Using exome sequencing to study adaptive evolution in non-human primates and human populations

Renee D. George

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Willie J. Swanson, Chair
Joshua M. Akey
Stanley Fields

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Using exome sequencing to study adaptive evolution
in non-human primates and human populations

Renee D. George

Chair of the Supervisory Committee:
Associate Professor Willie J. Swanson
Department of Genome Sciences

Positive selection promotes the fixation of functional genetic differences. Studying these adaptive differences can provide insights into how species adapt to their environment and develop their unique phenotypes. In this dissertation, I explore the effects of positive selection on the evolution of modern humans and non-human primates. I first describe current methods to detect positive selection and review how advances in next-generation sequencing technologies have improved our ability to detect adaptive events. I then present the first application of whole exome sequencing to non-human primates and describe new bioinformatic methods to assemble protein coding sequences for species without reference genomes. Using these coding sequences I scan the primate genome for genes experiencing positive selection and identify a novel class of adaptively evolving genes involved in keratinization. I then examine the effect of recurrent positive selection acting both in non-human primates and human populations using a large human exome data set from the NHLBI's Exome Sequencing Project. I find that genes with evidence for long-term positive selection in non-human primates also show diversity patterns in humans that are consistent with continued positive selection. I conclude by discussing the trade-offs between whole-genome and exome sequencing, the need for more quality reference genomes and the difficulties in distinguishing adaptive evolution from other non-selective biological processes.

TABLE OF CONTENTS

List of Figures ii

List of Tables iii

Chapter 1: Introduction 1

1.1 Methods for detecting positive selection from DNA sequences 2

1.2 Genome-wide scans for positive selection 4

1.3 Next-generation sequencing 5

Chapter 2: Positive selection in primate exomes 8

2.1 Introduction 9

2.2 Results/Discussion 10

2.3 Methods 15

Chapter 3: Recurrent positive selection in human populations 33

3.1 Introduction 34

3.2 Results 35

3.3 Discussion 40

3.4 Methods 42

Chapter 4: Conclusions and Future Directions 55

4.1 Importance of high quality reference genomes 55

4.2 Whole-genome *vs.* exome sequencing 57

4.2 Future directions for positive selection in primates 58

List of References 60

Appendix A: Supplemental Information for Chapter 2 72

Appendix B: Supplemental Information for Chapter 3 99

LIST OF FIGURES

Figure 2.1 Sequence capture and assembly of non-human primates 28

Figure 2.2 Read depth of targeted regions 29

Figure 2.3 Sequence differences and indel lengths in protein coding regions 30

Figure 2.4 Examples of gene loss in Old World monkeys 31

Figure 3.1 Phylogeny of primate species used to identify genes targeted by positive selection 50

Figure 3.2 Genomic map of genes under long-term positive selection in non-human primates 51

Figure 3.3 Patterns of human divergence and diversity at non-synonymous PSSs, non-synonymous non-PSSs and 4FD sites 52

LIST OF TABLES

Table 2.1 Sequence coverage of captured target 32

Table 3.1 Summary of human divergence and diversity analyses for PSGs and non-PSGs from 2,439 human individuals 53

Figure 3.2 Summary of polymorphism data for PSSs, non-PSSs and 4FD sites from 2,439 human individuals 54

ACKNOWLEDGEMENTS

Many people have greatly contributed to the research presented in this dissertation. My advisor, Willie Swanson has been an outstanding mentor and a constant provider of encouragement and support. I especially thank him for encouraging me to pursue my own research interests, being flexible with my personal life and always prioritizing my future career. I would also like to acknowledge the members of my committee, Josh Akey, Debbie Nickerson, Stan Fields and Bertil Hille, for guidance during my dissertation. In particular, Debbie Nickerson has provided me with ideas, resources and data access that are the foundation of the work described here. I have also appreciated the amusing and often valuable discussions I've had over the years with Josh Akey.

I thank past and present members of the Swanson lab for their support and feedback during graduate school: Geoff Findlay, Jan Aagaard, Joe Gaspar, Melody Palmer, Katrina Claw, Jen McCreight, Jenn Kohler, Joanna Kelley and Nathan Clark.

I would like to acknowledge my co-authors that directly contributed to the work in Chapters 2 and 3 of this dissertation. A very similar version of chapter 2 was published in *Genome Research* in the fall of 2011 and was co-authored with Graham McVicker, Rachel Diederich, Sarah Ng, Alex MacKenzie, Jay Shendure, Willie Swanson and Jim Thomas. Chapter 3 was co-authored with Josh Akey, Willie Swanson and the NHLBI Exome Sequencing Project and will soon be submitted for publication.

I will leave Genome Sciences with many life-long friends. I thank them for making the last 5 years an amazing experience. In particular, I thank Graham McVicker who has been my biggest supporter, best friend and mentor.

Finally, I would like to thank my parents for always believing in me.

Chapter 1
Introduction

Species are continually adapting to their ever-changing environments. When new mutations occur, they may increase in frequency within a population if they confer a fitness benefit to the organism. This rise in frequency and subsequent fixation of beneficial mutations within a population is referred to as positive natural selection. New mutations can also rise and fix in a population due to genetic drift, a neutral process in which the random assortment of alleles causes their frequencies to shift each generation, even in the absence of selection. Over time, positive selection and genetic drift can cause isolated populations to diverge and eventually form separate species (reviewed in [1]).

Understanding how natural selection and genetic drift contribute to the evolution of populations and species is an important area of research in evolutionary biology. By studying the targets of positive selection, the vast number of genetic differences between species can be reduced to a list of candidates that are likely to be functionally responsible for species-specific phenotypes. Here I summarize our current understanding of adaptive evolution in primates. I first discuss several methods to detect positive selection from multiple DNA sequences and focus on their application to humans and non-human primates. I then describe next-generation sequencing technologies and how they can improve the power and resolution with which we are able to identify regions of the genome targeted by positive selection.

1.1 Methods for detecting positive selection from DNA sequences

The majority of the human genome is thought to evolve neutrally [2], with a small fraction (~5%) under purifying selection [3,4]. A smaller fraction still is likely to be subject to positive selection. Statistical methods to detect positive selection typically identify genomic regions where patterns of polymorphism and/or divergence differ substantially from the background of putatively neutrally evolving sequence. In the following sections, I describe several commonly used methods to detect positive selection from comparisons of multiple DNA sequences.

1.1.1 Divergence between species

Mutations in protein coding regions are classified as non-synonymous if they alter the encoded protein's amino acid sequence, or as synonymous mutations if they leave the protein's amino acid sequence unchanged. One way natural selection can be detected is by comparing the rate of non-synonymous substitutions (d_N) to the rate of synonymous substitutions (d_S) between closely related species. In the absence of selection, synonymous and non-synonymous substitutions should accumulate at the same rate such that $d_N = d_S$ and $d_N/d_S = 1$. If there is purifying selection and non-synonymous sites are under functional constraint, then $d_N < d_S$ and $d_N/d_S < 1$. This is the case for most proteins. Alternatively, if there is positive selection and beneficial non-synonymous mutations frequently occur, then $d_N > d_S$ and $d_N/d_S > 1$. Statistical models whose parameters can be estimated by maximum likelihood have been developed to detect $d_N/d_S > 1$ across entire genes [5], at specific sites within genes [6-8] and on different lineages [9-11].

1.1.2 Variation within species

When an advantageous mutation rapidly rises in frequency and eventually becomes fixed in a population, linked variants also rise in frequency with it. This is referred to as a selective sweep and the result is a drastic reduction in neutral diversity that is linked to the selected site and an increase in the length of haplotype blocks [12]. Over time, diversity is gradually restored as new mutations occur and long haploytpe blocks are

broken down by genetic recombination [13]. Methods to detect positive selection acting within populations generally look for these signatures of selective sweeps.

A selective sweep alters the site frequency spectrum of polymorphisms within a population such that there is an excess of rare alleles (due to new mutations) and a higher frequency of derived alleles relative to neutral expectations. Statistical tests that are commonly used to detect these skews in the site frequency spectrum include Tajima's D [14], Fu and Li's D [15-17] and Fay and Wu's H [18]. Statistical models that use the site frequency spectrum to test for selection while controlling for demography and variation in the mutation rate have also been developed [19-21].

When a selected allele increases in frequency very quickly, recombination does not have time to decouple the selected allele from those surrounding it. As a result, in regions where a selective sweep has occurred, linkage disequilibrium (LD) is higher and the length of selected haplotype blocks are longer. These signatures are exploited by the long-range haplotype (LRH) test [22], the integrated haplotype score (iHS) [23] and other tests for positive selection [24-27]. These tests have the advantage that they are more sensitive to incomplete or partial sweeps where selection acts on standing neutral variation (rather than new mutations) or where the selected allele is not yet fixed in the population.

Positive selection can also increase allele frequency differentiation between populations [28-30]. This occurs, for example, when only one of the populations being examined experiences a selective sweep. The degree of population differentiation can be measured with the F_{ST} statistic, and this statistic has been used to identify putative targets of recent positive selection [31,32].

1.1.3 Variation within and between species

Several methods use both variation within a species and divergence between species to detect selection. These methods, which include the McDonald-Kreitman (MK) [33] and

HKA [34] tests, typically compare variation and divergence at one set of sites to another set of neutrally evolving sites. For example, the MK test uses synonymous sites within a gene of interest as a neutral reference and compares the ratio of non-synonymous to synonymous polymorphic sites (P_N/P_S) to the ratio of non-synonymous to synonymous substitutions (D_N/D_S). Under neutrality, these ratios should be the same, but under positive selection, the relative rate of non-synonymous divergence can exceed that of non-synonymous polymorphism such that $D_N/D_S > P_N/P_S$. The HKA test is similar to the MK test, but it uses an independent neutrally evolving locus rather than synonymous sites as its neutral reference.

1.2 Genome-wide scans for positive selection

1.2.1 Divergence between species

Genes with an elevated d_N/d_S ratio between species are thought to be under positive selection because they have a higher rate of fixed non-synonymous mutations, which tend to be functional, than fixed synonymous mutations, which tend to have no fitness effects. Elevated rates of non-synonymous substitutions can occur when there is recurrent positive selection that repeatedly drives functional variants to fixation. Power to detect these recurrent events depends on the total sequence divergence of the species used [35]. In primates, few reference genomes have been completed which limits the power to identify genes experiencing positive selection with d_N/d_S methods. Previous genome-wide scans for positive selection in primates [3,36-41] have been restricted to subsets of the currently available reference genomes, which are human [42], chimpanzee [38], orangutan [39], gorilla [41] and macaque [3]. Many of the genes identified by these scans are involved in fertility and defense against pathogens.

1.2.2 Variation within species

Many genome-wide scans for positive selection in humans have been performed using population-based statistical tests and either publically available genotype data (reviewed by [43]) or whole genome/exome data (*e.g.* [44,45]). Although each of these studies

identified a list of candidate regions targeted by positive selection, the overlap between studies is low suggesting a high rate of false positives or differences in power (reviewed by [43]). Additionally, there is some disagreement over the amount of the genome affected by positive selection. It has been estimated that ~10% of the human genome is linked to a selective sweep [21], but this estimate may instead reflect background selection, which is the reduction in neutral diversity at sites that are linked to loci under purifying selection [46-48]. Recent results that take into account background selection suggest that classic selective sweeps were rare in recent human evolution [49]. This has led to the proposal that most adaptation in humans may occur by "soft" or incomplete sweeps, potentially because selection acts on standing variation or on multiple loci simultaneously [50].

Despite the difficulties in pinpointing positive selection in the human genome, several loci stand out as clear targets of recent human adaptation. These include genes that interact with pathogens (*e.g. DARC* [51]) and pigmentation genes (*e.g. SLC24A5* [52] and *KITLG* [53]) and regulatory changes involved in lactase persistence [54,55].

1.3 Next-generation sequencing

Recent technological advances have drastically reduced sequencing time and costs. In 2001, the draft sequence of the human genome was published [42] using Sanger sequencing technology [56]. The finished human genome cost about $300 million to sequence, which is more than 500× the cost to sequence a human genome today [57].

The major advance of next-generation sequencing (NGS) technologies, is their ability to massively parallelize the sequencing reaction, usually by immobilizing millions of DNA templates onto a solid surface. Although NGS platforms differ in their sequencing biochemistry, most use a sequencing-by-synthesis method in which fluorescent nucleotides are incorporated into a DNA molecule one base at a time. After each cycle, the entire solid surface is imaged to identify the nucleotides incorporated into each molecule (reviewed in [58]).

Despite their cost and time savings, NGS platforms have several disadvantages. The read lengths for NGS platforms are much shorter than Sanger reads making it difficult or impossible to assemble long repetitive regions of genomes. Additionally, the error rates of NGS methods are higher than Sanger sequencing which necessitates sequencing more deeply. Several groups have developed clever strategies to overcome these limitations [59,60].

1.3.1 Exome sequencing

Sequencing costs can be cut even further by targeting a smaller fraction of the genome, such as the complete set of protein coding exons (the 'exome'). Coding exons or other target regions are most often isolated using targeted capture-by-hybridization technology. In this approach, genomic DNA is first hybridized to capture probes that are specific to the target regions and then isolated and purified by magnetic beads or another purification strategy. Other methods for target enrichment include multiplex PCR and capture by circularization (reviewed by [61]). Since the haploid human exome consists of only 30 Mb (1% of the entire genome), this substantially reduces the amount of sequencing necessary to ascertain protein coding variants. Additionally, by capturing only the exome we avoid sequencing the majority of the repetitive regions in the genome, which are difficult to map, assemble and call variants from.

1.3.2 Application to comparative and population genetic analyses

NGS has dramatically reduced the cost and increased the speed at which reference genomes for other species can be generated. It should be noted, however, that genomes that are assembled using only next-generation sequencing reads are generally of lower quality than those assembled from Sanger reads. In primates, a mixture of Sanger and NGS approaches have been used to complete the most recent reference genomes. The gorilla reference genome was assembled using a combination of Sanger and NGS reads [41], and the orangutan reference genome was assembled from only Sanger reads, but the project also used NGS to sequence multiple individuals from two sub-species [39]. The

addition of these two reference genomes has allowed the identification of many genes experiencing long-term and lineage-specific positive selection in primates [39,41].

Prior to NGS, genome-wide population genetic analyses relied on genotypes obtained from SNP microarrays (*e.g.* [62]). The data obtained from microarrays suffer from ascertainment bias because the SNPs that are genotyped by the arrays were discovered using a small number of individuals [63]. Furthermore, the limited number of probes on each array means that only a subset of (typically common) SNPs can be assayed in a single experiment. These issues have recently been alleviated by the adoption of NGS for genotyping. By sequencing the entire exome or genome of sampled individuals it is now possible to create a comprehensive catalog of variation that includes both rare and common polymorphism [44,45]. These new catalogs of variation promise to improve the power and accuracy with which selective events can be detected in populations.

Chapter 2
Positive selection in primate exomes

Comparison of protein coding DNA sequences from diverse primates can provide insight into these species' evolutionary history and uncover the molecular basis for their phenotypic differences. Currently, the number of available primate reference genomes limits genome-wide comparisons. Here we use targeted capture methods designed for human to sequence the protein coding regions, or exomes, of four non-human primate species (three Old World monkeys and one New World monkey). We are able to capture ~96% of coding sequences for these species, despite average sequence divergence of up to 4% from the human sequence probes. Using a combination of mapping and assembly techniques, we generated high quality full length coding sequences for each species. Both the number of nucleotide differences and the distribution of insertion and deletion (indel) lengths indicate that the quality of the assembled sequences is very high and exceeds that of most reference genomes. Using this expanded set of primate coding sequences, we performed a genome-wide scan for genes experiencing positive selection and identified a novel class of adaptively evolving genes involved in the production of keratin in the epithelial cells of skin, hair and nails. Interestingly, the genes we identify under positive selection also exhibit significantly increased allele frequency differences among human populations, suggesting that they play a role in both recent and long-term adaptation. We also identify several genes that have been lost on specific primate lineages, which illustrate the broad utility of this dataset for other evolutionary analyses. These results demonstrate the power of next-generation sequencing in comparative genomics and greatly expand the repertoire of available primate coding sequences.

2.1 Introduction

Comparative genomics is invaluable for the study of evolutionary processes such as mutation, selection and speciation [64]. In many cases, our power to detect evolutionary events is limited by the number of species with high quality genome sequences [65,66]. For example, power to detect positive selection depends on the total sequence divergence of the species being studied [65]. Additionally, for many evolutionary analyses, the sequences need to be of very high quality to limit the rate of false-positives [67,68]. Next-generation sequencing technologies have made sequencing new primate genomes more feasible, but it is still challenging to assemble these short reads into complete genomes.

New methods for targeted enrichment of the human exome allow high coverage sequence to be generated for the coding fraction of the genome [69-72]. These methods are currently used in human medical resequencing studies to identify causal genes in Mendelian disorders [73-75], but in principle can be extended to targeted sequencing of human orthologs in closely related species. Such an approach has the advantage that sequenced reads are limited to non-repetitive coding regions that are more easily assembled from short reads.

Here we use solution-based targeted capture [76] designed to human exons to sequence the exomes of three Old World monkeys and one New World monkey. We combine our high quality sequences with available primate reference genomes, and conduct a genome-wide scan for genes experiencing positive selection in primates. Our analysis has greater statistical power than previous scans for positive selection in primates [3,36-38,40], which were limited by a low number of species and low total sequence divergence [65]. Other studies, which used diverse mammals to identify targets of positive selection [77], are more powerful, but provide little information on more recent adaptation in primates.

We identify over 150 genes that show strong evidence of positive selection on the primate lineage, at least twice as many as previous studies with fewer species [3,40].

Many of the genes and gene classes we identify are in accordance with these previous scans (*e.g.* genes involved in defense and immunity), however we also find a number of novel adaptively evolving genes, most notably several genes involved in keratinization.

2.2 Results/Discussion

In total, we sequenced the exomes of three Old World monkeys (rhesus macaque, colobus monkey and vervet) and one New World monkey (tamarin) (Fig. 2.1A). For each species, we targeted 25.3 Mb of unique protein coding sequence from the Consensus Coding Sequence (CCDS) database [78] and generated, on average, 7.1 Gb of sequence per species with paired-end 76 bp reads. We aligned reads to the human reference genome and performed a local assembly of each target region (Fig. 2.1B, Table A1-2). Our approach differs from *de novo* assembly in that it retains information from the human reference genome, while still allowing for more diverged sequences than typical short read mapping techniques (see Methods for more details).

To assess capture, sequencing, and assembly quality, we compared our rhesus macaque exome to the macaque reference genome (rheMac2). We captured and mapped macaque sequences for greater than 96% of the target (Table 2.1 and Fig. 2.2) and surprisingly found only a low association between our ability to capture macaque sequences and the number of nucleotide differences ($R^2 = 0.0087$) or the number of indels ($R^2 = 0.0047$) in targeted regions. In fact, human capture efficiency is the most informative predictor of macaque capture efficiency, suggesting that unknown conserved sequence features predominate in determining capture efficiency (Table A3). We also compared targets that were successfully captured to those that failed to capture (less than 50% of bases covered by a read) despite having clear orthologs in the macaque genome (Table A4). The failed targets have comparable GC content (51.6% *vs.* 49.9%), slightly higher divergence (3.9% *vs.* 3.1%) and a substantially greater proportion of bases which were inserted or deleted (0.50% *vs.* 0.17%). In total, only a small number of targets failed to capture by this criteria (1,111 out of 155,707 targets with orthologs), and even targets up to 7% diverged from human are captured efficiently with a mean read depth greater than

60X (Fig. A1). Our ability to capture even the most divergent exons in macaque suggests that it will be possible to perform targeted capture of even more distantly related species.

We assembled high quality (≥Q40; error rate of $<10^{-4}$) macaque sequences for about 90% of our target (Table 2.1). These sequences are 2.24% different from the human reference genome, which corresponds almost precisely with sequence divergence in coding regions calculated from the macaque reference genome (Fig. 2.3A and Table A5). We estimate the pairwise differences of our assembly relative to the macaque reference genome to be about 0.10%, which agrees with a previous estimate of nucleotide diversity in Indian macaques (0.12%) [79]. These data indicate that the quality of our macaque exome is at least that of the macaque reference genome and demonstrate that we can generate high quality and accurate exome assemblies from short read data.

In the three other primates that we sequenced, between 95% and 97% of the targeted bases are covered by at least one read, and we generated high quality consensus sequence for between 86% and 90% of the target (Table 2.1). Once again, divergence had very little impact on our ability to capture exome sequences from non-human primates (Fig. 2.2). Even for the most divergent species, the tamarin, we captured greater than 96% and assembled greater than 88% of the target at high quality (Table 2.1).

We performed extensive filtering of our exome assemblies, because errors in sequencing, alignment, assembly or ortholog assignment may introduce false positives in comparative genomic analyses [67,68]. We removed sequences overlapping known segmental duplications in human [80,81], chimpanzee [81] and macaque [82], removed sequences with low read depth (<16×) and removed exons with very high levels of heterozygosity (which may reflect mis-assembly of paralogous sequences). We then removed exons and genes that had less than half of their sequence remaining after filtering. These filtering criteria exclude sequences that are more likely to be mis-assembled due to paralogous sequences (Text A1) and do not appear to be biased towards removing known rapidly evolving genes, such as those involved in reproduction or immunity (Text A2). For the exome assemblies, 61-72% of each species' targeted coding sequence was retained for

comparative analysis. This is comparable to the reference genome sequences, where we used less conservative filtering and retained 80-89% of the targeted coding sequences post-filtering.

We combined our four assembled exomes with coding sequences from the reference genomes of human, chimpanzee, orangutan and macaque and generated multiple sequence alignments for 16,707 genes. After filtering for high quality regions in common to all species, we calculated the average nucleotide divergence from the human reference genome to be 2.3% for Old World monkeys and 3.9% for the New World monkey (Fig. 2.3A and Table A5).

As a test of assembly quality, we examined the distribution of indel lengths with respect to the human reference genome (Fig. 2.3B). On average, 80% of indels from our exome assemblies have lengths that are multiples of three, an enrichment that is consistent with selection to preserve reading frame and remarkably similar to that of the macaque reference genome (Table A6). This enrichment is substantially higher than that seen in other human exome studies [73,83] or in the chimpanzee and orangutan reference genomes (Fig. 2.3B), which suggests that our exome assemblies are of higher quality. The increased rate of indel errors in the orangutan reference genome has also been previously noted [84].

From our coding sequence alignments, we filtered out sequences with too little sequence data, frameshifts or internal stop codons (Fig. A2) to obtain a highly confident set of 15,027 orthologs with sequence from at least three species. We then tested each of these orthologs for evidence of positive selection acting at any point during primate evolution using likelihood models that allow d_N/d_S to vary across codons [7,11]. The addition of our exome sequences increased the total branch length by 3-fold relative to analyses using just human, chimpanzee and macaque (median $S = 0.30$ *vs.* $S = 0.080$ nucleotide substitutions per codon), and should substantially increase our power to detect positive selection [65]. We find evidence of positive selection for 930 genes (nominal P-value < 0.05) without correcting for multiple testing, or a total of 157 at a false discovery rate (FDR) of 10% (Table A7-8).

We compared these 157 genes to a previous scan for positive selection in primates that identified 67 positively selected genes (at 10% FDR) using coding sequences from the human, chimpanzee and macaque genomes [3]. Of these 67 genes, we omitted 22 from our analysis because they were either not targeted or we could not confidently obtain sequences from at least three species, including the available primate reference genomes (Table A9). The remaining 45 genes rank significantly higher than other genes in our scan for positive selection ($p < 2.2 \times 10^{-16}$; two-sided Mann-Whitney U test) with 15 genes showing strong evidence for positive selection at an FDR of 10% and an additional 19 genes with a nominal $p < 0.05$. We thus identify 142 new candidates in our analysis. We find no evidence of positive selection for 11 of the genes identified by the previous analysis, which is likely due to differences in methodology, such as ortholog filtering, low quality sequence filtering or multiple sequence alignment. When we perform the same analysis using only human, chimpanzee and macaque sequences, we find evidence of positive selection for only 25 genes (at 10% FDR). Our approach is more conservative and should have fewer false positives due to low quality or mis-aligned sequences.

We identified several biological processes enriched for genes under positive selection (Table A10) using the Gene Ontology classification system [85]. In agreement with previous scans for positive selection in primates [3,37,40], several of the top categories are involved in immunity ("defense response" and "antigen processing and presentation"), sensory perception ("sensory perception of a chemical stimulus") and reproduction ("spermatogenesis" and "fertilization"). Among the genes that show the strongest evidence of positive selection are several that are known to be rapidly evolving (*e.g. PTPRC* [86], *PRM1* [87] and *APOBEC3G* [88]), and several with no previous evidence of positive selection in primates (*e.g. TF*, an iron transporter previously known to be under positive selection only in salmonids [89]).

