

Population History and Natural Selection Shape Patterns of Genetic Variation in 132 Genes

Joshua M. Akey^{1*}, Michael A. Eberle¹, Mark J. Rieder², Christopher S. Carlson², Mark D. Shriver³, Deborah A. Nickerson², Leonid Kruglyak^{1,4*}

1 Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, **2** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America, **3** Department of Anthropology, Pennsylvania State University, University Park, Pennsylvania, United States of America, **4** Howard Hughes Medical Institute, Chevy Chase, Maryland, United States of America

Identifying regions of the human genome that have been targets of natural selection will provide important insights into human evolutionary history and may facilitate the identification of complex disease genes. Although the signature that natural selection imparts on DNA sequence variation is difficult to disentangle from the effects of neutral processes such as population demographic history, selective and demographic forces can be distinguished by analyzing multiple loci dispersed throughout the genome. We studied the molecular evolution of 132 genes by comprehensively resequencing them in 24 African-Americans and 23 European-Americans. We developed a rigorous computational approach for taking into account multiple hypothesis tests and demographic history and found that while many apparent selective events can instead be explained by demography, there is also strong evidence for positive or balancing selection at eight genes in the European-American population, but none in the African-American population. Our results suggest that the migration of modern humans out of Africa into new environments was accompanied by genetic adaptations to emergent selective forces. In addition, a region containing four contiguous genes on Chromosome 7 showed striking evidence of a recent selective sweep in European-Americans. More generally, our results have important implications for mapping genes underlying complex human diseases.

Citation: Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, et al. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* 2(10): e286.

Introduction

Despite intense study and interest, a detailed understanding of the evolutionary and demographic forces that have shaped extant patterns of human genomic variation remains elusive. An important goal in studies of DNA sequence variation is to identify loci that have been targets of natural selection and thus contribute to differences in fitness between individuals in a population. Identifying regions of the human genome that have been subject to natural selection will provide important insights into recent human history (Sabeti et al. 2002; Tishkoff and Verrelli 2003), the function of genes (Akey et al. 2002), and the mechanisms of evolutionary change (Otto 2000), and it may also facilitate the identification of complex disease genes (Jorde et al. 2001; Nielsen 2001).

The neutral theory of molecular evolution (Kimura 1968; King and Jukes 1969), which posits that the majority of polymorphisms have no appreciable effects on fitness, has been integral to recent studies of natural selection. Specifically, the neutral theory makes explicit and quantitative predictions about the amount, structure, and patterns of sequence variation expected under neutrality, and serves as a null hypothesis by which to evaluate the evidence for or against selection in empirical data (Otto 2000; Nielsen 2001). Unfortunately, robust inferences of natural selection from DNA sequence data are difficult because of the confounding effects of population demographic history. For example, both positive selection and increases in population size lead to an excess of low-frequency alleles in a population relative to what is expected under a standard neutral model (i.e., a constant-size, randomly mating population at mutation-drift

equilibrium). Therefore, rejection of the standard neutral model usually cannot be interpreted as unambiguous evidence for selection.

One way out of this conundrum is to recognize that population demographic history affects patterns of variation at all loci in a genome in a similar manner, whereas natural selection acts upon specific loci (Cavalli-Sforza 1966; Przeworski et al. 2000; Andolfatto 2001; Nielsen 2001). Thus, by sampling a large number of unlinked loci throughout the genome, it is in principle possible to distinguish between selection and demography. For instance, Akey et al. (2002) recently used this approach to infer the presence of selection in a genome-wide collection of single nucleotide polymorphisms (SNPs). However, studies based on SNPs that were initially identified in a small sample and subsequently genotyped in a larger sample are not ideally suited for detecting selection, because ascertainment bias (i.e., a systematic bias introduced into a dataset because of the way in which the data were collected) complicates downstream analyses (Akey et al. 2003). However, DNA sequence data provides the opportunity to exhaustively catalog variation,

Received April 23, 2004; Accepted June 25, 2004; Published September 7, 2004
DOI: 10.1371/journal.pbio.0020286

Copyright: © 2004 Akey et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviations: FDR, false discovery rate; SNP, single nucleotide polymorphism

Academic Editor: Lon Cardon, University of Oxford

*To whom correspondence should be addressed. E-mail: jakey@fhcrc.org (JMA), leonid@fhcrc.org (LK)

evidence of natural selection and of inferring population demographic history.

