	1.0e+00
	CHD1L	898	0.19	0.22	11	2	17.98	2.2e-05	2.0e-02	7	8.1e-03	1.0e+00
	SOCS5	537	0.06	0.18	23	15	17.32	3.2e-05	2.5e-02	3.7	5.4e-02	1.0e+00
	YTHDF2	580	0.11	0.07	3	0	17.28	3.2e-05	2.5e-02	7	0.0e+00	5.7e-01
AGA Stem	*RNF213*	5208	0.22	0.43			66.4	0.0e+00	3.8e-12	35.2	3.0e-09	3.4e-06
	ATP11A	1192	0.08	0.09			24.66	6.8e-07	2.6e-03	0	1.0e+00	1.0e+00
	ADAM12	910	0.19	0.39			18.36	1.8e-05	5.3e-02	3.7	5.4e-02	1.0e+00
	REC8	547	0.31	0.39			15.62	7.7e-05	1.8e-01	10.94	9.4e-04	4.2e-01
	TCEAL4	359	0.53	0.67			14.38	1.5e-04	2.9e-01	0.1	7.5e-01	1.0e+00
	DENND3	1279	0.11	0.22			14	1.8e-04	3.0e-01	1.12	2.9e-01	1.0e+00
	EIF3A	1383	0.07	0.07			12.8	3.5e-04	4.0e-01	54.5	1.5e-13	3.0e-10
	MGAT4B	564	0.06	0.03			12.32	4.5e-04	4.3e-01	9.52	2.0e-03	5.8e-01
	FBXO39	443	0.08	0.41			11.98	5.4e-04	4.7e-01	5.08	2.4e-02	1.0e+00
	PAPD5	699	0.12	0.13			11.68	6.3e-04	4.7e-01	10.28	1.3e-03	4.9e-01
AGA Clade	*SUPT16H*	1048	0.02	0.04			47.3	6.1e-12	2.3e-08	0	1.0e+00	1.0e+00
	DNAH1	4330	0.09	0.1			32.8	1.0e-08	2.9e-05	0	1.0e+00	1.0e+00
	IQSEC1	1160	0.04	0.03			30.72	3.0e-08	6.9e-05	0	1.0e+00	1.0e+00
	FAT1	4592	0.09	0.14			28.2	1.1e-07	2.1e-04	31	2.6e-08	6.0e-05
	HPS3	1005	0.21	0.23			27.84	1.3e-07	2.1e-04	0.36	5.5e-01	1.0e+00
	DNAH2	4428	0.09	0.15			27.6	1.5e-07	2.1e-04	0.2	6.6e-01	1.0e+00
	RBBP4	426	0.03	0.06			27.36	1.7e-07	2.2e-04	0	1.0e+00	1.0e+00
	MOBKL3	226	0.04	0.18			24.14	9.0e-07	9.5e-04	4.4	3.6e-02	1.0e+00
	OTOF	1998	0.06	0.07			20.2	7.0e-06	5.6e-03	1.6	2.1e-01	1.0e+00
	DMRT3	473	0.08	0.17			20.12	7.3e-06	5.6e-03	0.3	5.8e-01	1.0e+00
Human BS	*RNPEP*	651	0.15	0.13	2	4	0.12	7.3e-01	1.0e+00	133.26	0.0e+00	0.0e+00
	KPNA7	517	0.3	0.3	2	2	-0.14	7.1e-01	1.0e+00	71.54	0.0e+00	0.0e+00
	KCNJ6	424	0.02	0	0	4	0	1.0e+00	1.0e+00	41.68	1.1e-10	4.1e-07
	CDC6	561	0.27	0.43	2	0	3.08	7.9e-02	7.7e-01	22.78	1.8e-06	4.7e-03
	SARM1	723	0.06	0.05	5	3	12.58	3.9e-04	2.5e-01	22.56	2.0e-06	4.7e-03
	SELPLG	413	0.53	0.89	35	10	0.6	4.4e-01	9.0e-01	20.76	5.2e-06	1.0e-02
	RHBDL3	429	0.05	0.45	2	1	0.42	5.2e-01	9.3e-01	19.76	8.8e-06	1.5e-02
	C20orf72	345	0.32	0.36	1	3	-1.08	3.0e-01	8.4e-01	18.64	1.6e-05	2.3e-02
	FAM120B	911	0.25	0.31	24	5	22.14	2.5e-06	7.3e-03	15.46	8.4e-05	1.1e-01
	GNAS	1038	0.37	0.45	7	5	0.06	8.1e-01	1.0e+00	15.1	1.0e-04	1.2e-01
Chimpanzee BS	*SETD2*	2565	0.17	0.25	5	11	-0.6	4.4e-01	9.3e-01	196.6	0.0e+00	0.0e+00
	FGD4	767	0.17	0.26	2	1	0.74	3.9e-01	9.1e-01	48.2	3.8e-12	2.2e-08
	PENK	268	0.16	0.18	0	0	0	1.0e+00	1.0e+00	44.42	2.6e-11	1.0e-07
	KIF1A	1800	0.06	0.02	2	8	2.76	9.7e-02	8.5e-01	37.6	8.7e-10	2.5e-06

Table B.1 (*continued on next page*)

