A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees

Rasmus Nielsen[1,2*], Carlos Bustamante[1], Andrew G. Clark[3], Stephen Glanowski[4], Timothy B. Sackton[3], Melissa J. Hubisz[1], Adi Fledel-Alon[1], David M. Tanenbaum[5], Daniel Civello[6], Thomas J. White[6], John J. Sninsky[6], Mark D. Adams[5¤], Michele Cargill[6]

1 Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America, **2** Center for Bioinformatics, University of Copenhagen, Denmark, **3** Molecular Biology and Genetics, Cornell University, Ithaca, New York, United States of America, **4** Applied Biosystems, Rockville, Maryland, United States of America, **5** Celera Genomics, Rockville, Maryland, United States of America, **6** Celera Diagnostics, Alameda, California, United States of America

Since the divergence of humans and chimpanzees about 5 million years ago, these species have undergone a remarkable evolution with drastic divergence in anatomy and cognitive abilities. At the molecular level, despite the small overall magnitude of DNA sequence divergence, we might expect such evolutionary changes to leave a noticeable signature throughout the genome. We here compare 13,731 annotated genes from humans to their chimpanzee orthologs to identify genes that show evidence of positive selection. Many of the genes that present a signature of positive selection tend to be involved in sensory perception or immune defenses. However, the group of genes that show the strongest evidence for positive selection also includes a surprising number of genes involved in tumor suppression and apoptosis, and of genes involved in spermatogenesis. We hypothesize that positive selection in some of these genes may be driven by genomic conflict due to apoptosis during spermatogenesis. Genes with maximal expression in the brain show little or no evidence for positive selection, while genes with maximal expression in the testis tend to be enriched with positively selected genes. Genes on the X chromosome also tend to show an elevated tendency for positive selection. We also present polymorphism data from 20 Caucasian Americans and 19 African Americans for the 50 annotated genes showing the strongest evidence for positive selection. The polymorphism analysis further supports the presence of positive selection in these genes by showing an excess of high-frequency derived nonsynonymous mutations.

Citation: Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3(6): e170.

Introduction

Genes, or regions of the genome, that have been affected by natural selection may show an excess of functionally important molecular changes, beyond what would be expected in the absence of selection. Genomic regions with such an excess of changes are said to have experienced positive selection, i.e., selection in favor of new genetic variants. The most common statistical technique for detecting positive selection takes advantage of the fact that mutations in coding regions of genes come in two classes: nonsynonymous mutations that change the resulting amino acid sequence of the protein and synonymous mutations, which do not change the encoded protein. An excess of nonsynonymous mutations over synonymous mutations, beyond what would be expected if the two types of mutations occur at the same rate, provides strong evidence for the past action of positive selection at the protein level. Using this logic, there have recently been numerous studies documenting positive selection in a variety of genes and organisms, including immune-response-related genes [1–3], viral genes [4–6], fertilization genes [7,8], and genes involved in sensory perception and olfaction in humans [9].

Clark et al. [10] compared 7,645 genes from humans to their orthologs from the chimpanzee and the mouse. For each gene, they tested if there was an excess of nonsynonymous substitutions on the evolutionary lineage leading to humans. They showed that there was an excess of putatively positively selected genes in several functional classes, including genes involved in sensory perception, olfaction, and amino acid catabolism. They also showed that human genes that have been targeted by positive selection are significantly more likely to harbor variation associated with known genetic diseases. We here report the results of an analysis of 20,361 human and chimpanzee genes (of which 6,630 later were eliminated in a very conservative quality control), which includes the 7,645 genes analyzed by Clark et al. [10]. While the objective of the study by Clark et al. [10] was to find genes that have experienced accelerated evolution on the human lineage, using the mouse as an outgroup, the aim of the current study is to find genes that have been targeted by positive selection at any point in time during the evolution of humans and chimpanzees, based on a larger set of genes. We use a likelihood ratio test to identify positive selection and do

Received September 30, 2004; Accepted March 14, 2005; Published May 3, 2005
DOI: 10.1371/journal.pbio.0030170

Abbreviations: (d_N/d_S) ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site; MWU, Mann-Whitney *U* test; PRF, Poisson random field; SNP, single nucleotide polymorphism

Academic Editor: Chris Tyler-Smith, Sanger Institute, United Kingdom

*To whom correspondence should be addressed. E-mail: rasmus@binf.ku.dk

¤ Current address: Department of Genetics, Case Western Reserve University, Cleveland, Ohio, United States of America

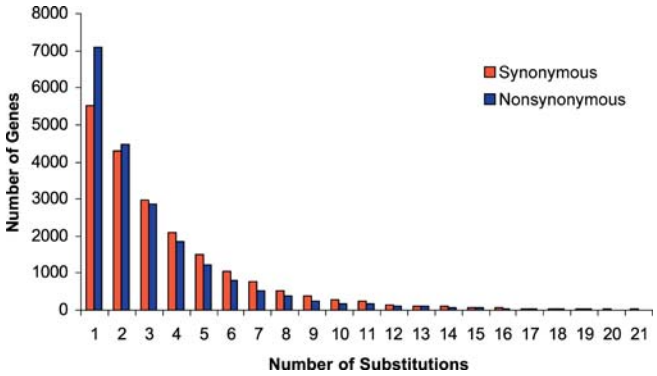

Figure 1. Distribution of Mutations

The figure shows the number of synonymous and nonsynonymous nucleotide differences in 13,731 human–chimpanzee orthologous gene pairs.

DOI: 10.1371/journal.pbio.0030170.g001

extensive simulations to find the appropriate critical values of the test. Positive selection is inferred if the ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site (d_N/d_S) is statistically significantly greater than one in a test of the neutral null hypothesis $d_N/d_S = 1$ [11,12]. The method used for detecting positive selection takes transition/transversion rate biases and unequal codon and amino acid frequencies into account. The test for positive selection applied in this study is a traditional test of d_N/d_S greater than one. It has more power than the test used in the Clark et al. study [10] if selection affects both the human and the chimpanzee lineages because it uses information from both lineages.

Results

Chimpanzee sequence was obtained by PCR using primers designed to flank exon sequence annotated in the human genome [10]. Our analysis begins with data from 20,361 coding regions, including 103,606 nucleotide differences and 403 indels among 17,687,331 aligned nucleotides. These numbers are significantly lower than the genome-wide averages [13,14], presumably due to selective constraints in the coding regions. The distributions of nonsynonymous and synonymous nucleotide differences among genes are shown in Figure 1. The average numbers of nonsynonymous and synonymous mutations per nucleotide site are 0.002578 and 0.003281, respectively. Eliminating reads without a hit to known genes in public databases (see Materials and Methods), there are 71,896 nucleotide differences in 13,731 genes. The remaining analysis is restricted to this set of genes. Among them, 5,574 were eliminated from the positive selection analysis because they had fewer than three mutations, and 797 were eliminated because the sequence was less than 50 bp long. Additionally, 45 genes were eliminated because they contained internal stop codons, presumably due to erroneous annotations or sequencing errors. Among the remaining 8,079 genes, 3,913 were also analyzed by Clark et al. [10].

The average level of sequence divergence was 0.60%, corresponding to a divergence level of 1.57% in silent sites. This figure matches well the level of divergence observed by Ebersberger et al. [14] for Chromosome 22 of 1.44% overall and 2.26% in CpG islands.

