# Hotspots of Biased Nucleotide Substitutions in Human Genes

**Jonas Berglund**[1], **Katherine S. Pollard**[2], **Matthew T. Webster**[1*]

[1] Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden, [2] Gladstone Institutes, University of California, San Francisco, California, United States of America

**Genes that have experienced accelerated evolutionary rates on the human lineage during recent evolution are candidates for involvement in human-specific adaptations. To determine the forces that cause increased evolutionary rates in certain genes, we analyzed alignments of 10,238 human genes to their orthologues in chimpanzee and macaque. Using a likelihood ratio test, we identified protein-coding sequences with an accelerated rate of base substitutions along the human lineage. Exons evolving at a fast rate in humans have a significant tendency to contain clusters of AT-to-GC (weak-to-strong) biased substitutions. This pattern is also observed in noncoding sequence flanking rapidly evolving exons. Accelerated exons occur in regions with elevated male recombination rates and exhibit an excess of nonsynonymous substitutions relative to the genomic average. We next analyzed genes with significantly elevated ratios of nonsynonymous to synonymous rates of base substitution ($d_N/d_S$) along the human lineage, and those with an excess of amino acid replacement substitutions relative to human polymorphism. These genes also show evidence of clusters of weak-to-strong biased substitutions. These findings indicate that a recombination-associated process, such as biased gene conversion (BGC), is driving fixation of GC alleles in the human genome. This process can lead to accelerated evolution in coding sequences and excess amino acid replacement substitutions, thereby generating significant results for tests of positive selection.**

Citation: Berglund J, Pollard KS, Webster MT (2009) Hotspots of biased nucleotide substitutions in human genes. PLoS Biol 7(1): e1000026. doi:10.1371/journal.pbio.1000026

## Introduction

Whole-genome comparisons have revealed hundreds of noncoding elements that are extremely conserved across mammals but show evidence for accelerated evolution along the human lineage [1–4]. One possible explanation for these human-accelerated regions (HARs) is the action of positive selection in the human lineage. However, HARs tend to have biased patterns of nucleotide substitution, dominated by AT → GC changes—referred to here as "weak-to-strong" (W→S) as they result in a replacement of a "weak" A:T bond with a "strong" G:C bond. This pattern is strongly discordant with the genomic average, where S→W substitutions predominate. Positive selection is not expected to generate such biased patterns of base substitution. Interestingly, HARs also have a propensity to occur in regions with high recombination rates. These biased substitution patterns could potentially be explained by variation in the pattern of mutation, localized selection for increased GC content, or by biased gene conversion (BGC), which is a recombination-associated molecular drive that favors fixation of W→S mutations [5,6] and has population dynamics similar to natural selection [7].

There is now strong evidence for an association between recombination and patterns of nucleotide substitutions in the human genome, suggesting that an excess of W→S base substitutions occur in regions of high recombination [8]. The evidence can be summarized as follows: first, patterns of substitution in human–primate genomic alignments correlate with human recombination rates [8–10]. Second, parts of mammalian and avian genomes subject to very high recombination rates, such as duplicated gene families [11–13] and the X-linked pseudoautosomal region [14], are both extremely GC-rich and have GC-biased substitution patterns. Third, GC content correlates with recombination in a wide range of eukaryotes [15–17]. Fourth, experiments on primate cell lines and yeast indicate a bias in repair mechanisms, which leads to mismatches being preferentially repaired to GC bases [15,18,19]. Fifth, GC-biased clustered substitutions have been observed close to human recombination hotspots and near telomeres [20], and these regions also tend to be more GC-rich [21].

Several studies have also reported a correlation between sequence divergence and recombination rate [22–24]. It is therefore possible that recombination could directly influence patterns of substitution, although it is unknown whether mutations generated by recombination are W→S biased. The fact that the proportion of W→S mutations leading to human single-nucleotide polymorphisms (SNPs) is discordant with the proportion leading to nucleotide substitutions on the human lineage [25] strongly suggests a bias in fixation rather than mutation processes. Further support comes from observations that W→S and S→W changes segregate at different frequencies on average in human populations, particularly in regions with elevated recombination rates [6,25,26] (but see [27]).

**Academic Editor:** Laurence D. Hurst, University of Bath, United Kingdom

**Received** June 18, 2008; **Accepted** December 11, 2008; **Published** January 27, 2009

**Abbreviations:** BGC, biased gene conversion; FET, Fisher's exact test; GO, gene ontology; HAR, human-accelerated region; LRT, likelihood ratio test; MK test, McDonald-Kreitman test; ML, maximum likelihood; PAR, pseudoautosomal region; W→S, weak-to-strong

* To whom correspondence should be addressed. E-mail: matthew.webster@imbim.uu.se

## Author Summary

Regions of the human genome that appear to evolve rapidly may have been under strong positive selection and could contain the genetic changes responsible for the uniqueness of our species. However, neutral (nonadaptive) evolutionary processes can give rise to signals that can be mistaken as signs of selection. In this article, we identify coding sequences that have undergone accelerated rates of change in humans, affecting the divergence of the proteins they encode. By analyzing patterns of molecular evolution in these genes and their distribution in the genome, we show that many protein-coding changes in the fastest-changing genes are not a result of selection operating on the genes, but instead result from biased fixation of AT-to-GC mutations. Our findings are consistent with a model of recombination-driven biased gene conversion. This leads to the provocative hypothesis that many of the genetic changes leading to human-specific characters may have been prompted by fixation of deleterious mutations.

The distribution of recombination events is highly variable along vertebrate chromosomes. Recombination is mainly restricted to short (<1 kb) hotspots [28], which are extremely short-lived over evolutionary time [29]. Clusters of W→S biased substitutions have been observed on a similar scale, and it is proposed that these are the result of biased fixation of GC alleles in recombination hotspots [20]. Recombination hotspots therefore could be responsible for localized lineage-specific bursts of W→S biased substitutions, which could contribute to human-accelerated evolution in conserved noncoding elements.

Biased substitution patterns could also potentially result from selection on GC content. Mammalian genomes exhibit variation in GC content on the scale of hundreds of kilobases, commonly referred to as the isochore structure [30,31]. A potential explanation for this variation is that some regions experience selection in favor of increased GC content due to increased thermal stability [30]. In addition to this, experimental evidence indicates that GC-rich genes may be expressed with greater efficiency than GC-poor genes [32], which could lead to selection in favor of increased GC content in expressed sequences. Selection is more efficient on regions of high recombination due to a reduction in Hill-Robertson interference [33], which could lead to increased fixation of W→S mutations in these regions due to selection.

Some protein coding sequences also show patterns of evolution that are consistent with high levels of recombination-associated fixation of W→S mutations. For example, the *Fxy* gene is found on the X-specific portion of the X chromosome in human, rat, and short-tailed mouse (*Mus spretus*). However, in the house mouse (*M. musculus*), this gene has been translocated so that only its 5′ portion now resides in the X-specific region, whereas its 3′ portion overlaps the pseudoautosomal region (PAR), which is subject to very high levels of recombination [14]. This translocation resulted in a massive increase in substitution rate in the PAR portion of the gene, including substitutions that cause amino acid replacements. Since the common ancestor of *M. spretus* and *M. musculus*, the *M. spretus* lineage has accumulated one replacement substitution in the 3′ portion and one replacement substitution in the 5′ portion. In contrast, the *M. musculus* lineage has accumulated no substitutions in the 5′ portion, but 28 replacement substitutions in the PAR-overlapping 3′ portion [5]. Furthermore, all 28 substitutions are W→S.

Substitution patterns in the HARs and the *Fxy* gene may indicate that recombination-associated biased fixation of W→S mutations can compete with purifying selection, leading to the accumulation of weakly deleterious variants [5]. It is possible that this effect could affect common tests for positive selection in coding sequences, although the predicted impact of W→S fixation bias on these tests has not been demonstrated. For example, a gene under the influence of W→S fixation bias on a particular lineage could potentially acquire an increased ratio of nonsynonymous to synonymous rates of base substitution ($d_N/d_S$). This could lead to a significant acceleration in $d_N/d_S$, which is commonly assumed to indicate positive selection [34]. Similarly, fixation of weakly deleterious variants could potentially lead to an excess of amino acid replacement substitutions compared with polymorphism. This could generate significant results for the McDonald-Kreitman [35] test of neutrality, which could also lead to false inference of positive selection on protein sequence.

We examined the possibility that biased fixation of W→S mutations could affect the evolution of human protein-coding regions across the genome by analyzing patterns of evolution in a genome-wide set of human-chimpanzee-macaque 1:1:1 orthologous genes [36]. We first identified individual protein-coding exons with evidence for accelerated rates of nucleotide substitution in the human lineage using a likelihood ratio test (LRT). We then characterized patterns of nucleotide substitution in these loci. We also tested whether fast-evolving genes are associated with recombination hotspots or regions of elevated recombination. Our findings are consistent with the hypothesis that a recombination-associated process has generated an increased rate of nucleotide substitutions on the human lineage within particular genes since the split with chimpanzee. Interestingly, these genes also have increased numbers of amino acid replacement substitutions. This observation motivated us to theoretically and empirically examine the relationship between W→S fixation bias and rates of nonsynonymous base substitution. Our results suggest that W→S fixation bias can generate significant results for tests designed to detect directional selection, including LRTs for accelerated $d_N/d_S$ [34] and McDonald-Kreitman tests [35], potentially leading to false inference of positive selection.

## Results

We analyzed a dataset of 10,238 genes for evidence of accelerated evolutionary rates. We first divided the genes into 84,784 exons and used a LRT to identify individual exons with evidence for an increased rate of nucleotide substitution on the human lineage, considering both synonymous and nonsynonymous substitutions. Using this approach, we were able to identify the effects of local increases in substitution rate over a scale of ~1 kb, which are likely to affect only single exons and nearby flanking sequence. Significance was evaluated by simulating 10,000 datasets based on the null model of no acceleration, and correcting for multiple testing using the method of Benjamini and Hochberg [37]. The entire dataset is presented in Table S1.

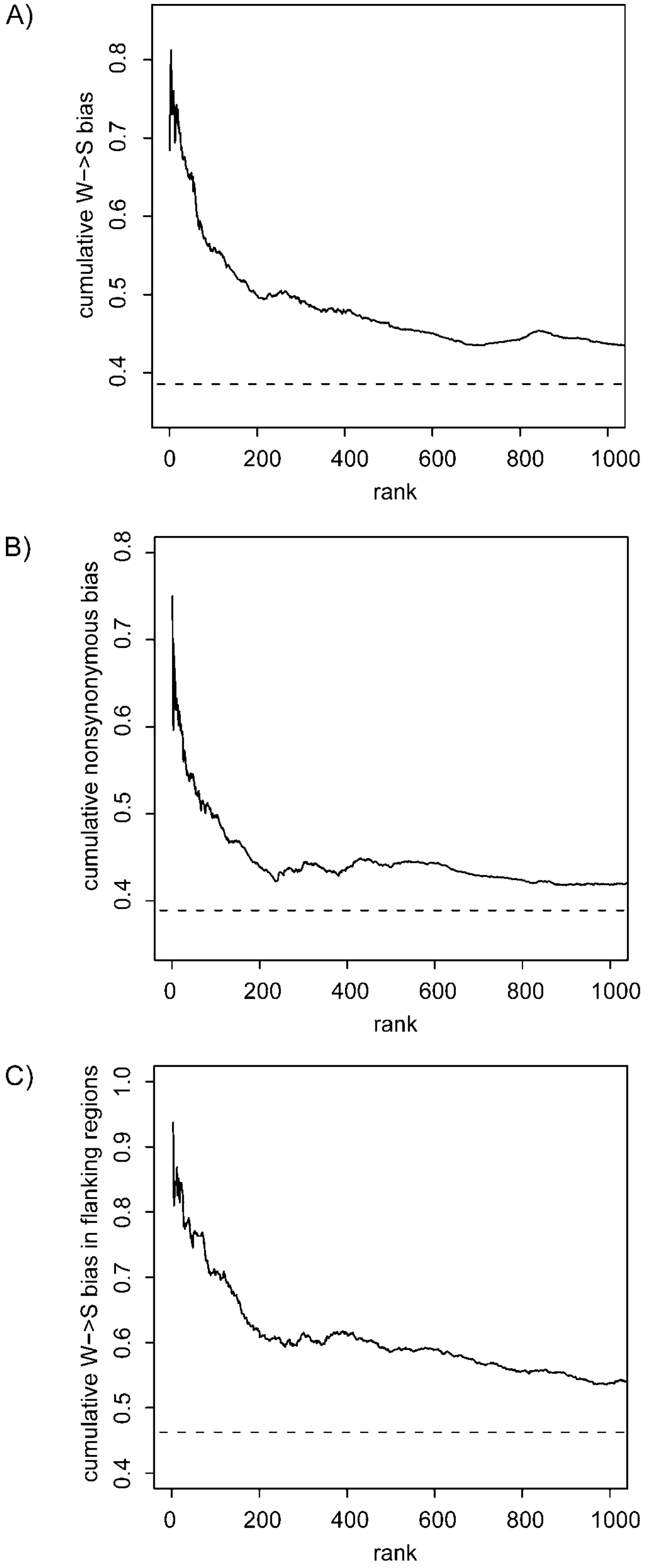

**Figure 1.** Cumulative Bias in Different Classes of Nucleotide Substitutions in the Exons with the Highest Degree of Acceleration on the Human Lineage

(A) The proportion of W→S substitutions compared to W→S and S→W substitutions on the human lineage.