Test	Gene	Len.	M0 dN/dS		Subst.		Branch-LRT			Branch-site LRT		
			Mam.	Pri.	N	S	LRT	p	FDR	LRT	p	FDR
	POLR1A	1721	0.12	0.21	17	4	22.96	1.7e-06	3.2e-03	27.08	2.0e-07	4.5e-04
	CELSR2	2924	0.04	0.05	4	8	5	2.5e-02	7.1e-01	26	3.4e-07	6.6e-04
	FAM102A	385	0.05	0.04	2	2	5.32	2.1e-02	6.9e-01	24.3	8.2e-07	1.4e-03
	EOMES	706	0.1	0.11	3	1	10.6	1.1e-03	2.7e-01	23.42	1.3e-06	1.9e-03
	C14orf179	214	0.27	0.37	0	3	-4.14	4.2e-02	7.7e-01	23.16	1.5e-06	1.9e-03
	PDE3B	1113	0.26	0.32	8	10	0.08	7.8e-01	1.0e+00	22.72	1.9e-06	2.2e-03
Gorilla BS	*SPTA1*	2420	0.25	0.39	23	18	1	3.2e-01	8.8e-01	218.8	0.0e+00	0.0e+00
	CUL3	769	0.02	0.01	0	2	-0.08	7.8e-01	1.0e+00	131.82	0.0e+00	0.0e+00
	MKRN2	417	0.11	0.14	1	3	0	1.0e+00	1.0e+00	82.54	0.0e+00	0.0e+00
	DDX43	649	0.21	0.29	4	3	0.22	6.4e-01	9.7e-01	78.2	0.0e+00	0.0e+00
	SCN9A	1990	0.12	0.13	3	11	-0.14	7.1e-01	9.8e-01	55.86	7.8e-14	1.3e-10
	TXNDC3	589	0.44	0.48	6	8	-1.3	2.5e-01	8.7e-01	40.48	2.0e-10	2.9e-07
	FER1L6	1858	0.17	0.21	3	16	-4.2	4.0e-02	7.3e-01	39.6	3.1e-10	4.0e-07
	PRDX4	272	0.14	0.14	1	1	0.46	5.0e-01	9.3e-01	33.36	7.7e-09	8.8e-06
	CACNA1A	2507	0.09	0.04	1	15	-0.6	4.4e-01	9.1e-01	32.4	1.2e-08	1.3e-05
	ZYG11A	760	0.31	0.49	9	3	2.72	9.9e-02	7.9e-01	30.88	2.7e-08	2.6e-05
AGA Stem BS	*FSIP2*	6908	0.47	0.58			-3.2	7.4e-02	8.9e-01	204	0.0e+00	0.0e+00
	OSGIN2	550	0.11	0.08			0	1.0e+00	1.0e+00	110.74	0.0e+00	0.0e+00
	ADAM29	821	0.52	0.38			1.9	1.7e-01	8.9e-01	90.82	0.0e+00	0.0e+00
	TCERG1L	587	0.22	0.39			2.74	9.8e-02	8.9e-01	76.66	0.0e+00	0.0e+00
	SSH2	1424	0.26	0.38			-1.4	2.4e-01	9.0e-01	65.4	1.0e-15	1.5e-12
	EIF3A	1383	0.07	0.07			12.8	3.5e-04	4.0e-01	54.5	1.5e-13	3.0e-10
	AHNAK	5891	0.31	0.27			-0.6	4.4e-01	9.4e-01	47.4	5.8e-12	9.5e-09
	PDHX	502	0.24	0.25			-3.9	4.8e-02	8.9e-01	44.02	3.2e-11	4.2e-08
	RNF213	5208	0.22	0.43			66.4	0.0e+00	3.8e-12	35.2	3.0e-09	3.4e-06
	CPB2	424	0.3	0.48			0.54	4.6e-01	9.5e-01	29.28	6.3e-08	6.6e-05
AGA Clade BS	*KIAA1370*	1077	0.27	0.41			-1.3	2.5e-01	7.4e-01	49.06	2.5e-12	9.5e-09
	QRICH2	1664	0.5	0.67			1.4	2.4e-01	7.2e-01	43.2	4.9e-11	1.4e-07
	FAT1	4592	0.09	0.14			28.2	1.1e-07	2.1e-04	31	2.6e-08	6.0e-05
	RG9MTD1	404	0.31	0.67			-0.86	3.5e-01	8.0e-01	28.36	1.0e-07	1.9e-04
	AHNAK	5891	0.31	0.27			-0.4	5.3e-01	8.8e-01	24.4	7.8e-07	1.3e-03
	SELPLG	413	0.53	0.89			0.04	8.4e-01	9.7e-01	22.82	1.8e-06	2.6e-03
	MUC6	2440	0.23	0.3			10.6	1.1e-03	1.4e-01	21.6	3.4e-06	4.3e-03
	ZNRF4	430	0.13	0.27			3.64	5.6e-02	5.0e-01	20.4	6.3e-06	7.3e-03
	GPR98	6307	0.22	0.31			20	7.7e-06	5.6e-03	20	7.7e-06	8.1e-03
	IFT140	1463	0.11	0.16			7.88	5.0e-03	2.5e-01	18.48	1.7e-05	1.7e-02
Human-Chimpanzee LRT$_{min}$	*FAT1*	4592	0.09	0.14	55	62	10.4					
	LOXHD1	1948	0.09	0.07	15	20	7.6					
	GPR98	6307	0.22	0.31	62	27	7.4					
	ZNF629	870	0.06	0.06	9	2	7.06					

Table B.1 (*continued on next page*)