Interestingly, we also found an excess of positively selected genes involved in the process of keratinization (Table A11). To our knowledge, none of these genes have been previously identified as targets of positive selection in primates, except for *IVL*, which

was recently shown to be subject to positive selection in human populations [90]. As keratinization is the cornification of outer epidermal cells in skin, hair and nails, these genes may be important for setting up physical barriers between the body and the outside world and could evolve rapidly in response to changing environments.

We also tested whether the genes that we find under positive selection in primates also show evidence for recent selection in human populations. The 157 genes with strong evidence for positive selection in primates have increased allele frequency differences between Europeans and Africans (mean F_{ST} = 0.0928) compared to the remaining genes (mean F_{ST} = 0.0710; $p < 2.2 \times 10^{-16}$; two-sided Mann-Whitney U test) [90]. We also tested the ten GO categories with the most significant enrichments for genes under positive selection, and found that three of the categories ("sensory perception of chemical stimulus", "oxidation reduction" and "sensory perception") have significantly higher levels of population differentiation (Table A12). Both observations are consistent with the idea that many of the genes under long-term positive selection in primates are also important in recent human adaptation.

In addition to the 157 positively selected genes identified across all branches of the primate phylogeny, 142 genes show evidence for positive selection on specific lineages at an FDR of 10% (Table A13-15). Of these, 28 overlap with the original set of 157 genes, bringing the total number of identified genes to 271. Two genes, *KRTAP4-5* and *CASP10*, have evidence for positive selection on more than one lineage. *KRTAP4-5* encodes a keratin-associated protein involved in the structure of hair fibers and shows evidence for positive selection on both the chimpanzee and the hominid branches. *CASP10* encodes an apoptosis-related caspase that appears to be adaptively evolving on both the Old World monkey and tamarin lineages. The number of genes identified on each branch ranges from zero on the human branch to 84 on the tamarin branch. The longer branches (*e.g.* tamarin) probably have more significant genes because they provide more power to detect positive selection [91].

To demonstrate how this dataset can be used for other types of evolutionary analyses, we identified genes that have been lost by either gene deletion or pseudogenization. For example, all or nearly all exons of the gene *GBP5* are missing in the three Old World monkeys, but are present in the New World monkey, the tamarin (Fig. 2.4A). Similarly *SNTN* and *CCL14* contain premature stop codons or frameshifts in all the Old World monkeys, yet not in tamarin (Fig. 2.4B-C). Sequences from the chimpanzee, orangutan and macaque reference genomes confirm the loss of these genes in the common ancestor of Old World monkeys.

We have demonstrated that solution-based hybrid capture is an efficient method for the sequencing of orthologs in other species. This method can be applied not just to non-human primate species, but to any species for which a closely related reference genome is available, and works well for coding sequences up to an average divergence of 7% (and possibly greater). In principle, this method is not limited to coding sequences, but can be extended by designing capture probes to other unique regions of the genome. We have used this method to identify 157 candidates for positive selection in primates and envision that this method can be applied to many other problems in molecular evolution and population genomics. By obtaining sequences from multiple individuals it will be possible to characterize patterns of genetic variation in numerous species. Such data will help answer many questions about selection, demography, mutation, gene loss and gene duplication.

2.3 Methods

2.3.1 Genomic DNA samples

Genomic DNA samples were obtained from Coriell Cell Repositories for a rhesus macaque (*Macaca mulatta*; NG07107), a colobus monkey (*Colobus angolensis*; PR00099), a vervet (*Chlorocebus aethiops*; PR00990), a tamarin (*Saguinus midas*; PR00550) and two HapMap human individuals (European-American NA12878 and East Asian NA18967).

2.3.2 Library Oligonucleotides and Adaptors

Oligonucleotides used in the libraray constuction were SLXA_Pair_For_Amp (AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC*T), SLXA_Pair_Rev_Amp (CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG ATC*T), Adapter_PE_Hi (ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TC*T) and Adapter_PE_Lo (/5Phos/GAT CGG AAG AGC GGT TCA GCA GGA ATG CCG AG). "*" refers to a phosphorothioate bond.

Adapter_PE_Hi and Adapter_PE_Lo were annealed to form Y-adaptors by incubating equimolar amounts at 95°C and then allowing them to cool to room temperature in a heat block.

2.3.3 Library Construction

Genomic DNA from each sample (3 µg) was sheared (Covaris AFA) in 85 µl elution buffer (Buffer EB,10 mM Tris-Cl, pH 8.5, Qiagen) using the settings: duty cycle 10%, intensity 5, and cycle/burst 200 for 600 sec. Fragmented DNA ends were repaired for 30 min at 20°C with 5 µl End Repair Enzyme Mix and 1X End Repair Reaction Buffer in a total volume of 100 µl (NEBNext End Repair Module, New England Biolabs) and eluted in 45 µl of water after clean-up. A-tails were then added to the end-repaired DNA at 70°C for 20 mins in a total volume of 100 µl (1X PCR buffer, 1.5 mM $MgCl_2$, 1 mM dATP and 5 units AmpliTaq DNA polymerase) and eluted in 38 µl of water after clean-up. Y-adaptors were ligated to the A-tailed fragments at 16°C for 20 min in a total volume of 50 µl (1X T4 DNA Ligase Buffer (Enzymatics), 240 units T4 DNA Ligase (Enzymatics) and 5 µl Y-adaptors (50 µM)) and eluted in 50 µl of water after clean-up. All clean-up steps were performed with 1.8X AmpureXP beads as directed by Agencourt.

The adaptor-ligated fragments were PCR amplified in four reactions per sample, each in a total volume of 40 µl (10 µl adaptor-ligated fragments, 1X iProof High Fidelity Master Mix (Bio-Rad) and 0.625 µM of both SLXA_Pair_For_Amp and

SLXA_Pair_Rev_Amp). The PCR conditions were: 96°C for 2 min, 16 cycles of 96°C for 20 sec, 65°C for 30 sec and 72°C for 45 sec, followed by a final 72°C for 5 min. The four reactions for each sample were then pooled, cleaned-up (PCR Purification Kit, Qiagen) and quantified (Nanodrop 8000 Spectrophotometer). Alex MacKenzie in the Shendure lab constructed these sequencing libraries.

2.3.4 Library Capture and Sequencing

Each library (1µg) was hybridized to SeqCap EZ Exome probes (v1.0, Nimblegen) according to manufacturer's protocols and blocked with 100 µl of 1 mg/ml human COT1 DNA (Invitrogen) and 10 µl of both SLXA_Pair_For_Amp (100 µM) and SLXA_Pair_Rev_Amp (100 µM). The hybridized library was captured and washed as directed by Nimblegen and eluted in 50 µl of water. The enriched library was PCR amplified in ten reactions per sample, each in a total volume of 50 µl (4 µl library, 1X iProof High Fidelity Master Mix and 0.625 µM of both SLXA_Pair_For_Amp and SLXA_Pair_Rev_Amp). The PCR conditions were: 98°C for 30 sec, 20 cycles of 98°C for 10 sec, 60°C for 30 sec and 72°C for 30 sec, followed by a final 72°C for 5 min. The ten reactions for each sample were then pooled and column purified (PCR Purification Kit, Qiagen). Alex MacKenzie in the Shendure lab performed this exome capture and purification.

One lane of 76 bp paired-end reads was generated for each sample on an Illumina Genome Analyzer IIx according to manufacturer's instructions by Charlie Lee in the Shendure lab.

2.3.5 Target description

All unique, well-annotated protein coding regions (including flanking regions for exons smaller than 200 bp) from the CCDS database (version 20080430) [78] and ~550 miRNAs were targeted by SeqCap EZ Exome probes. This resulted in 176,817 continuous captured genomic regions totaling to 34,108,810 bp. The 20080430 version of the CCDS database contains 164,367 protein coding genomic regions spanning

28,000,325 bp after merging regions with overlapping coordinates. Repetitive regions are excluded from the tiling probes, reducing the final protein coding target to 148,667 genomic regions for a total of 25,299,356 bp.

2.3.6 Merging overlapping paired-end reads

Although our genomic DNA was fragmented to an average size of 200 bp, a substantial fraction of our 76 bp paired-end reads overlapped their mate and could be merged into longer single reads. We aligned each pair of reads using a semi-global version of the Needleman-Wunsch [92] algorithm that constrained the alignment to the end of the left read and the start of the right read and used the following score scheme: *match* +1, *mismatch* -3, *gap* -5. If the alignment score was greater than or equal to 10, the reads were merged, with any mismatching positions masked to "N". When high scoring alignments contained gaps, both reads were discarded. Quality scores for the overlapping portion of the merged reads were calculated by summing the two independent quality scores, capping the maximum value at Q40. When alignments did not meet the score threshold, both reads were kept individually. If the semi-global alignment spanned the entire length of either read, a local Smith-Waterman alignment [93] was performed and only the aligned portion of both reads was kept. This prevented adaptor sequences from being included in the merged reads.

2.3.7 Mapping reads to the human genome

We used the human genome to guide local exome assemblies for each of the primate species listed above. Each merged read or pair of unmerged reads were mapped independently to the repeat masked human reference genome (hg18) using cross_match (v1.090518, http://www.phrap.org) with the parameters *–minscore* 25 *–minmatch* 12 *–maxmatch* 20. If a read mapped to more than one location, only the highest scoring match was kept and if there was no single highest scoring match, the read was discarded. Duplicate reads, which may result from PCR amplification or optical artifacts, were indentified as those that mapped to exactly the same chromosomal location in the same orientation. For these reads, only the one with the highest mean quality score was kept.

2.3.8 Assessment of capture efficiency

We assessed the capture efficiency of our method by comparing the sequenced macaque exome with its reference genome (rheMac2). We looked at the correlation between read depth and the number of nucleotide differences or indels and also built a linear model (Table A3) to identify sources of variability in captured target read depth. We used macaque read depth as our response variable and human read depth, number of nucleotide differences, number of indels, GC content and mappability as predictors. As data points we used 155,707 capture targets, for which we could identify orthologs in rheMac2 by best reciprocal BLAST hits (requiring scores to be at least 1.2× greater than the next best alignment). To separate capture and mapping efficiencies, we gave each base in our capture target a 'mappability' score determined from the depth of simulated 76 bp rheMac2 reads, which were aligned to human. Targets were discarded if they were less than 100 bp in length or if they contained a base with a rheMac2 quality score less than 40 (because this could affect the accuracy of our nucleotide difference and indel estimates). In total, 128,914 orthologous capture targets were used to fit the linear model.

2.3.9 Local assembly of mapped reads

Based on their mapped chromosomal locations, overlapping reads were partitioned into groups, which could be assembled independently. Overlap groups that contained more than 500 reads were split into equally sized subgroups to reduce computational time. Overlap groups and subgroups were assembled using phrap (v1.090518, http://www.phrap.org) with parameters that were previously optimized for short read assembly [59]: *–vector_bound* 0 *–forcelevel* 1 *–minscore* 12 *–minmatch* 10 *–indexwordsize* 8. Contigs from split overlap groups were further assembled into longer contigs with a second round of phrap using the same parameters. The final contigs were then mapped back to the repeat masked human reference genome using cross_match with the same parameters as above. We discarded contigs that mapped to a different chromosomal location than the individual reads and discarded contigs that did not map to one location uniquely (requiring the score of the best alignment to be at least 1.2× greater

than the next best alignment). These filters helped us reduce mis-assemblies caused by paralogous sequences.

Consensus calls and quality scores were determined from the contigs that overlapped the target sequences. When phrap created two or more contigs that mapped uniquely to the same chromosomal location, the contig with the highest cross_match score was used as the consensus. Target regions with no mapped contigs were assigned a base "N" with quality Q0, as were non-targeted regions that were present in the CCDS database. A fasta file containing these consensus contig sequences was then generated for each species. A summary of the sequence coverage and assembly of captured regions is found in Table 2.1 (whole captured target) and Table A16 (captured miRNAs).

2.3.10 Mapping unique reads to the assembled consensus

To identify heterozygous sites and assess the quality of the phrap assemblies, we re-mapped the paired-end reads to the assembled contigs. From our cross_match output, we identified uniquely mapping read pairs that aligned to one location with a score at least 1.2× higher than at any other location. We used only read pairs where both individual reads mapped uniquely, and replaced merged unique reads with the individual reads from which they came. We then mapped read pairs to the phrap assembled consensus using BWA 0.5.6 with default parameters for paired-end reads [94]. The alignments were sorted and filtered for duplicates using Picard 1.15 (http://picard.sourceforge.net) and a pileup file was generated with SAMtools 0.1.7a [95], which lists the bases of all the aligned reads for each position in the assembled consensus.

2.3.11 Identification of heterozygous sites

We called genotypes at all consensus sites using the observed bases and quality scores from the pileup file described above. We assigned a genotype quality score to each base using the independent error model described by [96]. For these calculations we used a prior probability of 0.001 of a site being heterozygous, and capped the base quality of individual reads at 30.

For comparison with our own data, we estimated nucleotide diversity in macaques from counts of previously identified segregating sites [79]. In this study a total of 1,476 SNPs were identified in 150,372 base pairs, which were sequenced in 38 Indian macaques and 9 Chinese macaques (94 chromosomes total). Of these SNPs, 486 were observed in both Indian and Chinese macaques and 386 were observed only in the Indian sample. From these numbers we estimated nucleotide diversity by calculating Watterson's population mutation rate estimator, θ [97] and dividing by the number of sequenced bases. Nucleotide diversity for the Chinese and Indian macaques was estimated to be 0.22% and 0.12% respectively.

2.3.12 Target masking

To avoid mis-assembly of paralogous sequences, we masked regions that we could not uniquely map human reads to. We simulated all possible human 76 bp reads (in one orientation), which overlapped the captured target and mapped them to the repeat masked human reference genome using cross_match (v1.090518, http://www.phrap.org) with parameters *–minscore* 68 *–minmatch* 12 *–maxmatch* 20. These parameters allowed reads to be mapped to all locations in the reference genome with one or two mismatches. We then tabulated the number of correctly and incorrectly mapped reads, for each base in the target. Target bases with less than 38 correctly mapped reads (half the expected 76) or more than 10 incorrectly mapped reads were considered 'unmappable'. We also masked coding sequences overlapping segmentally duplicated regions of the human [80,81], chimpanzee [81] or orangutan [82] genomes. This resulted in 1,400,787 bp (4.1%) masked due to potential segmental duplications and 1,711,106 bp (5.0%) masked due to low mappability for a total of 2,938,059 bp (8.6%) masked for downstream analyses.

2.3.13 Assembly quality filtering

From the pileup, we identified and filtered inconsistencies between the assembled consensus sequence and the individual reads. If the pileup consensus base disagreed with the phrap assembled base, we masked that base to an "N" with quality Q0 unless the

pileup contained 8 or more reads, in which case we changed that base to the pileup consensus and flagged it with quality Q1. If the pileup indicated a non-polymorphic insertion or deletion, suggesting an incorrectly placed indel in the phrap assembly, that region and the two flanking bases were masked to an "N" with quality Q0. For heterozygous sites, the pileup base with the majority of reads was used as the consensus, regardless of whether or not it matched the phrap assembled base, and given quality Q1. Exons with excess heterozygosity (≥3 heterozygous sites in any 20 bp window), which suggest paralogous assemblies, were removed completely. Exons with less than 16× read depth for more than half of their sequence were also removed completely.

2.3.14 CDS from exome assemblies

Coding sequences and quality scores were extracted from the quality filtered consensus sequence for each CCDS entry using human coordinates. Gaps were removed from the coding sequences so that the multiple sequence alignment program could place them. Exons or genes missing more than half of their sequence were completely masked or removed to avoid alignment errors. In total, there are 20,091 entries in the CCDS database (version 20080430). When coding sequences for multiple CCDS entries overlapped, only the entry with the longest sequence was kept, resulting in 16,707 unique coding sequences (27,492,897 bp).

2.3.15 CDS from reference genomes

Coding sequences were obtained from the publically available reference assemblies of human (hg18), chimpanzee (panTro2), orangutan (ponAbe2) and macaque (rheMac2). Pairwise whole genome alignments were downloaded from UCSC [98-100] for each of these species and filtered to be best-reciprocal and syntenic as described by McVicker *et al.* [48]. Bases overlapping segmentally duplicated regions of the human [80,81], chimpanzee [81] or orangutan [82] genomes were removed. Target sequences and quality scores were then extracted from the filtered alignments based on the human CCDS coordinates.

2.3.16 Multiple sequence alignments

Coding sequences for the combined set of species were aligned using PRANK (v0.100311) [101] with parameters *–t –F –twice –a –gapext*=0.8 *–kappa*=2.0 *–gaprate*=0.05 and a species tree representing the standard primate phylogeny [102]. Our assembled macaque sequences were included in addition to sequences from the macaque reference genome so that the quality of the two assemblies could be compared. This resulted in a total of 8 sequences in the multiple sequence alignments: hg18, panTro2, ponAbe2, rheMac2, macaque, vervet, colobus monkey and tamarin.

2.3.17 Multiple sequence alignment quality filtering

We extensively filtered low quality single nucleotide differences and indels to produce a set of high quality multiple sequence alignments. For each coding sequence alignment, we compared each non-human primate sequence to the human sequence and masked differences with quality scores less than Q40 in the non-human sequence to an "N". This includes heterozygous sites in both the non-human reference genome sequences and the exome sequences which have quality Q0 and Q1 respectively.

For simplicity, we refer to alignment gaps in the human sequence as insertions and alignment gaps in the non-human sequences as deletions, even though it is unclear what the exact events were or on which lineage they occurred. Indels were retained if they met any one of the following criteria: (1) a minimum quality score ≥Q40 within (insertions) or surrounding (deletions) the indel; (2) a minimum read depth ≥4 within or surrounding the indel; or (3) the presence of a high quality or high read depth indel in another non-human sequence (species confirmation). Indels not meeting one of these criteria were completely removed from all sequences of the alignment if they were insertions or masked if they were deletions.

2.3.18 Quality assessment of assemblies

We assessed the quality of our assemblies using the number of single nucleotide differences and distribution of indel lengths relative to the human reference genome. To

directly compare the number of nucleotide differences between each assembly, we limited our analysis to coding sites that were high quality (≥Q40) in all three of our non-human reference genomes and all four of our exome assemblies resulting in a total of 9,106,235 sites (36.0% of the coding target). We then calculated the proportion of high quality differences between each human reference base and its corresponding base in each non-human sequence. We note that these common high quality sites may be biased towards more conserved regions and thus, are likely to underestimate the average proportion of nucleotide differences between human and non-human coding sequences.

For calculating the proportion of high quality indels, we restricted our set of alignments to 4,637 genes containing greater than 75% high quality sequence in all species. To compute the length of each indel, we combined indels that were less than 15 bp apart in order to account for uncertainty in the alignment. For example, a 2 bp deletion followed closely by a 5 bp insertion would be considered a 3 bp insertion.

Additionally, we calculated the number of high quality nucleotide differences between our macaque exome assembly and the macaque reference genome from 14,924,161 coding sites (59.0%) to get an estimate of pairwise polymorphism (π) in Indian rhesus macaques. This number is likely to be an underestimate of π in Indian rhesus macaques because heterozygous sites are masked in both sequences.

2.3.19 Ortholog filtering for evolutionary analysis

For each orthologous gene set, we filtered out sequences from species that suggested sequencing/assembly errors, alignment errors or gene loss of function. 401 genes were completely removed because their CCDS status is currently listed as "withdrawn" (CCDS version 20100829), resulting in 16,303 genes before ortholog filtering. Of these genes, a species' sequence was removed if it: (1) contained less than 25% high quality sequence; (2) contained a premature stop codon more than 25 bp from the end of the gene sequence; or (3) contained a frameshift disrupting more than 15 bp of sequence. Stop codons within 25 bp of the end of the coding sequence and any sequence following them were masked,

as were frameshifted regions of less than 15 bp in length. 14 human coding sequences contained SECIS elements that direct internal UGA stop codons to be translated as selenocysteines. In these cases, the internal UGA was masked in each sequence of the alignment rather than throwing the gene out completely. Following this filtering, 15,037 genes contained sequence from three or more species and were used in downstream analyses of positive selection (Fig. A2).

2.3.20 Evolutionary analysis

We obtained likelihoods and d_N/d_S estimates for each orthologous set of genes using CODEML from the PAML 4.4 package [103]. Heterozygous and all missing or low quality sites were masked to "N" and treated as missing data with the cleandata=0 option. An unrooted phylogeny corresponding to the accepted relationships between the studied primates [102] was used in each analysis.

To test for selection acting at any point on the phylogeny, we compared a neutral model of $0 \leq d_N/d_S \leq 1$ (M7; model=1, NSsites=7) to a model of selection where an additional class of codons is allowed to have $d_N/d_S > 1$ (M8; model=1, NSsites=8) and performed a likelihood ratio test with a chi-square approximation to calculate *P*-values. *Q*-values [104] were used to set a significance threshold corresponding to an FDR of 10%. CODEML was run with multiple d_N/d_S starting points to ensure convergence of model parameter estimates. 10 genes failed to converge and were removed from the analysis resulting in 15,027 genes tested.

For tests of selection acting on individual lineages, we used CODEML's branch-site models (model=2, NSsites=2), which allow d_N/d_S to vary among codon sites and across branches of the phylogeny. For each branch, we compared a selection model, which allows a class of codons on that branch to have $d_N/d_S > 1$, to a neutral model, which constrains this additional class of sites to have $d_N/d_S = 1$ (fix_omega = 1, omega = 1). A likelihood ratio test was performed and *P*-values were computed from a 50:50 mixture of a chi-square distribution with 1 degree of freedom and a point mass at 0 [91]. The false

discovery rate was estimated with *Q*-values. Details on the numbers of genes tested and significant for each branch are provided in Table A13.

2.3.21 Gene ontology analysis

We used the following procedure to identify GO terms enriched for genes under positive selection. We first assigned each gene a UniProt identifier, using a list of CCDS to UniProt associations downloaded from BioMart [105]. We then assigned genes to GO terms using the UniProt identifiers and the human gene association file downloaded from http://geneontology.org (submission date 09/06/2010). Genes were also assigned to parent GO terms by propagating them up GO's hierarchy of "ISA" relationships.

We next ranked genes by their log-likelihood difference (between CODEML's M7 and M8 models) and performed one-sided Mann-Whitney U tests to determine whether genes in a given GO term ranked significantly higher than genes in a null distribution. The null distribution consisted of all genes assigned to GO terms, excluding those assigned to the term being tested. Following this procedure we reported the most significant GO term and removed its associated genes (to avoid reporting redundant or overlapping terms). This process was repeated iteratively until there remained no significant GO terms at the $p < 0.05$ level. In each iteration, only GO terms with at least 20 remaining genes were tested.

We used the same iterative procedure to identify GO terms enriched for "absent" genes (see Text A2). In this case, we compared counts of present and absent genes for the GO term being tested to those in all other GO terms and assessed significance using a one-sided Fisher's exact test.

2.3.22 Human population differentiation

We examined whether genes with evidence for positive selection in primates also have increased human population differentiation, which may result from recent positive selection. To estimate differentiation between European and African populations, we

assigned each gene in our filtered dataset an average F_{ST} value. These values were calculated using exome sequences from four African individuals and six European individuals in a study by Tennessen *et al* [90]. F_{ST} was calculated for 100 kb genomic regions that contained more than 500 bp of exonic sequence by averaging F_{ST} [31] across all exonic polymorphic sites. If a gene fell into multiple 100 kb regions, the lowest F_{ST} value was used. We tested whether the mean F_{ST} for genes with evidence of positive selection in primates (at 10% FDR) was different from the mean F_{ST} among the remaining genes using a two-sided Mann-Whitney U test. Additionally, we tested the top ten GO categories enriched for genes under positive selection in primates by comparing the mean F_{ST} of genes within the category to the mean F_{ST} of remaining genes with two-sided Mann-Whitney U tests.

2.3.23 Example of gene loss

To find an example of a deleted gene in our non-human primate samples, we considered exons with an average of at least 15 reads per base in human, but fewer than 5 reads per base in a non-human species as candidates for exon loss. Where possible, we confirmed these events in Old World monkeys using the macaque (rheMac2) reference genome. Rachel Diederich in the Thomas lab performed this analysis and is currently developing statistical methods for the detection of gene loss events genome-wide.