Seven hundred thirty-three of the 8,079 genes evolved with d_N/d_S greater than one, but only 35 had p-values less than 0.05, as determined by a likelihood ratio test of the null hypothesis of $d_N/d_S = 1$ against the alternative hypothesis of d_N/d_S greater than one. The number of significant genes at the 5% level, in this one-sided test, is lower than the nominal level because the vast majority of genes are conserved and evolve with d_N/d_S less than one. Nonetheless, after using Simes's improved Bonferroni procedure [15] we can, at the 5% significance level, reject the hypothesis that none of the genes are evolving with d_N/d_S greater than one. This also implies that a 5% false discovery rate set is nonempty. Even though the level of divergence between humans and chimpanzees is very low, there is statistically significant evidence for positive selection in the DNA sequences of these two species. Results for all genes are available in Dataset S1.

Biological Processes Affected by Positive Selection

To identify functional groups of genes with an over-representation of putatively positively selected genes, we used the PANTHER [16,17] classification of biological processes and a Mann-Whitney U test (MWU) based on the p-values from the likelihood ratio test (Table 1). The classification based on the MWU identifies categories of genes with small p-values from the likelihood ratio test. It is important to notice that genes that evolve approximately neutrally will tend to have smaller p-values than genes evolving under strong functional constraints. The classification based on the MWUs, therefore, does not provide unambiguous evidence for positive selection, but it provides a key to which groups harbors the most candidates for positive selection.

Immune-defense-related genes appear at the top of the list. It is not surprising that several of the genes experiencing most positive selection are involved in immune responses to viruses. Considering the speed at which many pathogens, such as viruses, evolve (e.g., [5]), a coevolutionary molecular arms race between pathogens and host cells might explain the presence of strong selection favoring new mutations in these genes. Other forces, including overdominant selection to diversify the spectrum of immune responses, may also cause positive selection in immune- and defense-related genes. Such explanations have previously been used to explain the presence of positive selection in the human major histocompatibility complex [18].

As in [10] we also identify genes involved in various forms of sensory perception, including olfaction and genes classified as "unknown biological function." Many of the genes with unknown biological function show sequence similarity with known transcription factors (data not shown). Much of the selection on sensory genes is driven by the selection on olfactory receptors previously found by Gilad et al. [9].

In contrast to Clark et al. [10], we also find that genes involved in spermatogenesis appear to have an excess of positively selected genes. The genes involved in spermatogenesis showing the strongest evidence for positive selection include several KRAB-containing zinc finger proteins that serve as repressors of transcription and are believed to be involved in determining the differentiation of pluripotent stem cells [19].

Expression Patterns and Positive Selection

We also categorized 3,464 of the 8,079 genes according to the tissue of expression in the Novartis Gene Expression Atlas

Table 1. Biological Process Categories with an Excess of Putatively Positively Selected Genes (Nominal p less than 0.05; MWU) among a Total of 133 Biological Process Categories

Biological Process	Number of Genes	*p*-Value
Immunity and defense	417	0.0000
T-cell-mediated immunity	82	0.0000
Chemosensory perception	45	0.0000
Biological process unclassified	3,069	0.0000
Olfaction	28	0.0004
Gametogenesis	51	0.0005
Natural killer-cell-mediated immunity	30	0.0018
Spermatogenesis and motility	20	0.0037
Inhibition of apoptosis	40	0.0047
Interferon-mediated immunity	23	0.0080
Sensory perception	133	0.0160
B-cell- and antibody-mediated immunity	57	0.0298

Note that the categories overlap; e.g., "T-cell-mediated immunity" is entirely nested within "Immunity and defense."

DOI: 10.1371/journal.pbio.0030170.t001

[20]. Because of the relatively small number of tissue-selective genes in our dataset (204) and the large number of tissues analyzed (28), many tissues had fewer than 20 tissue-selective genes, providing little statistical power for further subdivision. Therefore, we examined instead whether the tissue of maximal expression for a gene was correlated with positive selection, since high expression levels and importance in tissue function are often, but not always, correlated. The set of genes that have their maximal expression in the testes is the only one showing an excess of positive selection, after a Bonferroni correction for multiple tests (Table 2).

Genes with their maximal expression in the brain do not have an excess tendency toward positive selection. In fact, genes expressed in the brain seem to be among the most conserved genes with the least evidence for positive selection. MWUs, comparing genes with their maximal expression in the brain (83 genes) to all other genes, show that these genes tend to have significantly higher p-values of the likelihood ratio test for positive selection ($p = 0.035$), indicating high levels of selective constraint. Genes that are expressed in the brain at a level of twice the expression level found in blood show an even stronger tendency toward avoidance of positive selection ($p = 0.0002$). Although studies of gene expression in the brain tissue are complicated by low-abundance transcripts and heterogeneous specialized brain regions [21], the overall evidence points toward a deficiency of positively, or fast evolving, genes among those expressed in the brain. The causes for the cognitive differences may instead be sought in adaptive changes in just a few genes, in changes in gene expression [22], or in changes in copy number and/or organization of genes relating to cognitive function [23].

Dorus et al. [24] found that genes expressed in the nervous system showed a relative increase in the rate in primates relative to rodents when compared to housekeeping genes, but provided no direct evidence for positive selection on these genes. Nervous-system-specific genes appear to be so conserved that it is unlikely that direct evidence for positive selection will be discovered in this group of genes.

Table 2. Test for an Excess of Putatively Positively Selected Genes by Tissue Type

Tissue of Maximal Expression	Number of Genes	*p*-Value
Testis	247	0.0002
Thyroid	66	0.0287
Thymus	82	0.0599
Prostate	76	0.0902
Fetal_liver	114	0.1668
Salivary_gland	195	0.1696
Whole_blood	405	0.239
Heart	120	0.2906
Lung	64	0.3381
Trachea	47	0.3976
Liver	244	0.4468
Uterus	51	0.493
Adrenal_gland	70	0.5434
Spleen	134	0.5582
Pancreas	358	0.6063
Pituitary_gland	60	0.6493
Placenta	179	0.7566
Cortex	36	0.7696
Kidney	179	0.801
Amygdala	43	0.8398
Corpus_callosum	101	0.8909
Caudate_nucleus	36	0.8945
Thalamus	33	0.9018
Fetal_brain	201	0.912
Ovary	133	0.9295
Whole_brain	83	0.965
Cerebellum	93	0.9903
Spinal_cord	14	1

Small p-values (MWU; nominal p-values not corrected for multiple testing) indicate an excess of putatively positively selected genes in the tissue type.

DOI: 10.1371/journal.pbio.0030170.t002

Positive Selection in the X Chromosome

We also tested if any chromosomes show an excess of genes with evidence for positive selection. The only chromosome enriched in genes with small p-values from the likelihood ratio test for positive selection is the X chromosome ($p = 0.0049$; MWU). Several factors influence the contrast between the X and autosomes in tests of selection, including hemizygosity of the X in males, resulting in more effective selection against deleterious recessive and in favor of positive recessive mutations [25]. Male hemizygosity also results in mutations, with male-specific effects being more readily fixed by selection on the X [26]. This increased efficiency of selection for male-specific genes on the X may explain the excess of X-linked genes expressed in spermatogonia [27]. The observation that reproductive proteins generally evolve at a greater rate, coupled with the overrepresentation of male-specific genes on the X, could produce the excess positive selection seen on the X. However, after eliminating all genes with highest expression levels in the testis, or annotated as functioning in spermatogenesis, there is still an excess of putatively positively selected genes on the X chromosome ($p = 0.0131$; MWU). Thus, it appears that the elevated positive selection on the X is likely due to the general tendency of mutations to be recessive, regardless of their tendency to be male-limited in expression. Although other factors, such as an elevated male mutation rate [28], differences in the efficacy of genetic hitchhiking between autosomes and the X chromosome [29], and correlations

between recombination rate and divergence [30], may cause differences in variability and substitution rate between autosomes and the X chromosome, none of these factors alone can explain the excess of positively selected genes on the X chromosome.