(B) The proportion of nonsynonymous substitutions compared to all substitutions in each gene on the human lineage.

(C) The proportion of W→S substitutions compared to W→S and S→W substitutions on the human lineage in the flanking noncoding regions. Dashed lines represent averages for the entire dataset.

doi:10.1371/journal.pbio.1000026.g001

## Accelerated Exons Have Biased Patterns of Base Substitution

In total, 83 exons (in 82 genes) show evidence for acceleration in the human lineage with an expected false discovery rate (FDR) less than 5%. Henceforth, we refer to these 83 human-accelerated coding sequences as "accelerated exons." The mean length of the accelerated exons is 516.6 bp, which is higher than the mean length of exons in the entire dataset (167.6 bp). On average, the accelerated exons contain 7.73 substitutions on the human lineage, compared to a mean of 0.43 substitutions in the entire dataset. This suggests that the majority of exons are too short for an acceleration in evolutionary rate to be detectable by this method, given the evolutionary distance between human and chimpanzee.

Genes containing the accelerated exons have a mean of 13.3 human substitutions, compared with an average of 3.58 in all genes. Furthermore, there is a tendency for substitutions in these genes to be clustered: 11.0% of genes containing accelerated exons have a significantly nonuniform clustering of substitutions into the most diverged exon ($p <$ 0.05), compared to less than 1.1% of all genes (see Methods).

We estimated the pattern of substitution on the human and chimpanzee lineages by comparing the extant sequences with the maximum likelihood (ML) reconstructed ancestor using a codon model of substitution where the $d_N/d_S$ on the human branch was able to vary from the rest of the tree. The 83 accelerated exons demonstrate a bias for weak to strong (W→S) substitutions, with 326 W→S and 248 S→W base substitutions (W→S bias = 0.57; see Methods). This substitution pattern is significantly incongruent with the genome as a whole, where W→S bias = 0.39 (Fisher's exact test (FET) $p < 2.2 \times 10^{-16}$; bootstrap $p < 0.001$). This bias strongly affects the most accelerated exons and drops to close to the genomic average for exons with less evidence of acceleration (Figure 1A). Strikingly, the top 20 accelerated exons (Table 1) have 154 W→S compared to only 62 S→W substitutions (W→S bias = 0.71, FET $p < 2.2 \times 10^{-16}$; bootstrap $p < 0.001$). The positions of each nucleotide substitution on the human lineage in the genes containing the top 20 accelerated exons are shown in Figure 2. The excess of W→S substitutions can be clearly seen, and there is an obvious tendency for clusters of W→S substitutions to occur in single exons in particular genes.

We used the ML reconstructed human-chimpanzee ancestral sequences to infer the ancestral GC content of each exon. The accelerated exons have an average ancestral GC content of 0.53, whereas the top 20 accelerated exons have an average GC content of 0.54. In comparison, average ancestral GC content in all of the coding sequences in the dataset is 0.50. The elevated GC content of accelerated exons is observed at all three codon positions (unpublished data). Differences in GC content therefore cannot explain the differences in base substitution patterns between accelerated and nonaccelerated genes. If base substitution probabilities were constant across genomic regions, genes with higher GC content would be expected to have lower W→S substitution rates, when the opposite is actually observed. Our test for differences in the substitution patterns is therefore conservative.

The most accelerated exons also have a bias toward a greater number of nonsynonymous substitutions compared with the genomic average. The proportion of nonsynon-

**Table 1.** Patterns of Nucleotide Substitution in the Top 20 Accelerated Exons

| Rank | LRT | Gene | Exon | Chromosome | Subtelomeric | Length (bp) | Ancestral GC Content | Base Substitutions | | | | | | Nearest Hotspot (kb) | Recombination Rate | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | S→S | W→W | S→W | W→S | Non-Syn | Syn | | Average | Female | Male |
| 1 | 28.05 | *OR3A3* | 1 | 17 | 1 | 945 | 0.54 | 1 | 4 | 6 | 13 | 18 | 6 | 17.6 | 2.53 | 1.39 | 3.67 |
| 2 | 21.33 | *NAP1L4* | 6 | 11 | 0 | 132 | 0.38 | 0 | 0 | 0 | 10 | 3 | 7 | 56.2 | 2.38 | 2.03 | 2.72 |
| 3 | 19.80 | *OR4K17* | 1 | 14 | 0 | 1,029 | 0.41 | 0 | 0 | 6 | 12 | 10 | 8 | 6.9 | 1.05 | 1.16 | 0.93 |
| 4 | 19.42 | *KIF26B* | 1 | 1 | 1 | 718 | 0.72 | 0 | 0 | 0 | 17 | 17 | 0 | 8.3 | 3.42 | 2.63 | 4.20 |
| 5 | 17.00 | *CSTF2T* | 1 | 10 | 0 | 1,848 | 0.54 | 0 | 0 | 8 | 7 | 8 | 7 | 10.4 | 1.18 | 1.61 | 0.77 |
| 6 | 15.43 | *ODF1* | 2 | 8 | 0 | 430 | 0.50 | 1 | 0 | 4 | 6 | 5 | 6 | 1.4 | 1.74 | 2.95 | 0.54 |
| 7 | 15.15 | *C20ORF96* | 5 | 20 | 1 | 159 | 0.61 | 3 | 0 | 0 | 7 | 9 | 1 | 12.4 | 5.53 | 0.00 | 11.06 |
| 8 | 13.15 | *KCNV2* | 1 | 9 | 0 | 1,356 | 0.65 | 4 | 0 | 5 | 15 | 10 | 14 | 22.4 | 2.73 | 0.93 | 4.53 |
| 9 | 12.35 | *SLC35F5* | 10 | 2 | 0 | 65 | 0.43 | 0 | 0 | 0 | 5 | 3 | 2 | 79.6 | 1.31 | 0.98 | 1.64 |
| 10 | 12.29 | *CTRB1* | 1 | 16 | 0 | 52 | 0.63 | 2 | 0 | 3 | 2 | 5 | 2 | 7.0 | 0.61 | 0.86 | 0.36 |
| 11 | 11.89 | *WNK1* | 22 | 12 | 1 | 61 | 0.57 | 0 | 0 | 3 | 3 | 5 | 1 | 80.6 | 1.69 | 0.00 | 3.39 |
| 12 | 11.31 | *OR3A2* | 1 | 17 | 1 | 945 | 0.54 | 1 | 0 | 6 | 9 | 8 | 8 | 116.5 | 2.53 | 1.39 | 3.67 |
| 13 | 11.18 | *KRTAP19–7* | 1 | 21 | 0 | 189 | 0.58 | 0 | 0 | 8 | 5 | 9 | 4 | 100.6 | 1.11 | 0.94 | 1.28 |
| 14 | 10.64 | *ZC3H6* | 1 | 2 | 0 | 1,026 | 0.45 | 1 | 0 | 1 | 12 | 6 | 8 | 154.7 | 0.84 | 1.33 | 0.35 |
| 15 | 10.46 | *ZNF605* | 3 | 12 | 1 | 1,787 | 0.41 | 2 | 0 | 0 | 14 | 8 | 8 | 7.7 | 1.03 | 0.37 | 1.70 |
| 16 | 10.38 | *MYOM1* | 1 | 18 | 0 | 290 | 0.61 | 4 | 0 | 0 | 7 | 9 | 2 | 0.4 | 3.28 | 4.30 | 2.26 |
| 17 | 10.33 | *UGT3A1* | 5 | 5 | 0 | 232 | 0.50 | 1 | 1 | 3 | 3 | 6 | 2 | 77.3 | 0.80 | 1.61 | 0.00 |
| 18 | 9.95 | *TBC1D14* | 10 | 4 | 0 | 129 | 0.43 | 0 | 0 | 2 | 2 | 0 | 4 | 109.5 | 3.78 | 3.61 | 3.92 |
| 19 | 9.40 | *DLGAP2* | 11 | 8 | 1 | 216 | 0.60 | 0 | 1 | 0 | 5 | 1 | 5 | 17.9 | 3.00 | 0.79 | 5.23 |
| 20 | 9.03 | *CAP1* | 7 | 1 | 0 | 178 | 0.61 | 0 | 0 | 7 | 0 | 6 | 1 | 12.1 | 2.13 | 3.44 | 0.82 |
| Total | | | | | 7 | 11,787 | 0.54 | 20 | 6 | 62 | 154 | 146 | 96 | 45.0 | 2.13 | 1.62 | 2.65 |

doi:10.1371/journal.pbio.1000026.t001

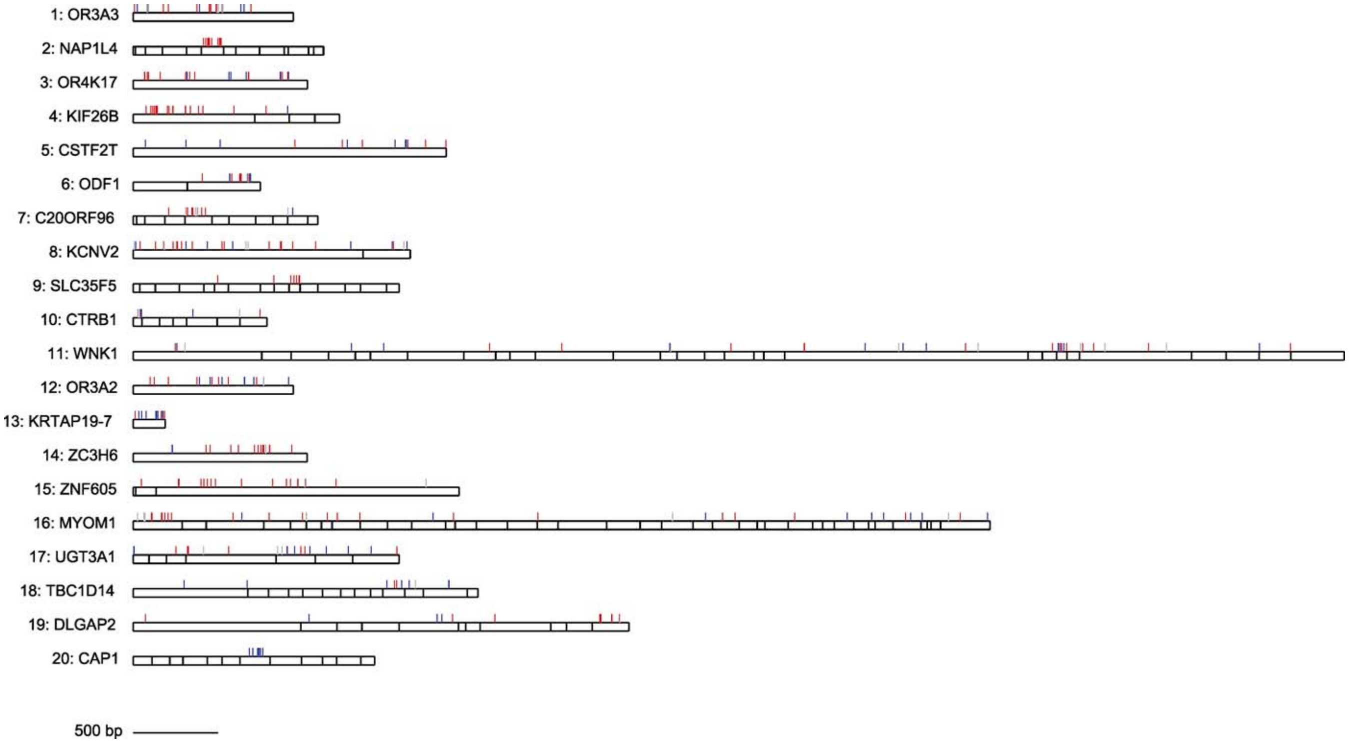

**Figure 2.** Genes Containing the Most Accelerated Exons

Exon boundaries are marked with black lines. S→W substitutions on the human lineage are marked with blue lines, W→S substitutions on the human lineage are marked with red lines, and all other substitutions on the human lineage are marked with grey lines.

doi:10.1371/journal.pbio.1000026.g002

ymous substitutions in the top 20 accelerated exons is 0.60 (146 nonsynonymous versus 96 synonymous), which is significantly higher than the proportion of 0.38 observed in the entire dataset (FET $p < 3.0 \times 10^{-11}$; bootstrap $p < 0.001$). This bias is most extreme for the most accelerated exons (Figure 1B). Nonsynonymous substitutions also have a tendency to exhibit a stronger W→S bias, particularly in the accelerated exons. The W→S bias of accelerated exons is 0.63 for nonsynonymous sites and 0.51 for synonymous sites (FET $p = 0.004$). In the top 20 accelerated exons, W→S bias is 0.75 for nonsynonymous sites and 0.66 for synonymous sites (FET $p = 0.129$). In the entire dataset, W→S bias is 0.41 for nonsynonymous sites and 0.37 for synonymous sites (FET $p = 2 \times 10^{-9}$).