Test	Gene	Len.	M0 dN/dS		Subst.		Branch-LRT			Branch-site LRT		
			Mam.	Pri.	N	S	LRT	p	FDR	LRT	p	FDR
	PARP3	541	0.18	0.29	11	0	6.74					
	DSTYK	930	0.1	0.15	4	0	5.88					
	SFRS8	952	0.1	0.09	8	6	5.6					
	C10orf118	899	0.16	0.23	8	0	5.48					
	ELF5	266	0.07	0.03	2	0	5.42					
	SGCA	388	0.14	0.19	10	2	5.38					
Gorilla-Human LRT_{min}	*IQSEC1*	1160	0.04	0.03	2	12	18.4					
	CANX	628	0.08	0.08	9	6	7.48					
	PLIN4	1358	0.3	0.22	83	73	7.2					
	DNAH2	4428	0.09	0.15	43	41	6.4					
	LACTB	548	0.15	0.15	7	0	6.4					
	PTK7	1079	0.06	0.1	13	11	6.2					
	LOXHD1	1948	0.09	0.07	18	25	6.2					
	ITIH3	891	0.15	0.15	11	3	6.04					
	CSMD1	3566	0.06	0.06	28	79	6					
	LGR6	968	0.12	0.13	14	9	5.9					
Gorilla-Chimpanzee LRT_{min}	*ATP11A*	1192	0.08	0.09	7	19	23.58					
	NKAP	416	0.22	0.18	2	1	11.48					
	EP400	3160	0.12	0.14	28	28	10					
	MAP3K9	1119	0.08	0.09	9	3	8.68					
	LOXHD1	1948	0.09	0.07	19	27	6.2					
	SMC1B	1236	0.16	0.39	12	0	5.7					
	LMO2	228	0.03	0.09	4	0	5.64					
	UNC45B	932	0.06	0.06	8	8	5.54					
	OTOF	1998	0.06	0.07	28	26	5.2					
	FGF3	240	0.11	0.08	6	2	4.9					
Gorilla-Human-Chimpanzee LRT_{min}	*LOXHD1*	1948	0.09	0.07	26	36	6.2					
	ITIH3	891	0.15	0.15	18	7	4.8					
	ELK4	432	0.16	0.26	7	0	4.22					
	FBXO24	619	0.13	0.2	12	4	3.64					
	ATAD2B	1459	0.1	0.07	8	6	3.18					
	PARP3	541	0.18	0.29	20	3	3.16					
	GPR98	6307	0.22	0.31	102	61	3					
	HSF1	530	0.08	0.06	6	3	2.94					
	FBXO2	299	0.06	0.06	3	0	2.94					
	UNC45B	932	0.06	0.06	12	16	2.82					

Table B.1 (*continued on next page*)

			M0 dN/dS		Subst.		Branch-LRT			Branch-site LRT		
Test	Gene	Len.	Mam.	Pri.	N	S	LRT	p	FDR	LRT	p	FDR

Table B.1: Genes with the stongest evidence for acceleration and positive selection in gorilla and the African great apes. Mam.—mammals; Pri—primates; N—the number of nonsynonymous substitutions; S—the number of synonymous substitutions; LRT—the likelihood ratio test statistic; FDR—the false discovery rate calculated using the Benjamini and Hochberg [1995] method.

Bibliography

After each reference, a list of pages from which the reference is cited is included in brackets.

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

Adachi J. and Hasegawa M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J Mol Evol*, **42**, 459–468.

Aguileta G., Refrégier G., Yockteng R., Fournier E. and 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.

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. *et al.* (2011). Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. *Nat Genet*, **43**, 735–7.

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

Alroy J. (1998). Cope's Rule and the Dynamics of Body Mass Evolution in North American Fossil Mammals. *Science*, **280**, 731–734.

Alroy J. (1999). The fossil record of North American mammals: Evidence for a Paleocene evolutionary radiation. *Systematic Biology*, **48**, 107–118.

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

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

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

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

Anisimova M., Nielsen R. and 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.

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

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. and Swine Genome Sequencing Consortium (2010). Pig genome sequence–analysis and publication strategy. *BMC Genomics*, **11**, 438.

Archibald J.D. and 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.

Armstrong P. (2006). Proteases and protease inhibitors: a balance of activities in host-pathogen interaction. *Immunobiology*, **211**, 263–81.

Arnheim N. and Calabrese P. (2009). Understanding what determines the frequency and pattern of human germline mutations. *Nat Rev Genet*, **10**, 478–488.

Ashburner M., Ball C., Blake J., Botstein D., Butler H., Cherry J., Davis A., Dolinski K., Dwight S., Eppig J., Harris M., Hill D., Issel-Tarver L., Kasarskis A., Lewis S. *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–9.

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

Axelsson E. and 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.

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

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.

Baer C.F., Miyamoto M.M. and Denver D.R. (2007). Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet*, **8**, 619–631.

Bakewell M., Shi P. and 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.

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

Beisswanger S. and 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.

Benjamini Y. and Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met*, **57**, 289–300.

Benner S.A., Cohen M.A. and Gonnet G.H. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J Mol Biol*, **229**, 1065–1082.

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

Bernig T., Taylor J.G., Foster C.B., Staats B., Yeager M. and Chanock S.J. (2004). Sequence analysis of the mannose-binding lectin (MBL2) gene reveals a high degree of heterozygosity with evidence of selection. *Genes Immun*, **5**, 461–476.

Bierne N. and Eyre-Walker A. (2003). The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics*, **165**, 1587–1597.

Bierne N. and Eyre-Walker A. (2004). The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol Biol Evol*, **21**, 1350–1360.

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

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. and Purvis A. (2007). The delayed rise of present-day mammals. *Nature*, **446**, 507–512.

Birney E., Stamatoyannopoulos J.A., Dutta A., Guigó R., Gingeras T.R., Margulies E.H., Weng Z., Snyder M., Dermitzakis E.T., Stamatoyannopoulos J.A., Thurman R.E., Kuehn M.S., Taylor C.M., Neph S., Koch C.M. *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

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

Boehler C., Gauthier L.R., Mortusewicz O., Biard D.S., Saliou J.M., Bresson A., Sanglier-Cianferani S., Smith S., Schreiber V., Boussin F. and Dantzer F. (2011). Poly(ADP-ribose) polymerase 3 (PARP3), a newcomer in cellular response to DNA damage and mitotic progression. *Proceedings of the National Academy of Sciences*, **108**, 2783–2788.