Analysis of the 50 Genes Showing Strongest Evidence for Selection

We studied the 50 genes with the highest likelihood ratios in greater detail to further characterize the causes of positive selection and examine error rates (Table 3). To investigate the degree to which our results might be influenced by sequencing errors, we compared the data for these genes with the public data available for the same genes. In the regions with overlap between the public data and our data there were a total of 327 mutations in the public data and 306 mutations in our data. This demonstrates that there is not an excess of (potentially artifactual) mutations in our data in the genes that show evidence for positive selection. While most of the 50 genes also show strong evidence for positive selection in the public data, six of the genes do not. HC19953, HC2758, HC6579, HC7761, HC8067, and HC9844 do not have d_N/d_S ratios larger than one in the public data. In most cases, the difference is caused by the fact that our database and the public database contain different regions of the genes. Not all regions of a gene are expected to be targeted by positive selection, but this does not challenge the evidence for positive selection in the regions of the genes included in this analysis. In any case, using the public data would not change the qualitative conclusions of the analysis of the genes presented here.

Immunity and Defense Genes Targeted by Positive Selection

The top 50 genes include many genes that we might a priori expect to be targets of positive selection, including four genes involved in olfaction *(OR2W1, OR5I1, OR2B2,* and *C20orf185)* and several genes involved in host–pathogen interactions, such as *CMRF35H, CD72* antigen, pre-T-cell antigen receptor α *(PTCRA), APOBEC3F,* and granzyme H *(GZMH).* Only one of these genes was among the 50 most significant entries in the Clark et al. [10] model 2 analysis. *APOBEC3F* encodes an antiviral factor that has previously been demonstrated to be under positive selection by Sawyer et al. [3] who note that this gene has been associated with anti-HIV activity.

Presumably, most of these genes have been targeted by positive selection throughout the primate and mammalian phylogeny. The widespread evidence for positive selection in immune-related genes confirms the hypothesis that much positive selection in the human and mammalian genomes may be driven by a coevolutionary arms race between host immune system and pathogens.

Spermatogenesis- and Apoptosis-Related Genes

The list also contains many testis- or sperm-specific genes including Protamine-1 *(PRM1),* which previously has been shown to be under positive selection [31], possibly due to sperm competition (but see [32] for an alternative explanation). Other sperm-specific genes on the list include *USP26, C15orf2, PEPP-2*, *TCP11, HYAL3,* and *TSARG1.* The inclusion of these genes in the list of the genes showing the strongest evidence for positive selection is consistent with the results, based on the PANTHER annotation and the Novartis expression data, of excess positive selection in sperm/testis-specific genes. The possible causes include sperm competition (e.g., [31]), sexual conflict (e.g., [7,8]), selection for reproductive isolation, pathogen-driven selection in the reproductive organs, and selection related to the occurrence of mutations causing segregation distortion.

We notice that at least one of these genes *(TSARG1)* is involved in apoptosis during spermatogenesis. Apoptosis of germ cells is conspicuous during normal spermatogenesis, eliminating up to 75% of the potential spermazoa [33–35], affecting cells both before and after the meiotic division [36]. It has been hypothesized that the main cause for the high rate of apoptosis during spermatogenesis is to maintain a proper cell-number ratio between maturing germ cells and Sertoli cells [35]. The natural process of elimination of germ cells by apoptosis creates a genomic conflict in which each individual germ cell will benefit from avoiding apoptosis, but apoptosis of a certain fraction of germ cells may be beneficial to the mature organism. New mutations occurring in cells during spermatogenesis, which reduces the probability of apoptosis, will be positively selected. This effect will be particularly strong for mutations in genes expressed after the meiotic division, potentially resulting in segregation distortion. A mutant with an even very small increase in the probability of escaping postmeiotic apoptosis will have a strong selective advantage. Compensatory mutations, reducing or eliminating the effect of the apoptosis avoidance mutation, may then later occur. These dynamics may lead to recurrent events of positive selection in genes affecting spermatogenesis apoptosis. The 40 genes in this study involved in inhibition of apoptosis show an excess of evidence for positive selection compared to other categories ($p = 0.0047$; see Table 2). Many of the genes showing most evidence for positive selection are known to be involved in either spermatogenesis, apoptosis, or both. For example, the apoptosis-related gene showing the strongest evidence for positive selection *(DFFA)* is an inhibitor of Fas-mediated apoptosis, which has been shown to be involved in apoptosis during spermatogenesis [36]. This may suggest that genomic conflict due to spermatogenesis apoptosis may be driving positive selection in many of the included genes.

Cancer-Related Genes

While we expected to find genes involved in olfaction, spermatogenesis, and immune defense among the 50 annotated genes showing the strongest evidence for positive selection, we were surprised to find a very large proportion of cancer-related genes, especially genes involved in tumor suppression, apoptosis, and cell cycle control. These genes include four putative tumor suppressors: *HYAL3, DFFA, PEPP-2* (note that both *HYAL3* and *PEPP-2* also appear to be involved in spermatogenesis), and *C16orf3,* another gene associated with tumor progression *(MMP26),* and a gene with unknown function but high similarity to melanoma-associated antigens *(FLJ32965).* In addition, there are several genes involved in apoptosis *(PPP1R15A, HSJ001348, TSARG1,* and *GZMH).* Given that many of the genes have very little functional information, it is surprising to find such a large proportion of genes that may be related to tumor development and control. The factors causing positive selection on these genes are unknown, but genes important in tumor