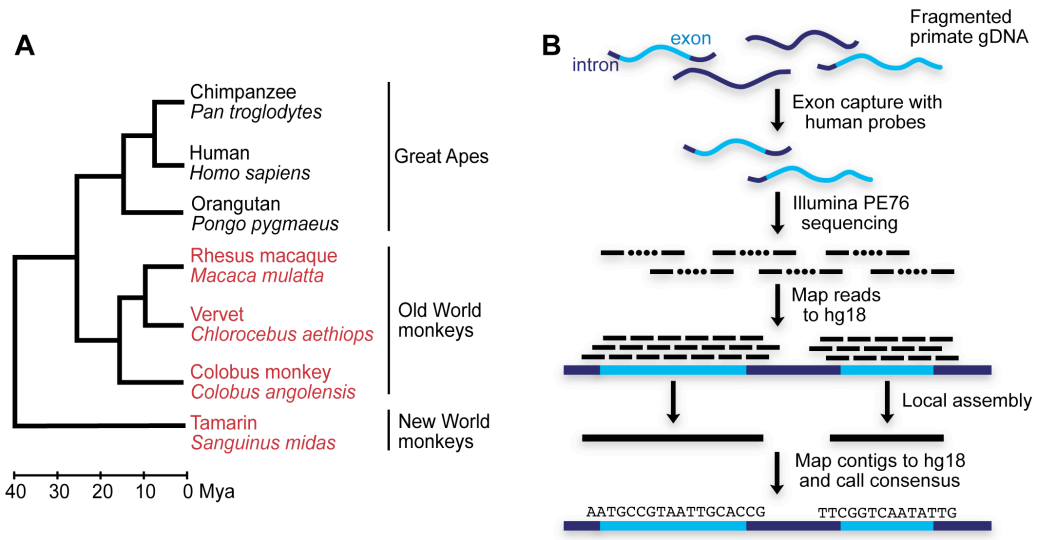

Figure 2.1 Sequence capture and assembly of non-human primates. (A) Phylogeny of primate species used in all analyses. Sequences from the species in red and black are from our exome assemblies and the publically available primate reference genomes, respectively. For rhesus macaque, we generated an exome assembly and compared it to the macaque reference genome to assess the accuracy of our capture and assembly method. The phylogenetic relationship of the species and estimates of their divergence dates are from Goodman, 1999 [102]. (B) Overview of sequencing, mapping and assembly pipeline. Primate genomic DNA is fragmented and protein coding regions are captured using a solution-based hybridization method and sequenced as 76 bp paired-end reads. Reads are mapped to the repeat masked human reference genome using cross_match (v1.090518, http://www.phrap.org). Reads with overlapping mapped chromosomal coordinates are partitioned into groups and assembled independently using phrap (v1.090518, http://www.phrap.org). The resulting contigs are mapped back to the repeat masked human reference genome and consensus bases are called from the highest scoring mapped contigs.

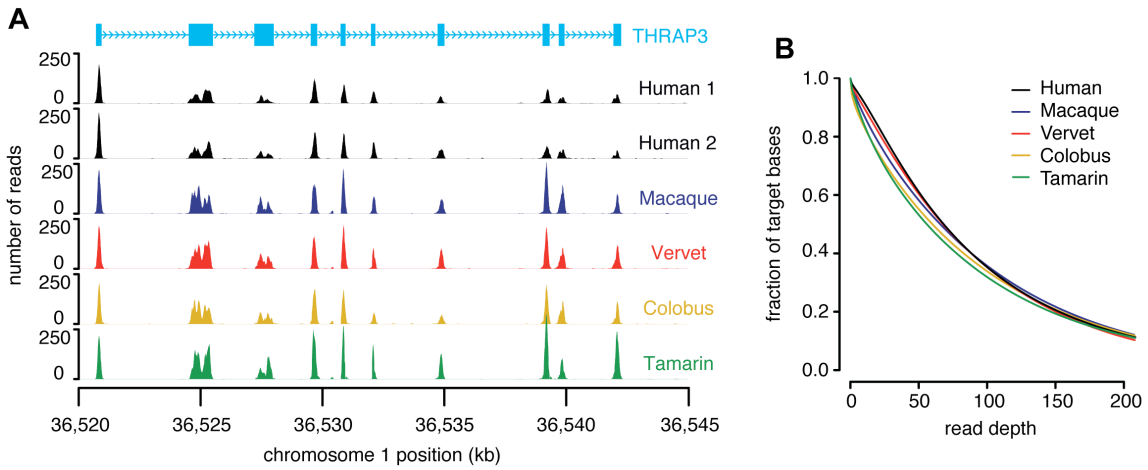

Figure 2.2 Read depth of targeted regions. (A) An example of sequence capture. Read depth for a region on chromosome 1 encompassing the gene *THRAP3* (CCDS405.1) from two human HapMap samples (Human 1: NA12878 and Human 2: NA18967) and four non-human primate samples (macaque, vervet, colobus and tamarin). (B) Cumulative coverage of all targeted bases from one lane of paired-end 76 bp reads mapped to the human reference genome using cross_match for human (NA12878), macaque, vervet, colobus and tamarin.

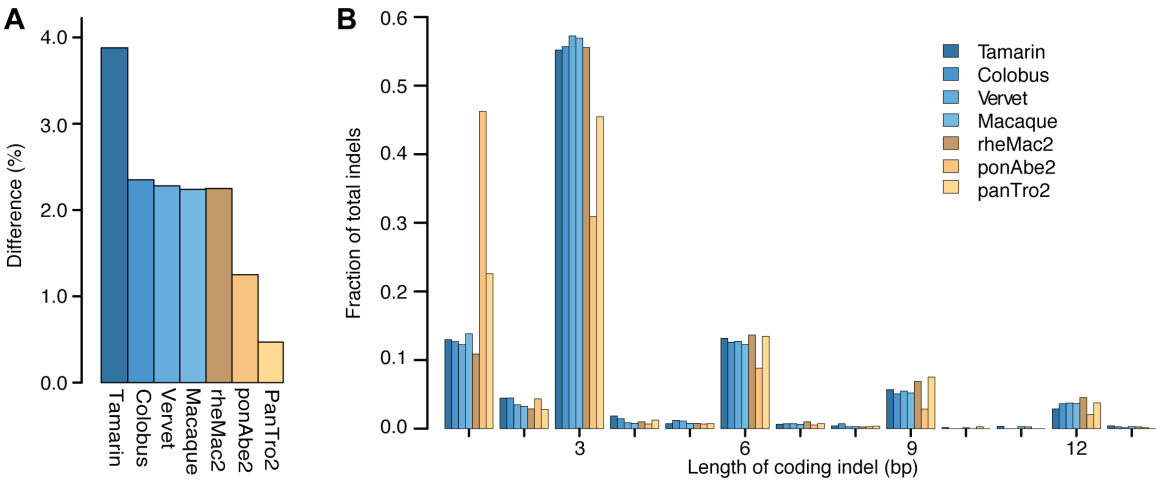

Figure 2.3 Sequence differences and indel lengths in protein coding regions. (A) Coding sequence differences relative to the human reference genome for each assembled exome and non-human primate reference genome, calculated from the 9,106,235 sites that are high quality in all species. (B) Distribution of coding indel lengths from the 4,637 gene alignments where at least 75% of sites have high quality sequence in all species. All indels are relative to the human reference genome. Low quality indels are not included unless their read depth is ≥4 or they are confirmed by a high quality indel in another species. Lengths from indels less than 15 bp apart are combined to account for uncertainty in the alignments. Indel lengths from the exome assemblies of macaque, vervet, colobus and tamarin are blue, while indel lengths from the reference genomes of chimpanzee (panTro2), orangutan (ponAbe2) and macaque (rheMac2) are yellow.

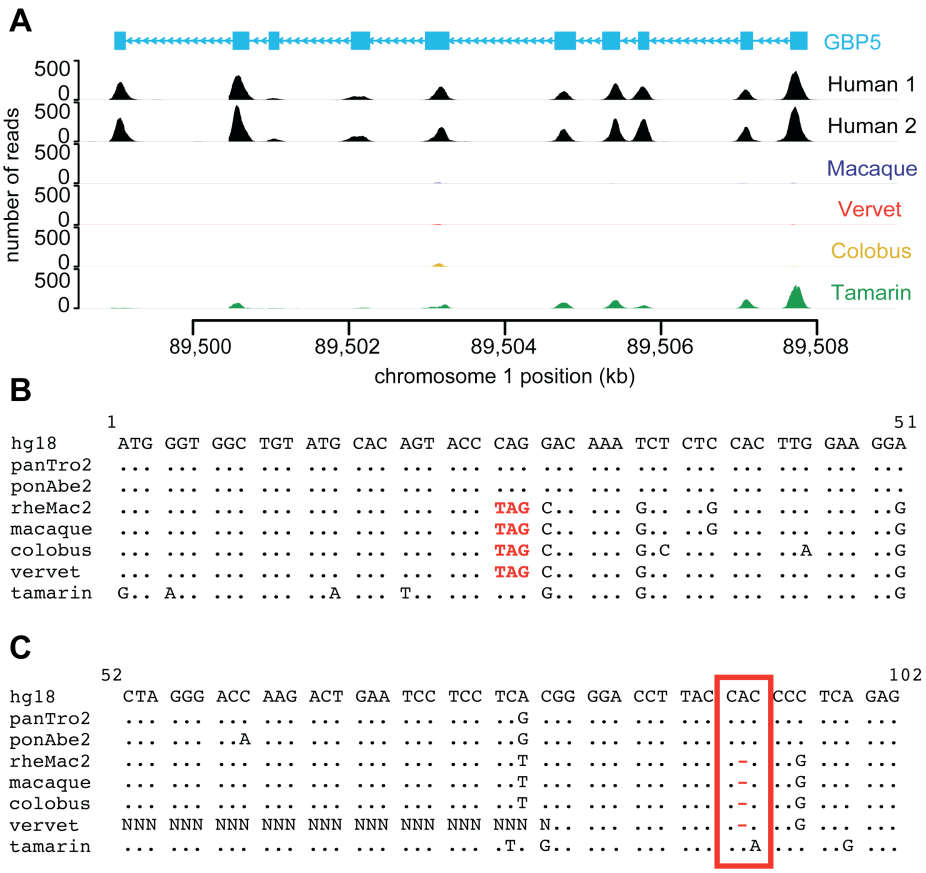

Figure 2.4 Examples of gene loss in Old World monkeys. (A) An example of a gene deletion detected by read depth differences between species. Read depth for *GBP5* (CCDS722.1) is high in human (Human 1: NA12878 and Human 2: NA18967) and tamarin, but absent in macaque, vervet and colobus. The absence of *GBP5* in the macaque exome sequences is supported by the macaque reference genome. (B) Shown is the beginning of the multiple sequence alignment for the gene *SNTN* (CCDS33779.1), which contains a premature stop codon in Old World monkeys. The substitution causing this premature stop codon is high quality (≥Q40) in macaque, vervet and colobus and confirmed by the macaque reference genome. (C) Shown is a portion of the multiple sequence alignment for the gene *CCL14* (CCDS32624.1), which contains a frameshift in Old World monkeys. The red box highlights a one base gap in macaque, vervet and colobus, which disrupts the reading frame of the latter half of the gene. This gap is surrounded by high quality (≥Q40) bases and also supported by the macaque reference genome.

Table 2.1 Sequence coverage of captured target. Summary of captured target sequence coverage for each non-human primate exome and two human HapMap exomes (Human 1: NA12878 and Human 2: NA18967). The total size of the captured target is 34,108,810 bp and includes all well annotated protein coding genes defined by the CCDS database (version 20080430) as well as regions flanking small exons and ~550 miRNAs. Listed for each exome are the number of bases in the target covered by at least one read, the number of bases assembled and the number of bases assembled with Phred consensus quality score ≥40 (Q40; 10^{-4} error rate).

Sample	≥1X coverage (bp)	≥1X coverage (%)	Consensus called (bp)	Consensus called (%)	≥Q40 consensus (bp)	≥Q40 consensus (%)	Avg. coverage
Human 1	33,533,729	98.3	32,539,921	95.4	32,232,123	94.5	82X
Human 2	33,508,928	98.2	32,368,476	94.9	31,891,724	93.5	92X
Macaque	32,995,459	96.7	31,407,528	92.1	30,656,733	89.9	88X
Vervet	33,091,493	97.0	31,268,911	91.7	30,649,250	89.9	86X
Colobus	31,938,787	93.6	30,185,890	88.5	29,298,814	85.9	85X
Tamarin	32,759,816	96.0	31,243,800	91.6	30,019,533	88.0	81X

Chapter 3

Recurrent positive selection in human populations

Positive selection that acts over long evolutionary time periods can be detected by an elevated number of protein coding sequence differences between species. An important question in evolutionary biology is whether positive selection repeatedly targets the same amino acids at different evolutionary time periods along the same lineage or along multiple independent lineages. To investigate this possibility, we tested whether genes that are rapidly evolving in non-human primates are also under positive selection in modern human populations. By comparing the protein coding sequences of 12 non-human primate species, we identified 577 genes containing codons with strong evidence for prior episodes of positive selection. We then looked for signatures of recent adaptive evolution in the same codons using a large exome data set of 2,439 individuals from the NHLBI's Exome Sequencing Project (ESP). Of the putatively positively selected sites (PSSs), 1.14% are fixed for non-synonymous mutations that occurred on the human lineage. This is far greater than the number of fixed non-synonymous mutations at other codons (0.16%) and even exceeds the number of fixed synonymous mutations at four-fold degenerate sites (0.47%). The PSSs have an elevated non-synonymous diversity (0.18%) compared to non-PSSs (0.04%) and synonymous diversity at neutral sites (0.10%). Variants at PSSs are skewed towards higher frequencies than those at non-PSSs, and the differences between PSSs and non-PSSs cannot be explained by regional variation in the mutation rate, hypermutable CpG dinucleotides, GC-biased gene conversion or flanking sequence contexts. These results are evidence that many amino acid sites that have previously undergone positive selection in primates have experienced similar selection pressure in the human lineage and continue to evolve adaptively.

3.1 Introduction

Positive selection increases the frequency and probability of beneficial alleles fixing within a population. Over long evolutionary periods, repeated episodes of positive selection can increase functional divergence between species. A high level of non-synonymous divergence compared to synonymous divergence may therefore signify that recurrent positive selection has acted on a locus (reviewed in [1]).

Genes that have undergone recurrent positive selection are interesting, not only because of their role in adaptation, but also because of their importance to human health. Many genes with the strongest evidence for recurrent positive selection are involved in important biological processes such as fertility and defense against pathogens. In these cases, the recurrent episodes of selection are hypothesized to result from adaptations in one system driving counter-adaptations in another system. For example, interacting proteins on the surfaces of sperm and egg cells are often rapidly evolving [106-112], possibly because of sperm competition and/or sexual conflict over the rate of fertilization [113-116], and many immune genes evolve rapidly in response to coevolutionary arms races with the pathogens they interact with [117-121].

Despite numerous scans to identify genes that have been targeted by recurrent positive selection, little is know about how such selection operates in primates. Studies in *Drosophila* suggest that recurrent positive selection operates across different species and populations [122-125], yet how important recurrent selection is in the recent evolution of human populations is less known. Enard *et al.* identified selective sweeps in a single human from reductions in heterozygosity and found coincident reductions in a chimpanzee, an orangutan and a macaque [126]. However, these sweeps may not overlap any more frequently than by chance [127]. Additionally, genes with evidence for long-term positive selection in primates show increased allele frequency differentiation (F_{ST}) between European American and African American human populations, which suggests that selection in human populations may mirror selection in other lineages or in ancestral populations [128].

Here, we further explore the effect of recurrent positive selection acting both in non-human primates and in current human populations. We identify genes with evidence of long-term positive selection in non-human primates and then ask whether these same genes show evidence for recent selection along the human lineage or in human populations. We find that genes with evidence for positive selection in non-human primates also show elevated rates of non-synonymous substitution on the human lineage and increased non-synonymous diversity in the human population. Using the increased resolution provided by a combination of 12 non-human primates and human polymorphism data from 2,439 individuals, we also find evidence suggesting that positive selection often targets the same codons repeatedly over different evolutionary periods.

3.2 Results

3.2.1 Identification of adaptively evolving genes in non-human primates

To investigate the recent evolutionary trajectory of genes with evidence for long-term or recurrent positive selection in the primate lineage, we first searched for adaptively evolving genes in non-human primates. We collected 15,192 coding sequences from 9 reference genomes and 3 exome assemblies [128], to create a data set containing sequences from 4 Apes (chimpanzee, gorilla, orangutan and gibbon), 4 Old World monkeys (rhesus macaque, baboon, vervet and colobus monkey), 2 New World monkeys (marmoset and tamarin) and 2 Prosimians (bushbaby and mouse lemur) (Fig. 3.1).

For each species, we filtered transcripts to minimize errors in alignment and orthology assignment. We also removed sequences containing frameshifts or premature stop codons, many of which are likely to be neutrally evolving pseudogenes. On average, each transcript in our data set has sequences from 8.9 species and 14,735 transcripts contain sequences from three or more species. Nucleotide sequence divergence between each of the species and the human reference suggests our aligned sequences are of high quality (Fig. B1; Table B1).

We tested these 14,735 transcripts for evidence of positive selection using d_N/d_S based likelihood models that can detect selection acting on a subset of codons [7,11]. Importantly, we withheld the human sequence from this analysis, so we could perform independent comparisons with molecular changes on the human lineage. We found evidence for adaptive evolution targeting 1619 of these transcripts (nominal $p < 0.05$), 716 of which remain significant after correcting for multiple testing (10% FDR) (Table B2). We refer to this set of genes with a 10% FDR as positively selected genes (PSGs). These genes are distributed across all of the human chromosomes (Fig. 3.2) and are enriched in the biological processes of defense and immune response, keratinization and spermatogenesis (Table B3).

3.2.2 Adaptively evolving genes show non-neutral patterns of diversity in human populations

We analyzed patterns of human variation at PSGs and non-PSGs in 2,439 individuals of European (n = 1351) and African (n = 1088) descent from the National Heart, Blood and Lung Institute (NHLBI) sponsored Exome Sequencing Project (ESP) [45]. After removing genes on the X chromosome, genes without polymorphic sites and genes without orthologous chimpanzee sequence (necessary for ancestral sequence reconstruction), 577 PSGs and 11,331 non-PSGs remained. The exons of the PSGs span 1,111,761 bp and were sequenced to an average read depth of 128.6×. The non-PSGs span 18,391,356 bp and were sequenced to a similar read depth (114.3×). In total, we identified 17,226 non-synonymous and 9,390 synonymous variants in PSGs (Table B4).

We hypothesized that positive selection acting recurrently on PSGs would increase the rate of fixation of beneficial mutations on the human lineage. We identified mutations that have become fixed in humans since their divergence from chimpanzees and found a significantly higher rate of non-synonymous fixed differences in PSGs ($d_N = 0.0019$) compared to non-PSGs ($d_N = 0.0010$; $p < 0.001$ by one-sided bootstrap). This difference remains significant even after using synonymous sites to correct for variation in the local substitution rate (Table 3.1).

Positive selection that favors new mutations increases the frequency of derived alleles within a population. If selection acts recurrently on the same genes, then PSGs might have both higher diversity and higher frequency derived alleles. To test this idea, we compared patterns of human polymorphism in PSGs to those from non-PSGs. At non-synonymous sites, the nucleotide diversity of PSGs ($\pi_N = 4.49 \times 10^{-4}$) is significantly higher than non-PSGs ($\pi_N = 2.71 \times 10^{-4}$; $p < 0.001$ by one-sided bootstrap). Neutral diversity at synonymous sites is slightly elevated for PSGs, but does not fully account for this observed difference in non-synonymous diversity (Table 3.1). In addition to this increase in functional diversity, the site frequency spectrum of non-synonymous derived alleles is shifted towards higher frequencies in PSGs than non-PSGs. This is reflected by a higher Tajima's D for PSGs (-2.18 vs -2.31; $p < 0.001$ by one-sided bootstrap). These observations suggest that recent positive selection has acted on genes that have already experienced episodes of positive selection in other primates.

3.2.3 The same amino acid sites show evidence of selection in non-human primates and human populations

To further explore the unusual patterns of human diversity and divergence at PSGs, we looked specifically at amino acid sites that are predicted to be adaptively evolving. We hypothesized that if positive selection targeted these sites multiple times during non-human primate evolution, many of them would be targeted again in recent human evolution. Potentially these sites could contribute disproportionately to the extreme human diversity and divergence of PSGs. We labeled nucleotides within PSGs as positively selected sites (PSSs) or non-positively selected sites (non-PSSs) depending on whether they were predicted to belong to the subset of codons evolving with $d_N/d_S > 1$ in non-human primates (posterior probability ≥ 0.95 for PSSs and posterior probability <0.5 for non-PSSs). Additionally, as putatively neutral sites, we used four-fold degenerate (4FD) sites in non-PSSs. Our final dataset includes 5,651 PSSs, 973,881 non-PSSs and 142,862 4FD sites (Table 3.2).

We used these data to ask whether the elevated rate of fixed differences at PSGs localizes specifically to PSSs. At PSSs, the rate of non-synonymous fixed differences (1.14%) is almost 7 times higher than the rate at non-PSSs (0.16%; $p < 0.001$ by one-sided bootstrap) (Tables B6-7; Fig. 3.3A) suggesting that these sites are responsible for the observed gene-wide patterns. The rate of non-synonymous fixed differences at PSSs also greatly exceeds the rate at 4FD sites (0.47%; $p < 0.001$ by one-sided bootstrap). Since 4FD sites are expected to evolve neutrally, the elevated rate at PSSs cannot be attributed to relaxed selective constraint.

We next examined nucleotide diversity at PSSs and non-PSSs to ask if the elevated diversity of PSGs is primarily due to selection acting on PSSs. A much larger fraction of PSSs are polymorphic (5.1%) compared to non-PSSs (2.5%; $p < 0.001$ by one-sided boostrap) and to 4FD sites (3.3%; $p < 0.001$ by one-sided bootstrap). Based on the degeneracy of the codons, we estimated the rate of non-synonymous and synonymous polymorphism in each class [129] (Table 3.2). The rates of synonymous and non-synonymous polymorphism at PSSs are 4.9% and 5.7% respectively, and greatly exceed those at non-PSSs (2.1% and 4.1% respectively; both $p < 0.001$ by one-sided bootstrap). Interestingly, the rate of synonymous polymorphism at PSSs is significantly greater than the rate at 4FD sites ($p < 0.001$ by one-sided bootstrap), even though both site types are expected to evolve neutrally (Table 3.2). This may indicate that PSSs have a higher cryptic mutation rate [130] or that mutations occur non-independently at PSSs.

In addition to having a larger fraction of sites that are polymorphic, PSSs also have higher nucleotide diversity (π). The non-synonymous nucleotide diversity at PSSs is $\pi_N = 1.8 \times 10^{-3}$, which is significantly greater than the non-synonymous nucleotide diversity at non-PSSs ($\pi_N = 0.40 \times 10^{-3}$; $p < 0.001$ by one-sided bootstrap) and nucleotide diversity at 4FD sites ($\pi = 1.0 \times 10^{-3}$; $p = 0.011$ by one-sided bootstrap) (Fig. 3.3C).

The higher nucleotide diversity at PSSs may be due to positive selection that increases the allele frequency of new (derived) alleles. This hypothesis is supported by the observation that non-synonymous θ_H (nucleotide diversity weighted by the frequency of

derived variants [18]) is also elevated at PSSs ($\theta_H = 3.4 \times 10^{-3}$) compared to non-PSSs ($\theta_H = 0.70 \times 10^{-3}$; $p = 0.002$ by one-sided bootstrap) and 4FD sites ($\theta_H = 2.0 \times 10^{-3}$; $p = 0.097$ by one-sided bootstrap) (Fig. 3.3D; Table B5). To more directly examine differences in allele frequencies at PSSs, we calculated the mean derived allele frequency (DAF) at these and other sites. Overall, PSGs have significantly higher mean DAFs than non-PSGs (Table 3.1) and within these genes the mean DAF at PSSs (5.2%) is higher than non-PSSs (mean DAF = 2.83%; $p = 0.018$ by one-sided bootstrap). While the mean DAF at PSSs is also higher than that at 4FD sites (mean DAF = 4.56%), this difference is not statistically significant ($p = 0.331$ by one-sided bootstrap) (Fig. 3.3E; Table B5).

Non-synonymous PSSs have a higher rate of fixation, higher diversity and a site frequency spectrum that is skewed towards higher derived allele frequencies compared to non-PSSs. While these results are consistent with recurrent positive selection, we also considered several non-selective explanations. To test whether hypermutable CpG dinucleotides [131] can explain the elevated diversity and rates of fixation at PSSs, we conservatively labeled all sites following a C or preceding a G as "CpG context sites". After re-performing our analysis using only non-CpG context sites, we still see a much higher rate of non-synonymous fixed differences at PSSs (0.93%) compared to non-PSSs (0.15%; $p < 0.001$ by one-sided bootstrap) or 4FD sites (0.43%; $p < 0.001$ by one-sided bootstrap). Nucleotide diversity and derived allele frequency also remain higher at PSSs, although these differences are not significant (Fig. B2; Table B6-7).