## Flanking Noncoding Sequences Also Exhibit Biased Patterns of Base Substitution

To determine whether the W→S substitution bias is confined to protein-coding regions, we analyzed the pattern of base substitution in noncoding regions within 100 bp flanking both sides of each accelerated exon. Around the top 20 accelerated exons, there are 60 W→S but only 11 S→W base substitutions (W→S bias = 0.85). This is significantly different from the flanking regions surrounding all exons in the dataset (W→S bias = 0.46; FET $p < 3.2 \times 10^{-11}$; bootstrap $p < 0.001$). The W→S bias in flanking sequences is strongest for the most accelerated exons (Figure 1C). Average ancestral GC content in the regions flanking the top 20 accelerated exons is 0.48, which is higher than the average of 0.44 in the entire dataset. This suggests that differences in GC content could not be responsible for the differences in patterns of base substitution.

We analyzed patterns of substitution in progressively larger windows of noncoding sequence surrounding each exon to determine the scale at which the W→S bias in substitution patterns exists (Figure 3). We found that there is a marked decrease in W→S bias with increasing distance from the accelerated exon. The W→S bias in 10 kb of noncoding sequence on both sides of the 20 most accelerated exons approaches the genomic average. These results suggest that the process generating the W→S bias in substitution patterns in accelerated exons acts on a regional level, rather than specifically on coding sequences. Furthermore, strongly W→S biased substitution patterns seem to be restricted to a scale of less than a few kilobases.

## Accelerated Exons Occur in Regions of Elevated Male Recombination

We investigated the recombination rates of the regions where accelerated exons reside. We find that the most accelerated exons tend to be found in regions with elevated male recombination rates. In the top 20 accelerated exons, the average male recombination rate is 2.65, which is significantly higher than the average of all exons in the dataset (0.92; bootstrap $p < 0.001$). By contrast, female recombination rate in the top 20 accelerated exons is 1.62, which is not significantly higher than the genomic average of 1.69 (bootstrap $p = 0.45$). Male recombination rate is therefore highly elevated in the most accelerated exons, whereas female recombination rate remains relatively constant (Figure 4A).

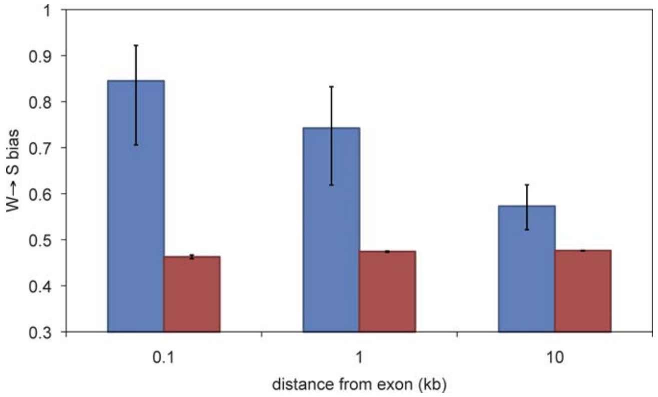

**Figure 3.** Patterns of Base Substitution in Flanking Regions

Average W→S bias of base substitution patterns in noncoding regions surrounding the top 20 accelerated exons (blue bars) compared with W→S bias in noncoding regions surrounding all exons in the dataset (red bars). 95% confidence intervals were estimated by bootstrapping with 1,000 replicates.

doi:10.1371/journal.pbio.1000026.g003

Accelerated exons show a slight tendency to occur near human recombination hotspots (Figure 4B). In the top 20 accelerated exons, average distance to a hotspot is 50.0 kb, compared to 54.8 kb in the entire dataset, although this difference is not significant (bootstrap $p = 0.16$). There is a highly significant tendency for accelerated exons to be found close to telomeres, where recombination rates are elevated in males [17]. Seven out of the top 20 accelerated exons are found in the last chromosome band. This proportion (0.35) is significantly higher than 0.075 observed in the entire dataset (bootstrap $p < 0.001$).

## Exons That Are Highly Diverged Compared to the Rest of the Gene Have Biased Substitution Patterns

We identified genes where one exon had greater divergence on the human lineage than the rest of the coding sequence, termed "relative divergence" (see Methods). Relative divergence was not calculated for exons with less than four substitutions to avoid bias from very short sequences. The 20 genes showing the highest relative divergence have most diverged exons with a significant excess of W→S substitutions compared with the entire dataset (FET $p = 0.00013$; Table 2). The spatial distribution of substitutions in the genes containing the top 20 most relatively diverged exons is presented in Figure S1. This pattern is also observed in the flanking regions of highly relatively diverged exons compared with the entire dataset (FET $p = 2.3 \times 10^{-9}$). Thus, when genes contain clusters of nucleotide substitutions in single exons, these substitutions tend to exhibit a W→S biased pattern. The observation that the patterns extend into flanking introns and intergenic sequence suggests that these patterns do not result from selection on protein coding sequence and that a regional bias in mutation or fixation of mutations is a more likely explanation.

## Patterns of Nucleotide Divergence and Polymorphism Are Discordant in Accelerated Exons

To distinguish between biases in patterns of mutation and fixation, we compared W→S bias in nucleotide substitutions with human SNPs from the HapMap project. Patterns of substitutions in accelerated exons are significantly more W→S biased than human SNPs (Table 3). These differences are highly significant for the top 20 accelerated exons (FET $p = 0.0015$) and for all 83 accelerated exons (FET $p = 0.00063$). These results suggest that either the W→S bias substitution patterns in these exons result from a mutation bias that is no longer active in the human population, or that the patterns result from a bias towards fixation of W→S mutations.

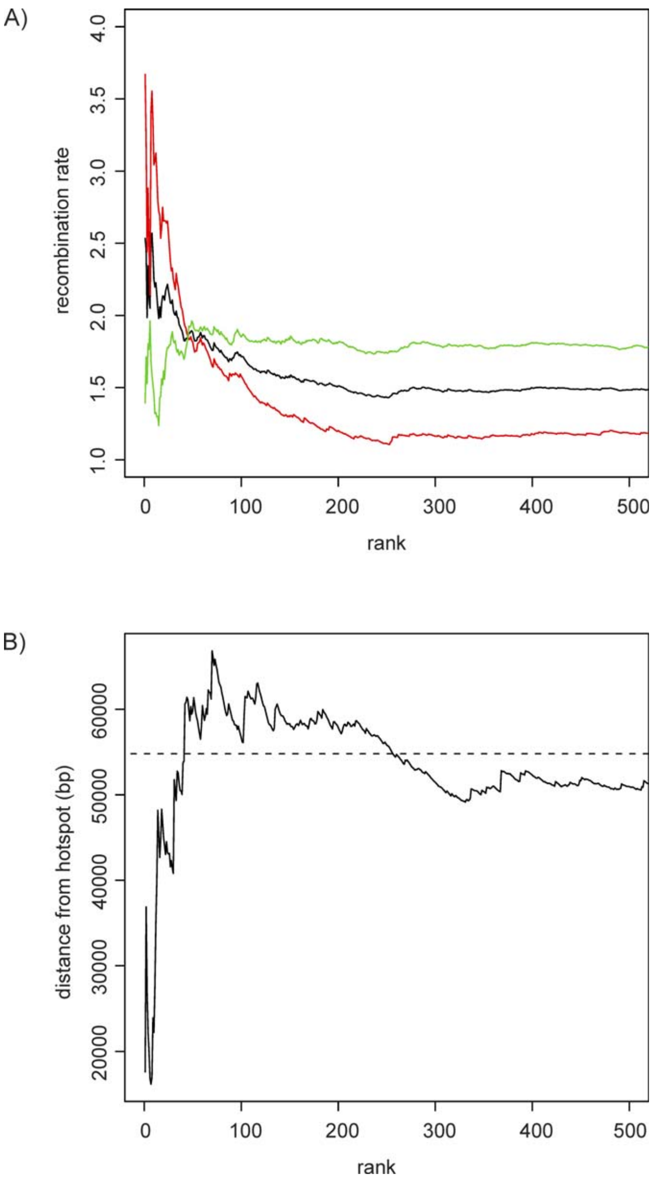

**Figure 4.** Cumulative Average Human Recombination Rate in the Regions Surrounding the Most Accelerated Exons

(A) Cumulative average male (red), female (green), and sex-averaged (black) recombination rates.

(B) Cumulative average distance to the nearest recombination hotspot. The dashed line represents the average for the entire dataset.

doi:10.1371/journal.pbio.1000026.g004

## Exons with Elevated $d_N/d_S$ in Humans Have Biased Substitution Patterns

We performed an additional LRT to identify individual exons with significantly elevated $d_N/d_S$ ratios on the human lineage. On average, we inferred 0.17 nonsynonymous and 0.26 synonymous substitutions per exon since the human-

**Table 2.** Patterns of Nucleotide Substitution in Exons with High Divergence Compared to the Remaining Exons of the Same Gene

| Category | Subset | Ancestral GC Content | S→S | W→W | S→W | W→S | W→S Bias |
|---|---|---|---|---|---|---|---|
| Exons | top 20 | 0.54 | 6 | 2 | 38 | 53 | 0.58 |
| | All | 0.52 | 2940 | 1093 | 11776 | 7239 | 0.38 |
| 100 bp flanking exons | top 20 | 0.48 | 3 | 0 | 6 | 43 | 0.88 |
| | All | 0.46 | 6720 | 4205 | 34962 | 30103 | 0.46 |

doi:10.1371/journal.pbio.1000026.t002

chimpanzee split, which suggests that estimates of $d_N/d_S$ are unreliable for most exons. We therefore restricted our analysis to exons with more than four substitutions inferred on the human lineage. Only 887 exons out of the entire dataset ($n = 84{,}784$) passed this criterion (1.0%). As shown in Table 4, exons with evidence for accelerated $d_N/d_S$ in humans tend to have more W→S biased patterns of nucleotide substitution. At the $p < 0.01$ (**) level, this is significant by FET ($p = 0.015$) but not bootstrap ($p = 0.104$). At the $p < 0.05$ (*) level, neither of the tests are significant (FET $p = 0.104$; bootstrap $p = 0.130$). Exons with accelerated $d_N/d_S$ on the human lineage therefore appear to be associated with W→S biased patterns of nucleotide substitution, although in the majority of exons, not enough substitutions have occurred to perform this test.

## Patterns of Base Substitution in Accelerated Exons Are Also Biased in Chimpanzee

The 83 accelerated exons have a significantly higher number of substitutions than average on the chimpanzee lineage. There is a base substitution in 0.0043 of sites on the chimpanzee lineage in the top 20 accelerated exons, compared to 0.0011 in all exons on the chimpanzee lineage (FET $p < 2.2 \times 10^{-16}$; bootstrap $p = 0.001$). This is an interesting observation, given that the LRT is designed to identify acceleration specifically on the human branch.

Accelerated exons show similar, although less pronounced, patterns of W→S biased substitutions in the chimpanzee lineage. The top 20 accelerated exons are inferred to have a W→S bias of 0.50, compared with 0.39 averaged across all exons. However, these values are not significantly different by FET ($p = 0.053$) or bootstrap ($p = 0.31$). The top 20 accelerated exons also have a greater-than-average proportion of nonsynonymous substitutions in the chimpanzee lineage, with 54 nonsynonymous and 30 synonymous substitutions, a bias of 0.64 toward nonsynonymous substitutions, compared with 0.42 in the entire dataset (FET $p < 5.35 \times 10^{-5}$; bootstrap $p = 0.002$). There is also a similar W→S bias in the noncoding regions flanking accelerated exons in the chimpanzee genome, with 18 W→S and 11 S→W substitutions (W→S bias = 0.62) compared with a W→S bias of 0.46 in all of the flanking sequences (FET $p = 0.093$; bootstrap $p = 0.146$). These results are consistent with previous studies suggesting that regional patterns of base substitution are correlated between human and chimpanzee [10,20,38].