Table 3. The Top 50 Genes Showing Evidence for Positive Selection

HC Name[a]	Gene Name	Function	NDHC[b]	SD HC[c]	NP H[d]	SP H[e]	HC LR[f]
HC208	*PRM1*	Substitutes for histones in sperm	9	0	0	2	10.12208
HC15768	*CMRF35H*	Leukocyte membrane antigen	13	0	0	0	9.262642
HC12140	*DGAT2L1*	Fatty acid synthesis (presumed)	10	1	2	0	6.625498
HC1860	*FLJ46156*	Unknown	10	1	4	3	6.401844
HC2436	*USP26*	Testis-specific expression	11	0	1	0	6.217652
HC3085	*C15orf2*	Testis-specific expression	18	2	12	4	6.093642
HC13803	*ABHD1*	Unknown	6	0	4	1	5.778402
HC11239	*SCML1*	Transcriptional repressor, embryonic development (presumed)	15	1	0	0	5.748762
HC3472	*OR2W1*	Olfactory receptor	8	0	2	1	5.702798
HC10799	*LOC389458*	Unknown	8	0	1	0	5.493604
HC7761	*APOBEC3F*	Antiretroviral factor	11	0	2	1	5.476024
HC19072	*MS4A12*	Unknown	8	0	1	1	5.36116
HC4477	*HYAL3*	Testis-specific expression, putative tumor suppressor	5	0	2	2	5.266036
HC7681	*FLJ32965*	Similar to melanoma-associated antigens (function unknown)	7	0	2	0	5.24997
HC8130	*LOC151534*	Function unknown	7	0	0	1	5.13903
HC3434	*MMP26*	Zinc-binding endopeptidase, tumor progression (presumed)	7	0	2	1	4.869976
HC7508	*KIAA0495*	Component of the cell membrane (by similarity)	6	0	1	1	4.67452
HC4613	*CD72*	Signaling in the immune system	5	0	2	0	4.516886
HC14419	*DFFA*	Inhibition of apoptosis, putative tumor suppressor	6	0	1	0	4.448548
HC11263	*KRN1*	Hair keratin	2	0	2	0	4.4187
HC8067	*TNKS1BP1*	Tankyrase-binding, multifunctional (presumed)	6	0	2	3	4.323594
HC19953	*RNPC4*	RNA-binding, pre-mRNA-splicing process (presumed)	6	0	3	3	4.283004
HC1586	*KRTAP19–1*	Keratin-associated protein 19–1	4	0	1	0	3.923226
HC18280	*HSJ001348*	Apoptosis, p53-induced	10	0	0	0	3.904888
HC3104	*HSA404617*	Unknown	5	0	0	1	3.748416
HC15059	*FLJ20489*	Unknown	5	0	0	2	3.653678
HC13738	*RPP38*	Component of RNase P	4	0	5	2	3.638472
HC2758	*FLJ35725*	Unknown	4	0	3	0	3.637784
HC4426	*PEPP-2*	Testis-homeobox gene, putative tumor suppressor	11	0	3	1	3.602598
HC18485	*PGR*	Progesterone receptor	11	0	2	3	3.51589
HC738	*MGC57858*	Unknown	3	0	2	0	3.208712
HC973	*GZMH*	Cell lysis	12	1	1	0	3.131548
HC4889	*TCP11*	Germ-cell development (presumed)	3	0	2	1	3.122404
HC17263	*C20orf185*	Possible carrier molecule for odorants	4	1	2	3	3.047252
HC18160	*PPP1R15A*	Growth arrest, DNA-damage inducible, apoptosis	9	2	5	5	3.007004
HC14000	*SLC22A4*	Cation transporter, susceptibility to rheumatoid arthritis	4	0	2	3	2.979532
HC16621	*GREAT (LGR8)*	Receptor for relaxin. Mutations may cause cryptorchidism	4	0	1	0	2.951438
HC11003	*LR8*	Unknown (expressed by a lung fibroblast subpopulation)	3	0	7	0	2.948436
HC16489	*HHLA1*	Unknown	3	0	0	0	2.569134
HC3738	*C16orf3*	Putative tumor suppressor	2	0	1	1	2.56533
HC2610	*CYSLTR2*	Anaphylactic reactions	9	1	5	0	2.47886
HC9844	*ASB11*	Cytokine signaling	3	0	2	0	2.448126
HC8169	*FLJ32743*	Unknown	3	0	2	2	2.39831
HC12857	*GDF3*	Putative regulator of cell growth and differentiation	2	0	1	1	2.261462
HC6579	*FLJ32844*	Unknown	2	0	3	0	2.2473
HC2300	*OR2B2*	Olfactory receptor	3	0	3	0	2.246402
HC1723	*MGC41945*	Unknown	12	2	5	0	2.184832
HC3892	*OR5I1*	Olfactory receptor	5	1	7	4	2.152316
HC4879	*PTCRA*	Pre-T-cell antigen receptor α	8	1	2	0	2.075898
HC16795	*TSARG1*	Spermatogenesis cell apoptosis	6	1	3	2	1.671884

[a]Reference number used in Dataset S1.
[b]Number of nonsynonymous differences between humans and chimps.
[c]Number of synonymous differences between humans and chimps.
[d]Number of nonsynonymous polymorphism in humans.
[e]Number of synonymous polymorphism in humans.
[f]Likelihood ratio from the likelihood ratio test of d_N/d_S equals one versus d_N/d_S is greater than one in the human–chimp alignment.
DOI: 10.1371/journal.pbio.0030170.t003

development and suppression may be positively selected due to other functional effects of the genes, particularly in immunity and defense or in spermatogenesis. Several of the genes involved in tumor suppression or progression show testis-specific expression, and models of genomic conflict may explain the presence of positive selection in these genes. It should be noted that there is no pattern of human-specific selection in these genes. The high number of nonsynonymous mutations in these genes is approximately evenly distributed between the human and the chimpanzee lineage (results not shown).

PAML Analysis

For each of the 50 genes, we searched public databases to find orthologous genes in other mammals. For 25 of the genes we were able to identify orthologs from mouse and rat, and for these 25 genes we estimated the d_N/d_S ratio of each lineage

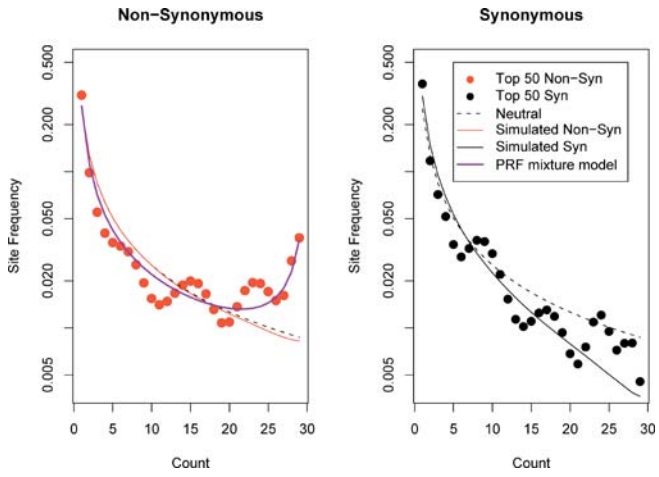

Figure 2. Frequency Spectra

The figure shows the frequency spectra of nonsynonymous (red) and synonymous (black) mutations among the 50 genes showing the strongest evidence for positive selection in the interspecific comparison. Also shown is the expectation from the standard neutral model, expectations from the neutral model taking the protocol used to select the 50 genes into account (see text), and from the prediction of the selection model. On the *x*-axis is the number of derived allele in a sample of size 30 chromosomes (Count), and on the *y*-axis is the proportion of sites expected in the sample with a particular frequency.

DOI: 10.1371/journal.pbio.0030170.g002

of the underlying phylogeny using PAML [37]. The d_N/d_S ratio was elevated ($p < 0.05$) in 5/25 cases in just the human lineage, in 5/25 cases in just the chimp lineage, in 8/25 cases in both lineages, and in 7/25 cases significant in neither lineage. These results show that the elevated d_N/d_S ratios are a consequence of positive selection in both the human and the chimpanzee lineage.

Population Genetic Analysis

To further investigate the effect of selection on the 50 genes showing the strongest evidence for positive selection, 20 European-American and 19 African-American individuals were sequenced for these genes. Forty-six of the genes contained intraspecific polymorphism, and there were a total of 55 synonymous polymorphisms and 116 nonsynonymous polymorphisms, showing that the d_N/d_S ratio is also relatively high in the polymorphism data.