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

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. and Bustamante C. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, **4**, e1000083.

Bradley B.J. (2008). Reconstructing phylogenies and phenotypes: a molecular view of human evolution. *J Anat*, **212**, 337–353.

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.

Brown W.M., Prager E.M., Wang A. and Wilson A.C. (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J Mol Evol*, **18**, 225–239.

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

Burgess R. and Yang Z. (2008). Estimation of Hominoid Ancestral Population Sizes under Bayesian Coalescent Models Incorporating Mutation Rate Variation and Sequencing Errors. *Mol Biol Evol*, **25**, 1979–1994.

Busto M. and Posada D. (2010). The Effect of Recombination on the Reconstruction of Ancestral Sequences. *Genetics*.

Bustos O., Naik S., Ayers G., Casola C., Perez-Lamigueiro M., Chippindale P., Pritham E. and de la Casa-Esperón E. (2009). Evolution of the Schlafen genes, a gene family associated with embryonic lethality, meiotic drive, immune processes and orthopoxvirus virulence. *Gene*, **447**, 1–11.

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

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

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

Camon E., Magrane M., Barrell D., Lee V., Dimmer E., Maslen J., Binns D., Harte N., Lopez R. and Apweiler R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, **32**, D262–D266.

Carroll S. (2003). Genetics and the making of Homo sapiens. *Nature*, **422**, 849–57.

Cartwright R.A. (2009). Problems and solutions for estimating indel rates and length distributions. *Mol Biol Evol*, **26**, 473–480.

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

Castillo-Davis C.I., Kondrashov F.A., Hartl D.L. and 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.

Castresana J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*, **17**, 540–552.

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

Chao K.M., Pearson W.R. and Miller W. (1992). Aligning two sequences within a specified diagonal band. *Computer applications in the biosciences : CABIOS*, **8**, 481–487.

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

Charlesworth J. and Eyre-Walker A. (2006). The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol*, **23**, 1348–1356.

Chen F. and Li W. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet*, **68**, 444–56.

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

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

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. *et al.* (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–3.

Clark A.G. and Civetta A. (2000). Evolutionary biology: Protamine wars. *Nature*, **403**, 261–263.

Clark A.G., Eisen M.B., Smith D.R., Bergman C.M., Oliver B., Markow T.A., Kaufman T.C., Kellis M., Gelbart W., Iyer V.N., Pollard D.A., Sackton T.B., Larracuente A.M., Singh N.D., Abad J.P. *et al.* (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**, 203–218.

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

Cock P., Fields C., Goto N., Heuer M. and 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.

Collins F.S. and McKusick V.A. (2001). Implications of the Human Genome Project for Medical Science. *JAMA*, **285**, 540–544.

Cooper G.M., Brudno M., Stone E.A., Dubchak I., Batzoglou S. and Sidow A. (2004). Characterization of Evolutionary Rates and Constraints in Three Mammalian Genomes. *Genome Research*, **14**, 539–548.

Cordaux R. and Batzer M.A. (2009). The impact of retrotransposons on human genome evolution. *Nat Rev Gen*, **10**, 691–703.

Cousins R. (2007). Annotated bibliography of some papers on combining significances or p-values. *ArXiV Preprint Archive:0902.0885*.

Crespi B.J. and Summers K. (2006). Positive selection in the evolution of cancer. *Biol Rev*, **81**, 407–424.

da Fonseca R., Kosiol C., Vinar T., Siepel A. and Nielsen R. (2010). Positive selection on apoptosis related genes. *FEBS Lett*, **584**, 469–76.

Darlington R.B. and Hayes A.F. (2000). Combining independent p-values: Extensions of the Stouffer and binomial methods. *Psychol Methods*, **5**, 496.

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.

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

Dayhoff M.O. and 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. and Arndt P.F. (2007). DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evolutionary Biology*, **7**, 191.

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

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

Dermitzakis E.T. and Clark A.G. (2001). Differential Selection After Duplication in Mammalian Developmental Genes. *Molecular Biology and Evolution*, **18**, 557–562.

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

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

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

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

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

Ebana Y., Ozaki K., Inoue K., Sato H., Iida A., Lwin H., Saito S., Mizuno H., Takahashi A., Nakamura T., Miyamoto Y., Ikegawa S., Odashiro K., Nobuyoshi M., Kamatani N. *et al.* (2007). A functional SNP in ITIH3 is associated with susceptibility to myocardial infarction. *Journal of Human Genetics*, **52**, 220–229.

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

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

Edvardson S., Jalas C., Shaag A., Zenvirt S., Landau C., Lerer I. and Elpeleg O. (2011). A deleterious mutation in the LOXHD1 gene causes autosomal recessive hearing loss in Ashkenazi Jews. *American Journal of Medical Genetics Part A*, **155**, 1170–1172.

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

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

Ellegren H. (2008). Comparative genomics and the study of evolution by natural selection. *Mol Ecol*, **17**, 4586–4596.

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

Enard W., Przeworski M., Fisher S.E., Lai C.S.L., Wiebe V., Kitano T., Monaco A.P. and Paabo S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language.

Endo T., Ikeo K. and Gojobori T. (1996). Large-scale search for genes on which positive selection may operate. *Mol Biol Evol*, **13**, 685–690.