**Table 3.** Patterns of Nucleotide Substitutions Compared with SNPs in Human-Accelerated Exons

| Category | Type | W→S | S→W | W→S Bias |
|---|---|---|---|---|
| Top 20 accelerated exons | substitutions | 154 | 62 | 0.71 |
| | SNPs | 8 | 14 | 0.36 |
| Significantly accelerated exons | substitutions | 326 | 248 | 0.57 |
| | SNPs | 28 | 50 | 0.36 |

doi:10.1371/journal.pbio.1000026.t003

## Genes with Elevated $d_N/d_S$ in Humans Have Biased Substitution Patterns

We next examined the relationship between rates of protein evolution and W→S substitution bias on the whole-gene level. We identified genes with significant evidence for accelerated nonsynonymous substitution rates on the human lineage using branch models of codon substitution and a LRT. We refer to these as "genes with accelerated $d_N/d_S$". Based on significance of rejection of a model with single $d_N/d_S$ ratio (codeml model 0) by a model where the human lineage had a separate $d_N/d_S$ (codeml model 2; see Methods), we defined three different levels of significance: $p < 0.001$ (***; 20 genes), $p < 0.01$ (**; 112 genes) and $p < 0.05$ (*; 485 genes). Table 5 shows the pattern of nucleotide substitution in genes at each level of significance. The distribution of these substitutions in the top 20 genes is presented in Figure S2. LRT statistics for all genes in our dataset are presented in Table S2.

Comparison of the substitution patterns in genes with accelerated $d_N/d_S$ to the entire dataset reveals a trend towards W→S biased substitution patterns, although this is not significant. However, when we restrict the analysis to the exon in each gene with the largest number of human substitutions per base, the substitution pattern in $d_N/d_S$ category *** exhibits a much higher W→S bias (0.68) than the average W→S bias for most diverged exons (0.39) in the entire dataset. This difference is highly statistically significant (FET $p = 7.1 \times 10^{-6}$; bootstrap $p = 0.0065$). The W→S bias in the most diverged exons of genes in $d_N/d_S$ category ** is less extreme (0.47), but still significantly different from the entire dataset (FET $p = 0.012$; bootstrap $p = 0.044$). In $d_N/d_S$ category *, the average W→S bias is lower (0.41), and not significantly different from the entire dataset. In addition to being the region with the most W→S biased substitution pattern, the most diverged exon in each gene also tends to have a larger proportion of nonsynonymous substitutions (Table 5).

We noticed that the gene *KIF26B*, kinesin family member 26B, contributes disproportionately to the number of substitutions in the genes in $d_N/d_S$ category ***. *KIF26B* is located in the last band of the q arm of human chromosome 1. Its most diverged exon (exon 1) contains 17 substitutions, which are all W→S. The most diverged exons in the

**Table 4.** Patterns of Nucleotide Substitution in Exons with Evidence for Elevated $d_N/d_S$ on the Human Lineage

| Significance Level | Number of Exons | Ancestral GC Content | S→S | W→W | S→W | W→S | W→S Bias |
|---|---|---|---|---|---|---|---|
| **-sig | 16 | 0.57 | 16 | 6 | 52 | 58 | 0.53 |
| *-sig | 58 | 0.55 | 62 | 17 | 213 | 177 | 0.45 |
| non-sig | 829 | 0.54 | 578 | 211 | 3,111 | 2,135 | 0.41 |
| total | 887 | 0.54 | 640 | 228 | 3,324 | 2,312 | 0.41 |

doi:10.1371/journal.pbio.1000026.t004

remaining 19 genes in $d_N/d_S$ category *** contain an average of just 2.6 substitutions. These exons have an average W→S bias of 0.55, compared with 0.68 when *KIF26B* is included, whereas the W→S bias among the most diverged exons is 0.39. The W→S bias in these 19 exons (with *KIF26B* excluded) is still significantly higher than average (FET $p = 0.035$; bootstrap $p = 0.019$). However, the evolution of *KIF26B* is particularly striking. It is the only gene that both contains one of the top 20 accelerated exons and has highly significant acceleration in $d_N/d_S$ (***). It also has a strongly W→S biased substitution pattern.

## No Evidence for W→S Bias in Noncoding Sequences Flanking Genes with Accelerated $d_N/d_S$

We examined patterns of base substitution in 100 bp of noncoding sequence flanking each side of all exons in the genes with accelerated $d_N/d_S$ (Table 6). There is no evidence that noncoding sequence nearby genes with accelerated $d_N/d_S$ at any of the levels of significance have W→S biased substitution patterns compared with genomic averages. This is the case when we analyze all the exons in genes with accelerated $d_N/d_S$ and when we analyze only the most diverged exon in each of these genes (tested by FET and bootstrap, unpublished data).

## Genes with Accelerated $d_N/d_S$ in Humans Are Closer to Recombination Hotspots

We examined whether genes with accelerated $d_N/d_S$ tend to be associated with regions of elevated recombination using a bootstrap test (Table 7). There is no significant tendency for accelerated genes at any of the $p$-value cutoffs to occur close to telomeres, or in regions of elevated male, female, or sex-averaged recombination. However, there is a highly significant ($p < 10^{-4}$) tendency for genes in $d_N/d_S$ significance category *** ($p < 0.001$) to be closer to recombination hotspots. This tendency is also significant for genes in $d_N/d_S$ category ** ($p = 0.032$) but not for those in $d_N/d_S$ category *.

## McDonald-Kreitman Tests for Selection

We examined patterns of nucleotide substitution in genes with evidence for an excess of amino acid replacement substitutions on the human lineage compared to human polymorphisms. These genes were identified using a modified version of the McDonald-Kreitman (MK) test [35] presented by Bustamante et al. [39]. A total of 3,878 genes in this dataset overlapped with the human-chimpanzee-macaque orthologues. Out of these, 20 show evidence for an excess of replacement amino substitutions at the $p < 0.01$ (**) level and 124 are significant at the $p < 0.05$ (*) level (Table 8). The proportion of W→S nucleotide substitutions is clearly elevated in these genes, and this is most pronounced in the most diverged exon of each gene. For the genes with the strongest excess of amino-acid substitutions (**), this increase is not significant for substitutions across the entire gene (FET $p = 0.055$; bootstrap $p = 0.11$), but is highly significant for substitutions in the most diverged exon (FET $p = 0.0012$; bootstrap $p = 0.008$). For all genes with evidence for an excess of amino-acid substitutions (*), the W→S bias is significant for substitutions across the entire gene (FET $p = 0.025$; bootstrap $p = 0.0446$), but not for substitutions in the most diverged exon (FET $p = 0.055$; bootstrap $p = 0.07$). MK tests based on HapMap SNP data (http://www.hapmap.org/) for all

**Table 5.** Patterns of Nucleotide Substitution in Genes with Human-Accelerated $d_N/d_S$

| Category | Significance Level | Number of Genes | $d_N/d_S$ | Ancestral GC Content | S→S | W→W | S→W | W→S | W→S Bias | Syn | Non-Syn | Non-Syn bias |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Genes | ***-sig | 20 | 0.77 | 0.56 | 16 | 2 | 73 | 60 | 0.45 | 45 | 106 | 0.70 |
| | **-sig | 112 | 1.47 | 0.53 | 67 | 22 | 340 | 242 | 0.42 | 133 | 538 | 0.80 |
| | *-sig | 485 | 0.97 | 0.53 | 251 | 80 | 1,299 | 833 | 0.39 | 672 | 1791 | 0.73 |
| | non-sig | 9,753 | 0.21 | 0.52 | 2,689 | 1,013 | 18,761 | 11,768 | 0.39 | 21,743 | 12,488 | 0.36 |
| | Total | 10,238 | 0.23 | 0.52 | 2,940 | 1,093 | 20,060 | 12,601 | 0.39 | 22,415 | 14,279 | 0.39 |
| Most diverged exons | ***-sig | 20 | - | 0.54 | 7 | 2 | 18 | 39 | 0.68 | 9 | 57 | 0.86 |
| | **-sig | 112 | - | 0.52 | 26 | 10 | 125 | 110 | 0.47 | 33 | 238 | 0.88 |
| | *-sig | 485 | - | 0.53 | 117 | 35 | 549 | 382 | 0.41 | 211 | 872 | 0.81 |
| | non-sig | 9,753 | - | 0.52 | 1,282 | 470 | 8,640 | 5,423 | 0.39 | 9872 | 5943 | 0.38 |
| | Total | 10,238 | - | 0.52 | 1,399 | 505 | 9,189 | 5,805 | 0.39 | 10083 | 6815 | 0.40 |

doi:10.1371/journal.pbio.1000026.t005

**Table 6.** Patterns of Nucleotide Substitution in Noncoding Regions Flanking Genes with Human-Accelerated $d_N/d_S$

| Category | Significance Level | Number of Genes | Ancestral GC Content | S→S | W→W | S→W | W→S | W→S Bias |
|---|---|---|---|---|---|---|---|---|
| 100 bp flanking all exons | ***-sig | 20 | 0.52 | 15 | 5 | 95 | 87 | 0.48 |
| | **-sig | 112 | 0.47 | 94 | 55 | 429 | 425 | 0.50 |
| | *-sig | 485 | 0.47 | 377 | 218 | 1868 | 1730 | 0.48 |
| | non-sig | 9753 | 0.46 | 6,343 | 3,987 | 33,094 | 28,373 | 0.46 |
| | Total | 10,238 | 0.46 | 6,720 | 4,205 | 34,962 | 30,103 | 0.46 |
| 100 bp flanking most diverged human exon | ***-sig | 20 | 0.52 | 1 | 2 | 16 | 10 | 0.38 |
| | **-sig | 112 | 0.48 | 7 | 6 | 60 | 42 | 0.41 |
| | *-sig | 485 | 0.47 | 51 | 16 | 226 | 189 | 0.46 |
| | non-sig | 9,753 | 0.46 | 856 | 500 | 4,325 | 3,631 | 0.46 |
| | Total | 10,238 | 0.46 | 907 | 516 | 4,551 | 3,820 | 0.46 |

doi:10.1371/journal.pbio.1000026.t006

genes in our dataset show similar patterns (unpublished data) but are subject to ascertainment bias and do not accurately reflect the true underlying SNP density. In summary, there is a clear association between excess amino acid replacement substitutions and W→S biased substitution patterns.

Table 9 shows the recombination rates of genes identified by the MK test. The genes with the strongest excess of nonsynonymous substitutions (**) are situated significantly closer to recombination hotspots (bootstrap $p = 0.026$) than the rest of the genes. For all significant genes, the mean distance to a hotspot is higher than average, although this is not significant ($p = 0.765$). There are no significant differences between the average or sex-specific recombination rates in genes with significant MK test values.

## Comparison of Fast-Evolving Genes Identified by Different Approaches

Figure 5 indicates the overlap between the main sets of fast-evolving genes we have identified. Fourteen genes with accelerated $d_N/d_S$ on the human lineage at the $p < 0.05$ level also contain accelerated exons. This is significantly higher than the 3.9 genes expected purely by chance (binomial test $p = 3.0 \times 10^{-5}$), although it is not surprising that genes with evidence for acceleration in the relative rate of nonsynonymous substitutions also show evidence for acceleration in evolutionary rates overall. Eleven of the genes with accelerated $d_N/d_S$ also show evidence for an excess of amino acid substitutions using the MK test [39] at the $p < 0.05$ level, which is larger than the 5.9 expected, but not significant ($p = 0.051$). Five of the genes with significant MK tests also contain accelerated exons, which is larger than the 2.0 expected, but not significant ($p = 0.054$). The three different tests therefore have a tendency to identify some of the same genes, but in general they appear to target genes with different evolutionary histories.

## Enrichment for Gene Ontology Categories

Out of the 82 genes containing accelerated exons, the gene ontology (GO) category "myosin complex" is enriched ($p = 0.0011$) due to the presence of five myosin complex protein coding genes (*MYOM1*, *MYO18B*, *MYO3B*, *MYH10*, and *MYH3*). The genes containing the top 20 accelerated exons contain three olfactory receptors (*OR3A3*, *OR3A2*, and *OR4K17*), which generates a significant enrichment for "neurological system process" ($p = 0.039$). The GO category "multicellular organismal process" is strongly enriched in all of the genes containing accelerated exons ($p = 5.09 \times 10^{-5}$), as well as those containing the top 20 accelerated exons ($p = 0.039$).

We tested the genes with accelerated $d_N/d_S$ for enrichment of particular GO categories. There is no evidence for enrichment for any GO category at the $p < 0.1$ level for any of the accelerated $d_N/d_S$ significance levels. We also tested for enrichment of GO categories within genes with significant MK tests at the $p < 0.05$ level (190 genes). There was a significant enrichment for "calcium ion binding" among genes with evidence of recent positive selection (21 genes, $p = 0.027$).