The distribution of allele frequencies within these genes, as summarized by the allele frequency spectrum, provides additional support for positive selection. The frequency spectrum (Figure 2) of synonymous polymorphisms does not deviate from the pattern expected under a standard neutral model [38]. However, this does not necessarily provide evidence for the adequacy of the standard neutral model, but may rather be caused by a cancellation of effects due to population growth, population subdivision, and linkage to selected mutations, low power due to the small sample size, or by an ascertainment bias described below. Other data in humans have shown an excess of rare derived alleles in synonymous sites, presumably caused by population growth [39,40]. In contrast, we find that nonsynonymous single nucleotide polymorphisms (SNPs) show evidence for an excess of high-frequency-derived alleles in these genes (Figure 2). The excess of high-frequency-derived nonsynonymous mutation supports the notion that these genes have been targeted by positive selection. An important caveat is that an ascertainment bias has been introduced because interspecific and intraspecific variability has been confounded when selecting genes with high d_N/d_S ratios. To assess the impact of this ascertainment bias, we simulated 1,000 new neutral datasets, each dataset consisting of 13,731 genes with a similar distribution of d_N/d_S ratios, mutation rates, and sequence lengths, as observed in the real data, and with both interspecific and intraspecific variation. From these datasets we selected the 50 genes with the largest d_N/d_S ratios, as in the selection procedure applied to the real data. There is a clear effect of the ascertainment bias on synonymous sites, but there is essentially no effect on nonsynonymous sites (Figure 2). The main effect of the ascertainment bias is to eliminate genes with many high-frequency-derived synonymous mutations. This shows that the excess of high-frequency-derived nonsynonymous mutations is not a result of the ascertainment bias.

In addition to selection, certain demographic factors, such as population bottlenecks and population subdivision [41,42], and/or incorrectly inferred ancestral states may also enrich the sample with apparent high-frequency-derived mutations. Przeworski [42] has previously reported an excess of high-frequency-derived mutations in human data. To investigate this possibility we compared the frequency spectrum in our data to the frequency spectrum of the genes in the Seattle SNP database (SeattleSNPs; http://pga.gs.washington.edu [01/10/03]). These data also consist of a mixture of declared African Americans and European Americans and should, therefore, comprise a suitable sample for comparison. With 24 out of 116 and 37 out of 360 nonsynonymous mutations of frequency greater than 50% in our data and the Seattle data, respectively, there is a significant excess of high-frequency-derived mutations in our data compared to the Seattle data ($p < 0.01$, chi-square test). The Seattle data shows a slight deficiency of nonsynonymous-derived mutations with frequency greater than 50%, primarily due to an excess of very low-frequency-derived mutations. These results strongly suggest that the pattern we observe is caused by ongoing positive selection and not by demographic effects. There are a total of 25/78 and 22/92 polymorphisms of frequency greater than 50% within the Caucasian and African-American groups, respectively. Analyzing each population separately gives an even more extreme excess of high-frequency-derived polymorphism, especially in the Caucasian population.

There is a very high variance in the ratio of divergence to polymorphism in these genes (Hudson-Kreitman-Aguadé test; p less than 0.05). While the overall ratio of divergence to polymorphism is around two (2.06), a few genes stand out as having particularly high levels of polymorphism. For example, one of the olfactory receptors, *OR5I1,* has six substitutions and 11 polymorphisms. This raises the possibility that positive selection in the olfactory receptors may be a type of balancing selection. One possibility is heterozygote advantage driven by selection to increase the repertoire of olfactory receptors.

Another gene with a low divergence to polymorphism ratio is *RPP38* (four substitutions and seven polymorphisms), which is a subunit of RNase P. *RPP38* is necessary for normal

processing of stable RNA in human cells, but it is also a target for antisera from systemic sclerosis patients. It is likely that the positive selection in this gene is caused by selection to avoid an autoimmune response. Such a hypothesis is plausible if the sequence pattern of RPP38 influences the likelihood of developing systemic sclerosis. This hypothesis can be tested using linkage or linkage disequilibrium studies.

Other genes show an apparent deficiency of polymorphisms. *SCML1* has 16 substitutions (of which 15 are nonsynonymous) and zero polymorphisms. Such a pattern is consistent with repeated selective sweeps driving divergence between species, while eliminating variation within species. *SCML1* is a repressor of expression of *Hox* genes and may play an important role in the control of embryonal development [43]. This gene may be a prime candidate for explaining developmental differences between humans and chimpanzees.

Poisson Random Field (PRF) Analysis

To further investigate the distribution of selection coefficients among mutations in these genes, we applied a PRF model [44]. In PRF models, the distribution of sample allele frequencies can be expressed as a function of the scaled selection coefficient, S, ($S = 2Ns$; N = population size, s = selection coefficient) acting on a mutation. We assumed that there were three types of mutations: negatively selected mutations (of frequency p_-), neutral mutations (of frequency p_0), and positively selected mutations (of frequency $p_+ = 1 - p_- - p_0$). We then estimated p_-, p_0, p_+, and the scaled selection coefficients of the mutations in the two selected categories (S_- and S_+) using maximum likelihood.

The maximum likelihood estimates of the parameters of the PRF model are $p_- = 0.748$, $p_0 = 0.172$, $p_+ = 0.080$, $S_- = -34.96$, and $S_+ = 267.11$; i.e., the estimated proportion of negatively selected mutations is approximately 75%, and the proportion of positively selected mutations is approximately 8%. The proportion of positively selected mutations is so high because we have analyzed the 50 genes showing the strongest evidence for positive selection among a very large pool of candidate genes. Likelihood ratio tests show that a model with three selected classes fits the nonsynonymous data significantly better than a model with fewer selective classes (see Materials and Methods). We conclude that the allelic distribution in nonsynonymous sites is best described by a mixture of neutral, positively selected and negatively selected mutations. In this case, our best estimate of the proportion of mutations in these genes that are neutral is less than 18%. The predicted frequency spectrum under the estimated selection model is shown in Figure 2.

The results of this analysis should be interpreted with some caution because the effects of linkage have been ignored. The effects of linkage would be to underestimate the selection coefficient and, possibly, to overestimate the number of mutations that have been targeted by selection [45]. As previously discussed, these types of inferences are also sensitive to the demographic assumptions of a panmictic population of constant size [41,42] and to the assumptions regarding unambiguous inference of the ancestral state from the chimpanzee. For these reasons, the exact values of the parameter estimates should not be overinterpreted, but may help suggest the magnitude of the selective forces necessary to explain the data in isolation.

Discussion

The statistical methods used for detecting selection have been the subject of debate over the past few years [46,47]. This debate has mainly focused on the validity of methods that model variation in the d_N/d_S ratio among sites. The current test does not model rate variation among sites and should, therefore, be uncontroversial. Unfortunately, this test may also have very low power.

To determine the power of the test, we conducted power simulations under parameter values estimated from the data (Figure 3). Notice first that the test does not result in excess significant results when $d_N/d_S = 1$, and results in very few falsely significant results when d_N/d_S is less than one. In fact, when $d_N/d_S = 1$ the power is lower than the nominal significance level because of the possibility of ties. However, the power increases steadily when d_N/d_S increases above one, and for a gene of length 500 codons, the test has more than 80% power when $d_N/d_S = 5$. For a functional gene, in which most sites are expected to be under functional constraints and evolve with d_N/d_S less than one, a significant value of the test is almost surely caused by positive selection. The fact that our data shows significant evidence for positive selection, when using this test with a correction for multiple tests, illustrates that positive selection can be detected from human–chimpanzee comparative data despite the very low levels of divergence.