Eyre-Walker A., Keightley P.D., Smith N.G.C. and Gaffney D. (2002). Quantifying the Slightly Deleterious Mutation Model of Molecular Evolution. *Mol Biol Evol*, **19**, 2142–2149.

Ezawa K., Oota S. and Saitou N. (2006). Genome-wide search of gene conversions in duplicated genes of mouse and rat. *Mol Biol Evol*, **23**, 927–940.

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

Felsenstein J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol*, **17**, 368–376.

Finarelli J.A. and Flynn J.J. (2006). Ancestral State Reconstruction of Body Size in the Caniformia (Carnivora, Mammalia): The Effects of Incorporating Data from the Fossil Record. *Syst Biol*, **55**, 301–313.

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. and Bateman A. (2010). The Pfam protein families database. *Nucleic Acids Res*, **38**, D211–22.

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

Fletcher W. and Yang Z. (2009). INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*, **26**, 1879–1888.

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

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. *et al.* (2011). Ensembl 2011. *Nucleic Acids Res*, **39**, D800–6.

Foster T. and Höök M. (1998). Surface protein adhesins of Staphylococcus aureus. *Trends Microbiol*, **6**, 484–8.

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

Garred P., Larsen F., Seyfarth J., Fujita R. and Madsen H.O. (2006). Mannose-binding lectin and its genetic variants. *Genes and Immunity*, **7**, 85–94.

Gibbs R., Weinstock G., Metzker M., Muzny D., Sodergren E., Scherer S., Scott G., Steffen D., Worley K., Burch P., Okwuonu G., Hines S., Lewis L., DeRamo C., Delgado O. *et al.* (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Gibbs R., Rogers J., Katze M., Bumgarner R., Weinstock G., Mardis E., Remington K., Strausberg R., Venter J., Wilson R., Batzer M., Bustamante C., Eichler E., Hahn M., Hardison R. *et al.* (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–34.

Goldman N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, **36**, 182–198.

Goldman N. and Yang Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, **11**, 725–736.

Goodall J. (1986). *The chimpanzees of Gombe: patterns of behavior*. Belknap Press of Harvard University Press, Cambridge, MA USA.

Grantham R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. *Science*, **185**, 862–864.

Grassly N.C. and Holmes E.C. (1997). A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol Biol Evol*, **14**, 239–247.

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

Gu X., Wang Y. and 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.

Guirao-Rico S. and Aguadé M. (2009). Positive selection has driven the evolution of the Drosophila insulin-like receptor (InR) at different timescales. *Mol Biol Evol*, **26**, 1723–32.

Hahn M. (2009). Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. *Journal of Heredity*, **100**, 605–617.

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

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

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

Harcourt A., Fossey D., Stewart K. and Watts D. (1980). Reproduction in wild gorillas and some comparisons with chimpanzees. *J Reprod Fertil Suppl*, **Suppl 28**, 59–70.

Hasegawa M., Kishino H. and Yano T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, **22**, 160–174.

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.

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

He X. and Zhang J. (2005). Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. *Genetics*, **169**, 1157–1164.

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

Heger A. and Ponting C. (2008). OPTIC: orthologous and paralogous transcripts in clades. *Nucleic Acids Res*, **36**, D267–70.

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

Hillier L.W., Miller W., Birney E., Warren W., Hardison R.C., Ponting C.P., Bork P., Burt D.W., Groenen M.A.M., Delany M.E., Dodgson J.B., Genome fingerprint map s., assembly:, Chinwalla A.T., Cliften P.F. *et al.* (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

Hobolth A., Christensen O., Mailund T. and Schierup M. (2007). Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet*, **3**, e7.

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

Hoffmann J.A., Kafatos F.C., Janeway C.A. and Ezekowitz R.A.B. (1999). Phylogenetic Perspectives in Innate Immunity. *Science*, **284**, 1313–1318.

Hokamp K., McLysaght A. and Wolfe K.H. (2003). The 2R hypothesis and the human genome sequence. *J Struct Func Genomics*, **3**, 95–110.

Hou Z., Romero R. and Wildman D. (2009). Phylogeny of the Ferungulata (Mammalia: Laurasiatheria) as determined from phylogenomic data. *Mol Phylogenet Evol*, **52**, 660–664.

Huang D.W., Sherman B.T. and Lempicki R.A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44–57.

Huang Y., Temperley N., Ren L., Smith J., Li N. and Burt D. (2011). Molecular evolution of the vertebrate TLR1 gene family—a complex history of gene duplication, gene conversion, positive selection and co-evolution. *BMC Evol Biol*, **11**, 149.

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. *et al.* (2007). Ensembl 2007. *Nucleic Acids Res*, **35**, D610–7.

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

Hughes A.L. (1999). *Adaptive evolution of genes and genomes*. Oxford University Press, New York.

Hughes A.L. and Nei M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–70.

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

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

Hwang D.G. and Green P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A*, **101**, 13994–14001.

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. *et al.* (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate protokaryotype. *Nature*, **431**, 946–57.

Jones D.T., Taylor W.R. and Thornton J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *CABIOS*, **8**, 275–282.

Jordan G. and Goldman N. (2011). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol*.

Jordan I.K., Wolf Y. and Koonin E. (2004). Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evolutionary Biology*, **4**, 22.

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

Jukes T. and Cantor C. (1969). Evolution of protein molecules, 21–132. In H. Munro, ed., *Mammalian Protein Metabolism*, Academic Press, New York.

Jukes T. and King J. (1979). Evolutionary nucleotide replacements in DNA. *Nature*, **281**, 605–6.

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

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

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

Kimura M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, **267**, 275–6.

Kimura M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, **16**, 111–120.