We next tested whether GC-biased gene conversion (BGC), which increases the transmission of strong alleles (S = G or C) over weak alleles (W = A or T) [132-135], can explain our results. To examine the influence of BGC on PSSs, we separated substitutions and polymorphisms that are potentially affected by BGC (W $\rightarrow$ S) from those that should be largely unaffected (S $\rightarrow$ W, W $\rightarrow$ W, S $\rightarrow$ S). For both categories of substitutions and polymorphisms, the relationship between PSSs, non-PSSs and 4FD sites remain the same (Fig. B3; Table B8-9). This indicates that the elevated rates of substitution and nucleotide diversity cannot be attributed to BGC.

Finally, we considered that PSSs could have a higher mutation rate due to their flanking sequence context (the two flanking nucleotides around each site) [136]. We calculated the genome-wide mean substitution rate between human and macaque for all 16 possible sequence contexts (Table B10) and used these values to estimate expected substitution rates for PSSs, non-PSSs and 4FD sites. While there are significant differences between the expected substitution rates for these site types, the differences are very small and cannot account for the greatly increased diversity and human divergence of non-synonymous PSSs (Fig. B4).

3.3 Discussion

We have found several lines of evidence that suggest that genes that have previously experienced long-term positive selection in non-human primates continue to be affected by positive selection in recent human populations. Positively selected genes (PSGs) are more diverse at functional sites, have higher nucleotide diversity (π_N), and have a site frequency spectrum skewed towards higher frequency derived alleles than non-PSGs. PSGs also have a higher rate of fixed non-synonymous mutations on the human lineage compared to non-PSGs. These patterns are localized to sites that are predicted to be positively selected in non-human primates and do not appear to be due to relaxed purifying selection, hypermutable CpG sites, GC-biased gene conversion or the sequence context of the flanking bases.

Potentially, other sources of mutation rate variation could contribute to the patterns that we observe. The substitution rate in mammalian genomes varies substantially over large scales (*e.g.* hundreds of kilobases) [137-139], but this variation is unlikely to explain our observations because we compare rates at PSSs and non-PSSs that are obtained from the same set of genes (PSGs). Another possibility is that complex local sequence contexts (beyond the immediate flanking bases that we correct for) cause cryptic mutation rate variation. Cryptic mutation rate variation has been proposed as an explanation for the excess co-occurrence of human and chimpanzee SNPs [130], but is poorly understood and is difficult to distinguish from recurrent positive selection. While it is possible that

cryptic mutation rate variation contributes to both elevated divergence and diversity at PSSs, it is unlikely to explain our results completely. Mutation rate variation should not influence the site frequency spectrum and we observe that derived alleles are skewed towards higher frequencies in PSGs (and within PSGs at PSSs compared to non-PSSs). Additionally, cryptic mutaiton rates are unlikely to be elevated in specific categories of genes, and we find that PSGs are enriched for genes involved in defense and keratinization.

We instead propose that these patterns of divergence and diversity at PSSs in humans are due to recurrent positive selection. The increased rate of fixed differences on the human lineage is particularly striking, and indicates that numerous selective events have affected these sites since the separation of human and chimpanzee lineages 4-6 million years ago [140-142]. The increase in human diversity at these sites is more subtle, but it suggests that a small fraction of PSSs are currently experiencing positive selection in modern human populations.

Many of the genes we identified as having experienced long-term adaptive evolution in non-human primates are thought to be involved in co-evolutionary arms races (reviewed in [1]), and it is likely that these arms races have continued along the human lineage. The increased diversity at PSSs could also indicate the action of balancing selection or diversifying selection that favors new mutations. In addition to increasing the diversity at the selected site, balancing selection is also predicted to increase diversity at linked neutral sites [143-145] and could explain the increase in synonymous site diversity we see at PSSs and non-PSSs. Long-term balancing selection does not predict an increase in fixed differences, but short-term episodes of balancing selection could increase the frequency of derived alleles such that they have a higher probability of fixation.

Finally, we want to emphasize the importance of using multiple approaches to study adaptive evolution acting on protein coding regions. By identifying signatures of selection from the divergence of non-human primate sequences as well as from patterns of polymorphism in human populations, we provide independent support that selection is

operating at these genes. If the PSGs we identified using non-human primate divergence simply represent the tail of a neutral distribution (a form of the "winner's curse"), then the divergence along the human lineage and the diversity in human populations should have resembled those from other genes.

3.4 Methods

3.4.1 Description of gene set

Our gene set consited of 33,854 protein-coding transcripts from the RefSeq database [146]. We downloaded RefSeq gene models for the hg19 build of the human reference genome from the UCSC genome browser (refGene.txt; downloaded 07/01/2011). We removed 150 transcripts from the Y chromosome and 1781 transcripts located on the "random" or on alternate haplotype chromosomes. We also removed 1,205 transcripts with duplicated RefSeq identifiers and 16,411 transcripts that either were not targeted for exome sequencing or that overlapped another transcript. When multiple transcripts overlapped, only the transcript with the highest fraction of bases targeted for exome sequencing was retained. Our final gene set contained 15,192 protein-coding transcripts, which corresponds to 25,160,340 bp of human sequence.

3.4.2 Non-human primates used

We obtained coding sequences for these 15,192 genes from 12 non-human primates that have publically available reference genomes or exome assemblies. These 12 species are distributed throughout the primate phylogeny and include chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), gibbon (*Nomascus leucogenys*), macaque (*Macaca mulatta*), baboon (*Papio hamadryas*), vervet (*Chlorocebus aethiops*), colobus monkey (*Colobus angolensis*), tamarin (*Sanguinus midas*), marmoset (*Callithrix jacchus*), bushbaby (*Otolemur garnettii*) and mouse lemur (*Microcebus murinus*). The phylogenetic relationship of these species is shown in Fig. 3.1.

3.4.3 CDS from reference genomes

We downloaded coding sequences for chimpanzee (panTro3), gorilla (gorGor3), orangutan (ponAbe2), gibbon (nomLeu1), rhesus macaque (rheMac2), baboon (papHam1), marmoset (calJac3), bushbaby (otoGar1) and mouse lemur (micMur1) from their publically available reference genomes [3,38,39,41,147]. For each of these species, we downloaded a pair of 'netted' and 'chained' whole-genome pair-wise alignments from the UCSC genome browser in April of 2012. Each pair of alignments consisted of one alignment that used hg19 as the reference and one alignment that used hg19 as the query [98]. We processed these alignments so that only best-reciprocal blocks remained and we removed short alignment blocks that fell below the following length thresholds: panTro3 and gorGor3: 2000 bp; ponAbe2 and nomLeu1: 1000 bp; papHam1, calJac3 and rheMac2: 250 bp; micMur1 and otoGar1: 100 bp. We grouped adjacent blocks together into 'chains' if they aligned to the same chromosome with a consistent orientation and ordering and then filtered out blocks in chains where the total length of blocks in the chain fell below the following thresholds: panTro3 and gorGor3: 2500 bp; papHam1 and nomLeu1: 2000 bp; micMur1 and ponAbe2: 1500 bp; calJac2 and rheMac2: 1000 bp; otoGar1: 500 bp. These thresholds were chosen because we found that they removed suspect alignment blocks with very high divergence, while still retaining most of the aligned sequences.

We extracted coding sequences for the RefSeq genes from the filtered pair-wise alignments using the human coordinates. Sites which were not present in the filtered alignments were assigned a base "N".

3.4.4 CDS from exome assemblies

For vervet, colobus monkey and tamarin, we used paired-end 76 bp exome sequencing reads to assemble their protein coding sequences [128]. We used a reference-guided assembly method that we previously described [128] to assemble these sequences, with the following modifications: (1) We used RefSeq instead of CCDS gene models; (2) We replaced the hg18 human reference with hg19; (3) After mapping the exome reads to the

human reference, we used paired-end rather than single-end read filters; and (4) We removed several of the read depth and segmental duplication filters because they were overly conservative.

3.4.5 Multiple sequence alignments

Coding sequences containing Ns for more than half of their sequence length were removed to avoid multiple sequence alignment and orthology assignment errors (Table B11). The remaining coding sequences from 13 primate species (including human) were then aligned using PRANK (v0.100311) [101] with parameters *-F -twice -a -gapext=0.8 -kappa=2.0 -gaprate=0.05*. Indels were removed or retained according to the criteria previously described [128] and a species' sequence was completely removed if it showed evidence of a loss of protein coding ability (i.e. their sequence contained a frameshift disrupting more than 15 bp of sequence or a premature stop codon >25 bp from the end of the sequence).

The number of transcripts remaining for each species is listed in Table B11. The Prosimians (bushbaby and mouse lemur) retained the least amount of sequence due to their genomes being sequenced to an average read depth of only 2× [147]. The exome assemblies are also missing a large fraction of sequence because many of the exons and transcripts defined by the RefSeq gene models were not targeted for capture. To assess the quality of our final multiple sequence alignments, we calculated the average nucleotide divergence for each species with human (Table B1; Fig. B1). The coding divergence estimates recapitulate the accepted species phylogeny (Fig. 3.1) and species within any clade outside of the Apes show similar divergences from human as other members of their clade. Additionally, there is no noticeable difference between the divergence estimates between human and either exome assembled species or species with reference genomes. Taken together, this suggests our coding sequence alignments are of high quality.

3.4.6 Analysis of positive selection in non-human primates

We were interested in identifying genes showing evidence of positive selection acting in non-human primates, so we removed the human sequence from all of our multiple sequence alignments. We then removed any genes that did not meet a minimum sequence requirement of three species, which reduced our set of genes from 15,192 to 14,735. We then tested each gene for evidence of positive selection using likelihood models that allow d_N/d_S to vary across codons from the CODEML program in the PAML 4.4 package [103]. For all analyses, an unrooted phylogeny corresponding to the accepted phylogenetic relationships between primates was used and Ns were treated as missing data using the cleandata=0 option. A neutral model where codons are only allowed to vary between $0 \leq d_N/d_S \leq 1$ (M7; model = 1, NSsites = 7) was compared to a model of positive selection where an additional class of codons can have $d_N/d_S > 1$ (M8; model = 1, NSsites = 8). A likelihood ratio test was then performed between these two models and *P*-values were calculated using a chi-square approximation. *Q*-values were then used to estimate a false discovery rate (FDR) [104]. Using a FDR threshold of 10%, we identified 716 positively selected genes (PSGs) and 14,019 non-positively selected genes (non-PSGs).

3.4.7 Description of human polymorphism data

For our human population data, we used filtered protein coding variant calls from 2,439 individuals sequenced by the National Heart, Lung, and Blood Institute (NHLBI) sponsored Exome Sequencing Project (ESP) [45]. This population included 1351 individuals of European ancestry and 1088 individuals of African ancestry. We also obtained unfiltered invariant sites and filtered them to closely match the filters already applied to the variant calls. For these invariant sites, we removed sites not targeted in all individuals, assigned an individual as missing at a particular site if their read depth (DP) was < 10 or their genotype quality (GQ) was < 30. Finally, we removed sites where more than 10% of individuals (≥25 individuals) were missing.

3.4.8 Polymorphism analysis of positively selected genes

To investigate the possibility of recent events of recurrent selection, we compared patterns of human polymorphism and divergence at PSGs to non-PSGs. Of the 716 PSGs and 14,019 non-PSGs, we excluded the following: (1) genes located on the X chromosome (47 for PSGs and 551 for non-PSGs); (2) genes where fewer than half of their bases were targeted for exome sequencing (9 for PSGs and 136 for non-PSGs); (3) genes missing sequence from chimpanzee so that the human-chimpanzee ancestral sequence is unable to be reconstructed (79 for PSGs and 1968 for Non-PSGs); and (4) genes with no polymorphic sites (3 for PSGs and 33 for non-PSGs). In total, this left 577 PSGs and 11,331 non-PSGs used for further analyses.

For our first analysis, we calculated the rate of fixation of synonymous (d_S) and non-synonymous (d_N) mutations on the human lineage for PSGs and non-PSGs. To determine whether a human base was ancestral or derived, we used our multiple sequence alignments that included at least the human and chimpanzee sequences to reconstruct the sequences of the human-chimpanzee common ancestor. For each internal node in our primate phylogeny, we reconstructed the ancestral sequences with an empirical Bayesian approach implemented in CODEML from the PAML 4.4 package [103] by setting the RateAncestor=1 option. For each site, we chose as the ancestral base the nucleotide with the maximum posterior probability (MAP). We then labeled bases that were invariant in our human population as derived if they differed from the reconstructed human-chimpanzee ancestral base. We further labeled these derived bases as non-synonymous or synonymous depending on whether or not the ancestral amino acid differed from the derived amino acid.

We then estimated the total numbers of potentially synonymous and potentially non-synonymous sites using the degeneracy of the codons as follows. For synonymous sites, we summed all 4-fold degenerate sites and one-third of all 2-fold degenerate sites, and for non-synonymous sites, we summed all non-degenerate sites and two-thirds of all 2-fold degenerate sites [129]. Finally, we computed d_N and d_S from the concatenated sequence of all PSGs and all non-PSGs.

For concatenated PSGs and non-PSGs, we then calculated the per site nucleotide diversity as the mean heterozygosity across all sites (π), synonymous sites (π_S) and non-synonymous sites (π_N). We also calculated Tajima's D values [14] for all sites, synonymous sites and non-synonymous sites. Finally, we examined the unfolded site frequency spectrum of derived non-synonymous and synonymous variants for PSGs and non-PSGs.

To assess the variance of our test statistics, we performed 1000 bootstrap replicates by re-sampling from the combined set of filtered PSGs and non-PSGs with replacement. For each bootstrap sample we recalculated the test statistics for concatenated PSGs and non-PSGs. We estimated 95% confidence intervals by taking the 2.5% and 97.5% percentiles of the bootstrap distributions as the lower and upper bounds. Our prior hypothesis was that PSGs would be more diverse or more diverged than non-PSGs and 4FD sites so we computed one-sided *P*-values as the proportion of replicates where the PSG mean was greater than the non-PSG mean or the 4FD site mean. Statistically significant differences between site frequency spectrums were assessed with two-sided Mann-Whitney U tests.

3.4.9 Polymorphism analysis of positively selected sites

We next examined human polymorphism and divergence at amino acid sites inferred to be adaptively evolving. For the 577 PSGs, we used the Bayes empirical Bayes method from CODEML's M8 model and a posterior probability (postPr) of ≥ 0.95 to identify codons containing positively selected sites (PSSs). Codons in these same PSGs with postPr < 0.5 were labeled as non-positively selected sites (non-PSSs). 47,887 sites from codons with weak evidence for being positively selected ($0.50 \leq$ postPr < 0.95) were not used in our analysis. We also defined four-fold degenerate (4FD) sites as a set of putatively neutrally evolving sites from non-PSSs. All three of these site classes were obtained from the same set of PSGs so that differences in the local mutation rate are minimized.

In all analyses, we focused on only non-synonymous PSSs and non-PSSs because they are the most likely to cause functional changes that selection might act on. For non-synonymous PSSs, non-synonymous non-PSSs and 4FD sites, we evaluated the following: the rate of fixed differences on the human lineage (d_N or d); nucleotide diversity (π_N or π); nucleotide diversity weighted by the frequency of derived variants (θ_H [18]); mean derived allele frequency (DAF); the unfolded site frequency spectrum of derived variants; and Tajima's D [14]. To assess variance, we performed 1000 bootstrap replicates by re-sampling PSGs with replacement, partitioning the sampled bases into PSS, non-PSS and 4FD site classes and then recalculating each test statistic. We generated 95% confidence intervals by taking the 2.5% and 97.5% percentiles of the bootstrap distrbution as the lower and upper bounds. We computed one-sided *P*-values as the proportion of replicates where the PSS mean was greater than the non-PSS mean or the 4FD site mean.

CpG sites have a higher mutation rate than other dinucleotides due to the deamination of methylated cytosines [148,149]. If CpG sites are found more often at PSSs than at non-PSSs, then the increase in diversity and divergence at PSSs could be the result of mutation rate differences rather than selection. To control for this, we classified all polymorphic and invariant sites as either part of a CpG or not. For polymorphic sites, if either allele could form a CpG, then it was labeled as one. For the analysis of fixed differences, we conservatively classified sites as part of a CpG-context if they were preceded by a C or followed by a G (*i.e.* Cp? or ?pG). Comparisons were then made between PSSs, non-PSSs and 4FD sites in CpG and non-CpG contexts.

In GC-biased gene conversion (BGC), weak (A or T) bases are more often converted to strong (G or C) bases, as mismatches in heteroduplexes are resolved during recombination [132-135]. This process favors the transmission of G and C alleles and increases their allele frequency within a population. If BGC affects PSSs more than non-PSSs, it could explain their differences in diversity and divergence. We compared rates of polymorphism and substitution at PSSs, non-PSSs and 4FD sites using the subset of rates that are potentially inflated by BGC. We considered substitution rates from weak to strong bases (W $\rightarrow$ S) as "BGC" rates and all other substitution rates as "non-BGC" rates.

Likewise polymorphisms were considered to be "BGC" polymorphisms if their ancestral base was either an A or T and their derived base was either a G or C.

We also considered whether sequence contexts other than CpGs could be responsible for the diversity and divergence patterns we see at PSSs. We calculated the expected substitution rate for PSSs, non-PSSs and 4FD sites based on their sequence composition. First, we used human-macaque whole genome alignments to calculate the substitution rate for each of 16 possible dinucleotide sequence contexts, where a dinucleotide sequence context is defined as the base on either side of a given site. For example, the dinucleotide sequence context of the central base in ACG is AC. We then scanned all possible triplets, discarding triplets where the human and macaque bases differed in either their first or last position, and calculated the substitution rate for each context as the fraction of differences between human and macaque at the context's central base (Table B10). Expected substitution rates for PSSs, non-PSSs and 4FD sites were then calculated by taking the mean substitution rate for contexts surrounding each base in PSSs, non-PSSs or 4FD sites. Differences in expected substitution rates were statistically assessed with *P*-values from one-sided bootstraps.

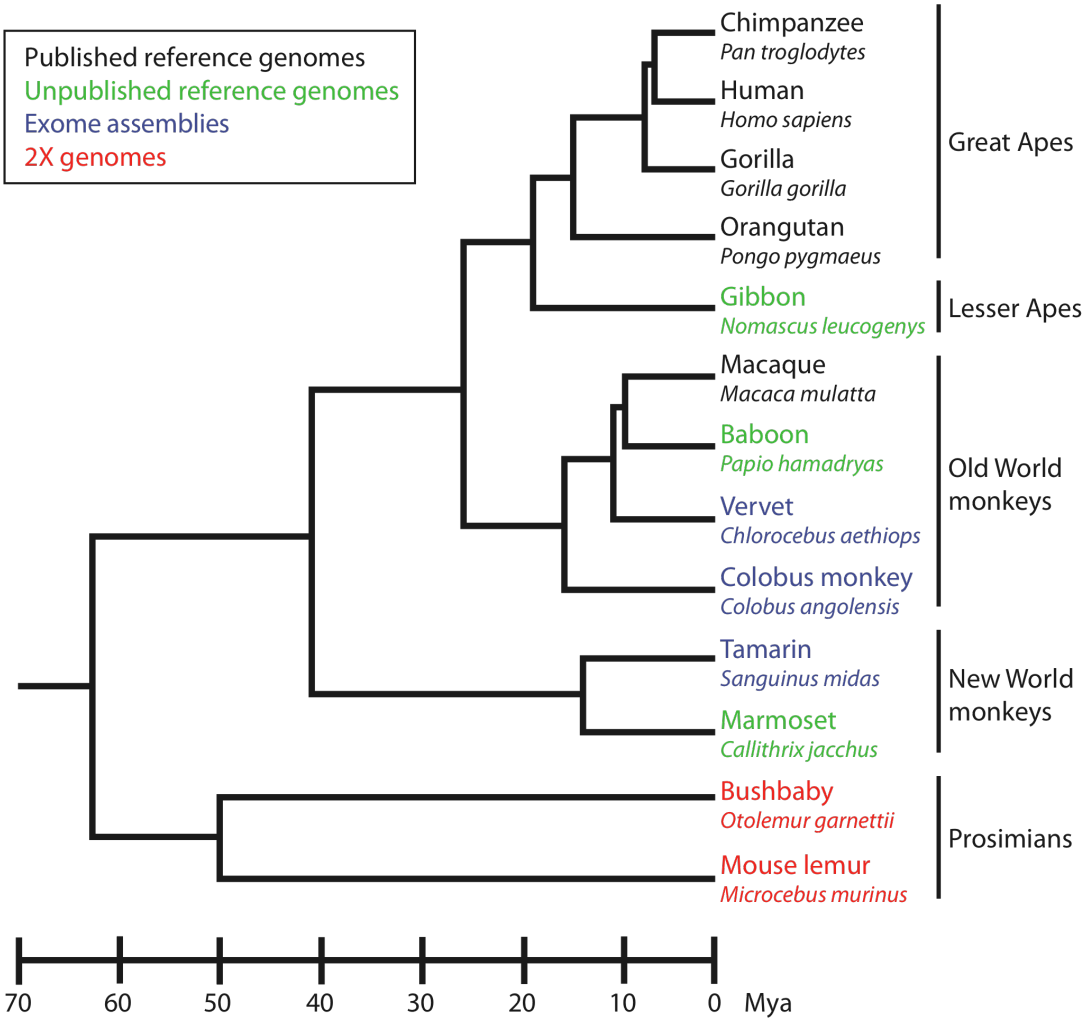

Figure 3.1 Phylogeny of primate species used to identify genes targeted by positive selection. Sequences from the species in red are from low coverage (2×) genomes [147], sequences in blue are from exome assemblies [128], and the remaining sequences are from ongoing or completed genome projects. Divergence times and topology are from Goodman, 1999 [102]. Note that the human sequence was not used to identify genes under positive selection.

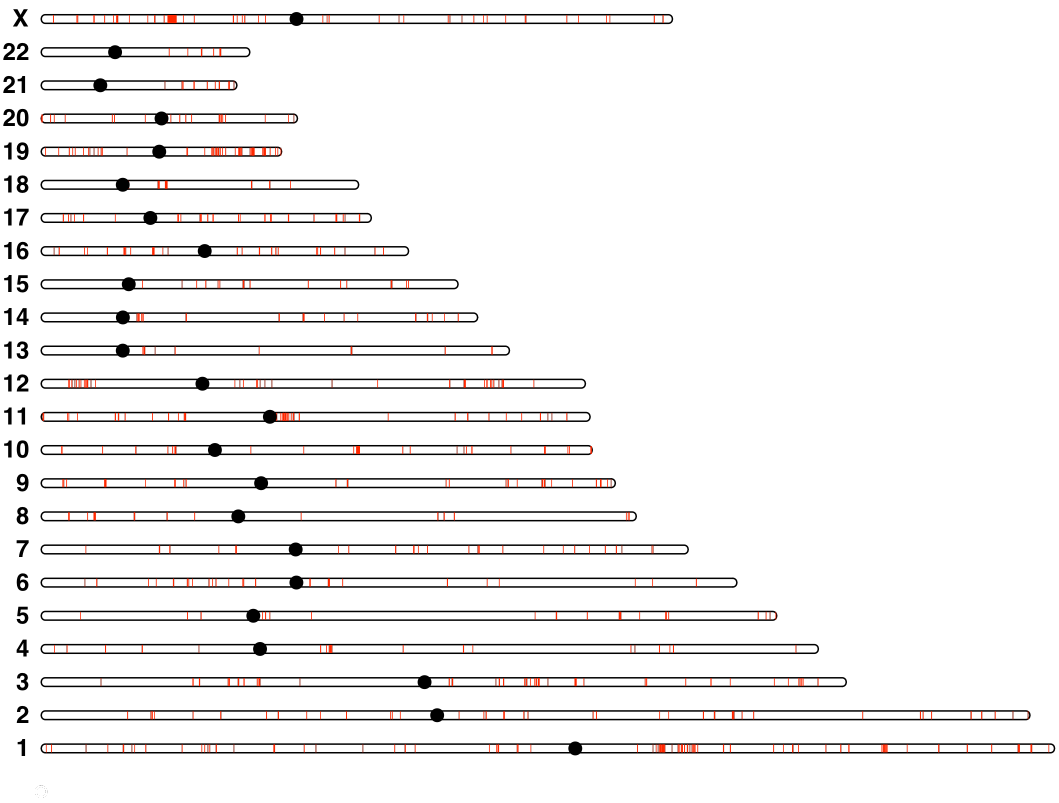

Figure 3.2 Genomic map of genes under long-term positive selection in non-human primates. The red lines indicate the genomic locations of 716 positively selected genes (PSGs) on 22 human autosomes and the human X chromosome. Even though PSGs are shown on human chromosomes, the human sequence was not used in tests of positive selection.