Correcting for Multiple Hypothesis Tests

In order to robustly correct for multiple hypothesis tests, the conventional practice of assuming no recombination when determining significance is not appropriate, because it results in conservative p values (Wall 1999) and hence decreases the statistical power to detect deviations from neutrality. Although recombination can easily be incorporated into coalescent simulations, in practice it is difficult to accurately estimate recombination rates, which vary substantially across the genome (Yu et al. 2001; McVean et al. 2004). To model the stochastic behavior and uncertainty in local rates of recombination, we reassessed the significance of Tajima's D , Fu and Li's D^* , Fu and Li's F^* , and Fay and Wu's H by coalescent simulations that incorporate recombination rates sampled from a $\text{Gamma}(2, 0.5 \times 10^{-8})$ distribution (see Materials and Methods). Finally, we corrected each statistic for multiple tests using the positive false discovery rate (FDR; Storey 2002) method, which determines the predicted proportion of "false positives" for the number of significant observations.

In the European-American sample, we observed 22 genes that were significant at a FDR of 5% (i.e., we expect approximately one false positive in this set of genes) for one or more tests of the allele frequency distribution (Tables 1 and S2). Thus, the number of significant genes in the European-American sample, after incorporating recombination and correcting for multiple tests, is very similar to the initial results where recombination was ignored and multiple tests

were not corrected for. However, in the African-American sample there were no genes significant at a FDR of 5% for any of the tests of the allele frequency distribution (unpublished data). This result is consistent with the relatively small number of significant genes that were initially found before correcting for multiple tests (Figure 1). Genes with the smallest FDR in African-Americans were *ABO*, *F2RL1*, and *IL17B*, which each had a FDR of 13.5% for Fu and Li's D^* .

Distinguishing between Selective and Demographic Forces

Although neutrality tests of the allele frequency distribution reveal many significant deviations, it is impossible to unambiguously interpret these data as evidence for natural selection, because the null model used to assess significance makes unrealistic assumptions about population demographic history. In principle, it is possible to distinguish between demography and selection, because demography affects all loci in the genome, whereas selection acts upon specific loci. Thus, by sampling a large number of loci dispersed throughout the genome, we can begin to construct a more realistic null hypothesis by which to evaluate the evidence for or against selection (Kreitman 2000).

To this end, we used the empirical data to explore four different demographic models (Figure 2A), which we could then use to account for demographic influences on tests of natural selection. For each model, we used coalescent theory to simulate data over a broad range of parameters and identified the particular combination of parameters that most closely matched summary statistics (average Tajima's D ,

Table 1. Significant Genes in European-Americans after Correcting for Multiple Tests

Gene	D	$p(D)$	D^*	$p(D^*)$	F^*	$p(F^*)$	H	$p(H)$
<i>ABO</i>	2.34	0.0002	1.73	0.0001	2.32	0.0001	-0.58	0.3717
<i>ACE2</i>	1.85	0.0058	1.63	0.0016	2.02	0.0024	-2.57	0.1474
<i>CYP4A11</i>	-0.90	0.2632	1.03	0.1570	0.41	0.5766	-13.60	0.0068
<i>DCN</i>	-1.99	0.0018	-3.88	0.0006	-3.83	0.0004	-7.69	0.0150
<i>EPHB6</i>	-2.02	0.0026	-4.09	0.0008	-4.00	0.0008	-4.18	0.0964
<i>F2</i>	-0.89	0.2288	0.53	0.5430	0.03	0.9722	-10.85	0.0168
<i>IL10RA</i>	-0.32	0.7206	-0.40	0.6624	-0.44	0.5902	-16.10	0.0060
<i>IL10RB</i>	1.89	0.0020	0.20	0.8458	0.97	0.1156	0.36	0.4174
<i>IL1A</i>	2.44	0.0004	1.40	0.0150	2.09	0.0016	3.57	0.7964
<i>IL1R2</i>	0.06	0.8552	1.00	0.0712	0.79	0.2364	-19.62	0.0192
<i>IRAK4</i>	-1.41	0.0304	-0.66	0.3324	-1.12	0.1736	-23.67	0.0014
<i>ITGA8</i>	-0.12	0.8686	0.67	0.2976	0.45	0.4764	-22.40	0.0130
<i>KEL</i>	-2.21	0.0001	-3.74	0.0022	-3.81	0.0016	-19.81	0.0002
<i>PLAUR</i>	-0.57	0.4382	-0.03	0.9848	-0.28	0.7182	-20.85	0.0141
<i>PPARA</i>	-0.77	0.2570	-0.86	0.2714	-0.99	0.2028	-13.92	0.0105
<i>SELP</i>	0.29	0.5612	0.41	0.4782	0.44	0.4828	-25.21	0.0189
<i>SFTPA1</i>	-1.35	0.0498	-0.60	0.4134	-1.05	0.1994	-38.02	0.0097
<i>SFTPA2</i>	-1.45	0.0402	-0.69	0.3264	-1.16	0.1678	-26.81	0.0132
<i>TNFRSF1B</i>	-0.52	0.3806	0.43	0.5636	0.10	0.9028	-30.68	0.0026
<i>TRPV5</i>	-2.07	0.0002	-4.98	0.0001	-4.69	0.0001	-32.86	0.0001
<i>TRPV6</i>	-2.74	0.0001	-6.06	0.0001	-5.80	0.0001	-45.43	0.0001
<i>VTN</i>	-0.43	0.7324	1.12	0.0738	0.71	0.4494	-8.30	0.0032