In the previous study by Clark et al. [10], an outgroup (mouse) was used to make inferences regarding human-specific processes. We have here analyzed a larger dataset but cannot, in general, distinguish between selection that is particular to the human evolutionary lineage and positive selection that tends to occur in both chimps and humans. While Clark et al. [10] found strongest evidence for positive selection in genes related to olfaction and sensory perception, we find the strongest evidence for positive selection in genes related to immunity and defense. The reason is probably that genes related to immunity and defense are targets for positive selection throughout the mammalian phylogeny, which the test used by Clark et al. [10] would not detect, whereas much of the selection on sensory perception and olfaction is likely to be specific to the distinct niches of humans and chimpanzees. Similar arguments may also explain why we obtain strong evidence for positive selection on genes related to spermatogenesis and inhibition of apoptosis, while Clark et al. [10] did not find any evidence for human-specific selection on genes related to spermatogenesis and apoptosis.

In this paper we analyzed population genetic data from the 50 genes showing most evidence for positive selection. An excess of high-frequency-derived nonsynonymous mutations in these data supports the conclusions that these genes are targeted by positive selection. Although some demographic models also may cause an excess of high-frequency-derived mutations [41,42], the excess observed in our data is beyond the level observed in other comparable human data.

The use of the population genetic data may also help suggest the mode of positive selection acting on the gene. For example, a developmental gene *(SCML1)* had 16 fixed substitutions and zero polymorphisms, suggesting repeated selective fixations, whereas and olfactory receptor had six substitutions and 11 polymorphisms consistent with the

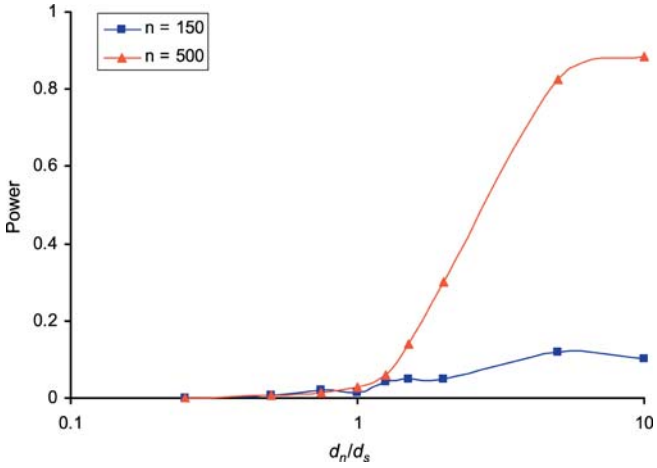

Figure 3. Power of the Likelihood Ratio Test for Positive Selection

The power is shown as a function of the proportion of the d_N/d_S ratio, and for sequence lengths *(n)* of 150 and 500 codons. Power is defined as the proportion of tests that are significant at the 5% level. Simulation parameters, including codon frequencies, transition/transversion bias, and divergence times, are equal to the values estimated from the data. Notice the logarithmic *x*-axis.

DOI: 10.1371/journal.pbio.0030170.g003

action of balancing selection. The combined use of comparative and population genetic data may help, not only to identify positive selection, but also to help narrow down possible models of positive selection. With the increased availability of both comparative genomic data and SNP data, we expect to see many future studies that take advantage of the availability of both types of data.

The discovery that many genes involved in spermatogenesis, apoptosis, and tumor suppression are positively selected may prompt further investigations into models of genomic conflict and other models predicting positive selection in these genes. In general, mutations that increase the expected number of functional sperm cells produced by a specific germ-line cell, such as mutations increasing the rate of cell division or decreasing the probability of apoptosis, will be favored. Such mutations will not necessarily increase the fitness of the mature organism, leading to a genomic conflict, in which selfish mutations causing avoidance of apoptosis are being counteracted by compensatory mutations in other loci. Many of the genes with evidence for positive selection encountered in this study play functional roles in cell cycle regulation, tumor suppression, apoptosis, or spermatogenesis. We suggest that a genomic conflict relating to the process of spermatogenesis may be responsible for much of the positive selection observed in this study. Because many of these genes are involved in inhibition of apoptosis, this may also explain the apparent excess of cancer-related genes targeted by positive selection. This raises the interesting prospect that the high prevalence of cancer in humans and other organisms may be related to selection for apoptosis avoidance in the germ line. Mutations that in general increase apoptosis avoidance will be selected in the germ line, but such mutations may at the same time increase the probability of cancer in somatic tissue. The relative high prevalence of cancer will, according to this hypothesis, be related to an evolutionary conflict between the selfish interests of a germ cell and selection at the level of mature organisms to decrease the cancer rate. We note that the fact that the same pathways (e.g., Fas-mediated apoptosis) are involved in the control of cancer and in apoptosis during spermatogenesis supports this hypothesis.

At present we cannot exclude an alternative hypothesis, such as pathogen-driven positive selection or sperm competition. Future functional and evolutionary studies of the genes suggested to be under positive selection by this study may help determine which of these alternative evolutionary models are most plausible.

Materials and Methods

DNA sequencing and alignment. Sequences of chimpanzee genes were obtained by PCR amplification of individual exons from a single western chimpanzee male. PCR products were directly sequenced on automated sequencers at Celera Genomics. Details of primer construction, DNA sequencing, and alignment were described in Clark et al. [10] and references therein. Chimpanzee sequences were obtained for both strands from PCR products, filtered to remove base calls with Phred scores less than 30. Genes that did not have a hit in the curated accessions (NM__ or NR__ series) in the REFSEQ 3.0 database (http://www.ncbi.nlm.nih.gov/RefSeq/) and for which the best hit did not map to the same genomic location in NCBI 34 build of the human genome, were omitted from the functional analysis, to minimize the chance of including nonfunctional genes/regions. This was done using BLAT v. 27 (http://www.genomeblat.com/genomeblat/index.asp). Regions of the alignment that were not in the REFSEQ database were eliminated from the analysis. However, we provide the full dataset (Dataset S2) for future exploration. Indels were identified after alignment as all nonterminal gaps that could not be attributed to low base-calling scores. All pairwise alignments are available in Phylip and FASTA format in Dataset S2.

Human polymorphisms were detected automatically from assembled sequencing traces using PolyPhred 4.0 [48] and RuleGen, a decision-tree-based method (S. Glanowski, unpublished data). Manual calls were employed if a potential SNP was not flagged by both programs. Validation of the automated pipeline using a set of several hundred manually called SNPs showed a sensitivity of 85% for all SNPs and up to 100% for SNPs with more than three minor alleles observed. Independent verification of several hundred SNPs using TaqMan assays indicated that validation rates of 95% for common SNPs and 90% for SNPs with only one minor allele were observed.