Kimura M. (1985). *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

Kimura M. and Ohta T. (1974). On some principles governing molecular evolution. *Proc Natl Acad Sci U S A*, **71**, 2848–2852.

King M. and Wilson A. (1975). Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–16.

Koonin E.V., Makarova K.S. and Aravind L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Ann Rev of Microbiol*, **55**, 709–742.

Kosakovsky Pond S.L., Posada D., Gravenor M.B., Woelk C.H. and Frost S.D.W. (2006). Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm. *Mol Biol Evol*, **23**, 1891–1901.

Kosiol C., Bofkin L. and Whelan S. (2006). Phylogenetics by likelihood: evolutionary modeling as a tool for understanding the genome. *J Biomed Inform*, **39**, 51–61.

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

Kosiol C., Vinar T., da Fonseca R., Hubisz M., Bustamante C., Nielsen R. and Siepel A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS Genet*, **4**, e1000144.

Kryazhimskiy S. and Plotkin J. (2008). The population genetics of dN/dS. *PLoS Genet*, **4**, e1000304.

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

Lander E., Linton L., Birren B., Nusbaum C., Zody M., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lartillot N. and Philippe H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*, **21**, 1095–1109.

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

Le S.Q., Lartillot N. and Gascuel O. (2008). Phylogenetic mixture models for proteins. *Philos T Roy Soc B*, **363**, 3965–3976.

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. *et al.* (2009). The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.

Lin Y.S., Hsu W.L., Hwang J.K. and Li W.H. (2007). Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol Biol Evol*, **24**, 1005–1011.

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. *et al.* (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803–819.

Lindblad-Toh K., Garber M., Zuk O., Lin M., Parker B., Washietl S., Kheradpour P., Ernst J., Jordan G., Mauceli E., Ward L., Lowe C., Holloway A., Clamp M., Gnerre S. *et al.* (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476–82.

Locke D., Hillier L., Warren W., Worley K., Nazareth L., Muzny D., Yang S., Wang Z., Chinwalla A., Minx P., Mitreva M., Cook L., Delehaunty K., Fronick C., Schmidt H. *et al.* (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, **469**, 529–33.

Lovelace L.L., Cooper C.L., Sodetz J.M. and Lebioda L. (2011). Structure of Human C8 Protein Provides Mechanistic Insight into Membrane Pore Formation by Complement. *J Biol Chem*, **286**, 17585–17592.

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

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

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

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

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

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

Margulies E., Vinson J., NISC Comparative Sequencing Program, Miller W., Jaffe D., Lindblad-Toh K., Chang J., Green E., Lander E., Mullikin J. and 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.

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. *et al.* (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res*, **17**, 760–74.

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

Marques-Bonet T., Ryder O.A. and Eichler E.E. (2009). Sequencing primate genomes: what have we learned? *Annu Rev Genomics Hum Genet*, **10**, 355–386.

Martin A.P. and Palumbi S.R. (1993). Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci U S A*, **90**, 4087–4091.

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

Martin-DeLeon P. (2006). Epididymal SPAM1 and its impact on sperm function. *Mol Cell Endocrinol*, **250**, 114–21.

Massingham T. and Goldman N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–62.

Massingham T., Davies L.J. and Liò P. (2001). Analysing gene function after duplication. *BioEssays*, **23**, 873–876.

McKerrow J.H., Sun E., Rosenthal P.J. and Bouvier J. (1993). The Proteases and Pathogenicity of Parasitic Protozoa. *Ann Rev Microbiol*, **47**, 821–853.

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

Messier W. and Stewart C. (1997). Episodic adaptive evolution of primate lysozymes. *Nature*, **385**, 151–4.

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

Mikkelsen T.S., Hillier L.W., Eichler E.E., Zody M.C., Jaffe D.B., Yang S.P., Enard W., Hellmann I., Lindblad-Toh K., Altheide T.K., Archidiacono N., Bork P., Butler J., Chang J.L., Cheng Z. *et al.* (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.

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

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

Monroe M.J. and Bokma F. (2010). Short communication: little evidence for Cope's rule from Bayesian phylogenetic analysis of extant mammals. *J Evol Biol*, **23**, 2017–2021.

Montgomery S., Capellini I., Venditti C., Barton R. and Mundy N. (2011). Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. *Mol Biol Evol*, **28**, 625–38.

Moran N.A., McCutcheon J.P. and Nakabachi A. (2008). Genomics and Evolution of Heritable Bacterial Symbionts. *Ann Rev Genet*, **42**, 165–190.

Morgan G.J. (1998). Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959–1965. *Journal of the History of Biology*, **31**, 155–178.

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

Muller J., Szklarczyk D., Julien P., Letunic I., Roth A., Kuhn M., Powell S., von Mering C., Doerks T., Jensen L. and 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.

Mundy N. (2007). Coloration and the genetics of adaptation. *PLoS Biol*, **5**, e250.

Murakami M., Taketomi Y., Girard C., Yamamoto K. and Lambeau G. (2010). Emerging roles of secreted phospholipase A2 enzymes: Lessons from transgenic and knockout mice. *Biochimie*, **92**, 561–82.

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

Muse S.V. and Gaut B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, **11**, 715–724.

Nabholz B., Glémin S. and Galtier N. (2008). Strong Variations of Mitochondrial Mutation Rate across Mammals–the Longevity Hypothesis. *Mol Biol Evol*, **25**, 120–130.

Nei M., Suzuki Y. and Nozawa M. (2010). The neutral theory of molecular evolution in the genomic era. *Ann Rev Genom Hum Genet*, **11**, 265–289.