D , D^* , F^* , and H denote Tajima's D , Fu and Li's D^* , Fu and Li's F^* , and Fay and Wu's H , respectively. Nominal p values determined from 10^4 coalescent simulations with recombination are shown in the column next to each statistic. The p values that are significant after correcting for multiple tests (FDR = 5%) are shown in bold.

DOI: 10.1371/journal.pbio.0020286.t001

Fu and Li's D^* , and Fu and Li's F^*) of the observed data. Of the four demographic models, the European-American data are most consistent with a bottleneck occurring approximately 40,000 y ago, which is nearly identical to a previously reported estimate (Sabeti et al. 2002). However, the confidence intervals for the observed summary statistics are broad, and various aspects of the data are also consistent with other models (Figure 2B). The African-American data are most consistent with either an exponential expansion or a relatively old and severe bottleneck (Figure 2). Similarly, using DNA sequence variation from ten unlinked, noncoding loci, Pluzhnikov et al. (2002) found that an African Hausa sample was consistent with a recent population expansion (although they did not consider bottleneck models).

We reestimated the significance of Tajima's D , Fu and Li's D^* , Fu and Li's F^* , and Fay and Wu's H in each population for each of the four demographic models using the best-fit parameter values. All simulations included recombination and correction for multiple tests using the FDR method (with a FDR of 5%) as described above. Population history can clearly have a profound effect on tests of natural selection (Figure 3A and 3B; see also Simonsen 1995; Przeworski 2002), and given the uncertainty in our knowledge of human demographic history, it is challenging to ascribe unusual patterns of genetic variation to either demography or selection. To address this problem, we identified genes whose statistical evidence for selection was robust to demographic history. We conservatively defined demographically robust selection genes as those that demonstrated significant evidence for selection in all five demographic models. We identified eight demographically robust selection genes in European-Americans, and zero in African-Americans (Figure 3C; Table 2). Thus, out of the 22 genes originally found to be significant (at a FDR of 5%) under a standard neutral model, our estimates suggest that demographic history can potentially account for approximately two-thirds of these observations.

Evidence for a Recent Selective Sweep on Chromosome 7q in European-Americans

One particularly interesting region of the genome is located at 7q and contains four contiguous demographically

robust selection genes (*EPHB6*, *TRPV6*, *TRPV5*, and *KEL*; Figure 4A). Collectively, the entire 115-kb region bears many of the hallmarks of a locus subject to a recent selective sweep: an excess of high-frequency-derived alleles (Figure 4B); an overall excess of rare polymorphisms, which results in an extreme skew of the site frequency spectrum reflected by sharply negative values of Tajima's D (Figure 4C); and a significant reduction in the amount of nucleotide diversity (Figure 4D). The signature of positive selection is seen only in European-Americans, suggesting that *EPHB6*, *TRPV6*, *TRPV5*, and/or *KEL* possess specific alleles that have conferred local adaptation to a unique environmental pressure in European-derived populations. Consistent with this hypothesis, we observed strong levels of population subdivision (Figure 4E) across the entire 115-kb region. The closest genes centromeric to *EPHB6* and telomeric to *KEL* are approximately 42 kb and 64 kb away, respectively, suggesting that one or more of these four genes is the target of selection. However, we have not surveyed patterns of DNA sequence variation outside of the region delimited by *EPHB6* and *KEL*, and therefore it is possible that the signature of selection extends even further. Based on the level of genetic variation on the putatively selected haplotype (see Materials and Methods), we can provide a rough estimate of the time back to the selective sweep as approximately 10,000 y ago. Although this number should be interpreted cautiously, it suggests that selection operated recently.