**Table 7.** Recombination Rate in Genes with Human-Accelerated $d_N/d_S$

| Significance Level | Number of Genes | Number Telomeric | Nearest Hotspot (kb) | Recombination Rate (cM/Mb) | | |
|---|---|---|---|---|---|---|
| | | | | Average | Female | Male |
| ***-sig | 20 | 1 | 15.9 | 1.46 | 1.67 | 1.25 |
| **-sig | 112 | 8 | 32.7 | 1.21 | 1.53 | 0.86 |
| *-sig | 485 | 38 | 40.0 | 1.24 | 1.55 | 0.90 |
| non-sig | 9,753 | 726 | 41.7 | 1.34 | 1.69 | 0.96 |
| Total | 10,238 | 764 | 41.6 | 1.33 | 1.68 | 0.96 |

doi:10.1371/journal.pbio.1000026.t007

**Table 8.** Patterns of Nucleotide Substitution in Genes with Evidence for Positive Selection Based on MK Tests

| Category | Significance Level | Number of Genes | Ancestral GC Content | S→S | W→W | S→W | W→S | W→S Bias |
|---|---|---|---|---|---|---|---|---|
| **Genes** | **-sig | 20 | 0.51 | 6 | 9 | 74 | 66 | 0.47 |
| | *-sig | 124 | 0.50 | 67 | 30 | 388 | 297 | 0.43 |
| | non-sig | 3,754 | 0.51 | 1,220 | 482 | 8,533 | 5,417 | 0.39 |
| | Total | 3,878 | 0.51 | 1,287 | 512 | 8,921 | 5,714 | 0.39 |
| **Most diverged exons** | **-sig | 20 | 0.51 | 2 | 1 | 21 | 34 | 0.62 |
| | *-sig | 124 | 0.50 | 22 | 10 | 167 | 138 | 0.45 |
| | non-sig | 3,754 | 0.51 | 535 | 208 | 3,533 | 2,281 | 0.39 |
| | Total | 3,878 | 0.51 | 557 | 218 | 3,700 | 2,422 | 0.40 |

doi:10.1371/journal.pbio.1000026.t008

## Minimal Effect of Ancestral Misinference

Another cause of apparent W→S substitutions could be misinference of ancestral bases. This is particularly problematic at CpG sites, where multiple CpG mutations on different lineages (always S→W) could give a false inference of the reverse (W→S) substitution due to homoplasy. In order to quantify this, we performed BLAST searches against three additional closely related primate genomes (gorilla, orangutan, and baboon) for the genes containing the top 20 accelerated exons, and the genes with strongest evidence ($p <$ 0.001, ***) for accelerated $d_N/d_S$. All of the bases in both of these datasets were alignable to at least one of the three primates. We were able to align 95% of bases to at least two of the species and 84% of bases to all three species.

We compared the human-chimpanzee ancestral bases, inferred from the human-chimpanzee-macaque alignments by ML, to the orthologous bases in the additional primate genomes. We did not identify a single case where the ML inferred ancestral base of a human-specific substitution was incongruent with the orthologous base in the most closely related primate species. This indicates the effect of ancestral misinference is negligible in the human-accelerated sequences.

## Modeling the Effect of a W→S Fixation Bias on the $d_N/d_S$ Ratio

Our empirical results suggest that genes with elevated rates of nucleotide substitution have been affected by a fixation bias in favor of W→S mutations. Such a bias could be caused by BGC or directional selection on GC content. We also observe W→S biased substitution patterns in genes with (a) accelerated $d_N/d_S$ ratios on the human lineage, and (b) elevated numbers of nonsynonymous changes in substitution versus polymorphism data in MK tables. This suggests that a W→S fixation bias could potentially generate an increased rate of nonsynonymous compared with synonymous substitutions.

To investigate this issue, we used a theoretical model of the interaction between a W→S fixation bias and purifying selection in an ideal Wright-Fisher population with parameter values determined empirically from the exon dataset. We assume that a W→S fixation bias can be modeled using a selection coefficient, as demonstrated by Nagylaki [7]. We estimated the pattern of mutation in genes with different ancestral GC contents by analyzing substitutions in 4-fold degenerate (4d) sites. We used these to estimate the relative number of mutations expected in each mutational class, given the ancestral sequences. We then calculated the probability of fixation of each mutation, based on the selective coefficient and effective population size ($N_e$; assumed to be 10,000). The selective coefficients were determined by combining a bias (*f*) that favors fixation of all W→S mutations and loss of all S→W mutations with a distribution of negative fitness effects on nonsynonymous mutations (*c*) derived by Eyre-Walker et el. [40]. We calculated the predicted substitution rate in each mutational class from the product of mutation and fixation probabilities.

As expected, increasing the fixation bias (*f*) in favor of W→S mutations results in a W→S bias in the pattern of substitution (Figure 6A). This effect is most pronounced for GC-poor genes, whose substitution patterns have a higher degree of W→S mutational bias in the absence of a W→S fixation bias. A significant effect of *f* on the W→S bias can be observed once $f > 1/4N_e$ ($2.5 \times 10^{-5}$), which is the approximate selection coefficient required for a new mutation under selection to have a higher probability of fixation than a neutral mutation. For values of $f > 10^{-4}$, the W→S bias approaches 1.

**Table 9.** Recombination Rate in Genes with Evidence for Positive Selection Based on MK Tests

| Significance Level | Number of Genes | Number Telomeric | Nearest Hotspot (kb) | Recombination Rate (cM/Mb) | | |
|---|---|---|---|---|---|---|
| | | | | Average | Female | Male |
| **-sig | 20 | 1 | 24.8 | 1.16 | 1.40 | 0.86 |
| *-sig | 124 | 11 | 42.8 | 1.13 | 1.60 | 0.88 |
| non-sig | 3,754 | 239 | 37.7 | 1.34 | 1.70 | 0.96 |
| Total | 3,878 | 250 | 37.9 | 1.34 | 1.70 | 0.95 |

doi:10.1371/journal.pbio.1000026.t009

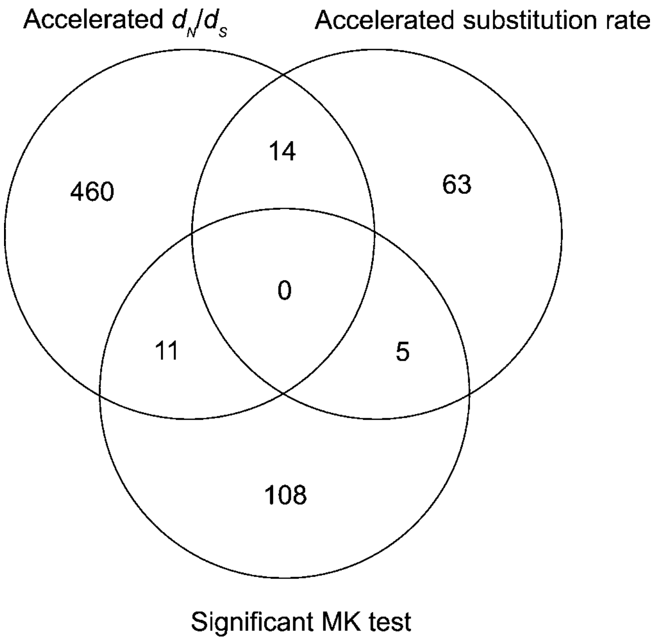

**Figure 5.** Venn Diagram Showing Overlap between Three Different Subsets of Fast-Evolving Genes

Genes with evidence for accelerated $d_N/d_S$ on the human lineage based on a LRT p<0.05 using a chi-squared test are shown in one circle. Genes containing exons with evidence for accelerated evolutionary rate in humans based on simulations with a FDR $p < 0.05$ are in the second circle. Genes with evidence for a significant McDonald-Kreitman test [39] with $p < 0.05$ are in the third circle.

doi:10.1371/journal.pbio.1000026.g005

Perhaps more surprisingly, at values of $f > 10^{-4}$, W→S fixation bias is also predicted to increase the $d_N/d_S$ ratio (Figure 6B). Hence, when $f$ is high it appears to override the effects of negative selection, leading to an increased proportion of nonsynonymous fixations. Interestingly, the effect of $f$ is much more pronounced on GC-rich genes, and can potentially lead to $d_N/d_S > 1$ (typically assumed to indicate positive selection). This phenomenon can be explained by an observation that W→S mutations in GC-rich genes have a greater probability of occurring in nonsynonymous sites. In GC-poor genes (ancestral GC = 0.3–0.4), we predict 47% of new W→S mutations to be nonsynonymous, whereas in GC-rich genes (ancestral GC = 0.6–0.7), we predict this proportion to be 66%. This difference is likely due to synonymous sites in GC-rich genes already being saturated with W→S substitutions. The effect of a W→S fixation bias is predicted to have a similar effect on genes under different levels of selective constraint (Figure S3).

## Discussion

We have identified a large number of exons with significantly accelerated evolutionary rates on the human lineage. The pattern and distribution of nucleotide substitutions in these exons suggests that a recombination-associated process, such as BGC [6] or strong localized selection in favor of increased GC content, is responsible for accelerated substitution rates in fast-evolving exons. The main support for this hypothesis is (a) an excess of W→S substitution in accelerated exons, (b) similar patterns of W→S bias in noncoding sequence flanking these exons, and (c) an enrichment of accelerated exons in regions of elevated male recombination and near the ends of chromosome arms. Accelerated exons also have an excess of replacement amino acid substitutions on the human lineage, which suggests that the process governing their evolution can compete with purifying natural selection.

### W→S Biased Evolution of Protein-Coding Sequences

Homologous recombination events between a pair of chromosomes that are heterozygous at a particular locus can lead to the formation of a heteroduplex DNA molecule during meiosis. The BGC hypothesis proposes that when the heteroduplex contains a weak/strong (AT/GC) mismatch, this is preferentially repaired to the strong allele [6]. This implies that weak/strong heterozygotes transmit more GC than AT alleles to the next generation, particularly at loci in regions of high recombination. Theoretical modeling has shown that this leads to biased fixation of W→S substitutions, with dynamics similar to natural selection [7], consistent with the patterns of evolution we observe in accelerated exons.

An alternative hypothesis is that directional selection for increased GC content has been operating on accelerated exons and their flanking regions. Clusters of highly expressed genes have been found to occur in regions of elevated GC content [41,42], and it is therefore possible that GC content has a direct effect on gene expression [43]. Hence, another explanation for our findings is that selection for increased gene expression has driven a local increase in GC content in the accelerated exons. mRNA structure and isochore GC content [30] are other possible sources of selective pressure. Without deeper understanding of the role of GC content in gene expression and chromosome evolution, it is difficult to hypothesize why selection on GC content would affect single exons and their flanking sequences rather than chromosomal domains or spliced transcripts.

Across the entire genome, clusters of human and chimpanzee nucleotide substitutions have a significant tendency to be W→S biased and to occur in regions of elevated recombination [20]. Such clusters are also observed in conserved noncoding elements with evidence for acceleration in humans (HARs) [1,5]. Here we have shown that protein-coding sequences are also subject to this unusual phenomenon. Exons with elevated substitution rates in humans exhibit a striking excess of W→S biased substitutions compared with all exons in the dataset. These substitutions also show a strong tendency to be clustered in single exons, rather than the entire gene. Similar W→S biased patterns can be observed in surrounding noncoding sequence, decaying sharply to background levels with increasing distance from the accelerated exon, so that W→S bias in 10 kb on each side of a accelerated exons approaches the genome average.

The observation that W→S biased substitution patterns extend into surrounding noncoding sequence strongly suggests that natural selection acting at the protein level could not be responsible for the biased fixation of GC alleles. However, it appears that clusters of W→S substitutions in accelerated exons are extremely localized to within a few kilobases around the exon. We also observe that when genes have evidence for clustering of base substitutions in a single exon, these substitutions exhibit a strong W→S bias. Furthermore, there is a strong discordance between levels of W→S bias in polymorphism compared with divergence,

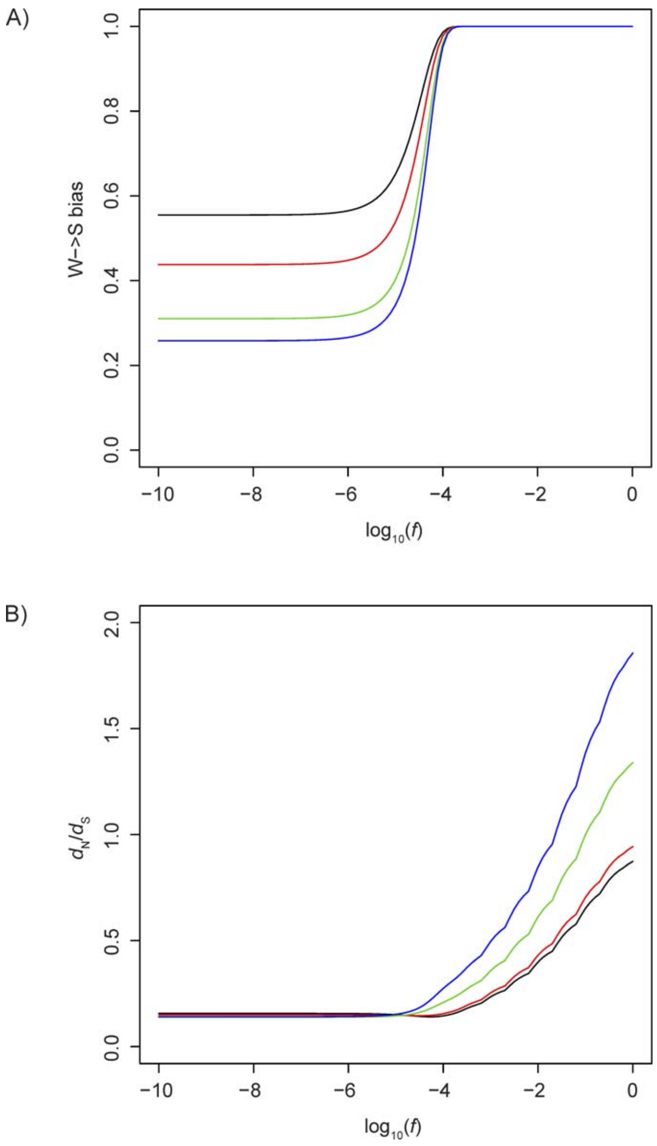

**Figure 6.** Predicted Effect of a Bias Towards Fixation of W→S Mutations, Considered as a Selective Coefficient (*f*), on Coding Sequences under Purifying Selection

(A) Effect of *f* on the pattern of nucleotide substitutions
(B) Effect of *f* on the $d_N/d_S$ ratio
The line colors represent ancestral GC content (0.3–0.4, black; 0.4–0.5, red; 0.5–0.6, green; 0.6–0.7, blue).
doi:10.1371/journal.pbio.1000026.g006

indicative of a bias towards fixation of W→S mutations. All of these observations are consistent with the action of BGC or localized selection in favor of increased GC content driving the evolution of the human-accelerated coding sequences we have identified.