Nembaware V., Crum K., Kelso J. and Seoighe C. (2002). Impact of the presence of paralogs on sequence divergence in a Set of mouse-human orthologs. *Genome Research*, **12**, 1370–1376.

Nicholson D. and Thornberry N. (1997). Caspases: killer proteases. *Trends Biochem Sci*, **22**, 299–306.

Nickel G., Tefft D., Goglin K. and Adams M. (2008). An empirical test for branch-specific positive selection. *Genetics*, **179**, 2183–93.

Nielsen R. (2005). Molecular signatures of natural selection. *Ann Rev Genet*, **39**, 197–218.

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

Nielsen R. and 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.

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. and Cargill M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*, **3**, e170.

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

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

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

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

North B., Curtis D. and Sham P. (2002). A note on the calculation of empirical p-values from Monte Carlo procedures. *Am J Hum Genet*, **71**, 439–41.

Notredame C. (2007). Recent Evolutions of Multiple Sequence Alignment Algorithms. *PLoS Computational Biology*, **3**, e123.

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

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

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

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

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

Ohbayashi T., Irie A., Murakami Y., Nowak M., Potempa J., Nishimura Y., Shinohara M. and Imamura T. (2011). Degradation of fibrinogen and collagen by staphopains, cysteine proteases released from *Staphylococcus aureus*. *Microbiology*, **157**, 786–792.

Ohno S. (1970). *Evolution by gene duplication*. Springer-Verlag, New York.

Ohta T. (1992). The Nearly Neutral Theory of Molecular Evolution. *Annu Rev Ecol Syst*, **23**, 263–286.

Olszewski M.A., Gray J. and Vestal D.J. (2006). In silico genomic analysis of the human and murine guanylate-binding protein (GBP) gene clusters. *J Interfer Cytok Res*, **26**, 328–352.

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

Paten B., Herrero J., Beal K., Fitzgerald S. and Birney E. (2008a). Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res*, **18**, 1814–28.

Paten B., Herrero J., Fitzgerald S., Beal K., Flicek P., Holmes I. and Birney E. (2008b). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res*, **18**, 1829–43.

Penn O., Privman E., Landan G., Graur D. and Pupko T. (2010). An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol*, **27**, 1759–1767.

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

Pollard K., Hubisz M., Rosenbloom K. and Siepel A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, **20**, 110–21.

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

Popadin K., Polishchuk L., Mamirova L., Knorre D. and 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.

Privman E., Penn O. and Pupko T. (2011). Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol Biol Evol*.

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. *et al.* (2008). The amphioxus genome and the evolution of the chordate karyotype. **453**.

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

Ratnakumar A., Mousset S., Glémin S., Berglund J., Galtier N., Duret L. and 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.

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

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

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

Rouleau M., Saxena V., Rodrigue A., Paquet E., Gagnon A., Hendzel M., Masson J., Ekker M. and Poirier G. (2011). A key role for poly(ADP-ribose) polymerase-3 in ectodermal specification and neural crest development. *PLoS One*, **6**, e15834.

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. *et al.* (2008). TreeFam: 2008 Update. *Nucleic Acids Res*, **36**, D735–40.

Sato H., Taketomi Y., Isogai Y., Miki Y., Yamamoto K., Masuda S., Hosono T., Arata S., Ishikawa Y., Ishii T., Kobayashi T., Nakanishi H., Ikeda K., Taguchi R., Hara S. *et al.* (2010). Group III secreted phospholipase A2 regulates epididymal sperm maturation and fertility in mice. *J Clin Invest*, **120**, 1400–14.

Sawyer S.L., Wu L.I., Emerman M. and Malik H.S. (2005). Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. *Proc Natl Acad Sci U S A*, **102**, 2832–2837.

Schierup M.H. and Hein J. (2000). Consequences of Recombination on Traditional Phylogenetic Analysis. *Genetics*, **156**, 879–891.

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

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

Seyfarth J., Garred P. and Madsen H.O. (2005). The "involution" of mannose-binding lectin. *Human Molecular Genetics*, **14**, 2859–2869.

Sharp P.M. (1997). In search of molecular darwinism. *Nature*, **385**, 111–112.

Shriner D., Nickle D., Jensen M. and Mullins J. (2003). Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genetical Res*, **81**, 115–21.

Sidak Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*, 626–633.

Siepel A. (2009). Phylogenomics of primates and their ancestral populations. *Genome Res*, **19**, 1929–41.

Siepel A. and Haussler D. (2004). Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. *Mol Biol Evol*, **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. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034–50.

Sipos B., Massingham T., Jordan G. and Goldman N. (2011). PhyloSim–Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC Bioinformatics*, **12**, 104.

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

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

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. *et al.* (2010). The Evolution of Maximum Body Size of Terrestrial Mammals. *Science*, **330**, 1216–1219.

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

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

Sodeinde O., Subrahmanyam Y., Stark K., Quan T., Bao Y. and Goguen J. (1992). A surface protease and the invasive character of plague. *Science*, **258**, 1004–1007.

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

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

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

Studer R. and Robinson-Rechavi M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet*, **25**, 210–6.

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

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. and Lavin M.F. (2007). Senataxin, defective in ataxia oculomotor apraxia type 2, is involved in the defense against oxidative DNA damage. *J Cell Biol*, **177**, 969–979.

Swanson W.J., Nielsen R. and Yang Q. (2003). Pervasive Adaptive Evolution in Mammalian Fertilization Proteins. *Mol Biol Evol*, **20**, 18–20.

Takahata N. and Satta Y. (1997). Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc Natl Acad Sci U S A*, **94**, 4811–5.

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

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

Taudien S., Ebersberger I., Glöckner G. and Platzer M. (2006). Should the draft chimpanzee sequence be finished? *Trends Genet*, **22**, 122–5.