Discussion

In summary, we have found that both population demographic history and natural selection shaped patterns of DNA sequence variation in the 132 genes studied here. By studying multiple unlinked loci dispersed throughout the genome, we were able to develop a rigorous computational approach to distinguish between the confounding effects of natural selection and demographic history on patterns of genetic variation. Using this strategy, we found that approximately two-thirds of the genes that were initially significant could be accounted for by population demographic history. Thus, our analyses clearly demonstrate the importance of considering

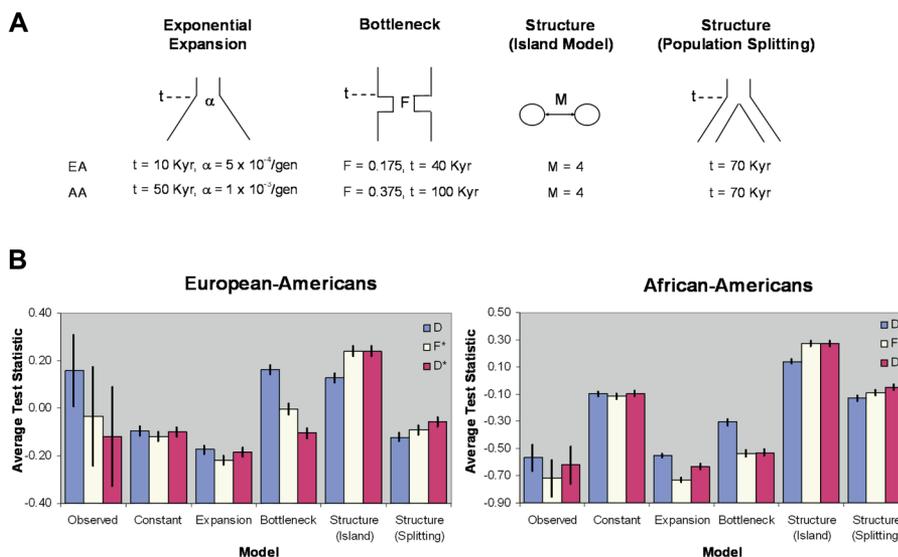


Figure 2. Summary of the Four Demographic Models Considered in Each Population

(A) Schematic diagram of each demographic model and its associated parameters (see Materials and Methods for details). Parameter values that match the observed data most closely for European-Americans (EA) and African-Americans (AA) are shown below the diagrams.

(B) Average and 95% confidence intervals of Tajima's D (blue bars), Fu and Li's D^* (red bars), and Fu and Li's F^* (pale yellow bars) for the observed data and each demographic model (using the parameters that most closely match the empirical data). Results from the standard neutral model (Constant) are also shown.

DOI: 10.1371/journal.pbio.0020286.g002

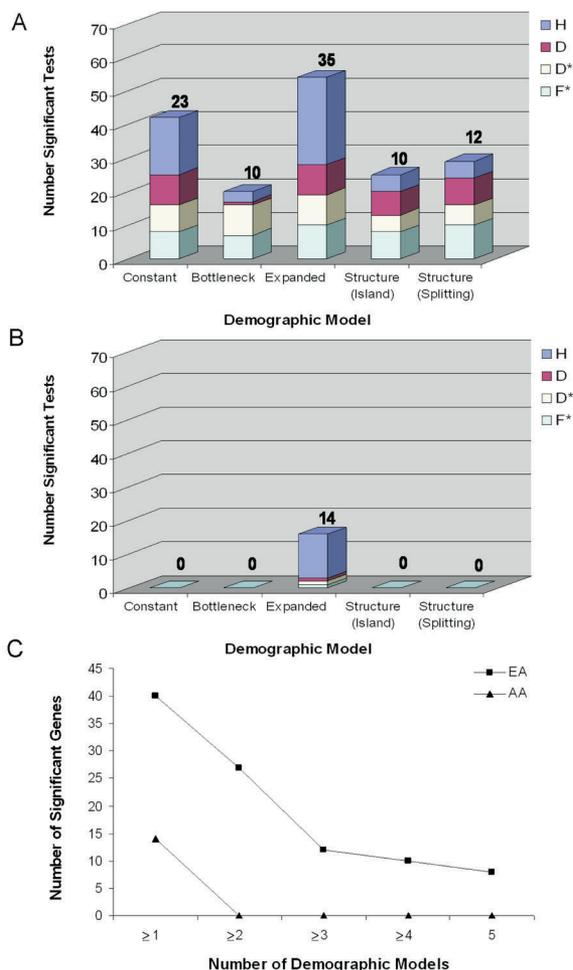


Figure 3. The Influence of Demographic History on Tests of Selection

(A and B) The significance of observed values of Tajima's D (red), Fu and Li's D* (pale yellow), Fu and Li's F* (pale blue), and Fay and Wu's H (dark blue) were reassessed for each best-fit demographic model in European-Americans (A) and African-Americans (B). Results from the standard neutral model (Constant) are shown for comparison. The number of significant genes for each demographic model is noted above each category in (A) and (B). For example, there were a total of 19 significant test statistics across all four tests of neutrality assuming a bottleneck model for Europeans, which define ten unique genes. Therefore, each gene is supported by approximately two (19/10) tests of neutrality.