Likelihood ratio tests. For each human–chimpanzee orthologous gene pair, a likelihood ratio test of the hypothesis of an equal d_N/d_S ratio was performed using a codon-based likelihood model (see [12] for such tests). The test was performed as a one-sided test of the hypothesis H_0: $d_N/d_S = 1$ versus the alternative of H_A: d_N/d_S greater than one. To reduce the computational burden, the transition/transversion rate ratio was first estimated for genes in high-GC-content regions and low-GC-content regions separately. This parameter was then considered fixed for the remainder of the analysis. Because many of the sequence pairs showed very little divergence, the usual asymptotic assumptions of a chi-square distribution of the likelihood-ratio test statistic would not have been appropriate. Instead, simulations were performed to determine the appropriate distribution of the test statistic. The simulations were performed under the empirical distribution of the divergence time and other parameter estimates assuming $d_N/d_S = 1$. The distribution of the test statistic, conditional on the observed number of nucleotide differences between the sequences (for each gene), was then determined. One of the advantages of using the conditional distribution is that the distribution becomes more robust to violations of the assumptions regarding the nuisance parameters, particularly the divergence times, and this will allow us to exclude genes with very little variability while maintaining the right size of the test. Genes with fewer than three nucleotide differences, or with fewer than 50 aligned codons, were excluded from the analysis.

Functional analysis. The functional annotation was performed as in [10], using the PANTHER database [16,17]. Throughout, excesses of positively selected genes in a category were tested, using an MWU comparing the distribution of *p*-values obtained from the likelihood ratio tests in genes included in the category to the distribution of such values in genes not included in the category. Genes with fewer than three nucleotide differences between humans and chimpanzees were excluded from the identification of categories with an excess of putatively positively selected genes. The MWU does not in itself

demonstrate that the evolution of a particular category of genes is affected by positive selection, but it shows that the category contains more evidence for positive selection than other genes in the study. Because genes of short sequence length are less likely to show strong evidence for positive selection, but are more likely to show spurious evidence for positive selection, the MWU (or any other categorization) may be affected by different sequence lengths in different categories. The reason for using an MWU, instead of reporting overall *p*-values for a category after correction for multiple testing, is that such an approach would be strongly influenced by just one or a few genes. However, correction for multiple testing reveals significant positive selection in several categories, including the immune and defense and the spermatogenesis category.

Expression data. Expression data from normal human tissues were obtained from the Novartis Gene Expression Atlas ([20]; http://wombat.gnf.org/index.html); 6,741 gene symbols could be matched unambiguously to the human–chimp alignments. All negative expression and values less than 20 were coded as 20. Tissue selectivity was determined by averaging probe expression values across samples and replicate tissues. In total, 61 samples were collapsed into 28 tissues. A probe was classified as tissue selective if it was expressed in only one tissue at a value of 200 or higher, and all other tissues were less than 100. Probes were then collapsed into genes. A gene was classified as tissue selective if at least one of its probes showed specificity. The tissue of maximal expression was determined by identifying the probe and sample ($n = 61$) with the highest expression value that was greater than 20 (85% of the genes had values greater than 200). Probes were then collapsed into genes. Tissue expression was determined by averaging the sample replicates ($n = 28$). A gene was considered expressed in a tissue if its expression value was greater than 200.

PAML analysis. To obtain orthologous sequences for the 50 annotated genes with the highest likelihood ratios, we downloaded "Unique Best Reciprocal Hits" between human and mouse and human and rat from the Ensembl Web site (http://www.ensembl.org/). Sets of human, mouse, rat, and chimpanzee sequences were translated and aligned using ClustalW [49]. Codon alignments were generated using the ClustalW alignments as a guide, then manually checked. Partial sequences covering less than 80% of the human sequence were eliminated, and ambiguously aligned regions were masked before analysis. The underlying phylogeny was assumed to be ([chimpanzee, human], [mouse, rat]) for all genes.

The lineage-specific analysis was done in PAML [37] by allowing two values of the d_N/d_S ratio along the lineages of the phylogeny, one for the human lineage and one for all other lineages. To test if the d_N/d_S ratio was different on the human lineage, we then compared the maximum likelihood value in this model to the maximum likelihood value obtained, assuming the d_N/d_S ratio was constant among lineages. If two times the log likelihood ratio was larger than 3.84, we rejected the model of constant d_N/d_S ratio at the 5% significance level. This analysis was then repeated using the chimpanzee lineage as the focal lineage instead of the human lineage.

Calculating the frequency spectrum. Because of missing data for many polymorphisms, the frequency spectrum in a sample of size 30 is reported. The frequency spectrum was calculated in a sample of size 30 as

$$p_{i,30} = k^{-1} \sum_{j=1}^{k} \frac{\binom{f_j}{i}\binom{n_j - f_j}{30 - i}}{\binom{n_j}{30}} \quad (1)$$

where $p_{i,30}$ is the frequency of SNPs with derived alleles that exist in i copies in a sample of size 30, n_j is the chromosomal sample size of the jth SNP, f_j is the frequency of the derived allele for the jth SNP, k is the number of SNPs, and $\binom{i}{j} = 0$ if i is less than j. The polarity of the mutation was determined using the chimpanzee sequence as outgroup.

Analysis of ascertainment bias. To assess the impact of the ascertainment scheme in the tests that contrast human polymorphism data to the human–chimp divergence, new datasets were simulated, using standard neutral coalescence simulations (e.g., [38]). Each simulated dataset generated one chimp sequence and 78 human sequences for each of the 13,731 genes. For each simulated gene, one human sequence was randomly chosen and compared to the chimp sequence using a chi-square statistic for the goodness-of-fit test of $d_N/d_S = 1$. The 50 genes with largest chi-square statistic among genes with d_N/d_S greater than one were selected for population genetic analysis. This scheme was repeated 1,000 times to investigate the effect of the ascertainment protocol of the 50 genes. The parameters of the simulations were estimated from the data, using the observed distribution of sequence lengths, and synonymous-site mutation rate and humans–chimp divergence time estimated from the concatenated data. The distribution of d_N/d_S ratios among genes was estimated assuming the d_N/d_S ratios follow a γ distribution among genes, keeping the synonymous rate constant among them.

Power analysis. To analyze the power of the test for positive selection, we simulated pairs of sequences and performed likelihood ratio tests of H_0: d_N/d_S equals one versus d_N/d_S is greater than one for each sequence pair. The simulations were done using the average value of synonymous sequence divergence observed in the data, while nonsynonymous divergence was varied. For more details regarding such simulations, see, e.g. [50].

PRF analysis. Assume nonlethal mutations enter a population of constant size $2N$ according to a Poisson process and are assigned to one of three categories: neutral ($S = 0$), positively selected with selection coefficient S_+, and negatively selected with selection coefficient S_-, according to probabilities p_0, p_+, and p_- (where $p_0 + p_+ + p_- = 1$). Furthermore, assume mutations evolve independently. It follows from standard population genetic theory, the total law of probability, and the rules of conditional probability that the probability of an SNP being found at frequency i out of n chromosomes under this scheme [44] is

$$P(X = i \mid n, p_+, p_-, p_0, S_+, S_-) = \frac{p_+F(i,n,S_+) + p_-F(i,n,S_-) + p_0\frac{1}{i}}{\sum_{j=1}^{n-1} p_+F(j,n,S_+) + p_-F(j,n,S_-) + p_0\frac{1}{j}} \quad (2)$$

where $F(i,n,S)$ is given by

$$F(i,n,S) = \int_0^1 \binom{n}{i} x^i (1-x)^{n-i} \frac{1 - e^{-2S(1-x)}}{1 - e^{-2S}} \frac{1}{x(1-x)} dx \quad (3)$$

The likelihood of observing counts $x_1, x_2, \ldots, x_S$ where S is the total number of segregating sites out of $n_1, n_2, \ldots, n_S$ chromosomes is, thus,

$$L(p_+, p_-, p_0, S_+, S_- \mid x, n) = \prod_{j=1}^{S} P(x_j = i \mid n_j, p_+, p_-, p_0, S_+, S_-) \quad (4)$$

The maximum likelihood value and the maximum likelihood parameter estimates can then be obtained by numerically maximizing this function with respect to the parameters. Likelihood ratio tests can be constructed by constraining certain of the parameters to take on particular values. For example, setting $p_0 = 1$ defines a model with no selected mutations. Likewise, setting $p_0 + p_- = 1$ defines a model that allows negative selection, but no positive selection.