It should be noted that not all accelerated exons have W→S biased substitution patterns. In particular, adenylate cyclase-associated protein 1 (CAP1) on human chromosome 1p34 contains a cluster of seven substitutions that are all S→W in exon 7. Six of these substitutions are nonsynonymous. None of the substitutions appear to be the result of CpG hypermutability. Furthermore, there is no evidence of a local increase in base substitution rate in the noncoding regions flanking exon 7; there are no substitutions on the human lineage within 100 bp of noncoding sequence on each side of this exon. It is possible that human-specific positive selection has contributed to acceleration in evolutionary rate in this exon. However, the biased pattern of base substitution suggests a bias in the pattern of mutation or fixation has also contributed, although the cause of this bias is unclear. Our method for analyzing accelerated evolutionary rates in single exons could potentially be a promising approach for identifying genes involved in human-specific adaptations.

## The Effect of Recombination Hotspots on Genome Evolution

Accelerated exons show a highly significant tendency to occur in regions of the human genome with elevated male recombination, consistent with the results of Dreszer et al. [20]. A correlation between the proportion of W→S substitutions and male recombination rate in humans is also observed in studies using human-chimpanzee-macaque comparisons across the genome [10], and in patterns of evolution in Alu repeats [9]. In addition, we find a significant enrichment of accelerated exons close to telomeres, similar to what has been observed for clusters of base substitutions in general and for HARs. Furthermore, like HARs [1,2], accelerated exons tend to be closer to recombination hotspots.

Most recombination in humans is restricted to short ($<1$ kb) hotspots where recombination rates are $>10\times$ the genomic average [28]. If BGC were able to generate nucleotide substitutions, we would expect them to be concentrated in these regions. Selection in favor of increased GC content would also be expected to be more efficient in these regions, due to a reduction in Hill-Robertson interference [33]. Recombination hotspots are believed to arise rapidly and become rapidly extinguished, potentially because of the "hotspot conversion paradox" [44]. Hotspot turnover is suggested by differences in the location of hotspots between human and chimpanzee [45,46], effects of certain allelic variants on recombination rate in known hotspots [47,48], and theoretical modeling [49]. In contrast, large-scale recombination rates are correlated between humans and chimpanzees [46]. Assuming that a strong W→S fixation bias is associated with recombination hotspots, then localized clusters of W→S substitutions would be expected to occur in bursts, mirroring the rapid turnover of recombination hotspots.

Due to their ephemeral nature, it is extremely difficult to measure the impact of hotspots in a particular genomic region since the human-chimpanzee ancestor. The two measures we used are large-scale recombination rates estimated from pedigree data [17] and a map of human recombination hotspots generated by analyzing patterns of linkage disequilibrium in the HapMap dataset. Estimates of the regional recombination rate and distance from the nearest hotspot of a particular genomic location are only rough indicators of the average density of recombination hotspots in that region since the human-chimpanzee split. Another problem is that we only have measures of when recombination events are resolved as crossovers and cannot directly measure the frequency or length of gene conversion events. Hence, although the clusters of substitutions we observe in accelerated exons are consistent with the action of an intense W→S fixation bias in recombination hotspots, we cannot reconstruct or implicate the locations of specific ancient hotspots.

We also observe W→S biased substitution patterns in the chimpanzee lineage for exons that are accelerated in humans, although the degree of bias is weaker. This suggests that high rates of recombination could also affect these exons in the chimpanzee lineage. Analysis of nucleotide substitutions across the entire human and chimpanzee genomes show similar patterns [10]. These findings are consistent with conservation of the average density of hotspots between human and chimpanzee, which could affect patterns of substitution in both lineages. It should be noted that we do not expect W→S fixation bias due to selection or BGC to be specific to the human lineage, and we would generally expect different exons to be accelerated on the chimpanzee lineage.

It is unclear why male recombination shows a stronger correlation with clusters of W→S biased substitutions (and with the pattern of W→S biased substitution across the genome) than does female recombination. This difference is not predicted by a model of selection on GC content modulated by Hill-Robertson effects on the efficacy of selection. It is possible that there is a stronger correspondence between the rate of crossovers and gene conversion in males than in females, which would cause the strength of BGC to correlate more strongly with male recombination.

The site frequency spectrum should be altered in regions where a W→S fixation bias is occurring, due to an increase in frequency of GC alleles. Several studies have demonstrated that GC alleles segregate at elevated frequencies across the genome in human populations, and that this effect is stronger near recombination hotspots [21,26,50]. However, it has been suggested that these findings may have been influenced by a systematic error in inferring the ancestral state of SNPs [27]. An alternative theory for how recombination could generate W→S substitutions is by a direct mutagenic effect [23,51]. However, neither ancestral misinference nor a mutagenic effect of recombination can explain why the excess of GC alleles at elevated frequency increases in the vicinity of recombination hotspots [10]. In particular, the allele frequency distribution at a recently arisen recombination hotspot should be skewed towards low frequency GC alleles if recombination generated W→S mutations, which is the opposite of what is observed. A recent model of the effect of BGC on the pattern of base substitution [10] is a good fit to observed patterns across the human genome, suggesting that BGC—rather than a direct effect of recombination—can account for the patterns of molecular evolution observed in the human-accelerated coding regions we have identified. It is also possible that a reduction in Hill-Robertson interference in regions of high recombination could result in clusters of W→S biased substitutions due to selection. However, it is currently unclear whether this process would result in the observed variation in W→S bias and substitution rate over short physical distances.

### W→S Fixation Bias Drives Amino Acid Replacement Substitutions

An important finding of this study is that biased fixation of W→S mutations can drive replacement amino acid substitutions. In addition to W→S biased substitution patterns, accelerated exons exhibit an excess of nonsynonymous to synonymous changes compared with the genomic average. This is consistent with previous suggestions that BGC may compete with purifying selection, resulting in the fixation of deleterious mutations [5]. This process may have occurred in noncoding HARs, which are generally extremely highly conserved between mammalian species other than humans, and probably play important functional roles (e.g., in regulation of gene expression). An increase in mutation rate is not expected to increase the proportion of nonsynonymous to synonymous changes, as it would be expected to remove the same proportion of nonsynonymous changes in regions of high or low mutation rates. Natural selection on amino acid sequence is also not expected to generate the W→S biased substitution patterns we observe in accelerated exons, which extend into noncoding flanking sequence.

Using theoretical modeling, we have shown that W→S fixation bias is predicted to increase the $d_N/d_S$ ratio under realistic assumptions regarding the strength of bias and distribution of negative fitness effects on nonsynonymous mutations. This effect is observable with a fixation bias corresponding to a selective coefficient $>10^{-4}$ (assuming $N_e = 10{,}000$). An important simplifying assumption made by our model is that the W→S fixation bias can be modeled in an identical way to positive selection [7]. In reality there is likely to be a complex interaction between the two processes. In particular, the effects of selection may extend to distant linked sites, whereas the effects of BGC are likely to be confined to short conversion tracts. Further work is necessary to fully understand this interaction. Our model uses the method of Li [52] to predict the $d_N/d_S$ ratio, whereas we used a codon-based ML method [53] to estimate this ratio from our alignments. Although the two methods may give slightly different estimates of $d_N$ and $d_S$ under certain scenarios, we do not expect this discrepancy to influence our prediction that a W→S fixation bias increases the $d_N/d_S$ ratio. One assumption of codon models of substitution is that codon frequencies are at equilibrium, which is unlikely to be true at loci where a fixation bias is operating. However, the choice of model is not likely to lead to misinference of ancestral bases at short genetic distances, such as between human and chimpanzee.

The strength of BGC depends on the rate of formation of heteroduplex DNA, the size of the hetoroduplex tracts, and the strength of the repair bias, all of which are difficult to estimate empirically. By comparing a large number of studies in yeast, Birdsell [15] has estimated a that GC/AT mismatches are repaired to GC with a bias of about 1.5. Recombination rates of >50 cM/Mb are predicted to be common in localized (1–2 kb) hotspots in the human genome [28], and even higher rates have been observed in individual hotspots [54]. We cannot predict how often a particular site in a recombination hotspot will be involved in biased repair from these figures. However, it does not seem unrealistic that fixation biases $>10^{-4}$ could occur at particular sites due to BGC if they are regularly included in recombination events in hotspots.

### Function of Genes Containing Accelerated Exons

We observed enrichment of certain GO categories in genes containing accelerated exons. Olfactory receptor proteins were overrepresented in the top 20 accelerated exons, and the genes containing the 83 accelerated exons were enriched for myosin complex genes. It is notable that these genes belong to superfamilies with many paralogs. In addition to allelic gene conversion events between homologues, gene conversion also occurs between duplicated paralogous genes. Previous studies have suggested that gene conversion between paralogs generates W→S biased patterns of base substitution in gene families [11,12]. Hence it is possible that extremely high levels of BGC between paralogs has contributed to the accelerated evolution and biased patterns of substitution we observe. Gene conversion is likely to occur more frequently between physically linked gene duplicates [11]. It is interesting to note that two of the fast-evolving olfactory receptors, *OL3A3* and *OL3A2*, lie within 200 kb of each other on the last band of chromosome 17p, close to a number of other olfactory receptors. Although none of the myosin complex genes containing accelerated exons occur on the same chromosome, *MYH3* lies within a cluster of myosin family genes spanning 300 kb on chromosome 17p13.

Olfactory receptors are part of the largest supergene family in mammals. Several of these genes are believed to be under positive selection in humans, although disproportionately large numbers have become pseudogenes in humans compared with chimpanzee [55]. It is possible that excess fixations caused by BGC in human olfactory receptors are tolerated because purifying selection is relaxed in these genes in humans, and in some cases this has resulted in pseudogenization. However, it is also possible that previous reports of an enrichment of olfactory receptors amongst genes with accelerated $d_N/d_S$ on the human lineage [56] could be influenced by BGC. Because our dataset is constructed from 1:1:1 orthologues, we cannot observe the effects of gene conversion between recent duplicates. However, BGC between closely related paralogs could potentially have a major influence on their evolution.

## Evolution of *KIF26B*

One gene, kinesin family member 26B (*KIF26B*), shows a particularly striking pattern of evolution on the human lineage. The most accelerated exon of this gene is exon 1, which is 718 bp long and has 17 substitutions on the human lineage. All of these substitutions are W→S and nonsynonymous. The remaining three exons lie $>$ 9 kb away from the first and have just two substitutions (one W→S and one S→W) within 500 bp. This gene is also the only one with a highly significantly accelerated $d_N/d_S$ (***) that also contains one of the top 20 accelerated exons.

The molecular evolution of *KIF26B* in primates strongly parallels the *Fxy* gene in rodents. The 3′ portion of *Fxy* has been translocated into the highly recombining pseudoautosomal region (PAR) in *M. musculus*, whereas in the closely related *M. spretus* the entire gene is nonrecombining. This translocation coincides with a massive increase in the substitution rate in the 3' end of *Fxy*; *M. musculus* has 28 nonsynonymous substitutions, all of which are W→S, compared with only one nonsynonymous substitution in *M. spretus* [14]. The human substitutions at *KIF26B* are also almost exclusively nonsynonymous (18 out of 19) and the ML inferred $d_N/d_S$ ratio for the human branch is 1.75. This extreme substitution pattern may indicate the involvement of both positive selection and BGC. However, *KIF26B* has an extremely high GC content (0.72), and our theoretical modelling suggests that a W→S fixation bias can potentially result in $d_N/d_S > 1$ in such GC-rich genes in the absence of positive selection. This is because synonymous sites are more saturated with W→S substitutions in GC-rich genes so that W→S mutations have a greater probability of occurring in nonsynonymous sites, which could also explain why the W→S bias is greater in nonsynonymous sites in accelerated exons. It is possible that positive selection can promote compensatory amino acid replacement substitutions after deleterious mutations become fixed due to BGC, but these substitutions would not all be expected to be W→S, as observed at the *KIF26B* locus.