(C) The distribution of the number of significant genes across the five demographic models in European-Americans and African-Americans. For example, in European-Americans, 40 genes were significant in at least one of the demographic models, and 27 genes were significant in at least two of the demographic models. DOI: 10.1371/journal.pbio.0020286.g003

both neutral and nonneutral forces when interpreting DNA sequence variation.

An interesting feature of our data is that the majority of deviations from neutrality, and all of the demographically robust selection genes, are not shared between the two population samples, suggesting that local adaptation has played an important role in recent human evolutionary history. Consistent with this observation, several possible examples of local adaptation in humans have previously been reported (Stephens et al. 1998; Rana et al. 1999; Hollox et al. 2001; Tishkoff et al. 2001; Currat et al. 2002; Fullerton et al. 2002; Gilad et al. 2002; Hamblin et al. 2002; Rockman et al.

Table 2. Demographically Robust Selection Genes in European-Americans

Gene	Chromosome	Biological Process
<i>IL1A</i>	2	Immunity and defense
<i>EPHB6</i>	7	Signal transduction
<i>KEL</i>	7	Protein metabolism and modification
<i>TRPV5</i>	7	Transport
<i>TRPV6</i>	7	Transport
<i>ABO</i>	9	Protein metabolism and modification
<i>DCN</i>	12	Signal transduction
<i>ACE2</i>	X	Protein metabolism and modification

Biological Process terms were assigned using the Panther classification scheme (Thomas et al. 2003). DOI: 10.1371/journal.pbio.0020286.t002

2003). We hypothesize that the stronger signature of selection in the European-derived population may reflect the exposure of non-African populations to novel and evolutionarily recent selective pressures (e.g., unique dietary, climatic, and cultural environments) as modern humans migrated out of Africa and spread throughout the world. In contrast, the African-derived population may have experienced fewer evolutionarily recent selective forces. Theoretical studies have demonstrated that the power to detect a selective sweep is generally greatest if it occurred less than approximately 0.1 N_e generations ago (i.e., approximately 20,000–25,000 y ago [Kim and Stephan 2000; Przeworski 2002]), which is consistent with our hypothesis that signatures of selection in European-Americans reflect recent selective events. However, it is important to note that we have surveyed less than 1% of all human genes, and many of the genes that we did analyze are involved in mediating inflammatory and immune responses; thus our results may not be representative of the genome at large. Interestingly, Glinka et al. (2003) found that European-derived populations of *Drosophila melanogaster* demonstrated abundant evidence for recent selective sweeps, whereas African populations did not, which is strikingly similar to our results in humans.

An alternative explanation for why we observed fewer significant results in African-Americans than in European-Americans is that African-Americans are an admixed population (Parra et al. 1998), and the admixture process may mask the signature of selection. However, simulation studies in which we constructed an artificially admixed European-American sample with African-American chromosomes resulted in an increase in significant genes relative to the observed data (unpublished data). Therefore, to the extent that our simulations recapitulate the dynamics of the admixture process in African-Americans, admixture is unlikely to explain the discrepancies between the two samples.

It is important to point out that some genes that do not meet our rigorous definition of a high-confidence selection gene may have nonetheless been targets of selection, such as *ABO* in African-Americans (Table S2). In this initial survey we have elected to be conservative and identify genes that possess the strongest signatures of selection. Ultimately, it will