Tavaré S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some mathematical questions in biology: DNA sequence analysis*, Lectures on mathematics in the life sciences, American Mathematical Society.

Taylor W. (1986). The classification of amino acid conservation. *J Theor Biol*, **119**, 205–18.

Thomas P.D., Campbell M.J., Kejariwal A., Mi H., Karlak B., Daverman R., Diemer K., Muruganujan A. and Narechania A. (2003). PANTHER: A Library of protein families and subfamilies indexed by function. *Genome Res*, **13**, 2129–2141.

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

Thompson J.D., Higgins D.G. and Gibson T.J. (1994b). 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.

Torgerson D.G., Kulathinal R.J. and Singh R.S. (2002). Mammalian Sperm Proteins Are Rapidly Evolving: Evidence of Positive Selection in Functionally Diverse Genes. *Mol Biol Evol*, **19**, 1973–1980.

Uddin M., Goodman M., Erez O., Romero R., Liu G., Islam M., Opazo J.C., Sherwood C.C., Grossman L.I. and Wildman D.E. (2008). Distinct genomic signatures of adaptation in pre- and postnatal environments during human evolution. *Proceedings of the National Academy of Sciences*, **105**, 3215–3220.

Uzzell T. and Corbin K.W. (1971). Fitting Discrete Probability Distributions to Evolutionary Events. *Science*, **172**, 1089–1096.

Van de Peer Y., Maere S. and Meyer A. (2009). The evolutionary significance of ancient genome duplications. *Nat Rev Gen*, **10**, 725–732.

Varmus H. (2010). Ten years on—the human genome and medicine. *New Engl J Med*, **362**, 2028–2029.

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

Verdu P., Barreiro L.B., Patin E., Gessain A., Cassar O., Kidd J.R., Kidd K.K., Behar D.M., Froment A., Heyer E., Sica L., Casanova J.L., Abel L. and Quintana-Murci L. (2006). Evolutionary insights into the high worldwide prevalence of MBL2 deficiency alleles. *Human Molecular Genetics*, **15**, 2650–2658.

Verga Falzacappa M.V., Segat L., Puppini B., Amoroso A. and Crovella S. (2004). Evolution of the mannose-binding lectin gene in primates. *Genes and Immunity*, **5**, 653–661.

Vigilant L. and Bradley B.J. (2004). Genetic variation in gorillas. *Am J Primatol*, **64**, 161–172.

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

Villa P., Kaufmann S. and Earnshaw W. (1997). Caspases and caspase inhibitors. *Trends Biochem Sci*, **22**, 388–93.

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

Wang J., Gonzalez K., Scaringe W., Tsai K., Liu N., Gu D., Li W., Hill K. and Sommer S. (2007). Evidence for mutation showers. *Proc Natl Acad Sci U S A*, **104**, 8403–8.

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

Warnecke T. and 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.

Warren W.C., Hillier L.W., Marshall Graves J.A., Birney E., Ponting C.P., Grutzner F., Belov K., Miller W., Clarke L., Chinwalla A.T. and et al. (2008). Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, **453**, 175–183.

Waterston R., Lindblad-Toh K., Birney E., Rogers J., Abril J., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P., Antonarakis S., Attwood J., Baertsch R., Bailey J., Barlow K. *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–62.

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

Werck-Reichhart D. and Feyereisen R. (2000). Cytochromes P450: a success story. *Genome Biol*, **1**, REVIEWS3003.

Wetterstrand K. (2011). DNA sequencing costs: data from the NHGRI large-scale genome sequencing program. Accessed November 20, 2011 from http://www.genome.gov/sequencingcosts/.

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

Whelan S. and Goldman N. (1999). Distributions of Statistics Used for the Comparison of Models of Sequence Evolution in Phylogenetics. *Mol Biol Evol*, **16**, 1292.

Whelan S. and Goldman N. (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol Biol Evol*, **18**, 691–699.

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

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

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

Wilks S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, **9**, 60–62.

Wilson D. and Reeder D. (2005). *Mammal Species of the World: A Taxonomic and Geographic Reference*. v. 1, Johns Hopkins University Press.

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

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

Wong W., Yang Z., Goldman N. and Nielsen R. (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, **168**, 1041–51.

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

Wright S. (1931). Evolution in Mendelian Populations. *Genetics*, **16**, 97–159.

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

Yang W., Bielawski J. and Yang Z. (2003). Widespread Adaptive Evolution in the Human Immunodeficiency Virus Type 1 Genome. *J Mol Evol*, **57**, 212–221.

Yang Z. (????). *Computational Molecular Evolution (Oxford Series in Ecology and Evolution)*. Oxford University Press, New York.

Yang Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol*, **39**, 306–314.

Yang Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol*, **11**, 367–72.

Yang Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*, **15**, 568–73.

Yang Z. (2005). The power of phylogenetic comparison in revealing protein function. *Proc Natl Acad Sci U S A*, **102**, 3179–3180.

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

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

Yang Z. and Nielsen R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, **17**, 32–43.

Yang Z. and Nielsen R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, **19**, 908–917.

Yang Z. and Swanson W. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol*, **19**, 49–57.

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

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

Yang Z., Wong W.S.W. and Nielsen R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol*, **22**, 1107–1118.

Young M., Wakefield M., Smyth G. and Oshlack A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*, **11**, R14.

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

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

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

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

Zhang J., Nielsen R. and 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.

Zhang L. and Li W.H. (2005). Human SNPs reveal no evidence of frequent positive selection. *Mol Biol Evol*, **22**, 2504–2507.

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

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

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