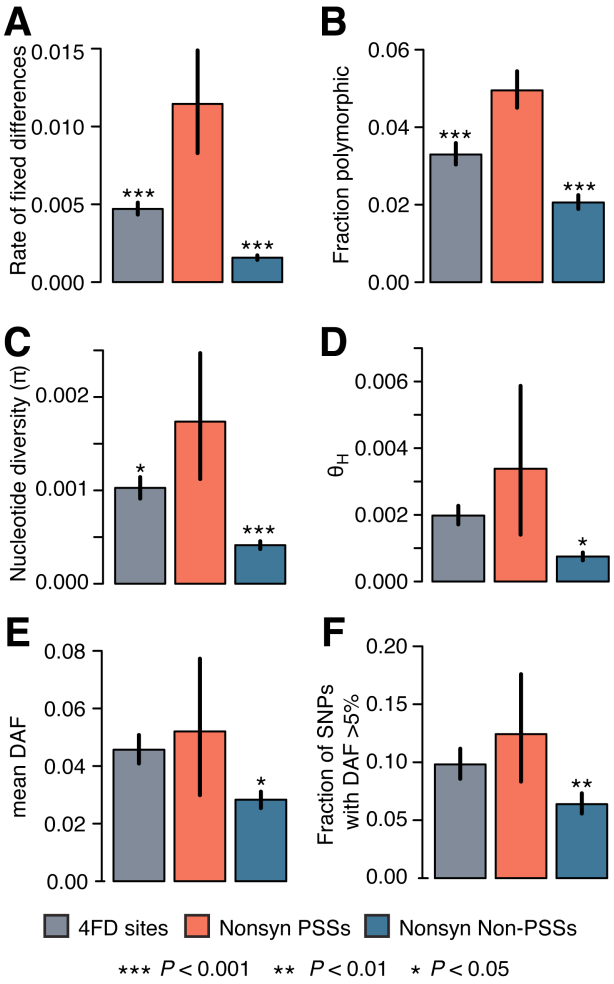

Figure 3.3 Patterns of human divergence and diversity at non-synonymous PSSs, non-synonymous non-PSSs and 4FD sites. We identified positively selected sites (PSSs) and non-PSSs in 577 positively selected genes (PSGs). We then calculated summary statistics of non-synonymous divergence and non-synonymous diversity at each of these site classes. We also calculated the same statistics for synonymous divergence and diversity at four-fold degenerate (4FD) sites, which are likely to be neutrally evolving. Asterisks indicate significant differences with the non-synonymous rate at PSSs from one-sided bootstrap *P*-values. (A) Rate of fixed differences on the human lineage since divergence with chimpanzee. (B) Fraction of total sites that are polymorphic in 2,439 human individuals. (C) Nucleotide diversity, π. (D) Nucleotide diversity weighted by the frequency of derived variants, θ_H [18]. (E) Mean derived allele frequency (DAF) of polymorphic sites. (F) Fraction of polymorphic sites with a DAF >5%.

Table 3.1 Summary of human divergence and diversity analyses for PSGs and non-PSGs from 2,439 human individuals. In each analysis in this table, we compared the summary statistics computed from positively selected genes (PSGs) with those computed from non-positively selected genes (non-PSGs). Summary statistics were computed for all sites, non-synonymous sites, and four-fold degenerate (4FD) sites. *P*-values were computed using one-sided bootstraps.

Test		PSGs	Non-PSGs	*P*-value
Rate of fixed differences	all	0.0028	0.0020	< 0.001
	non-synonymous (d_N)	0.0019	0.0010	< 0.001
	synonymous (d_S)	0.0060	0.0056	0.032
	4FD	0.0049	0.0047	0.133
d_N/d_S	--	0.309	0.174	< 0.001
Fraction polymorphic	all	0.0259	0.0232	0.030
	non-synonymous	0.0215	0.0185	0.004
	synonymous	0.0415	0.0397	0.223
	4FD	0.0334	0.0336	0.501
Nucleotide diversity (π)	all	0.0006	0.0004	< 0.001
	non-synonymous (π_N)	0.0004	0.0003	< 0.001
	synonymous (π_S)	0.0012	0.0010	< 0.001
	4FD	0.0010	0.0009	0.032
π_N/π_S	--	0.367	0.271	< 0.001
θ_H	all	0.0012	0.0008	< 0.001
	non-synonymous	0.0009	0.0004	< 0.001
	synonymous	0.0024	0.0021	0.019
	4FD	0.0020	0.0018	0.095
Mean DAF	all	0.0353	0.0260	< 0.001
	non-synonymous	0.0308	0.0183	< 0.001
	synonymous	0.0434	0.0387	0.004
	4FD	0.0451	0.0405	0.024
Fraction of SNPs with DAF > 5%	all	0.0766	0.0589	< 0.001
	non-synonymous	0.0675	0.0589	< 0.001
	synonymous	0.0935	0.0830	0.055
	4FD	0.0967	0.0887	0.155
Tajima's D	all	-2.18	-2.31	< 0.001
	non-synonymous	-2.25	-2.41	< 0.001
	synonymous	-2.04	-2.15	< 0.001
	4FD	-2.01	-2.10	0.010

Table 3.2 Summary of polymorphism data for PSSs, Non-PSSs and 4FD sites from 2,439 human individuals. We labeled 979,532 nucleotides from 577 positively selected genes (PSGs) as positively selected sites (PSSs; posterior probability ≥ 0.95) or non-positively selected sites (non-PSSs; posterior probability < 0.5). Additionally, subset of non-PSSs were labeled as four-fold degenerate (4FD) sites if they fell within four-fold degenerate codon positions. For each type of site, this table shows the total number of bases, the number of bases that are polymorphic in 2,439 human individuals and the number of bases that are fixed for a derived allele since the divergence of humans and chimpanzees. The total numbers of synonymous and non-synonymous sites were estimated using the degeneracy of the codons [129].

Site type	Function	Total bp	Variable sites		Fixed differences	
			bp	%	bp	%
PSSs	all	5651	289	5.11	61	1.08
	synonymous	1281	73	5.70	11	0.86
	non-synonymous	4370	216	4.94	50	1.14
Non-PSSs	all	973881	24357	2.50	2454	0.25
	synonymous	213686	8745	4.09	1257	0.59
	non-synonymous	760195	15612	2.05	1197	0.16
4FD sites	all	142862	4701	3.29	673	0.47

Chapter 4
Conclusions and Future Directions

In this dissertation, I developed methods to capture, sequence and assemble the protein coding regions (exomes) of species that do not have reference genomes. To demonstrate the utility of this approach, I created high quality exome assemblies for several non-human primates. After combining these data with existing reference genomes, and a new large-scale human exome data set, I explored the role of recurrent adaptive evolution in primates and human populations. In this chapter, I comment on the importance of high quality reference genomes and the utility of exome sequencing. I also discuss the implications of some of my results and propose directions for further investigation.

4.1 Importance of high quality reference genomes

To answer many fundamental evolutionary questions, we require a large number of sequences from diverse species to achieve sufficient statistical power [65,66]. Increasing the number of genome sequences improves statistical power to detect evolutionary events such as positive selection, but the quality of the sequences used is also critically important. Low quality sequences are more difficult to analyze and can bias estimates or create false positives [68].

Since the completion of the human genome reference sequence in 2003 [150], genome sequences have been assembled for a considerable number of mammalian species. While the quality of the human genome sequence is very high, the quality of the other mammalian reference genomes is quite variable and is, in almost all cases, considerably lower. The human genome was assembled from Sanger-sequenced BAC clones, but to

reduce costs, most subsequent projects adopted a whole-genome shotgun strategy. Recently, the shift from Sanger sequencing to next-generation sequencing (NGS) has also impacted the quality of assembled genomes. Ideally all genomes would be sequenced and assembled to a high standard of quality, but there exists a trade-off between sequencing many species at low quality and sequencing a few species at high quality.

To study the processes of mutation and selection broadly across genomes, the NHGRI approved the sequencing of 29 mammalian genomes to 2× coverage. This project was completed in 2011 [147], but as a result of their low coverage, these assembled genomes are less accurate and less complete than those with higher read depth. Using these assemblies, it is difficult to distinguish genes from pseudogenes and impossible to estimate rates of lineage-specific adaptation or gene loss [151].

These issues presented challenges in my analyses of positive selection in primates. The majority of protein coding sequences obtained from species with 2× genomes were discarded (>70%) due to incomplete coverage or frameshifts. The rates of frameshifting mutations in these genomes are much higher than expected, which indicates that most of them are likely to be sequencing or assembly errors. Given this high assembly error rate, I did not attempt to identify lineage-specific bursts of positive selection because the number of false positives would have been unacceptably high.

High quality reference genomes are required to gain a complete and accurate picture of adaptive evolution acting in primates and other species. For many low coverage genomes, there are plans to sequence them more deeply. The added sequencing depth will come from NGS technologies [152,153], however current NGS read lengths are still not able to span long repetitive regions or resolve large segmental duplications [151]. In the future, NGS technologies will likely generate reads that are similar in length and quality to Sanger reads. It will then be possible to assemble high quality reference genomes quickly, completely and accurately.

4.2 Whole-genome *vs.* exome sequencing

As sequencing costs plummet, the shelf life of exome sequencing is often called into question. By focusing sequencing efforts on only the known protein coding regions of the genome, we miss variants and substitutions in regulatory regions, as well as unannotated genes and many structural variants. At what point does it make sense to switch from exome sequencing to whole genome sequencing to avoid losing this non-coding information?

While whole genome sequencing identifies variants in regulatory regions, the functional consequences of these variants are hard to interpret. Variants in protein coding regions are easier to understand, because those that change the encoded amino acid sequence are much more likely to affect the protein's function. Research to understand the functional effect of variants in regulatory regions (*e.g.* "promoter bashing or "enhancer bashing") is ongoing [154,155], but protein coding variation remains easier to decipher.

Comparisons between whole genome and whole exome studies have found that exome sequencing identifies more variants in protein coding regions because of its higher read depth [156]. Some regions of the genome are more difficult to sequence due to sequence characteristics such as GC content and secondary structure. While these issues affect both exome and whole genome sequencing, exome sequencing can target more probes to problematic regions, making them more likely to be sequenced.

Currently, it costs about 6× more to sequence a whole genome than to sequence a whole exome with the same coverage [157]. This cost difference is misleading, however, because it does not consider the additional sequencing time, data storage and data processing that is required for whole genome sequencing. The true cost difference is therefore closer to 15:1 [157].

For these reasons, exome sequencing will likely remain the less expensive and more interpretable strategy for several years. For a fixed budget, the lower cost of exome

sequencing allows researchers to sequence more individuals or more species than whole genome sequencing.

4.3 Future directions for positive selection in primates

The work in this dissertation points to several directions for future study and raises some key questions about adaptive evolution. In the last couple of years, the number of primate species with available protein coding sequences has greatly increased due to the release of several new primate reference genomes and the exome assemblies that I described in Chapter 2. Despite the increase in protein coding data, we are still statistically underpowered to detect positive selection in primates. The median tree length for the currently available primates (human, chimpanzee, orangutan, macaque, vervet, colobus monkey and tamarin) is 0.08 nucleotide substitutions per codon, yet simulations have shown that our power to detect positive selection is maximized at 1.1 nucleotide substitutions per codon [65]. To reach this level of sequence divergence, many more primate species need to be targeted for sequencing, ideally ones from diverse clades of the phylogeny.

The methods I developed for sequencing and assembling protein coding regions from species without reference genomes are broadly applicable. These methods could be utilized to sequence more primate species as described above or could be applied to species from other parts of the tree of life (providing a reference genome from a closely related species is available). Recently, new methods for detecting copy number variants from exome sequencing reads have been developed [158], which make it possible to study copy number changes using the exome data I described in Chapter 2. Additionally, these methods could be extended to sequence the protein coding regions of many individuals from a single species, rather than one individual from multiple species. This would be useful for studying very recent adaptation in other species.

One area of research that warrants more investigation is distinguishing different types of selection from each other and distinguishing selection from non-selective biological

processes. One such process is GC-biased gene conversion, which increases the transmission of strong alleles (G and C) over weak alleles (A and T) during recombination [131,134,135]. Current methods to detect positive selection between species or within populations do not take GC-biased gene conversion into account, and it is likely that many genes affected by this process have been falsely identified as adaptively evolving [159]. I correct for biased gene conversion in my analysis of recurrent positive selection in humans in Chapter 3, but this process was not taken into account when I identified positively selected genes in Chapter 2.

It is also difficult to distinguish amino acid sites that are under recurrent positive selection in humans from sites that have a high cryptic mutation rate [130]. Understanding the complex sequence contexts responsible for variation in the cryptic mutation rate is key to distinguishing between these two hypotheses. An analysis of neutral variation should also help in this regard, because positively selected sites are expected to reduce linked neutral variation, while neutral sites with higher mutation rates are not.

LIST OF REFERENCES

1. Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39: 197-218.

2. Kimura M (1983) The neutral theory of molecular evolution. Cambridge, UK: Cambridge University Press.

3. Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316: 222-234.

4. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

5. Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol (Amst) 15: 496-503.

6. Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA 94: 7712-7718.

7. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929-936.

8. Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16: 1315-1328.

9. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568-573.

10. Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46: 409-418.

11. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449.

12. Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 23-35.

13. Wiehe TH, Stephan W (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster. Mol Biol Evol 10: 842-854.

14. Tajima F (1989) The effect of change in population size on DNA polymorphism. Genetics 123: 597-601.

15. Fu YX (1996) New statistical tests of neutrality for DNA samples from a population. Genetics 143: 557-570.

16. Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915-925.

17. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693-709.

18. Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

19. Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765-777.

20. Nielsen R, Williamson S, Kim Y, Hubisz M, Clark AG et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15: 1566-1575.

21. Williamson SH, Hubisz M, Clark AG, Payseur BA, Bustamante C et al. (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3: e90.

22. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.

23. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. Plos Biol 4: e72.

24. Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513-1524.

25. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

26. Tang K, Thornton KR, Stoneking M (2007) A new approach for using genome scans to detect recent positive selection in the human genome. Plos Biol 5: e171.

27. Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci USA 103: 135-140.

28. Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res 70: 155-174.

29. Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175-195.

30. Slatkin M, Wiehe T (1998) Genetic hitch-hiking in a subdivided population. Genet Res 71: 155-160.

31. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. Evolution 38: 1358-1370.

32. Wright S (1950) Genetical structure of populations. Nature 166: 247-249.

33. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-654.

34. Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.

35. Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164: 1229-1236.

36. Bakewell MA, Shi P, Zhang J (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci USA 104: 7489-7494.

37. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A et al. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302: 1960-1963.

38. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.

39. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV et al. (2011) Comparative and demographic analysis of orang-utan genomes. Nature 469: 529-533.

40. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. Plos Biol 3: e170.

41. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I et al. (2012) Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169-175.

42. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

43. Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711-722.

44. Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

45. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64-69.

46. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289-1303.

47. Hudson RR, Kaplan NL (1995) Deleterious background selection with recombination. Genetics 141: 1605-1617.

48. McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 5: e1000471.

49. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A et al. (2011) Classic selective sweeps were rare in recent human evolution. Science 331: 920-924.

50. Pritchard JK, Di Rienzo A (2010) Adaptation - not by sweeps alone. Nat Rev Genet 11: 665-667.

51. Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70: 369-383.

52. Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL et al. (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310: 1782-1786.

53. Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA et al. (2007) cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. Cell 131: 1179-1189.

54. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF et al. (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74: 1111-1120.

55. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31-40.

56. Sanger F (1988) Sequences, sequences, and sequences. Annu Rev Biochem 57: 1-28.

57. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46.

58. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26: 1135-1145.

59. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J (2010) Parallel, tag-directed assembly of locally derived short sequence reads. Nat Methods 7: 119-122.

60. Kitzman JO, MacKenzie AP, Adey A, Hiatt JB, Patwardhan RP et al. (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol 29: 59-63.

61. Turner EH, Ng SB, Nickerson DA, Shendure J (2009) Methods for genomic partitioning. Annu Rev Genomics Hum Genet 10: 263-284.

62. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E et al. (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307: 1072-1079.

63. Clark AG, Hubisz M, Bustamante C, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15: 1496-1502.

64. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424: 788-793.

65. Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18: 1585-1592.

66. Eddy SR (2005) A model of the statistical power of comparative genome sequence analysis. Plos Biol 3: e10.

67. Fletcher W, Yang Z (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol 27: 2257-2267.

68. Mallick S, Gnerre S, Muller P, Reich D (2009) The difficulty of avoiding false positives in genome scans for natural selection. Genome Res 19: 922-933.

69. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D et al. (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4: 903-905.

70. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27: 182-189.

71. Tewhey R, Warner JB, Nakano M, Libby B, Medkova M et al. (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 27: 1025-1031.

72. Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Methods 6: 315-316.

73. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276.

74. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42: 30-35.

75. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR et al. (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106: 19096-19101.

76. Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ et al. (2010) Whole exome capture in solution with 3 Gbp of data. Genome Biol 11: R62.

77. Kosiol C, Vinař T, Da Fonseca R, Hubisz M, Bustamante C et al. (2008) Patterns of Positive Selection in Six Mammalian Genomes. PLoS Genet 4: e1000144.

78. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M et al. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19: 1316-1323.

79. Hernandez RD, Hubisz M, Wheeler DA, Smith DG, Ferguson B et al. (2007) Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. Science 316: 240-243.

80. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061-1067.

81. Cheng Z, Ventura M, She X, Khaitovich P, Graves T et al. (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437: 88-93.

82. Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z et al. (2009) A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457: 877-881.

83. Pelak K, Shianna KV, Ge D, Maia JM, Zhu M et al. (2010) The characterization of twenty sequenced human genomes. PLoS Genet 6:

84. Meader S, Hillier LW, Locke D, Ponting CP, Lunter G (2010) Genome assembly quality: assessment and improvement using the neutral indel model. Genome Res 20: 675-684.

85. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

86. Filip LC, Mundy NI (2004) Rapid evolution by positive Darwinian selection in the extracellular domain of the abundant lymphocyte protein CD45 in primates. Mol Biol Evol 21: 1504-1511.

87. Wyckoff GJ, Wang W, Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. Nature 403: 304-309.

88. Zhang J, Webb DM (2004) Rapid evolution of primate antiviral enzyme APOBEC3G. Hum Mol Genet 13: 1785-1791.

89. Ford MJ (2001) Molecular evolution of transferrin: evidence for positive selection in salmonids. Mol Biol Evol 18: 639-647.

90. Tennessen JA, Madeoy J, Akey JM (2010) Signatures of positive selection apparent in a small sample of human exomes. Genome Research 20: 1327-1334.

91. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472-2479.

92. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443-453.

93. Smith T, Waterman M (1981) Identification of common molecular subsequences. J Mol Biol 147: 195-197.

94. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

95. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

96. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18: 1851-1858.

97. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256-276.

98. Chiaromonte F, Yap VB, Miller W (2002) Scoring pairwise genomic sequence alignments. Pac Symp Biocomput 115-126.

99. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA 100: 11484-11489.

100. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R et al. (2003) Human-mouse alignments with BLASTZ. Genome Res 13: 103-107.

101. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci USA 102: 10557-10562.

102. Goodman M (1999) The genomic record of Humankind's evolutionary roots. Am J Hum Genet 64: 31-39.

103. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.

104. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100: 9440-9445.

105. Smedley D, Haider S, Ballester B, Holland R, London D et al. (2009) BioMart--biological queries made easy. BMC Genomics 10: 22.

106. Civetta A (2003) Positive selection within sperm-egg adhesion domains of fertilin: an ADAM gene with a potential role in fertilization. Mol Biol Evol 20: 21-29.

107. Gao Z, Garbers DL (1998) Species diversity in the structure of zonadhesin, a sperm-specific membrane protein containing multiple cell adhesion molecule-like domains. J Biol Chem 273: 3415-3421.

108. Lee YH, Ota T, Vacquier VD (1995) Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol Biol Evol 12: 231-238.

109. Metz EC, Palumbi SR (1996) Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. Mol Biol Evol 13: 397-406.

110. Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 20: 18-20.

111. Swanson WJ, Vacquier VD (1995) Extraordinary divergence and positive Darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. Proc Natl Acad Sci USA 92: 4957-4961.

112. Swanson WJ, Yang Z, Wolfner MF, Aquadro CF (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci USA 98: 2509-2514.

113. Gavrilets S (2000) Rapid evolution of reproductive barriers driven by sexual conflict. Nature 403: 886-889.

114. Parker GA, Partridge L (1998) Sexual conflict and speciation. Philos Trans R Soc Lond, B, Biol Sci 353: 261-274.

115. Rice WR, Holland B (1997) The enemies within: intergenomic conflict, interlocus contest evolution (ICE), and the intraspecific Red Queen. Behav Ecol Sociobiol 41: 1-10.

116. Stockley P (1997) Sexual conflict resulting from adaptations to sperm competition. Trends Ecol Evol (Amst) 12: 154-159.

117. Burrows JM, Bromham L, Woolfit M, Piganeau G, Tellam J et al. (2004) Selection pressure-driven evolution of the Epstein-Barr virus-encoded oncogene LMP1 in virus isolates from Southeast Asia. J Virol 78: 7131-7137.

118. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335: 167-170.

119. Lynn DJ, Lloyd AT, Fares MA, O'Farrelly C (2004) Evidence of positively selected sites in mammalian alpha-defensins. Mol Biol Evol 21: 819-827.

120. Sawyer SL, Emerman M, Malik H (2004) Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. Plos Biol 2: E275.

121. Yeager M, Hughes AL (1999) Evolution of the mammalian MHC: natural selection, recombination, and convergent evolution. Immunol Rev 167: 45-58.

122. Andolfatto P (2007) Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Genome Res 17: 1755-1762.

123. Bachtrog D (2008) Similar rates of protein adaptation in Drosophila miranda and D. melanogaster, two species with different current effective population sizes. BMC Evol Biol 8: 334.

124. Jensen JD, Bachtrog D (2010) Characterizing recurrent positive selection at fast-evolving genes in Drosophila miranda and Drosophila pseudoobscura. Genome Biol Evol 2: 371-378.

125. Macpherson JM, Sella G, Davis JC, Petrov DA (2007) Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. Genetics 177: 2083-2099.

126. Enard D, Depaulis F, Roest Crollius H (2010) Human and non-human primate genomes share hotspots of positive selection. PLoS Genet 6: e1000840.

127. Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20: 393-402.

128. George RD, McVicker G, Diederich R, Ng SB, MacKenzie AP et al. (2011) Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. Genome Res 21: 1686-1694.

129. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418-426.

130. Hodgkinson A, Ladoukakis E, Eyre-Walker A (2009) Cryptic variation in the human mutation rate. Plos Biol 7: e1000027.

131. Ehrlich M, Wang RY (1981) 5-Methylcytosine in eukaryotic DNA. Science 212: 1350-1357.

132. Duret L, Arndt PF (2008) The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet 4: e1000071.

133. Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10: 285-311.

134. Eyre-Walker A (1993) Recombination and mammalian genome evolution. Proc Biol Sci 252: 237-243.

135. Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet 23: 273-277.

136. Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci USA 101: 13994-14001.

137. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-562.

138. Gaffney DJ, Keightley PD (2005) The scale of mutational variation in the murid genome. Genome Res 15: 1086-1094.

139. Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ et al. (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res 13: 13-26.

140. Benton MJ, Donoghue PC (2007) Paleontological evidence to date the tree of life. Mol Biol Evol 24: 26-53.

141. Satta Y, Hickerson M, Watanabe H, O'hUigin C, Klein J (2004) Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and BAC end sequences. J Mol Evol 59: 478-487.

142. Yang Z, Rannala B (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol 23: 212-226.

143. Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. Genetics 120: 831-840.

144. Kaplan NL, Darden T, Hudson RR (1988) The coalescent process in models with selection. Genetics 120: 819-829.

145. Nordborg M (1997) Structured coalescent processes on different time scales. Genetics 146: 1501-1514.

146. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40: D130-5.

147. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476-482.

148. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in Escherichia coli. Nature 274: 775-780.

149. Wang RY, Kuo KC, Gehrke CW, Huang LH, Ehrlich M (1982) Heat- and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. Biochim Biophys Acta 697: 371-377.

150. Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.

151. Green P (2007) 2x genomes--does depth matter? Genome Research 17: 1547-1549.

152. Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16: 545-552.

153. Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferriera S et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci USA 103: 11240-11245.

154. Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D et al. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol 27: 1173-1175.

155. Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB (2009) Stepwise modification of a modular enhancer underlies adaptation in a Drosophila population. Science 326: 1663-1667.

156. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R et al. (2011) Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29: 908-914.

157. Biesecker LG, Shianna KV, Mullikin JC (2011) Exome sequencing: the expert view. Genome Biol 12: 128.

158. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M et al. (2012) Copy number variation detection and genotyping from exome sequence data. Genome Research .

159. Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N et al. (2010) Detecting positive selection within genomes: the problem of biased gene conversion. Philos Trans R Soc Lond, B, Biol Sci 365: 2571-2580.

Appendix A

Supplemental Information for Chapter 2

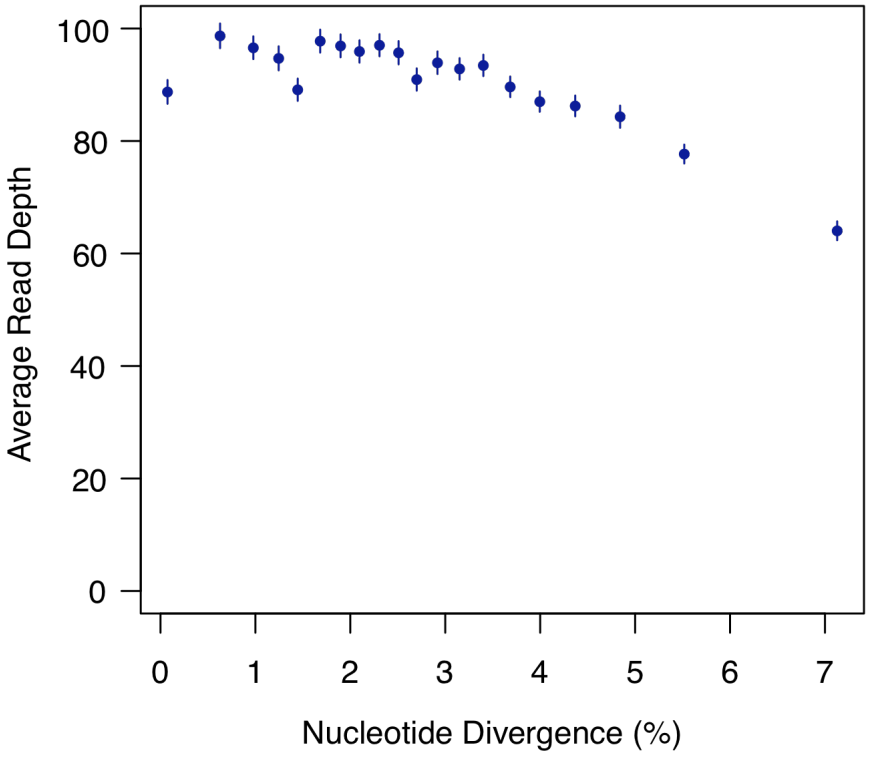

Figure A1 Average macaque read depth of captured targets versus human-macaque sequence divergence. We calculated sequence divergence between the human and macaque reference genome for 134,401 orthologous targets, which contained no indels. We placed targets into 20 bins of equal size based on their human-macaque sequence divergence and then calculated the mean macaque read depth of the targets within each bin.

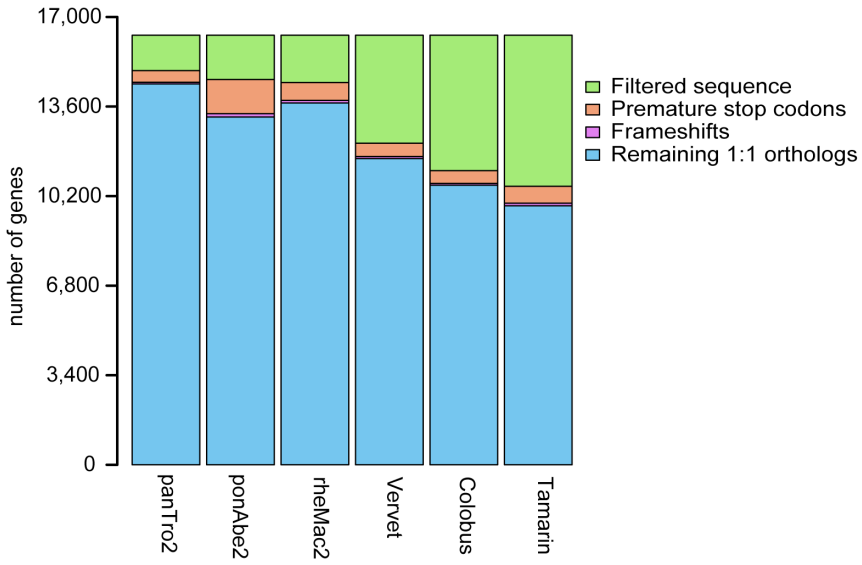

Figure A2 Filtering of orthologous gene alignments. This figure is a summary of the ortholog filtering described in the methods.

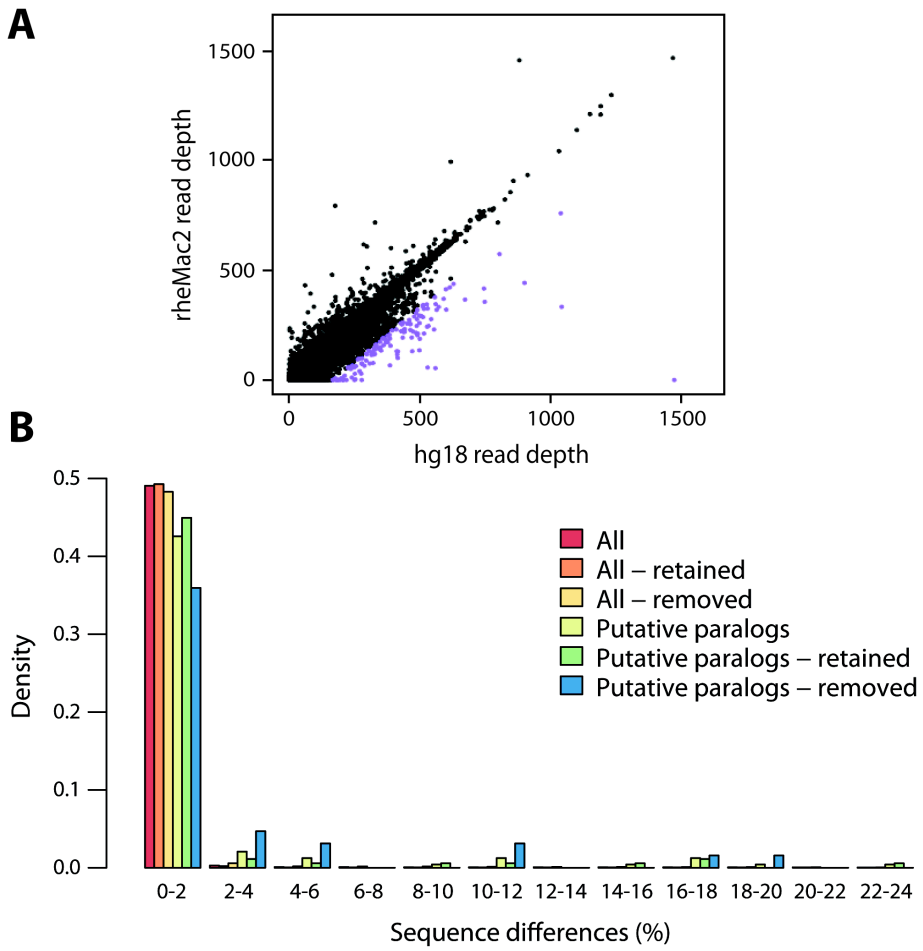

Figure A3 Identification of putative paralogous targets and nucleotide differences between assembled and reference macaque sequences. (A) We identified putative paralogous captured targets that may be susceptible to mis-assembly by comparing the depth of macaque reads mapped to the macaque reference genome (rheMac2) to the depth of the same reads mapped to the human reference genome (hg18). We consider as putative paralogs (purple), the 137 targets where the hg18 read depth is at least two standard deviations greater than the rheMac2 read depth. (B) Histogram of nucleotide differences between the macaque assembled exome and the macaque reference genome for 153,546 targeted regions that we assembled and uniquely mapped to the macaque reference genome using cross_match (v1.090518, http://www.phrap.org). Targets were further categorized by whether they were putative paralogs and by whether they are retained or removed following filtering for segmental duplications, extreme heterozygosity and missing sequence (see Methods for more details).

Table A1 Summary of read merging and mapping. Summary of read merging and mapping for each non-human primate exome and two human HapMap exomes (Human 1: NA12878 and Human 2: NA18967). Listed for each exome are the number of paired-end 76 bp reads (PE76) generated, the number of overlapping read pairs merged into single longer reads, the number of read pairs discarded due to gaps in the overlapping portion, the total number of individual reads after read merging, the number of individual reads discarded due to low quality (>10% Ns), the number of reads uniquely mapped to the repeat masked human reference genome with cross_match (v1.090518, http://www.phrap.org) and the number of uniquely mapped reads remaining (and used for assembly of exomes) after filtering for PCR and optical duplicates.

Sample	PE76 read pairs	Merged PE76 read pairs	Discarded PE76 read pairs	Total individual reads	Discarded low quality	Uniquely mapped	After duplicate filtering
Human 1	37,520,750	20,945,392	142,933	53,810,242	628,917	47,615,608	40,589,744
Human 2	47,628,042	25,921,661	337,984	68,658,455	2,821,253	52,136,229	47,538,474
Macaque	46,373,786	22,949,875	356,160	69,085,377	929,405	55,206,587	44,741,572
Vervet	47,568,368	20,979,776	301,513	73,553,934	1,300,948	57,365,966	45,683,191
Colobus	46,834,936	22,969,487	357,702	69,984,981	1,732,508	55,369,898	43,484,635
Tamarin	46,044,421	21,469,749	390,156	69,838,781	1,429,809	52,555,753	42,067,032

Table A2 Assembly statistics. Summary of phrap assemblies for each non-human primate exome and two human HapMap exomes (Human 1: NA12878 and Human 2: NA18967). Listed for each exome are the total number of groups of overlapping reads, the number of overlap groups containing more than one read that were assembled using phrap (v1.090518, http://www.phrap.org), the number of assembled contigs, the average length of the assembled contigs, the number of assembled contigs that mapped uniquely to the repeat masked human reference genome with cross_match (v1.090518, http://www.phrap.org) and the number of discarded contigs that mapped to a different location than their individual reads (off location contigs).

Sample	Overlap groups	Groups assembled	Contigs	Avg. contig length (bp)	Mapped contigs	Off location contigs
Human 1	1,988,564	660,218	590,549	214	570,036	13,284
Human 2	2,644,570	981,000	902,670	170	884,566	16,244
Macaque	1,611,560	593,794	592,954	215	570,219	30,016
Vervet	1,631,204	606,930	604,158	224	578,120	28,221
Colobus	1,565,107	637,026	642,079	207	617,462	29,884
Tamarin	1,485,380	569,441	727,059	210	684,602	33,741

Table A3 Linear regression model of macaque read depth. To assess what factors influence captured target read depth, we fit a linear model using observations from 128,914 captured target regions. The response variable, macaque read depth, is in units of reads/base. $R^2 = 0.674$.
β – Slope estimate. Standard errors of the slope estimates are in parentheses.
β* – Normalized slope estimate. Predictor variables are normalized to have mean 0 and standard deviation 1 so that the slope estimates are comparable. The response variable, macaque read depth, is not normalized. Standard errors of the slope estimates are in parentheses.
P-value – *P*-value from a two-sided *t*-test with null $\beta = 0$.
Macaque mappability – The number of 76 bp simulated macaque reads uniquely mapped to each base in the human target sequence.

Predictor	β	β*	***P*-value**
(Intercept)	39.7 (1.21)	92.1 (0.13)	$< 2.0 \times 10^{-16}$
Human read depth (reads/base)	0.93 (0.0019)	67.2 (0.14)	$< 2.0 \times 10^{-16}$
Nucleotide differences (%)	-7.72 (0.080)	-13.3 (0.14)	$< 2.0 \times 10^{-16}$
Indels (%)	-4.18 (0.22)	-2.54 (0.13)	$< 2.0 \times 10^{-16}$
Macaque mappability (reads/base)	-0.060 (0.0093)	-0.89 (0.14)	8.46×10^{-11}
GC content (%)	7.10 (1.29)	0.80 (0.14)	3.87×10^{-08}

Table A4 Genomic features of captured and not captured macaque targets. We examined the following genomic features for the 155,707 human targets with best-reciprocal orthologs in the macaque genome: the number of nucleotide differences between human and macaque, the number of indel bases between human and macaque and the GC content. A target was considered "captured" if more than half of the human targeted bases were covered by at least one sequencing read.

	No. targets	**No. bases**	**Differences (%)**	**Indel bases (%)**	**GC (%)**
Captured targets	154,596	29,479,474	3.07	0.169	49.9
Not captured targets	1,111	162,361	3.92	0.500	51.7

Table A5 High quality coding sequence differences relative to the human reference genome. Coding sequence differences relative to the human reference genome calculated from high quality (≥Q40) sequence for each assembled exome and for the non-human primate reference genomes of chimpanzee (panTro2), orangutan (ponAbe2) and rhesus macaque (rheMac2). Coding sequences are non-overlapping transcripts (longest transcript retained from overlapping transcripts) from the 20080430 version of the CCDS database [78] totaling 27,583,228 bp. 2,655,850 bp of this sequence was not targeted by our capture method and contributes to the difference between the number of assembled exome bases and the number of bases from the non-human primate reference genomes. Heterozygous sites and sites overlapping known segmental duplications were excluded from all species' sequences. Exons with excess heterozygosity, low read depth or more than half their sequence filtered were removed from the exome assemblies. Common ≥Q40 sites – 9,106,235 sites that are high quality in all species.

	All sites			Common ≥Q40 sites	
Species	**≥Q40 consensus (bp)**	**≥Q40 differences (bp)**	**Difference (%)**	**≥Q40 differences (bp)**	**Difference (%)**
panTro2	23,904,584	128,495	0.54	42,482	0.47
ponAbe2	23,139,970	329,972	1.43	113,938	1.25
rheMac2	21,925,828	574,515	2.62	204,446	2.25
Macaque	17,459,375	439,792	2.52	203,818	2.24
Vervet	18,185,119	465,882	2.56	208,018	2.28
Colobus	16,197,614	428,758	2.65	213,885	2.35
Tamarin	15,346,639	642,281	4.19	353,732	3.88

Table A6 Summary of high quality coding indel lengths. This table summarizes the number of indels with lengths that are multiples of three (3*n*) in gene alignments that contain greater than 75% high quality sequence in all species. Low quality indels were only included if their read depth was sufficiently high (≥4) or they were confirmed in another species. Indels less than 15 bp apart were combined to account for uncertainty in the alignment.
Gaps in human – regions of the alignments causing a gap in the human sequence; appears as an insertion in the other species.
Gaps in other species – regions of the alignments causing a gap in the other sequence; appears as a deletion in the other species.

	Gaps in human			**Gaps in other species**		
Species	**Total**	**3*n***	**3*n* (%)**	**Total**	**3*n***	**3*n* (%)**
panTro2	152	121	79.6%	183	123	67.2%
ponAbe2	442	234	52.9%	471	207	43.9%
rheMac2	452	384	85.0%	487	391	80.3%
Macaque	373	290	77.7%	417	341	81.8%
Vervet	376	309	82.2%	449	357	79.5%
Colobus	359	295	82.2%	480	370	77.1%
Tamarin	584	487	83.4%	904	690	76.3%

Table A7 Numbers of genes showing evidence of positive selection at several FDR thresholds.

1% FDR	5% FDR	10% FDR
52	93	157

Table A8 Complete list of genes tested for evidence of positive selection in primates ranked by significance. Table is provided as a separate supplemental file. Shown for each gene is the number of species, the nominal *P*-value from a chi-square approximated likelihood ratio test between CODEML's M7 and M8 models [103], the estimated false discovery rate calculated by *Q*-values [104] and the average F_{ST} [90].
dNdS_REF – A '1' indicates a gene previously identified as subject to positive selection from the Rhesus Macaque Sequencing and Analysis Consortium [3], which used a similar d_N/d_S method and sequences from human, chimpanzee and macaque.

Table A9 Genes identified by the Rhesus Macaque Genome Sequencing and Analysis Consortium, but not identified by our study. Shown for each gene is the nominal *P*-value determined from a similar analysis using human, chimpanzee and macaque sequences [3] (*), the nominal *P*-value and *Q*-value [104] from our analysis and the number of species used in our analysis. Genes are ranked by their significance in the other study. The "Note" column indicates genes that were not tested in our analysis because we could not confidently obtain enough sequences (1), the models in CODEML did not converge (2), or they were not targeted by our probe set (3). In total, the other study identified 67 genes under positive selection at an FDR of 10%, 15 of which we also identified at the same FDR threshold.

CCDS	Gene name	Chr	*P*-value*	*P*-value	*Q*-value	No. species	Rank	Note
41683.1	*KRTAP5-8*	11	6.20E-16	7.45E-03	3.01E-01	3	361	
42614.1	*LILRB1*	19	7.20E-14			1		1
11896.1	*DSG1*	18	1.10E-10	4.52E-03	2.33E-01	3	280	
14217.1	*MAGEB6*	X	5.30E-08			2		1
7831.1	*MRGPRX4*	11	5.60E-08			2		1
	COL6A5 (FLJ35880)	3	1.70E-07					3
	LOC442247	6	3.80E-07					3
5007.1	*CGA*	6	1.20E-06	1.01E-02	3.52E-01	7	408	
	KRTAP5-4	11	2.70E-06					3
12231.1	*ICAM1*	19	2.70E-06	2.24E-03	1.61E-01	5	201	
	NA_1024667	1	4.50E-06					3
	CCDC45	17	4.90E-06					3
4931.1	*CRISP1*	6	1.60E-05	1.66E-03	1.33E-01	7	184	
7973.1	*FAM111A*	11	2.80E-05					2
12891.1	*LAIR1*	19	3.10E-05			1		1
4854.1	*TREM1*	6	6.30E-05					2
7972.1	*FAM111B*	11	1.30E-04	4.52E-03	2.33E-01	3	285	
3785.1	*DCHS2 (AK123368)*	4	1.30E-04			1		1
31672.1	*C11orf87 (LOC399947)*	11	1.30E-04	1.0	9.87E-01	7	8,821	
31685.1	*CD3E*	11	1.30E-04	2.73E-02	5.79E-01	7	675	
1385.1	*CFH*	1	1.50E-04	6.47E-03	2.87E-01	4	347	
	TCRA	14	1.50E-04					3
	OTTHUMT00000004245	1	1.50E-04					3
6507.1	*IFNA8*	9	1.50E-04	9.07E-02	9.87E-01	6	1,276	
	TGOLN2	16	1.80E-04					3
31168.1	*PDSS1*	10	1.80E-04	3.33E-01	9.87E-01	7	2,811	
7753.1	*HBB*	11	2.00E-04	5.49E-01	9.87E-01	5	3,919	
7854.1	*SLC6A5*	11	2.30E-04			1		1
2717.1	*ZNF197*	3	3.40E-04	1.83E-02	4.83E-01	3	542	

42475.1	*ZNRF4*	19	3.90E-04	8.19E-01	9.87E-01	4	5,813	
33522.1	*MRPL39*	21	4.00E-04	4.0E-04	9.87E-01	6	12,698	
5042.1	*COQ3 (RP11-98I9.1-002)*	6	4.00E-04	1.83E-01	9.87E-01	7	1,922	
32792.1	*CEP192 (AB051446)*	18	4.30E-04			1		1
31080.1	*FAM36A*	1	4.30E-04	1.66E-02	4.55E-01	4	528	
33747.1	*LTF*	3	4.90E-04	3.03E-03	1.89E-01	4	226	
	CCDC129	7	4.90E-04					3
2932.1	*CPOX*	3	4.90E-04	2.74E-03	1.81E-01	7	219	
8840.1	*KRT78*	12	5.00E-04	9.05E-01	9.87E-01	7	6,635	
5806.1	*OPN1SW*	7	5.80E-04	5.50E-02	8.33E-01	7	959	
13983.1	*APOBEC3C*	22	5.90E-04	2.24E-02	5.27E-01	5	623	
	RP1-321E8.3-001	X	6.00E-04					3
	KIAA1731 (AB051518)	11	6.80E-04					3
3419.1	*FGFBP2 (KSP37)*	4	6.80E-04	3.33E-01	9.87E-01	6	2,750	
3278.1	*AHSG*	3	6.90E-04					2
4470.1	*HUS1B*	6	6.90E-04	4.08E-02	7.17E-01	4	831	
434.1	*NDUFS5*	1	7.10E-04	1.11E-02	3.71E-01	5	440	
3143.1	*TM4SF1*	3	7.30E+04	6.74E-03	2.87E-01	7	346	
	ADAM32	8	7.60E-04					3
11799.1	*DCXR*	17	8.10E-04	1.11E-02	3.71E-01	7	433	
5959.1	*DEFB1*	8	8.30E-04	1.11E-03	1.04E-01	7	158	
35001.1	*C9orf11*	9	8.30E-04	2.73E-01	9.87E-01	5	2,479	
10276.1	*C15orf39*	15	8.30E-04	3.68E-01	9.87E-01	6	3,022	

Table A10 GO categories enriched for genes predicted to be under positive selection. 13,838 of 15,027 genes tested for positive selection were assigned to UniProt identifiers and used to identify GO categories enriched for genes predicted to be under positive selection. Shown are the numbers of genes assigned to each biological process category, the numbers of genes in a null distribution consisting of all genes except those assigned to the term being tested, the numbers of tests performed and the nominal *P*-values from a one-sided Mann-Whitney U test. Bolded *P*-values are significant after a conservative Bonferroni correction for multiple testing ($p < 0.05$). Only categories with a nominal *P*-value less than 0.05 are reported.

GO ID	Biological process	No. tests	Genes in category	Genes in null	*P*-value
GO:0006952	defense response	1,681	476	10,888	**4.03E-14**
GO:0031424	keratinization	1,526	36	10,412	**3.50E-09**
GO:0007606	sensory perception of chemical stimulus	1,523	259	10,376	**2.40E-08**
GO:0055114	oxidation reduction	1,517	510	10,117	5.00E-05
GO:0019882	antigen processing and presentation	1,445	33	9,607	7.80E-05
GO:0046483	heterocycle metabolic process	1,441	227	9,574	7.96E-04
GO:0015698	inorganic anion transport	1,364	41	9,347	6.76E-04
GO:0030193	regulation of blood coagulation	1,363	24	9,306	8.76E-04
GO:0044243	multicellular organismal catabolic process	1,346	22	9,282	9.95E-04
GO:0007586	digestion	1,337	29	9,260	1.72E-03
GO:0042254	ribosome biogenesis	1,333	23	9,231	1.87E-03
GO:0007283	spermatogenesis	1,331	214	9,208	2.36E-03
GO:0007600	sensory perception	1,291	208	8,994	1.26E-03
GO:0032368	regulation of lipid transport	1,270	24	8,786	2.22E-03
GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation	1,255	34	8,762	1.75E-03
GO:0006974	response to DNA damage stimulus	1,219	247	8,728	3.28E-03
GO:0006955	immune response	1,177	137	8,481	4.95E-03
GO:0008544	epidermis development	1,146	69	8,344	7.06E-03
GO:0009566	fertilization	1,124	25	8,275	1.25E-02
GO:0060249	anatomical structure homeostasis	1,121	45	8,250	1.72E-02
GO:0032886	regulation of microtubule-based process	1,098	30	8,205	8.38E-03
GO:0006869	lipid transport	1,088	70	8,175	1.77E-02
GO:0006725	cellular aromatic compound metabolic process	1,076	23	8,105	1.14E-02
GO:0022610	biological adhesion	1,072	396	8,082	1.54E-02
GO:0030198	extracellular matrix organization	1,028	36	7,686	4.17E-03
GO:0031960	response to corticosteroid stimulus	1,016	45	7,650	9.87E-03
GO:0006732	coenzyme metabolic process	992	50	7,605	1.15E-02
GO:0001775	cell activation	983	75	7,555	2.01E-02
GO:0010562	positive regulation of phosphorus metabolic process	966	36	7,480	1.33E-02

GO:0008643	carbohydrate transport	937	39	7,444	1.74E-02
GO:0034097	response to cytokine stimulus	931	31	7,405	1.70E-02
GO:0007010	cytoskeleton organization	915	227	7,317	4.39E-02
GO:0016042	lipid catabolic process	857	69	7,090	2.76E-02
GO:0055085	transmembrane transport	849	343	7,021	4.65E-02

Table A11 Genes assigned to the GO biological process category, "keratinization", ranked by statistical evidence of positive selection. Listed below are 36 of the 41 genes assigned to the "keratinization" category in GO (GO:0031424) that were tested for positive selection. The majority of these genes reside in a cluster on chromosome 1. Shown for each gene are the nominal *P*-value from a chi-square approximated likelihood ratio test between CODEML's M7 and M8 models [103], the corresponding *Q*-value [104] and the overall rank. For the top 6 (bolded *Q*-values), there is statistical evidence for positive selection at an FDR of 10%.