## W→S Fixation Bias Can Influence Tests of Positive Selection

We identified genes with evidence for increased rates of $d_N/d_S$ on the human lineage using an LRT. We find that the most diverged exons in these genes have significantly W→S biased substitution patterns. This pattern is not seen in the most diverged exons of genes overall, indicating that W→S fixation bias may affect the evolution of the genes with the strongest evidence for accelerated rates of amino acid substitutions in humans. The strongest signal for W→S bias occurs in the exons with the largest proportion of amino acid replacement substitutions. These observations suggest that a W→S fixation bias could contribute to elevated levels of $d_N/d_S$ and possibly lead to false inference of positive selection at the protein level.

Genes with accelerated $d_N/d_S$ differ from accelerated exons in several ways. First, we do not observe significantly W→S biased substitution patterns in the noncoding regions associated with genes with accelerated $d_N/d_S$. Nonetheless, the significantly elevated GC content of these regions suggests that they may have experienced W→S biased substitution patterns previously. Second, compared to accelerated exons, genes with accelerated $d_N/d_S$ show a weaker association between substitution rates and local recombination rates. These genes are significantly enriched close to human recombination hotspots, but they are not enriched in regions of elevated recombination (measured from the DECODE recombination map [17]) or in distal chromosome regions. These findings suggest that while the signature of recombination-associated W→S fixation bias is observable within genes with elevated $d_N/d_S$ in humans, they have been affected to a lesser extent than the accelerated exons. Alternatively, W→S biased substitution patterns in genes with accelerated $d_N/d_S$ may result from different evolutionary processes. For example, one factor that could contribute to the relationship between recombination hotspots and $d_N/d_S$ is the higher efficiency of natural selection in regions of high recombination, due to a reduction in the strength of Hill-Robertson interference [33].

We also observed an increase in W→S bias in genes with an excess of amino acid replacement substitutions relative to human polymorphism identified using the modified MK test [35] of Bustamante et al. [39]. The largest W→S bias occurs in genes with the most significant values of the test. W→S bias is particularly marked in the most diverged exon of each gene. These genes are also significantly enriched close to human recombination hotspots, although their average recombination rates are not significantly different from the genomic average. This pattern is consistent with a W→S fixation bias driving additional weakly deleterious amino acid substitutions, mainly restricted to single exons. Our observations are consistent with the past existence of transient recombination hotspots in these regions, which are, in general, no longer actively driving fixation of GC alleles in the human population. However, as we do not observe a strong correlation between present day recombination rates and significant MK test results, it is possible that the W→S fixation bias is unrelated to recombination in these genes.

Overlap between genes identified by the two tests of positive selection is not significantly higher than expected by chance, which is indicative that they identify genes with different evolutionary histories. The $d_N/d_S$ LRT identifies genes with increased $d_N/d_S$ ratios on the human lineage. These genes would not necessarily exhibit a significant excess of amino acid substitutions relative to polymorphism, unless they were under the influence of strong positive selection, or experienced recent shifts in their mode of evolution. In contrast, the MK test compares patterns of divergence and polymorphism, but does not detect whether the rate of protein evolution has changed on the human lineage. The accelerated exons do not all show a strong excess of nonsynonymous changes, which explains the limited overlap between them and genes identified by tests of selection. It is notable that all three tests identify coding sequences with W→S biased patterns of substitution.

The effect of W→S fixation bias at a particular locus is likely to depend on a variety of factors, including the time scale and intensity of a W→S drive and the locus-specific interaction with natural selection. All of these factors are predicted to vary between loci, likely due to stochastic variation in the strength and location of recombination hotspots over time. It is therefore not surprising that signals of a W→S fixation bias can be found using a variety of tests for increased evolutionary rates. Importantly, our results suggest that a W→S fixation bias, rather than positive selection on protein function, could be responsible for generating significant tests for selection in some genes, which cause us to urge care in the interpretation of these tests.

## Conclusion

We have presented evidence that protein-coding sequences with accelerated rates of evolution in humans have significantly biased patterns of nucleotide substitutions. These results are consistent with a strong effect of W→S fixation bias on the evolution of the most rapidly evolving coding exons in our genome. This process may have led to the increased fixation of replacement amino acid changes on the human lineage, and may bias tests of positive selection.

## Materials and Methods

**Data.** We analyzed a dataset of 10,376 alignments of 1:1:1 human-chimpanzee-macaque orthologous genes presented in the rhesus macaque genome paper [36] and available from http://compgen.bscb.cornell.edu/orthologs/. The alignments consisted of a filtered dataset of orthologous genes derived from known human protein coding genes identified from the RefSeq [57], Vega [58], and UCSC known gene [59] annotations. Genes with poor syntenic relationships, incomplete alignments, frame-shift indels, changes in exon-intron structure, and evidence for recent duplications had all been excluded from the dataset [36]. Annotation files for all of the genes were downloaded from Biomart (http://www.ensembl.org/biomart/martview/) and UCSC (http://genome.ucsc.edu/). These were used to identify the exon boundaries in all of the alignments and their location in the hg18 human genome sequence assembly. A small number of genes ($<1\%$) were excluded due to poorly matching gene annotation data. We finally excluded alignments with ten or more bases in runs of mismatches of three or more between any of the sequences, resulting in a dataset of 10,238 genes.

Human recombination rates, based on the DECODE map [17] were obtained from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables). Positions of human recombination hotspots as inferred from coalescent analysis of large-scale SNP genotyping data were downloaded from the HapMap website (http://www.hapmap.org/downloads/). All analyses were based upon the hg18 human genome assembly, aligned to the chimpanzee (panTro2) and the macaque genome (rheMac2). The liftover tool, available from the UCSC website, was used to convert human annotation to the hg18 assembly where needed (http://hgdownload.cse.ucsc.edu/downloads.html).

We also constructed alignments of the noncoding sequence flanking each exon in the coding dataset. To do this, we first obtained pairwise chained and netted blastz alignments of the hg18 versus panTro2 assemblies and the hg18 versus rheMac2 assemblies from the UCSC website (http://hgdownload.cse.ucsc.edu/downloads.html). These were converted into a human-chimpanzee-macaque alignment for each human chromosome using tools from the multiz package [60]. Finally, we masked all exons and extracted flanking sequence on both sides of each exon using the gene annotation files.

**Analysis of nucleotide substitutions.** We used the phast package [61] to analyze the rate of nucleotide substitution individually for each exon sequence alignment, implementing the general time-reversible (REV) model. We compared a model with relative branch lengths (i.e., relative substitution rates) equal to those from a genome-wide model to a model where the human branch is longer (i.e., has an accelerated substitution rate). We used an LRT to identify exons with statistically significant substitution rate acceleration on the human branch [1].

We used the codeml program of PAML [62] with F3x4 codon frequencies and the Goldman and Yang [53] model of codon substitution to infer the pattern of synonymous and nonsynonymous nucleotide substitutions at each gene on the human and chimpanzee branches of the tree under two models usingML. We first used the one-ratio model, where the $d_N/d_S$ ratio was fixed along all lineages of the tree. We compared this with the two-ratio model, were $d_N/d_S$ was allowed to vary along the human lineage. Sequences with accelerated $d_N/d_S$ on the human branch were identified with an LRT. We repeated this analysis on the whole-gene alignments and on individual exon alignments. For the single exon analysis, codons that overlapped between two consecutive exons were removed from the alignments.

We compared the ML reconstructed human-chimpanzee ancestral sequence from the two-ratio model with the human sequence to determine the pattern of substitutions along the human lineage. Substitutions were divided into four different classes: strong-to-strong (S→S), strong-to-weak (S→W), weak-to-strong (W→S), and weak-to-weak (W→W). "Weak" designates A or T base pairs, which are bound by only two hydrogen bonds. "Strong" designates C or G base pairs, which are bound by three hydrogen bonds. We defined the W→S bias of a genomic region as follows: W→S bias $= n_{W\rightarrow S} / (n_{W\rightarrow S} + n_{S\rightarrow W})$, where $n_{W\rightarrow S}$ and $n_{S\rightarrow W}$ are the number of W→S and S→W substitutions, respectively. Substitutions were also classified as nonsynonymous or synonymous. Multiple substitutions in the same codon were taken into account using the same criteria as single substitutions. These were inferred by ML to account for only 0.9% of substitutions. Substitutions in noncoding alignments surrounding each exon, and the GC content of the ancestral sequences, were inferred using parsimony. Human-specific substitutions were identified and assigned to the four different categories above by determining the ancestral state of each site using the macaque sequence.

We analyzed whether there was a tendency for nucleotide substitutions to cluster using a statistic that captures the relative substitution rate in the most diverged exon of a gene compared to the overall substitution rate in the gene:

$$T = \frac{\max_{i\in I}\{s_i/l_i\}}{\sum_{i\in I} s_i \Big/ \sum_{i\in I} l_i},$$

where $i \in I$ indexes the exon, $s_i$ is the number of single base substitutions in exon $i$, and $l_i$ is the length of exon $i$ (in bases). Genes with large values of $T$ have an exon that has a higher substitution rate than expected given the overall rate of substitutions across exons in the gene. For each gene, we conducted a simulation to assess the statistical significance of the observed value of $T$. Fixing the exon boundaries at the observed positions, we uniformly placed the observed number of substitutions $\sum_{i\in I} s_i$ at random sites across the gene and calculated the value of $T$. Repeating this substitution assignment 1,000 times provides a null distribution for the statistic $T$, under the assumption that the substitution process is uniform. An empirical $p$-value can be calculated as the proportion of the 1,000 null $T$ values that exceed the observed $T$ value.

**Identifying accelerated evolution in humans.** We estimated the total number of substitutions of each type for every gene, dividing substitutions into individual exons. We also calculated the GC content of each exon, the distance to the nearest recombination hotspot, the recombination rate, and whether each gene is in the last chromosome band. We ranked each exon according to its degree of acceleration in evolutionary rate on the human lineage, taking all substitutions into account, using the LRT statistic with the REV model of nucleotide substitution. Significance was estimated by simulating 10,000 datasets from the null model and calculating the LRT statistic for each exon. The $p$-value for each exon was estimated as the number of simulated LRTs that exceed the observed value. These p-values were adjusted for multiple testing using the FDR controlling method of Benjamini & Hochberg [37].

We also classified genes according to evidence of a significantly different $d_N/d_S$ along the human lineage. We divided genes into three levels of significance based on comparing the LRT statistic to the chi-square distribution: $p < 0.001$ (***), $p < 0.01$ (**), and $p < 0.05$ (*). We ranked exons according to their level of "relative divergence," using the statistic $T$ defined above. Significant differences in the patterns of nucleotide substitution in the most accelerated exons and genes in all categories were identified using FET and by bootstrapping each exon with 10,000 replicates. The FET assumes that each substitution is an independent data point, whereas the bootstrap test considers each exon independently thereby accounting for correlation between substitutions within an exon.

**Comparison with human polymorphism.** We identified human SNPs within our alignments using data from the HapMap project. We obtained the position and alleles of >15 million SNPs on the hg18 human genome build using the HapMart tool available at http://hapmart.hapmap.org/BioMart/martview. We then determined the ancestral allele of each SNP that overlapped one of our alignments by comparison with the chimpanzee base at that position. To minimize errors due to ancestral misinference, only biallelic SNPs where one allele matched both the chimpanzee and macaque sequence were included in the analysis.

We identified genes in our dataset with evidence for an excess of amino acid replacement substitutions relative to human polymorphism based on the genome scan of Bustamante et al. [39]. This analysis estimated the selection coefficient from MK contingency tables [35] of polymorphism and divergence at synonymous and nonsynonymous sites. The posterior probability of the selection coefficient was used to estimate the probability that a gene is under positive selection (has an excess of amino acid replacement substitutions). Only genes that appeared in our dataset and in the dataset of Bustamante et al. [39] were included in the analysis. We compared patterns of nucleotide substitutions between genes with evidence for positive selection with the entire dataset using FET and bootstrapping each gene with 10,000 replicates. The FET assumes that each substitution is an independent data point, whereas the bootstrap test considers each gene independently.

**Enrichment for GO categories.** We tested whether genes containing exons with significantly accelerated rates of base substitution, genes with significantly accelerated $d_N/d_S$ ratios, and genes with significant MK tests were enriched for particular GO categories. We performed these analyses using GOstat [63], available at the website http://gostat.wehi.edu.au/.

**Homology searches.** To detect potential incidences of ancestral misidentification in our dataset, we performed BLAST searches of the most accelerated genes and exons in our alignments against the NCBI trace archives from three other primate sequences: gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus abelii*), and baboon (*Papio hamadryas*). These sequences were aligned to the existing alignments and used to identify human-chimpanzee mismatches where the ML inferred ancestral base was incongruent with the orthologous base in the additional species. In cases where bases were not in concordance between the three additional species, the base from the species most closely related to human and chimpanzee was compared with the inferred ancestral base.