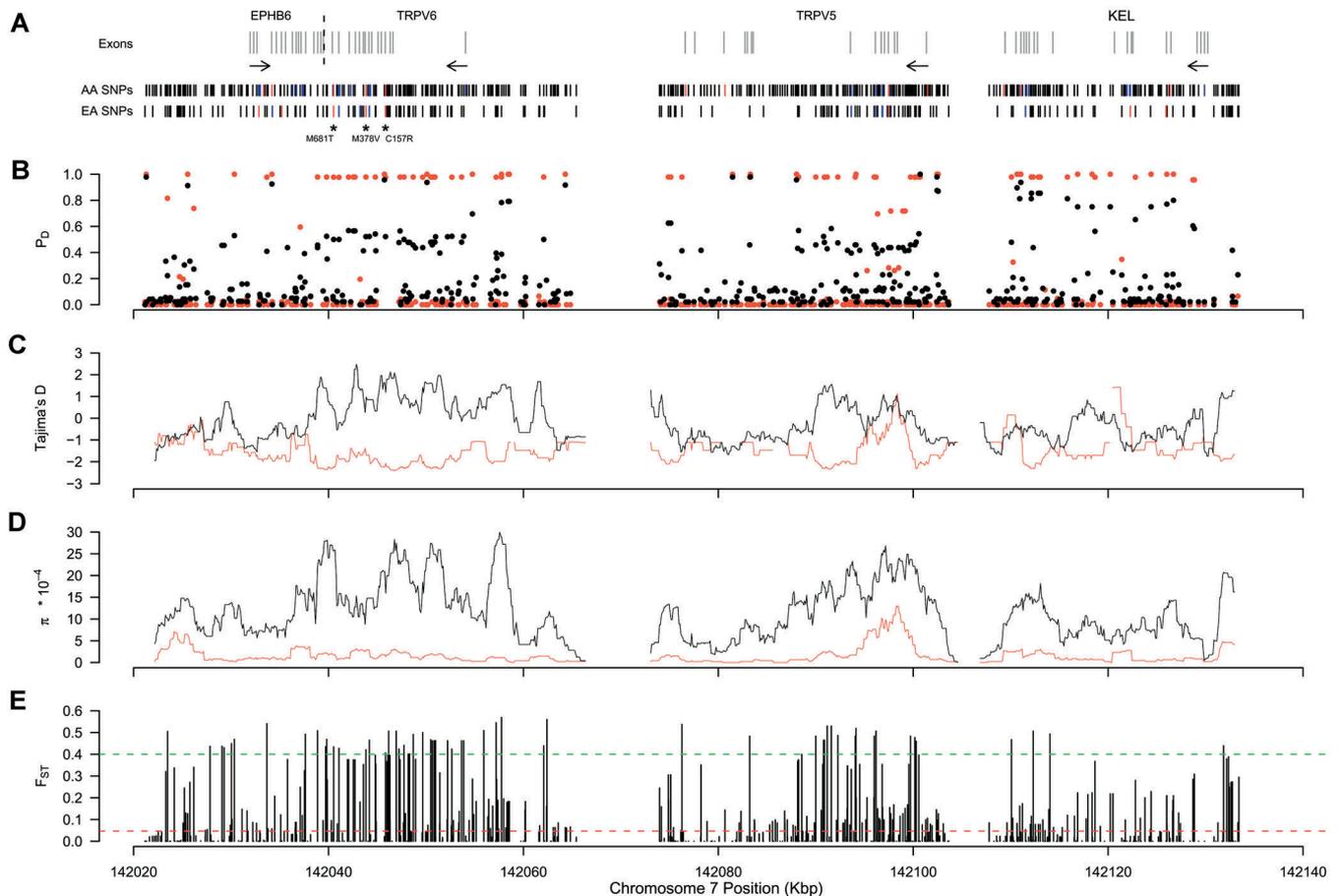


Figure 4. A Strong Signature of Positive Selection Spanning 115 kb on Chromosome 7q

(A–D) Exons for *EPHB6*, *TRPV6*, *TRPV5*, and *KEL* are shown as gray vertical lines. A dashed black line indicates the boundary between *EPHB6* and *TRPV6* exons, which are approximately 1 kb apart. Transcriptional orientation is indicated by the arrows below exon positions. SNPs found in European-Americans and African-Americans are shown below. Noncoding, synonymous, and nonsynonymous SNPs are denoted as black, blue, and red vertical bars, respectively. The positions of three nonsynonymous SNPs in *TRPV6* are shown with asterisks. For each of the resulting nonsynonymous amino acid changes, the most frequent amino acid in European-Americans is given first. The frequency of derived alleles, P_D (B), sliding window plots of Tajima's D (C), and nucleotide diversity, π (D), are shown across the entire region. Gaps in the sliding window plots indicate positions where sequence data were not obtained. In (B–D), European- and African-American data are shown in red and black, respectively.

(E) The distribution of F_{ST} across the 115-kb region. The average F_{ST} for all SNPs across the 132 genes is shown as a dashed red line. The dashed green line indicates the threshold for significantly ($p < 0.01$) large values of F_{ST} , determined by coalescent simulations.

DOI: 10.1371/journal.pbio.0020286.g004

be necessary to confirm our results in geographically diverse populations (a more comprehensive sampling of African populations is particularly needed), as well as in replicate samples of the populations we studied, and to functionally characterize the suspected targets of selection.

Recently, Clark et al. (2003) presented an evolutionary analysis of 7,645 orthologous human-chimp-mouse gene trios by looking for accelerated rates of synonymous and nonsynonymous nucleotide substitution in either the human or the chimp lineages. In total, 50 genes overlap between our dataset and theirs (Table S3), including three demographically robust selection genes (*TRPV6*, *EPHB6*, and *DCN*; see Table 2). All three of the demographically robust selection genes also demonstrate statistically significant evidence ($p < 0.05$) of accelerated evolution in either the human (*TRPV6* and *EPHB6*) or chimp (*DCN*) lineage. In addition, Clark et al. (2003) found evidence for accelerated evolution in seven genes along the human lineage that did not demonstrate

evidence for selection in our dataset (Table S3). This observation may simply reflect either false negatives in our analysis or false positives in Clark et al. (2003). However, it is important to note that the statistical methods and data used to detect selection in Clark et al. (2003) (divergence between species) are quite different from our methods (polymorphism within species), so completely overlapping results are not expected. More specifically, the analyses of Clark et al. (2003) will preferentially detect selective events between species, whereas our analyses will preferentially identify selection operating within species. In other words, these two methods are complimentary and may potentially detect selection operating over different time scales. In this respect, it is particularly interesting that the genes we identified as possessing the strongest evidence for recent selection in one human population also show evidence of selection in the human or chimp lineage following their divergence (Clark et al. 2003).