CCDS	Gene Name	Chr	Description	No. Species	*P*-value	*Q*-value	Rank
30866.1	*SPRR2E*	1	small proline-rich protein 2E	7	1.13E-07	**1.28E-04**	13
30867.1	*SPRR2F*	1	small proline-rich protein 2F	6	4.56E-07	**3.98E-04**	17
1030.1	*IVL*	1	involucrin	4	6.81E-07	**5.32E-04**	19
1015.1	*LCE3C*	1	late cornified envelope 3C	6	2.04E-05	**6.73E-03**	44
30865.1	*SPRR2B*	1	small proline-rich protein 2B	6	6.13E-05	**1.62E-02**	56
11737.1	*EVPL*	17	envoplakin	7	3.71E-04	**5.73E-02**	95
1032.1	*SPRR1A*	1	small proline-rich protein 1A	6	1.11E-03	1.04E-01	159
1020.1	*LCE2B*	1	late cornified envelope 2B	4	2.24E-03	1.61E-01	204
1033.1	*SPRR3*	1	small proline-rich protein 3	6	2.48E-03	1.70E-01	213
30870.1	*LOR*	1	loricrin	4	2.47E-02	5.56E-01	638
1031.1	*SPRR4*	1	small proline-rich protein 4	6	2.47E-02	5.56E-01	643
1021.1	*LCE2A*	1	late cornified envelope 2A	3	3.34E-02	6.50E-01	747
1014.1	*LCE3D*	1	late cornified envelope 3D	5	8.21E-02	9.87E-01	1,226
1026.1	*LCE1C*	1	late cornified envelope 1C	4	1.35E-01	9.87E-01	1,573
30864.1	*SPRR2D*	1	small proline-rich protein 2D	6	1.65E-01	9.87E-01	1,825
33435.1	*TGM3*	20	transglutaminase 3	7	2.02E-01	9.87E-01	1,974
1024.1	*LCE1E*	1	late cornified envelope 1E	5	3.01E-01	9.87E-01	2,595
10526.1	*PPL*	16	periplakin	4	3.01E-01	9.87E-01	2,645
1034.1	*SPRR2A*	1	small proline-rich protein 2A	6	4.07E-01	9.87E-01	3,208
1011.1	*LCE5A*	1	late cornified envelope 5A	4	4.49E-01	9.87E-01	3,392
30863.1	*SPRR1B*	1	small proline-rich protein 1B (cornifin)	6	4.49E-01	9.87E-01	3,393
8835.1	*KRT2*	12	keratin 2	5	6.07E-01	9.87E-01	4,160
1025.1	*LCE1D*	1	late cornified envelope 1D	4	6.70E-01	9.87E-01	4,685
1019.1	*LCE2C*	1	late cornified envelope 2C	4	6.70E-01	9.87E-01	4,704
1017.1	*LCE3A*	1	late cornified envelope 3A	5	6.70E-01	9.87E-01	4,726
1016.1	*LCE3B*	1	late cornified envelope 3B	7	8.19E-01	9.87E-01	5,592
1013.1	*LCE3E*	1	late cornified envelope 3E	7	9.05E-01	9.87E-01	6,225

1018.1	*LCE2D*	1	late cornified envelope 2D	4	9.05E-01	9.87E-01	6,289
1027.1	*LCE1B*	1	late cornified envelope 1B	5	9.05E-01	9.87E-01	6,314
43777.1	*SHARPIN*	8	SHANK-associated RH domain interactor	4	9.05E-01	9.87E-01	6,650
9622.1	*TGM1*	14	transglutaminase 1	6	1.0	9.87E-01	7,436
12606.1	*CNFN*	19	cornifelin	6	1.0	9.87E-01	9,232
1023.1	*LCE1F*	1	late cornified envelope 1F	4	1.0	9.87E-01	9,754
30868.1	*SPRR2G*	1	small proline-rich protein 2G	5	1.0	9.87E-01	9,950
31777.1	*PPHLN1*	12	periphilin 1	6	1.0	9.87E-01	10,567
1028.1	*LCE1A*	1	late cornified envelope 1A	3	1.0	9.87E-01	12,504

Table A12 GO categories enriched for genes with increased human population differentiation. The top ten GO categories enriched for genes under positive selection in primates were also tested for higher levels of population differentiation, as measured by F_{ST} between European and African populations [90]. Shown are the number of genes assigned to each biological process category, the number of genes in a null distribution consisting of all genes except those assigned to the term being tested, the average F_{ST} for each group and the nominal *P*-values from a one-sided Mann-Whitney U test. Bolded *P*-values are significant after a Bonferroni correction for multiple testing ($p < 0.05$).

GO ID	Biological process	Genes in cat.	Genes in null	F_{ST} in cat.	F_{ST} in null	*P*-value
GO:0006952	defense response	448	13,809	0.0778	0.0710	0.0161
GO:0031424	keratinization	36	14,221	0.128	0.0711	0.0550
GO:0007606	sensory perception of chemical stimulus	253	14,004	0.0961	0.0708	**2.97E-13**
GO:0055114	oxidation reduction	496	13,761	0.0840	0.0707	**8.24E-04**
GO:0019882	antigen processing and presentation	37	14,220	0.103	0.0711	0.0574
GO:0015698	inorganic anion transport	41	14,216	0.0733	0.0712	0.394
GO:0046483	heterocycle metabolic process	259	13,998	0.0619	0.0714	0.858
GO:0030193	regulation of blood coagulation	35	14,222	0.0664	0.712	0.567
GO:0044243	multicellular organismal catabolic process	23	14,234	0.0791	0.0712	0.0975
GO:0007600	sensory perception	491	13,766	0.0837	0.0708	**5.80E-07**

Table A13 Numbers of genes showing evidence for positive selection on lineages of the phylogeny. Shown for each branch of the primate phylogeny (Fig. 2.1A) are the sequences required to test that branch, the number of genes tested and the numbers of genes showing evidence for positive selection at several FDR thresholds and at a nominal $p < 0.05$. *P*-values were computed from a 50:50 mixture of a chi-square distribution with 1 degree of freedom and a point mass at 0 [91]. Note that the branch labeled as "tamarin" combines both the branch leading to tamarin as well as the branch leading to the common ancestor of Apes and Old World moneys.
OWM – sequence required from any single Old World monkey: macaque (rheMac2), vervet or colobus.
APE – sequence required from any single Great Ape: human (hg18), chimpanzee (panTro2) or orangutan (ponAbe2).

			No. significant genes			
Branch	**Sequences required**	**No. genes tested**	**1% FDR**	**5% FDR**	**10% FDR**	***P*-value < 0.05**
panTro2	panTro2, hg18, 1 other	14,250	2	3	6	203
hg18	hg18, panTro2, 1 other	14,230	0	0	0	254
panTro2, hg18	panTro2, hg18, ponAbe2	12,316	0	5	6	256
ponAbe2	ponAbe2, hg18 or panTro2, 1 other	12,906	2	4	8	338
panTro2, hg18, ponAbe2	hg18 or panTro2, ponAbe2, 1 OWM, 1 other	8,863	3	3	8	223
rheMac2	rheMac2, vervet, 1 other	11,005	3	4	4	173
vervet	vervet, rheMac2, 1 other	11,003	1	3	7	168
rheMac2, vervet	rheMac2, vervet, colobus	9,609	0	0	0	126
colobus	colobus, rheMac2 or vervet, 1 other	10,695	2	4	8	245
rheMac2, vervet, colobus	rheMac2 or vervet, colobus, 1 APE, tamarin	9,018	5	5	14	264
tamarin	tamarin, 1 OWM, 1 APE	9,906	6	12	84	708

Table A14 Complete lists of genes tested for evidence of positive selection acting on individual lineages ranked by significance. Tables are provided as a separate supplemental file with a worksheet for each individual branch. For each gene are shown the number of species, the length of the sequence, the nominal *P*-value computed from a 50:50 mixture of a chi-square distribution with 1 degree of freedom and a point mass at 0 [91] and the estimated false discovery rate calculated by *Q*-values [104].

Table A15 Top 5 GO categories enriched for genes predicted to be under lineage specific positive selection for each lineage. Genes were assigned to UniProt identifiers and used to identify GO categories enriched for genes predicted to be under positive selection along specific lineages. Shown are the number of genes assigned to each biological process category, the number of genes in a null distribution consisting of all genes except those assigned to the term being tested, the number of tests performed and the nominal *P*-values from a one-sided Mann-Whitney U test. Bolded *P*-values are significant after a conservative Bonferroni correction for multiple testing ($p < 0.05$).

GO ID	Biological process	No. tests	Genes in category	Genes in null	*P*-value
panTro2					
GO:0007606	sensory perception of chemical stimulus	1,651	189	10,348	**1.45E-14**
GO:0050900	leukocyte migration	1,644	50	10,159	**8.42E-06**
GO:0031424	keratinization	1,606	35	10,109	5.74E-05
GO:0006323	DNA packaging	1,603	23	10,074	8.18E-05
GO:0044242	cellular lipid catabolic process	1,602	70	10,051	5.38E-04
hg18					
GO:0007606	sensory perception of chemical stimulus	1,648	188	10,329	**1.59E-12**
GO:0000723	telomere maintenance	1,640	24	10,141	1.00E-03
GO:0022411	cellular component disassembly	1,632	32	10,117	1.13E-03
GO:0031424	keratinization	1,628	34	10,085	2.39E-03
GO:0007126	meiosis	1,625	54	10,051	5.14E-03
panTro2, hg18					
GO:0006952	defense response	1,538	383	8,968	**7.22E-08**
GO:0007276	gamete generation	1,405	203	8,585	**2.11E-07**
GO:0031424	keratinization	1,366	31	8,382	6.60E-05
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	1,363	31	8,351	1.05E-04
GO:0007606	sensory perception of chemical stimulus	1,361	97	8,320	8.06E-04
ponAbe2					
GO:0007608	sensory perception of smell	1,566	103	9,374	**1.43E-13**
GO:0031424	keratinization	1,558	32	9,271	**1.10E-07**
GO:0042742	defense response to bacterium	1,554	82	9,239	5.27E-05
GO:0031214	biomineral tissue development	1,500	23	9,157	7.43E-03
GO:0010562	positive regulation of phosphorus metabolic process	1,496	107	9,134	1.14E-02
panTro2, hg18, ponAbe2					
GO:0006952	defense response	1,313	263	6,645	**1.47E-06**
GO:0006818	hydrogen transport	1,212	35	6,382	1.08E-03
GO:0050866	negative regulation of cell activation	1,204	28	6,347	1.60E-03
GO:0051606	detection of stimulus	1,172	51	6,319	1.50E-03
GO:0046942	carboxylic acid transport	1,159	81	6,268	1.48E-03
rheMac2					
GO:0022904	respiratory electron transport chain	1,464	32	8,182	**5.40E-10**
GO:2000021	regulation of ion homeostasis	1,457	47	8,150	**1.55E-05**
GO:0006952	defense response	1,395	355	8,103	5.94E-05
GO:0007606	sensory perception of chemical stimulus	1,296	102	7,748	3.89E-05

GO:0006399	tRNA metabolic process	1,292	79	7,646	9.01E-04
vervet					
GO:0007608	sensory perception of smell	1,463	84	8,180	**1.63E-06**
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	1,459	22	8,096	**1.75E-05**
GO:0006952	defense response	1,457	368	8,074	6.12E-04
GO:0007205	activation of protein kinase C activity by G-protein coupled receptor protein signaling pathway	1,337	21	7,706	5.72E-04
GO:0000075	cell cycle checkpoint	1,322	57	7,685	1.75E-03
rheMac2, vervet					
GO:0006952	defense response	1,366	321	7,167	**4.52E-06**
GO:0015698	inorganic anion transport	1,264	35	6,846	1.37E-03
GO:0006935	chemotaxis	1,259	55	6,811	2.26E-03
GO:0018130	heterocycle biosynthetic process	1,230	37	6,756	2.84E-03
GO:0050818	regulation of coagulation	1,212	20	6,719	5.20E-03
colobus					
GO:0007606	sensory perception of chemical stimulus	1,426	98	7,958	**3.50E-06**
GO:0003018	vascular process in circulatory system	1,419	29	7,860	1.15E-03
GO:0006968	cellular defense response	1,401	32	7,831	1.48E-03
GO:0043038	amino acid activation	1,397	32	7,799	1.45E-03
GO:0048609	reproductive process in a multicellular organism	1,394	230	7,767	1.61E-03
rheMac2, vervet, colobus					
GO:0006952	defense response	1,326	265	6,774	**9.5E-08**
GO:0042129	regulation of T cell proliferation	1,223	22	6,509	2.30E-04
GO:0007586	digestion	1,197	20	6,487	2.78E-03
GO:0071103	DNA conformation change	1,187	28	6,467	3.11E-03
GO:0009451	RNA modification	1,183	27	6,439	2.96E-03
tamarin					
GO:0006955	immune response	1,380	245	7,436	**6.32E-10**
GO:0006631	fatty acid metabolic process	1,291	133	7,191	**9.10E-07**
GO:0007155	cell adhesion	1,232	349	7,058	**8.72E-06**
GO:0031099	regeneration	1,175	40	6,709	1.80E-04
GO:0001775	cell activation	1,151	80	6,669	1.25E-04

Table A16 Sequence coverage of captured miRNAs. Summary of targeted miRNA sequence coverage for each non-human primate exome and two human HapMap exomes (Human 1: NA12879 and Human 2: NA18967). The total size of the captured miRNA target is 48,075 bp and includes ~550 miRNAs. Listed for each exome are the number of bases in the target covered by at least one read, the number of bases assembled and the number of bases assembled with Phred consensus quality score ≥40 (Q40; 10^{-4} error rate).

Sample	≥1X coverage (bp)	≥1X coverage (%)	Consensus called (bp)	Consensus called (%)	≥Q40 consensus (bp)	≥Q40 consensus (%)	Avg. coverage
Human 1	43,530	90.6	42,995	89.4	42,336	88.1	91X
Human 2	43,585	90.7	43,312	90.1	42,297	88.0	102X
Macaque	42,086	87.5	40,815	84.9	39,732	82.6	94X
Vervet	42,579	88.6	41,022	85.3	40,063	83.3	93X
Colobus	41,530	86.4	39,955	83.1	38,730	80.6	88X
Tamarin	41,278	85.9	39,070	81.3	36,860	76.7	84X

Table A17 Nucleotide differences and indels between the assembled macaque exome and the macaque reference genome. We calculated the number of nucleotide differences and indels between our assembled macaque targeted sequences and the macaque reference genome for the 153,546 assembled targets that uniquely mapped to the macaque genome using cross_match (v1.090518, http://www.phrap.org). Targets are further categorized by whether they are retained or removed following filtering for segmental duplications, low read depth, extreme heterozygosity and missing sequence (see Methods for more details), and by whether they were putative paralogous targets (see Fig. A3). Note that the percent difference for all targets is higher than that reported in the main text (0.253% *vs.* 0.10%) because we used all captured targets which include flanking intronic and miRNA sequences in addition to coding sequences.

	No. targets	**No. bases**	**Differences (%)**	**Indels (%)**
All targets	153,546	27,746,320	0.253	0.0252
All retained targets	122,411	22,921,623	0.228	0.0187
All removed targets	31,135	4,824,697	0.376	0.0560
Putative paralogs	121	21,789	1.55	0.142
Putative retained paralogs	89	16,100	1.07	0.118
Putative removed paralogs	32	5,689	2.90	0.211

Table A18 GO categories enriched for genes that are absent from the macaque exome assembly. 15,827 of 16,707 genes were assigned to UniProt identifiers and used to identify GO categories enriched for genes that are absent from the macaque exome assembly. Shown for each category is the number of genes absent and present in that category, the number of genes absent and present excluding genes in that category and the *P*-value and odds ratio (OR) from a one-sided Fisher's exact test. Bolded *P*-values are significant after a conservative Bonferroni correction for multiple testing ($p < 0.05$). Only categories with a nominal *P*-value less than 0.05 are reported.

GO ID	Biological process	No. tests	Cat. absent genes	Cat. present genes	Absent genes	Present genes	*p*-value	OR
GO:0007608	sensory perception of smell	2,130	146	213	2,775	9,470	**5.3E-14**	2.34
GO:0006355	regulation of transcription, DNA dependent	2,121	681	1632	2,629	9,257	**5.2E-14**	1.47
GO:0031424	keratinization	1,688	21	19	1,948	7,625	**7.1E-06**	4.33
GO:0034340	respose to type I interferon	1,683	22	22	1,927	7,606	**1.1E-05**	3.95
GO:0016339	calcium-dependent cell-cell adhesion	1,674	11	9	1,905	7,584	5.9E-04	4.86
GO:0007586	digestion	1,669	20	31	1,894	7,575	1.3E-03	2.58
GO:0006952	defense response	1,662	120	344	1,874	7,544	1.4E-03	1.40
GO:0048609	multicellular organismal reproductive process	1,522	77	204	1,754	7,200	1.1E-03	1.55
GO:0034728	nucleosome organization	1,462	23	40	1,677	6,996	1.2E-03	2.40
GO:0031023	microtubule organizing center organization	1,457	12	16	1,654	6,956	3.6E-03	3.15
GO:0051258	protein polymerization	1,455	15	23	1,642	6,940	3.0E-03	2.76
GO:0006275	regulation of DNA replication	1,445	15	27	1,627	6,917	8.5E-03	2.36
GO:0070507	regulation of microtubule cytoskeleton organization	1,434	12	19	1,612	6,890	8.4E-03	2.70
GO:0042384	cilium assembly	1,423	9	13	1,600	6,871	1.4E-02	2.97
GO:0016579	protein deubiquitination	1,417	10	16	1,591	6,858	1.6E-02	2.69
GO:0006351	transcription, DNA-dependent	1,414	32	85	1,581	6,842	1.6E-02	1.63
GO:0003013	circulatory system process	1,399	17	37	1,549	6,757	1.7E-02	2.00
GO:0034470	ncRNA processing	1,374	14	122	1,532	6,720	1.7E-02	1.51
GO:0050909	sensory perception of taste	1,362	11	21	1,490	6,598	2.4E-02	2.32
GO:0007275	multicellular organismal development	1,357	73	254	1,479	6,577	4.4E-02	1.28
GO:0007565	female pregnancy	1,308	11	24	1,406	6,323	4.2E-02	2.06

Text A1 Identification of paralogous sequences and evaluation of their impact on exome assemblies.

Genes that have duplicated to become paralogs in other lineages are susceptible to mis-assembly because reads from multiple genomic locations may map to a single location in the human genome. To identify problematic paralogous targets we mapped macaque reads to both the human (hg18) and macaque (rheMac2) reference genomes. We then compared the depth of human target sequences to that of 155,707 orthologous target sequences in the macaque genome. As putative paralogs, we identified 137 targets with substantially higher read depth in human compared to macaque (Fig. A3A).

To evaluate the impact these putative paralogous targets have on the macaque exome assembly, we compared assembled target sequences to the macaque reference genome sequence. Targets that are mis-assembled will have more nucleotide differences and indels compared to correctly assembled targets (which will have some differences due to polymorphisms in macaque). The assembled putative paralogous targets have a ~6-fold increase in the number of nucleotide differences (1.6% *vs*. 0.25%) and indels (0.14% *vs*. 0.025%) compared to the entire set of assembled targets (Table A17 and Fig. A3B). This indicates that the putative paralogous targets are enriched for mis-assembled sequences.

We performed several post-assembly filtering steps to reduce the amount of mis-assembled sequences from paralogs, such as removing targets that overlap known segmental duplications or that have high levels of heterozygosity. The targets that are removed by filtering have a higher proportion of nucleotide differences (0.38% *vs*. 0.23%) and indels (0.056% *vs*. 0.019%) compared to those that are retained. For the subset of targets that are putative paralogs, filtering reduces the proportion of nucleotide differences (from 2.9% to 1.1%) and indels (from 0.21% to 0.12%) (Table A17 and Fig. A2B). This demonstrates that our filtering steps remove targets that are more likely to have assembly errors. The nucleotide differences for the retained putative paralogs remain high, however, suggesting that a small fraction of our final exome assemblies contain paralogous assembly errors.

Text A2 Types of genes absent from positive selection analyses.

To better understand the types of genes that failed to capture, assemble or pass filters, we labeled human genes with macaque orthologs as "absent" if they had no macaque sequence following filtering. We then identified gene ontology (GO) terms that were enriched for absent genes (Table A18). The GO terms with the most significant enrichments are "sensory perception of smell" and "regulation of transcription, DNA-dependent". This may indicate that genes in large families (such as olfactory receptors or zinc-finger transcription factors) are difficult to assemble or are preferentially removed by our filtering procedure. Another of the most significant categories, "keratinization", is also enriched for genes predicted to be under positive selection (Table A9), suggesting that some rapidly evolving genes may be excluded from our analysis.

We checked whether any of the GO terms enriched for absent genes are related to immune response or reproduction, because genes involved in these processes are known to be rapidly evolving. The terms "response to type I interferon", "defense response" and "multicellular organismal reproductive success" are enriched for absent genes, but only the first category is significant after Bonferroni correction for the number of tests performed (Table A18). These results indicate some fraction of rapidly evolving genes may be excluded from our positive selection analysis, but most are likely to be retained.

Appendix B

Supplemental Information for Chapter 3

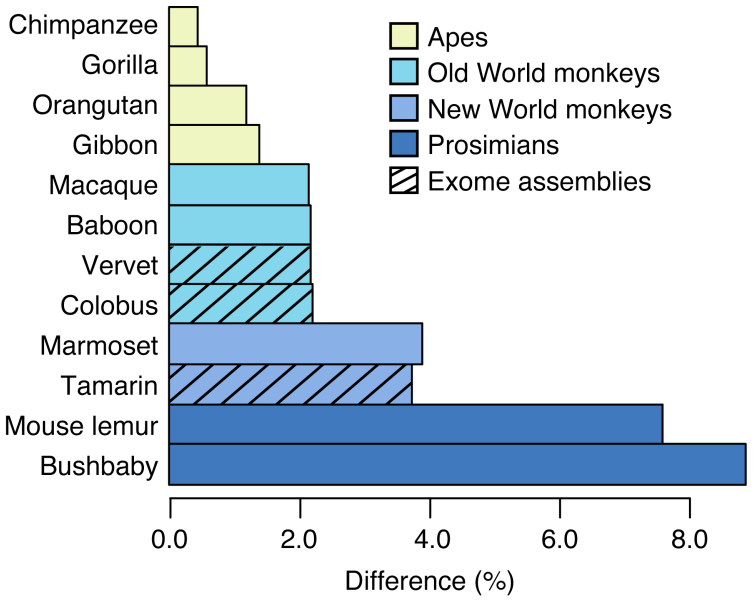

Figure B1 Nucleotide differences between human and other non-human primates. Sites were restricted to the 376,613 bp that is present in all 12 non-human primates. Coding sequence divergence was calculated between each species and the human reference genome (hg19). Bars are colored by clades and hatches indicate species whose coding sequences were extracted from assembled exome reads.