**Modeling the effect of a W→S fixation bias on coding sequence evolution.** Nagylaki [7] demonstrated that BGC can be modeled using a selection coefficient. We therefore assume that BGC and a W→S fixation bias due to selection can be modeled in the same way. We modeled the effect of a W→S fixation bias by applying the inferred neutral mutation rate, a realistic distribution of negative fitness effects on nonsynonymous sites, and a range of values of a selection coefficient that favors fixation of W→S mutations and loss of S→W mutations, to the inferred ancestral sequences. We first estimated the pattern of neutral mutation on the human lineage using the inferred pattern of substitution at fourfold degenerate sites from our codeml analysis with the two-ratio model (see above). The mutation pattern was estimated separately within four categories of ancestral GC content (0.3–0.4, 0.4–0.5, 0.5–0.6, 0.6–0.7). The relative probability of every possible mutation at each base in each GC content category was calculated as follows for each of the 12 possible single base mutations. The A→C mutation is shown as an example:

$$p_{AC} = n_{AC}/(n_{AA} + n_{AC} + n_{AG} + n_{AT})$$

where $n_X$ is the number of sites in category X and X is a mutational type (e.g., $n_{AC}$ is the number of sites with A→C mutations).

We next concatenated the ancestral sequences of all genes in each GC content category. For each concatenated sequence, we calculated the expected relative rate of synonymous and nonsynonymous mutations in each of the following mutational classes (W→S, S→W, W→W, S→S) by adapting the method of Li [52] as follows. First, the number of nondegenerate ($L_0$), 2-fold degenerate ($L_2$), and 4-fold degenerate sites ($L_4$) in the ancestral sequence were counted. A site is nondegenerate if all possible mutations at that site are nonsynonymous, 2-fold degenerate if one of the possible mutations is synonymous, and 4-fold degenerate if all possible changes are synonymous. The one possible case of a 3-fold degenerate site is treated as 2-fold degenerate. Each possible mutation at every site in the ancestral sequence was then classified as a transition or a transversion. The relative numbers of expected transitional ($S_i$) and transversional ($V_i$) mutations ($i$ = 0, 2, 4) at each type of site in the entire sequence were then calculated separately for W→S, S→W, W→W, and S→S mutations as the sum of the relative probability of each possible transition and transversion at each site. Finally, we calculated the expected relative rate of synonymous and nonsynonymous mutations separately for each of the four mutational classes ($u_{WS(syn)}$, $u_{SW(syn)}$, $u_{WW(syn)}$, $u_{SS(syn)}$, $u_{WS(nonsyn)}$, $u_{SW(nonsyn)}$, $u_{WW(nonsyn)}$, $u_{SS(nonsyn)}$), using the values of $S_i$, $V_i$, and $L_i$ to estimate $d_N$ and $d_S$ according to Li [52]:

We simulated the selective coefficient ($s$) for each of the eight classes of mutation based on the combined effects of a W→S fixation bias ($f$) and selective constraint ($c$). The values of $s$ for synonymous changes are: $s_{WS(syn)} = f$, $s_{SW(syn)} = -f$, $s_{WW(syn)} = 0$, and $s_{SS(syn)} = 0$. The W→S fixation bias alters the probability of fixation of W→S and S→W mutations, but W→W and S→S mutations are assumed to evolve completely neutrally. The values of $s$ for nonsynonymous changes are: $s_{WS(nonsyn)} = f - c$, $s_{SW(nonsyn)} = -f - c$, $s_{WW(nonsyn)} = -c$ and $s_{SS(nonsyn)} = -c$. Selective constraint on a nonsynonymous mutations depends on a distribution of negative fitness effects, which is commonly modeled using a gamma distribution. We therefore sampled $c$ from a descretized gamma distribution with shape parameter 0.23 and mean 0.0425, assuming an effective population size ($N_e$) of 10,000. This distribution was inferred by Eyre-Walker et al. [40] to be a good fit to the distribution of fitness effects of SNPs segregating in the human population, and is in good concordance with other studies (e.g., [64]). We considered a range of values of $f$ between $10^{-10}$ and 1.

We calculated the probability of fixation separately for mutations in each of the 8 classes using the following equation derived by Kimura [65]:

$$p = (1 - e^{-2s})/(1 - e^{-4Ns})$$

where $N$ is the population size of an ideal Wright-Fisher population, which we assume to be 10,000. The predicted relative substitution rates ($K$) at each mutational class were then calculated by multiplying their probability of fixation, $P$, by their relative mutation rates, $u$. We calculated K for each value of $f$ (between $10^{-10}$ and 1), separately for each ancestral GC content category. We used these rates to calculate $d_S$, $d_N$, and W→S bias by summing across the different mutational categories.

## Supporting Information

**Figure S1.** Genes Containing the Most Relatively Diverged Exons

Exon boundaries are marked with black lines. S→W substitutions on the human lineage are marked with blue lines, W→S substitutions on the human lineage are marked with red lines and all other substitutions on the human lineage are marked with grey lines. Relative divergence is calculated as described in the methods.

Found at doi:10.1371/journal.pbio.1000026.sg001 (268 KB EPS).

**Figure S2.** Top 20 Genes with Strongest Evidence for Accelerated $d_N/d_S$ Based on LRT

Exon boundaries are marked with black lines. S→W substitutions on the human lineage are marked with blue lines, W→S substitutions on the human lineage are marked with red lines and all other substitutions on the human lineage are marked with grey lines.

Found at doi:10.1371/journal.pbio.1000026.sg002 (262 KB EPS).

**Figure S3.** The Predicted Effect of Different Levels of Constraint on the Relationship between W→S Fixation Bias and $d_N/d_S$

A selective coefficient ($f$) is used to represent the W→S fixation bias. The level of constraint is fixed at the following values: (a) $2 \times 10^{-5}$, (b) $4 \times 10^{-5}$, (c) $8 \times 10^{-5}$, (d) $1.6 \times 10^{-4}$. The line colors represent ancestral GC content (0.3–0.4, black; 0.4–0.5, red; 0.5–0.6, green; 0.6–0.7, blue).

Found at doi:10.1371/journal.pbio.1000026.sg003 (294 KB EPS).

**Table S1.** LRT Statistics for Human Acceleration and Pattern of Substitution on the Human Lineage for All Exons

Found at doi:10.1371/journal.pbio.1000026.st001 (9.13 MB CSV).

**Table S2.** LRT for Accelerated $d_N/d_S$ on Human Branch Compared with Constant $d_N/d_S$ across the Tree for All Genes

Found at doi:10.1371/journal.pbio.1000026.st002 (327 KB CSV).

## Acknowledgments

We thank Adam Siepel for providing alignments. We thank Jeff Wall, Mike Zody, Elinor Karlsson, and Kerstin Lindblad-Toh for helpful discussions.

**Author contributions.** KSP and MTW conceived and designed the experiments. JB, KSP, and MTW performed the experiments and analyzed the data. KSP and MTW wrote the paper.

**Funding.** JB and MTW were funded by The Swedish Research Council and Erik-Philip Sörensens Stiftelse. KSP was supported by National Institutes of Health (NIGMS) grant #GM82901-01A1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests.** The authors have declared that no competing interests exist.

**References**

1. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, et al. (2006) Forces shaping the fastest evolving regions in the human genome. PLoS Genet 2: e168. doi:10.1371/journal.pgen.0020168
2. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, et al. (2006) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443: 167–172.
3. Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, et al. (2007) Fast-evolving noncoding sequences in the human genome. Genome Biol 8: R118.
4. Prabhakar S, Noonan JP, Paabo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. Science 314: 786.
5. Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet 23: 273–277.
6. Galtier N, Piganeau G, Mouchiroud D, Duret L (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159: 907–911.
7. Nagylaki T (1983) Evolution of a finite population under gene conversion. Proc Natl Acad Sci U S A 80: 6278–6281.
8. Meunier J, Duret L (2004) Recombination Drives the Evolution of GC-Content in the Human Genome. Mol Biol Evol 21: 984–990.
9. Webster MT, Smith NG, Hultin-Rosenberg L, Arndt PF, Ellegren H (2005) Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. Mol Biol Evol 22: 1468–1474.
10. Duret L, Arndt PF (2008) The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet 4: e1000071. doi:10.1371/journal.pgen.1000071
11. Galtier N (2003) Gene conversion drives GC content evolution in mammalian histones. Trends Genet 19: 65–68.
12. Backstrom N, Ceplitis H, Berlin S, Ellegren H (2005) Gene conversion drives the evolution of HINTW, an ampliconic gene on the female-specific avian W chromosome. Mol Biol Evol 22: 1992–1999.
13. Kudla G, Helwak A, Lipinski L (2004) Gene conversion and GC-content evolution in mammalian Hsp70. Mol Biol Evol 21: 1438–1444.
14. Montoya-Burgos JI, Boursot P, Galtier N (2003) Recombination explains isochores in mammalian genomes. Trends Genet 19: 128–130.
15. Birdsell JA (2002) Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol 19: 1181–1197.
16. Fullerton SM, Bernardo Carvalho A, Clark AG (2001) Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol 18: 1139–1142.
17. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, et al. (2002) A high-resolution recombination map of the human genome. Nat Genet 31: 241–247.
18. Brown TC, Jiricny J (1989) Repair of base-base mismatches in simian and human cells. Genome 31: 578–583.
19. Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454: 479–485.
20. Dreszer TR, Wall GD, Haussler D, Pollard KS (2007) Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. Genome Res 17: 1420–1430.
21. Spencer CCA, Deloukas P, Hunt S, Mullikin J, Myers S, et al. (2006) The influence of recombination on human genetic diversity. PLoS Genet 2: e148. doi:10.1371/journal.pgen.0020148
22. Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet 72: 1527–1535.
23. Lercher MJ, Hurst LD (2002) Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet 18: 337–340.
24. Filatov DA (2004) A gradient of silent substitution rate in the human pseudoautosomal region. Mol Biol Evol 21: 410–417.
25. Webster MT, Smith NG, Ellegren H (2003) Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Mol Biol Evol 20: 278–286.
26. Webster MT, Smith NG (2004) Fixation biases affecting human SNPs. Trends Genet 20: 122–126.
27. Hernandez RD, Williamson SH, Bustamante CD (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24: 1792–1800.
28. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321–324.
29. Coop G, Przeworski M (2007) An evolutionary view of human recombination. Nat Rev Genet 8: 23–34.
30. Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, et al. (1985) The mosaic genome of warm-blooded vertebrates. Science 228: 953–958.
31. Eyre-Walker A, Hurst LD (2001) The evolution of isochores. Nat Rev Genet 2: 549–555.
32. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol 4: e180. doi:10.1371/journal.pbio.0040180
33. Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res 8: 269–294.
34. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568–573.
35. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652.
36. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, et al. (2007) Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. Science 316: 222–234.
37. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc [Ser B] 57: 289–300.
38. Smith NG, Webster MT, Ellegren H (2002) Deterministic mutation rate variation in the human genome. Genome Res 12: 1350–1356.
39. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153–1157.
40. Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173: 891–900.
41. Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, et al. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res 13: 1998–2004.
42. Hurst LD, Pál C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5: 299–310.
43. Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD (2003) A unification of mosaic structures in the human genome. Hum Mol Genet 12: 2411–2415.
44. Boulton A, Myers RS, Redfield RJ (1997) The hotspot conversion paradox and the evolution of meiotic recombination. Proc Natl Acad Sci U S A 94: 8058–8063.
45. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, et al. (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308: 107–111.
46. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, et al. (2005) Fine-scale

recombination patterns differ between chimpanzees and humans. Nat Genet 37: 429–434.

47. Jeffreys AJ, Neumann R (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat Genet 31: 267–271.
48. Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P (2005) Human recombination hot spots hidden in regions of strong marker association. Nat Genet 37: 601–606.
49. Coop G, Myers SR (2007) Live hot, die young: transmission distortion in recombination hotspots. PLoS Genet 3: e35.
50. Duret L, Semon M, Piganeau G, Mouchiroud D, Galtier N (2002) Vanishing GC-rich isochores in mammalian genomes. Genetics 162: 1837–1847.
51. Haldane JBS (1948) The rate of mutation of human genes. Proceedings of the 8th International Congress of Genetics. Heriditas Suppl 35: 267–273.
52. Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36: 96–99.
53. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736.
54. Kauppi L, Jeffreys AJ, Keeney S (2004) Where the crossovers are: recombination distributions in mammals. Nat Rev Genet 5: 413–424.
55. Gilad Y, Man O, Glusman G (2005) A comparison of the human and chimpanzee olfactory receptor gene repertoires. Genome Res 15: 224–230.
56. 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.
57. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61–65.
58. Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, et al. (2005) The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res 33: D459–465.
59. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, et al. (2006) The UCSC known genes. Bioinformatics 22: 1036–1046.
60. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14: 708–715.
61. 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.
62. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24: 1586–1591.
63. Beissbarth T, Speed TP (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20: 1464–1465.
64. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4: e1000083. doi:10.1371/journal.pgen.1000083
65. Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47: 713–719.