The strongest signature of selection that we observed occurs on Chromosome 7q in European-Americans. The signature of selection extends for at least 115 kb and spans the genes *EPHB6*, *TRPV6*, *TRPV5*, and *KEL*. To our knowledge, this is the largest footprint of selection that has been described in the human genome, and likely reflects the combination of strong and recent selective pressures and reduced recombination in this region (the average ratio of genetic to physical distance, cM/Mb, is approximately 0.68 according to the deCode map). Based on our current data it is impossible to identify which gene (or perhaps genes) has been the target (or targets) of selection. However, *TRPV6* is a particularly interesting candidate, as it possesses three non-synonymous amino acid substitutions (C157R, M378V, and M681T) that are each nearly fixed for the derived allele in European-Americans, show significant frequency differences between European-Americans and African-Americans, and are located in the most significant regions of both Tajima's D and reduced nucleotide diversity (Figure 4). The program PolyPhen (Ramensky et al. 2002) predicts that the C157R replacement may alter protein structure. Recently, *TRPV6* was shown to be up-regulated in prostate cancer (Wissenbach et al. 2001), and a susceptibility locus for aggressive prostate cancer was mapped to the *TRPV6* region (7q31–33; Paiss et al. 2003). These observations, combined with the large difference in disease prevalence between Europeans and African-Americans (Crawford 2003), make *TRPV6* a strong candidate gene for prostate cancer susceptibility and/or aggressiveness.

TRPV6, as well as *TRPV5*, constitute the rate-limiting step in kidney, intestine, and placenta calcium absorption (Nijenhuis et al. 2003; van de Graaf et al. 2003). Interestingly, Northern European populations have very high frequencies of the lactase persistence allele (*LCT*P*; Hollox et al. 2001), which allows digestion of fresh milk throughout adulthood. It is widely accepted that strong selection has driven *LCT*P* to high frequency in Northern Europeans, beginning sometime after the domestication of animals approximately 9,000 y ago (Feldman and Cavalli-Sforza 1989; Hollox et al. 2001; Bersaglieri et al. 2004). What has been debated, however, is the specific selective advantage conferred by lactase persistence (Holden and Mace 1997). Our finding that *TRPV6* and/or *TRPV5* have been under strong selective pressure in Northern Europeans suggests that increased calcium absorp-

tion may have been the driving force behind selection for lactase persistence, which was originally hypothesized by Flatz and Rotthauwe (1973). Although additional studies are clearly needed, our results provide additional insight into the molecular mechanisms of adaptation to a new dietary niche (i.e., high-lactose diets).

More generally, our results have several implications for mapping genes underlying complex human diseases. Specifically, four of the high-confidence selection genes have been implicated in various complex diseases (Table 3). If genes underlying complex diseases have experienced differential selective pressures, then this could in part explain the failure of many studies to replicate disease associations across populations (Florez et al. 2003; Moore 2003). Finally, our data are consistent with the notion that variation in genes that was once beneficial may have become detrimental in the environmental and cultural milieu of contemporary human populations, akin to the "thrifty gene" hypothesis for type II diabetes (Neel 1962).

Materials and Methods

DNA samples and sequencing. Human DNAs were obtained from the Coriell Institute (Camden, New Jersey, United States). We analyzed DNA from 24 African-Americans from the Human Variation Panel, African-American Panel of 50 (HD50AA) and DNA from 23 European-Americans derived from various CEPH pedigrees. We also sequenced each gene in a common chimpanzee (*Pan troglodytes*) to determine the derived allele for Fay and Wu's H test. These data were generated under the auspices of the SeattleSNPs Program for Genomic Applications, which resequences candidate genes involved in inflammatory processes in humans. In general, we resequenced the complete genomic region for each gene, including introns and approximately 2 kb 5' of the gene and 1 kb 3' of the gene using Big-Dye terminator chemistry on an ABI 3700 or ABI 3730XL (Applied Biosystems, Foster City, California, United States). For several exceptionally large genes, such as *FL3AI*, less than complete coverage was obtained (see Table S1). All variants occurring once in the sample were confirmed with an additional sequencing run. Further experimental details and all of the raw data can be found at our website (<http://pga.gs.washington.edu/>).