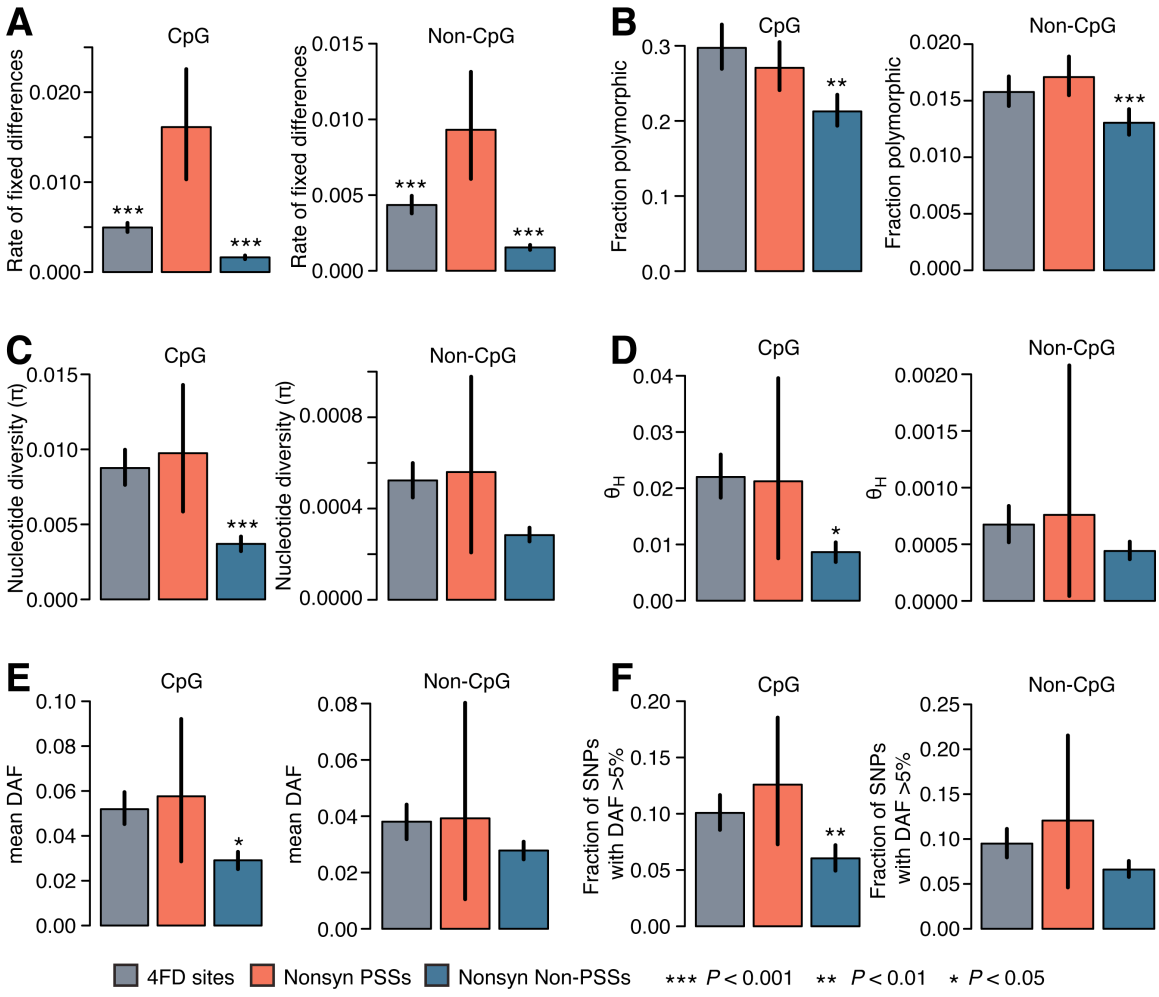

Figure B2 Patterns of human divergence and diversity at CpG and non-CpG sites for non-synonymous PSSs, non-synonymous non-PSSs and 4FD sites. The data are as described for Fig. 3.3, except that sites were additionally classified by their CpG context. For polymorphic sites, if either allele was part of a CpG, it was classified as such. (A) Rate of fixed differences on the human lineage since its divergence with chimpanzee. For this analysis, we conservatively considered sites as potentially part of a CpG if they were preceded by a C or followed by a G. (B) Fraction of total sites that are polymorphic in 2,439 human individuals. (C) Nucleotide diversity, π. (D) Nucleotide diversity weighted by the frequency of derived variants, θ_H [18]. (E) Mean derived allele frequency (DAF) of polymorphic sites. (F) Fraction of polymorphic sites with a DAF >5%.

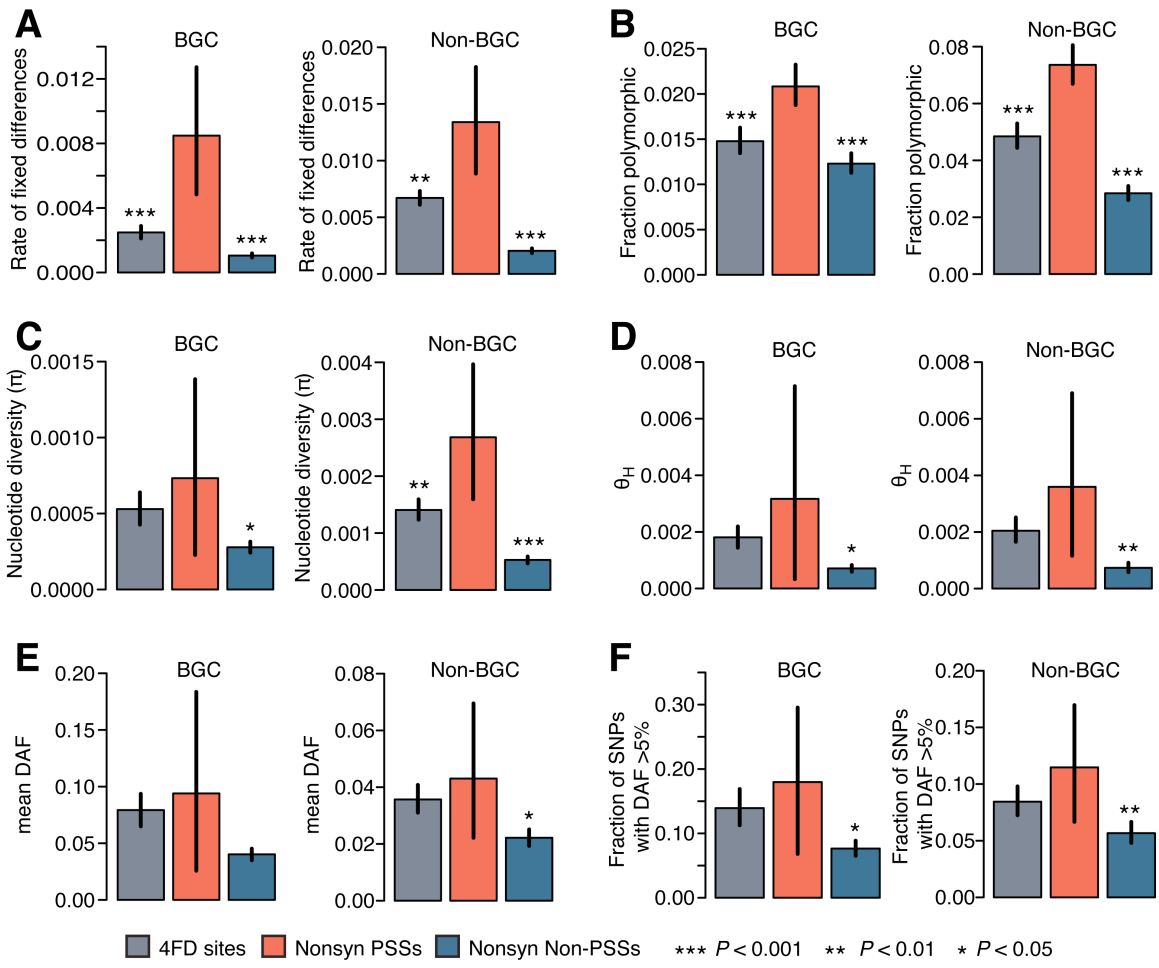

Figure B3 Patterns of human divergence and diversity at BGC and non-BGC sites for non-synonymous PSSs, non-synonymous non-PSSs and 4FD sites. These data are as described for Fig. 3.3, except that rates were calculated separately for types of substitutions and polymorphisms that either are or are not potentially inflated by GC-biased gene conversion (BGC). Rates were considered to be "BGC" rates if they changed a weak (A or T) ancestral base to a strong (C or G) derived base. (A) Rate of fixed differences on the human lineage since its divergence from chimpanzee. (B) Fraction of total sites that are polymorphic in 2,439 human individuals. (C) Nucleotide diversity, π. (D) Nucleotide diversity weighted by the frequency of derived variants, θ_H [18]. (E) Mean derived allele frequency (DAF) of polymorphic sites. (F) Fraction of polymorphic sites with a DAF >5%.

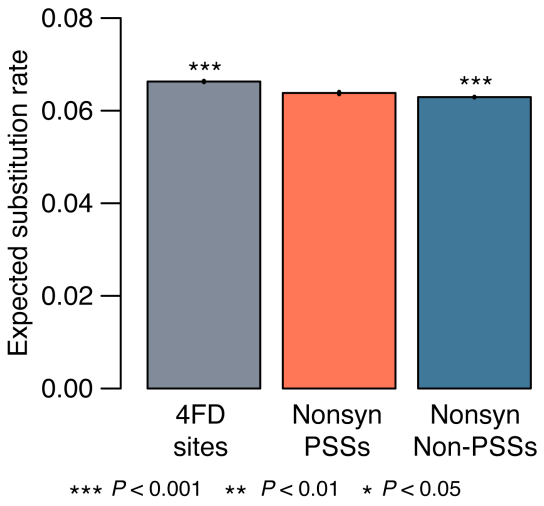

Figure B4 Expected substitution rates of PSSs, non-PSSs and 4FD sites. We calculated the expected non-synonymous and synonymous substitution rates at positively selected sites (PSSs; red), non-positively selected sites (non-PSSs; blue) and four-fold degenerate sites (4FD sites; grey), using the two flanking nucleotides at each base. Genome-wide estimates for substitution rates between human and macaque for different sequence contexts are listed in Table B10. We tested for significant differences from the expected rate of non-synonymous substitutions at PSSs using *P*-values from one-sided bootstraps.

Table B1 Coding sequence divergence between human and other non-human primates. For each of our 12 non-human primates, we compared sequences from the filtered transcripts with the human reference sequence and calculated nucleotide divergence (Table B11 contains the number of transcripts remaining for each species). Codons containing Ns were removed, as were codons containing more than one substitution. This table shows the number of codons used for each comparison, the number of codons containing non-synonymous, synonymous or four-fold degenerate (4FD) differences and pairwise nucleotide divergence with human calculated from all (d), non-synonymous (d_N), synonymous (d_S) and four-fold degenerate (d_{4FD}) sites.

Species	Total codons	Nonsyn diffs	Syn diffs	4FD diffs	d	d_N	d_S	d_{4FD}
Chimpanzee	6816436	42824	67610	39804	0.0054	0.0027	0.0148	0.0125
Gorilla	7071472	54390	93229	54836	0.0070	0.0033	0.0197	0.0166
Orangutan	5283840	75959	142849	83856	0.0138	0.0062	0.0403	0.0339
Gibbon	5663751	92998	171740	99844	0.0156	0.0070	0.0456	0.0382
Rhesus macaque	6830384	158996	337485	197416	0.0242	0.0100	0.0737	0.0617
Baboon	6101978	138819	302142	177476	0.0241	0.0098	0.0736	0.0617
Vervet	6239358	131939	306683	178204	0.0234	0.0091	0.0734	0.0611
Colobus monkey	5623519	121172	288532	168571	0.0243	0.0093	0.0762	0.0633
Marmoset	5681756	220739	459780	264686	0.0399	0.0167	0.1213	0.1005
Tamarin	5157508	158137	403818	235424	0.0363	0.0132	0.1165	0.0968
Mouse lemur	1921831	116836	282630	160825	0.0693	0.0261	0.2197	0.1790
Bushbaby	1214801	81020	200882	110457	0.0774	0.0285	0.2510	0.2008

Table B2 Numbers of genes showing evidence of positive selection at several FDR thresholds.

1% FDR	5% FDR	10% FDR
278	543	716

Table B3 GO categories enriched for genes predicted to be under positive selection in non-human primates. Each gene tested for positive selection was assigned to a UniProt identifier and used to identify Gene Ontology (GO) categories enriched for genes predicted to be under positive selection. Only GO categories with a minimum of 20 genes were tested. This table shows the numbers of genes assigned to each biological process category, the numbers of genes in a null distribution consisting of all genes except those assigned to the term being tested, the numbers of tests performed and the nominal *P*-values from a one-sided Mann-Whitney U test. Bolded *P*-values are significant after a conservative Bonferroni correction for multiple testing ($p < 0.05$). Only categories with a nominal *P*-value less than 0.05 are reported.

GO ID	Biological process	No. Tests	Genes in category	Genes in null	*P*-value
GO:0006952	defense response	1178	282	6859	**1.58E-14**
GO:0031424	keratinization	1080	27	6577	**3.84E-05**
GO:0006955	immune response	976	112	5927	2.76E-04
GO:0006974	response to DNA damage stimulus	999	176	6103	3.19E-04
GO:0007283	spermatogenesis	1078	157	6550	3.70E-04
GO:0055114	oxidation reduction	1040	290	6393	3.71E-04
GO:0003008	system process	869	496	5519	4.46E-04
GO:0007017	microtubule-based process	943	109	5815	4.54E-04
GO:0042180	cellular ketone metabolic process	921	187	5706	4.83E-04
GO:0006508	proteolysis	783	290	4992	1.80E-03
GO:0007218	neuropeptide signaling pathway	786	31	5023	2.29E-03
GO:0016042	lipid catabolic process	738	34	4702	2.31E-03
GO:0035023	regulation of Rho protein signal transduction	602	23	4079	4.24E-03
GO:0007155	cell adhesion	732	267	4668	7.60E-03
GO:0016052	carbohydrate catabolic process	691	23	4401	7.94E-03
GO:0006935	chemotaxis	687	26	4378	1.50E-02
GO:0051707	response to other organism	666	20	4352	1.51E-02
GO:0008643	carbohydrate transport	664	24	4332	1.53E-02
GO:0022403	cell cycle phase	598	83	4056	1.58E-02
GO:0006915	apoptosis	637	154	4233	1.99E-02
GO:0015674	di-, tri-valent inorganic cation transport	590	39	3973	2.01E-02
GO:0002694	regulation of leukocyte activation	659	27	4260	2.12E-02
GO:0006814	sodium ion transport	663	48	4308	2.19E-02
GO:0051259	protein oligomerization	583	44	3934	4.02E-02
GO:0007599	hemostasis	568	22	3890	4.95E-02
GO:0042327	positive regulation of phosphorylation	554	24	3844	4.98E-02

Table B4 Summary of polymorphism data for PSGs and non-PSGs from 2,439 human individuals. We identified 577 positively selected genes (PSGs) at an FDR of 10% from protein coding comparisons of 12 non-human primate species. We labeled the 11,331 filtered genes above the FDR threshold as non-positively selected genes (non-PSGs). Genes were filtered if they were on the X chromosome, had missing chimpanzee sequence, no polymorphic sites or if less than half their sequence was targeted for human exome capture. Sites from these genes were only included in our analyses if polymorphism data was present for > 90% of individuals. This table shows the number of synonymous, non-synonymous and four-fold degenerate (4FD) sites in PSGs and non-PSGs, the number of polymorphic sites in 2,439 individuals and the number of sites with fixed differences on the human branch. The total numbers of synonymous and non-synonymous sites were estimated using the degeneracy of the codons [129].

		Total bp	Variants		Fixed differences	
			bp	%	bp	%
PSGs	all	1027752	26616	2.59	2836	0.28
	synonymous	226040	9390	4.15	1350	0.60
	non-synonymous	801712	17226	2.15	1486	0.19
	4FD	151664	5069	3.34	738	0.49
Non-PSGs	all	14076999	326357	2.32	28368	0.20
	synonymous	3123766	124112	3.97	17631	0.56
	non-synonymous	10953233	202245	1.85	10737	0.10
	4FD	2126208	71452	3.36	9907	0.47

Table B5 Summary of human divergence and diversity analyses for non-synonymous PSSs, non-synonymous non-PSSs and 4FD sites. We calculated summary statistics of non-synonymous divergence and non-synonymous diversity for positively selected sites (PSSs) and non-positively selected sites (non-PSSs), as well as synonymous divergence and diversity for four-fold degenerate (4FD) sites. These summary statistics included: the rate of fixed differences on the human lineage, the fraction of total bases that are polymorphic, the nucleotide diversity (π), the nucleotide diversity weighted by the frequency of derived alleles (θ_H) [18], the mean derived allele frequency (DAF), the fraction of polymorphic sites with a DAF > 5% and Tajima's D [14]. We tested for significant differences from PSSs with *P*-values from one-sided bootstraps.

Test	Nonsyn PSSs	Nonsyn Non-PSSs		4FD sites	
	Mean	Mean	*P*-value	Mean	*P*-value
Rate of fixed differences	0.0114	0.0016	< 0.001	0.0047	< 0.001
Fraction polymorphic	0.0494	0.0205	< 0.001	0.0329	< 0.001
Nucleotide diversity (π)	0.0018	0.0004	< 0.001	0.0010	0.011
θ_H	0.0034	0.0007	0.002	0.0020	0.097
Mean DAF	0.0520	0.0283	0.018	0.0456	0.331
Fraction of SNPs with DAF > 5%	0.1250	0.0639	0.006	0.0981	0.147
Tajima's D	-1.82	-2.27	0.001	-1.99	0.151

Table B6 Summary of polymorphism data at CpG and non-CpG sites for PSSs, non-PSSs and 4FD sites. The data are as described for Table 3.2 except that sites were additionally classified by their CpG context. Polymorphic sites were labeled as a CpG if either allele could form a CpG. Fixed differences were conservatively labeled as a CpG if they were preceded by a C or followed by a G.

		Total bp	Variants		Fixed differences	
			bp	%	bp	%
CpG Sites						
PSS	all	1951	15	0.77	27	1.38
	synonymous	577	1	0.17	5	0.87
	non-synonymous	1374	14	1.02	22	1.60
Non-PSS	all	367762	3564	0.97	1057	0.29
	synonymous	105705	1541	1.46	629	0.60
	non-synonymous	262057	2023	0.77	428	0.16
4FD	all	84376	993	1.18	418	0.50
Non-CpG sites						
PSS	all	3700	274	7.41	34	0.92
	synonymous	704	72	10.23	6	0.85
	non-synonymous	2996	202	6.74	28	0.93
Non-PSS	all	606119	20793	3.43	1397	0.23
	synonymous	107982	7204	6.67	628	0.58
	non-synonymous	498137	13589	2.73	769	0.15
4FD	all	58486	3708	6.34	255	0.44

Table B7 Summary of human divergence and diversity analyses at CpG and non-CpG sites for non-synonymous PSSs, non-synonymous non-PSSs and 4FD sites. The data are as described for Table B5 except that CpG sites and non-CpG sites were analyzed independently.

Test	Nonsyn PSSs	Nonsyn Non-PSSs		4FD sites	
	Mean	Mean	*P*-value	Mean	*P*-value
CpG Sites					
Rate of fixed differences	0.0161	0.0016	< 0.001	0.0050	< 0.001
Fraction polymorphic	0.2707	0.2126	0.001	0.2972	0.885
Nucleotide diversity (π)	0.0097	0.0037	< 0.001	0.0088	0.328
θ_H	0.0212	0.0086	0.035	0.0220	0.57
Mean DAF	0.0576	0.0291	0.024	0.0519	0.383
Fraction of SNPs with DAF > 5%	0.1259	0.0604	0.009	0.1007	0.172
Tajima's D	-1.78	-2.34	< 0.001	-2.03	0.078
Non-CpG Sites					
Rate of fixed differences	0.0093	0.0015	< 0.001	0.0043	< 0.001
Fraction polymorphic	0.0171	0.0130	< 0.001	0.0158	0.110
Nucleotide diversity (π)	0.0006	0.0003	0.084	0.0005	0.448
θ_H	0.0008	0.0004	0.332	0.0007	0.484
Mean DAF	0.0392	0.0278	0.312	0.0381	0.521
Fraction of SNPs with DAF > 5%	0.1206	0.0661	0.104	0.0950	0.287
Tajima's D	-1.75	-2.23	0.029	-1.93	0.273

Table B8 Summary of polymorphism data at BGC and non-BGC sites for PSSs, non-PSSs and 4FD sites. The data are as described for Table 3.2 except that sites potentially affected by GC-biased gene conversion (BGC) are separated from those that are predominately unaffected.

		Total bp	Variants		Fixed differences	
			bp	%	bp	%
BGC Sites						
PSS	all	2768	62	2.24	21	0.76
	synonymous	652	18	2.76	3	0.46
	non-synonymous	2116	44	2.08	18	0.85
Non-PSS	all	497655	6600	1.33	786	0.16
	synonymous	104383	1773	1.70	373	0.36
	non-synonymous	393272	4827	1.23	413	0.11
4FD	all	71093	1048	1.47	177	0.25
Non-BGC sites						
PSS	all	2871	218	7.59	37	1.29
	synonymous	625	53	8.48	7	1.12
	non-synonymous	2246	165	7.35	30	1.34
Non-PSS	all	475072	16656	3.51	1615	0.34
	synonymous	108562	6251	5.76	866	0.80
	non-synonymous	366510	10405	2.84	749	0.20
4FD	all	71561	3461	4.84	480	0.67

Table B9 Summary of human divergence and diversity analyses at BGC and non-BGC sites for non-synonymous PSSs, non-synonymous non-PSSs and 4FD sites. The data are as described for Table B5 except that sites potentially affected by GC-biased gene conversion were analyzed independently from those that are unaffected.

Test	Nonsyn PSSs	Nonsyn Non-PSSs		4FD sites	
	Mean	Mean	*P*-value	Mean	*P*-value
BGC Sites					
Rate of fixed differences	0.0085	0.0011	< 0.001	0.0025	< 0.001
Fraction polymorphic	0.0208	0.0123	< 0.001	0.0148	< 0.001
Nucleotide diversity (π)	0.0007	0.0003	0.047	0.0005	0.284
θ_H	0.0032	0.0007	0.049	0.0018	0.207
Mean DAF	0.0940	0.0402	0.081	0.0793	0.373
Fraction of SNPs with DAF > 5%	0.1798	0.0765	0.040	0.1392	0.300
Tajima's D	-1.63	-2.21	0.012	-1.86	0.264
Non-BGC Sites					
Rate of fixed differences	0.0134	0.0020	< 0.001	0.0067	0.002
Fraction polymorphic	0.0736	0.0284	< 0.001	0.0485	< 0.001
Nucleotide diversity (π)	0.0027	0.0005	< 0.001	0.0014	0.007
θ_H	0.0036	0.0007	0.003	0.0020	0.144
Mean DAF	0.0430	0.0222	0.018	0.0357	0.292
Fraction of SNPs with DAF > 5%	0.1148	0.0567	0.007	0.0844	0.123
Tajima's D	-1.77	-2.31	0.002	-2.04	0.062

Table B10 Human-macaque substitution rates for all possible sequence contexts. We defined the sequence context of a given site by the two nucleotides flanking it. We estimated a substitution rate for each possible sequence context from whole-genome alignments of human and macaque. This table shows, for each context, the number of differences between human and macaque, the total number of sites with that context and the resulting estimate of the substitution rate.

Context	Diffs.	Total sites	Subs. rate
AA	37247	635919	0.0586
AC	25426	435275	0.0584
GT	25624	431978	0.0593
AG	34756	469773	0.0740
CC	32780	497279	0.0659
CA	26416	409031	0.0646
CG	43696	484736	0.0901
TT	36706	628903	0.0584
GG	32499	495249	0.0656
GC	26887	461432	0.0583
AT	31064	498507	0.0623
GA	24818	482909	0.0514
TG	26284	405665	0.0648
CT	34382	467994	0.0735
TC	24901	482315	0.0516
TA	26640	518413	0.0514

Table B11 Transcripts removed due to missing sequence or loss of protein function. Our set of genes consisted of 15,192 transcripts defined by the RefSeq database [146]. This table shows the number of transcripts that were removed for each non-human primate due to too much missing or undefined sequence (>50% of transcript sequences are "N" bases), frameshifts that disrupt more than 15 bp of sequence and premature stop codons that are more than 25 bp from the end of the sequence.

Species	Transcripts removed due to…			Transcripts removed	% removed	Transcripts remaining
	Undefined bp	Frameshift	Premature stop			
Chimpanzee	813	1100	155	2068	13.61	13124
Gorilla	858	607	129	1594	10.49	13598
Orangutan	395	3519	180	4094	26.95	11098
Gibbon	651	2782	202	3635	23.93	11557
Rhesus macaque	635	1204	193	2032	13.38	13160
Baboon	1085	1532	145	2762	18.18	12430
Vervet	1478	644	124	2246	14.78	12946
Colobus monkey	2259	652	123	3034	19.97	12158
Marmoset	607	2924	245	3776	24.86	11416
Tamarin	2781	907	129	3817	25.13	11375
Mouse lemur	3797	5673	91	9561	62.93	5631
Bushbaby	3060	7839	133	11032	72.62	4160

VITA

Renee D. George was born in St. Louis, Missouri. She graduated from Parkway Central High School in 2001 and then moved to Seattle to attend the University of Washington where she studied Biochemistry. Following her undergraduate studies, she worked at NanoString Technologies, a biotech startup in Seattle. In 2007, she returned to the University of Washington as a graduate student. In 2012, she received a Doctor of Philosophy in Genome Sciences. In her free time, Renee enjoys knitting, running, rock climbing and playing with her dog, Niña.