This analysis assumes that mutations are independent. Because of linkage and the possibility of epistasis, the independence assumption may not be met by the data. However, a full analysis taking the correlation among SNPs into account is not computationally feasible. Fortunately, the average correlation is low between SNPs because they have been sampled among 50 genes distributed throughout the genome. The effect of the correlation among SNPs on this analysis should, therefore, be minimal.

The maximum log likelihood value for the full model is −234.19. However, the maximum log likelihood values for models assuming only neutral mutations, or a single class of selected mutations, are −243.82 and −240.88, respectively. Under the independence assumption, both of these simpler models can be rejected against the model with three classes of mutations, using a likelihood ratio test ($p = 0.0006$ and $p = 0.004$).

Supporting Information

Dataset S1. Results File

Found at DOI: 10.1371/journal.pbio.0030170.sd001 (3.1 MB XLS).

Dataset S2. Alignment File

Found at DOI: 10.1371/journal.pbio.0030170.sd002 (9.8 MB ZIP).

Accession Numbers

The sequence analyzed in this study has been submitted to GenBank (http://www.ncbi.nlm.nih.gov/Genbank/).

Acknowledgments

The data from this paper were obtained from more than 18 million sequencing reads obtained from the Celera Genomics sequencing center in Rockville, Maryland. We especially acknowledge the technical contributions of J. Duff, C. Evans, S. Ferriera, C. Forbes, C. Gire, B. Murphy, M. A. Rydland, B. Small, and G. Wang.

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

Author contributions. SG, DMT, DC, TJW, JJS, MDA, and MC conceived and designed the experiments and performed the experiments. RN, CB, AGC, TBS, MJH, and AFA analyzed the data and contributed reagents/materials/analysis tools. RN, CB, AGC, MDA, and MC wrote the paper. ∎

References

1. Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13: 685–690.
2. Hughes AL (1997) Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. Mol Biol Evol 14: 1–5.
3. Sawyer, SL, Emerman, M, Malik, HS (2004) Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol 2: e275. DOI:10.1371/journal.pbio.0020275
4. Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci U S A 94: 7712–7718.
5. Bush RM, Fitch WM, Bender CA, Cox NJ (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol 16: 1457–1465.
6. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–936.
7. Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2000) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Natl Acad Sci U S A 98: 7375–7379.
8. Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 20: 18–20.
9. Gilad Y, Segre D, Skorecki K, Nachman MW, Lancet D, et al. (2000) Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nat Genet 26: 221–224.
10. 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.
11. Simes RJ, (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 751–754.
12. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335: 167–170.
13. Yang Z, Bielawski J (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503.
14. Ebersberger I, Metzler D, Schwarz C, Pääbo S (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet 70: 1490–1497.
15. Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, et al. (2004) DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature 429: 382–388.
16. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, et al. (2003) PANTHER: A browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res 31: 334–341.
17. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: A library of protein families and subfamilies indexed by function. Genome Res 13: 2129–2141.
18. Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124: 967–978.
19. Yang VW (1998) Eukaryotic transcription factors: Identification, characterization and functions. J Nutr 128: 2045–2051.
20. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, et al. (2002) Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A 99: 4465–4470.
21. Mirnics K, Pevsner J (2004) Progress in the use of microarray technology to study the neurobiology of disease. Nat Neurosci 7: 434–439.
22. Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, et al. (2002) Intra- and interspecific variation in primate gene expression patterns. Science 296: 340–343.
23. Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, et al. (2004) Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol 2: e207. DOI:10.1371/journal.pbio.0020207
24. Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, et al. (2004) Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. Cell 199: 1027–1040.
25. Charlesworth B, Coyne JA, Barton NH (1987) The relative rates of evolution of sex chromosomes and autosomes. Am Nat 130: 113–146.
26. Torgerson DG, Singh RS (2003) Sex-linked mammalian sperm proteins evolve faster than autosomal ones. Mol Biol Evol 20: 1705–1709.
27. Wang PJ, McCarrey JR, Yang F, Page DC (2001) An abundance of X-linked genes expressed in spermatogonia. Nat Genet 27: 422–426.
28. Makova KD, Li WH (2002) Strong male-driven evolution of DNA sequences in humans and apes. Nature 416: 624–626.
29. Betancourt AJ, Kim Y, Orr HA (2004) A pseudohitchhiking model of X vs. autosomal diversity. Genetics 168: 2261–2269.
30. 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.
31. Wyckoff GJ, Wang W, Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. Nature 403: 304–309.
32. Clark AG, Civetta A (2000) Evolutionary biology: Protamine wars. Nature 403: 261–263.
33. Allan DJ, Harmon BV, Kerr JFR (1987) Cell death in spermatogenesis. In: Potten CS, editor. Perspective on mammalian cell death. London: Oxford University Press. pp. 229–258.
34. Sinha Hikim AP, Wang C, Leung A, Swerdloff RS (1995) Involvement of apoptosis in the induction of germ cell degeneration in adult rats after gonadotropin-releasing hormone antagonist treatment. Endocrinology 136: 2770–2775.
35. Rodriquez I, Ody C, Araki K, Garcia I, Vassalli P (1997) An early and massive wave of germinal cell apoptosis is required for the development of functional spermatogenesis. EMBO J 16: 2262–2270.
36. Francavilla S, D'Abrizio P, Cordeschi G, Pelliccione F, Necozione S, et al. (2002) Fas expression correlates with human germ cell degeneration in meiotic and post-meiotic arrest of spermatogenesis. Mol Hum Reprod 8: 213–220.
37. Yang Z (1997) PAML: A program for package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 15: 555–556.
38. Hudson RR (1990) Gene genealogies and the coalescent process. In: Harvey PH, Partridge L, editors. Oxford surveys in evolutionary biology, Volume 7. New York: Oxford University Press. pp. 1–44.
39. Polanski A, Kimmel M (2003) New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165: 427–436.
40. Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166: 351–372.
41. Nielsen R (2001) Statistical tests of neutrality in the age of genomics. Heredity 86: 641–647.
42. Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 60: 1179–1189.
43. van de Vosse E, Walpole SM, Nicolaou A, van der Bent P, Cahn A, et al. (1998) Characterization of *SCML1*, a new gene in Xp22, with homology to developmental polycomb genes. Genomics 49: 96–102.
44. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.
45. Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. Genetics 159: 1779–1788.
46. Zhang J (2004) Frequent false detection of positive selection by the likelihood method with branch-site models. Mol Biol Evol 21: 1332–1339.
47. Wong W, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168: 1041–1051.
48. Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 25: 2745–2751.
49. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
50. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32–43.

Note Added in Proof

The version of this paper that was first made available on 3 May 2005 has been replaced by this, the definitive, version.