Data analysis. We calculated the following summary statistics of nucleotide variation for each gene: $\theta_W = Sa_p$, where S is the number of segregating sites, $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$, and n is the sample size (Watterson 1975); $\hat{\pi} = \sum_{i=1}^S h_i$, where h_i is an unbiased estimate of nucleotide

diversity for the i th segregating site (see equation 12 in Tajima 1989) and η_S , which is the number of singletons (Fu and Li 1993). From these statistics we calculated several tests of the standard neutral model including Tajima's D (Tajima 1989), Fu and Li's D* (Fu and Li 1993), Fu and Li's F* (Fu and Li 1993), and Fay and Wu's H statistic (Fay and Wu 2000). In calculating Fu and Li's F*, we used the formulas provided in Simonsen et al. (1995), which correct a typographical error in the original description of the method (Fu and Li 1993). For a discussion of the similarities and differences of Tajima's D, Fu and Li's D*, Fu and Li's F*, and Fay and Wu's H, see Fu and Li (1993), Simonsen et al. (1995), and Przeworski (2002).

We initially assessed the significance of these statistics by comparing the observed values to 10^4 coalescent simulations (Hudson 1983), conditional on the observed sample size and number of segregating sites, assuming a standard neutral model with no recombination. Coalescent simulations were performed using the program ms (obtained from R. Hudson's Web site [<http://home.uchicago.edu/~rhudson1/source.html>]). In order to correct for multiple tests, we repeated the coalescent simulations as described above, but included recombination. Following Pluzhnikov et al. (2002), for each of the 10^4 coalescent realizations, we sampled the recombination rate from a Gamma(2, 0.5×10^{-8}) distribution whose expectation equals the average genome-wide recombination rate of 10^{-8} /generation (Hamblin et al. 2002). The positive FDR method was used to correct for multiple hypothesis tests using the software QVALUE (Storey 2002; <http://faculty.washington.edu/~jstorey/qvalue/>).

Table 3. Disease Associations with Demographically Robust Selection Genes

Gene	Disease	Reference
<i>ACE2</i>	Hypertension	Crackower et al. (2002)
<i>DCN</i>	Renal disease	De Cosmo et al. (2002)
<i>IL1A</i>	Nasal polyposis	Karjalainen et al. (2003)
	Atopy	Karjalainen et al. (2002)
	Systemic sclerosis	Kawaguchi et al. (2003)
<i>ABO</i>	Alzheimer's disease	Combarros et al. (2002)
	Norwalk virus infection	Hutson et al. (2002)
	Asthma	Ronchetti et al. (2001)

DOI: 10.1371/journal.pbio.0020286.t003

We quantified the allele frequency differences between the European- and African-American samples by the statistic F_{ST} as described in Akey et al. (2002). All of the analyses described above excluded insertion/deletion polymorphisms, but their inclusion does not affect any of our conclusions (unpublished data). We assigned PANTHER Biological Process terms (Thomas et al. 2003) to each gene.

We estimated the time since the selective sweep for the Chromosome 7q region in European-Americans by analyzing the amount of nucleotide diversity that has accumulated on the selected haplotype as described in Rozas et al. (2001). We assumed that *TRPV6* is the target of selection and the selected haplotype is defined by the C157R, M378V, and M681T polymorphisms. If mutations are Poisson-distributed, the expected number of segregating sites in a genealogy is $E[S] = \mu E[T]$, where S , μ , and T denote segregating sites, neutral mutation rate of the locus, and total branch length of the genealogy, respectively. Assuming a star-shaped genealogy, $E[T] = n \times t$, where n is the number of selected haplotypes. Thus, the time back to the selective sweep, t , can be estimated by $S/(n\mu)$. For *TRPV6* in European-Americans, $n = 45$ (i.e., 45 out of 46 haplotypes carry C157, M378, and M681), $S = 11$, and $\mu = 2.5 \times 10^{-9}$.

Demographic modeling. We assessed the impact of demographic history on the robustness of the statistical tests of neutrality by using coalescent theory to simulate data under four different population histories, including a bottleneck, exponential expansion, population structure according to an island model that allows symmetric migration between demes, and population structure assuming population splitting with no subsequent migration. For each model we simulated data under a wide variety of parameters by conditioning on the observed sample size and θ_w for each population. The bottleneck model is specified by the parameters F (the inbreeding coefficient) and t (the time in years measured from the present) at which the bottleneck occurred. Values of F and t considered were $F = [0.05, 0.075, \dots, 0.40]$ and $t = [10,000, 20,000, \dots, 100,000]$. The exponential expansion model is determined by the parameters α (the growth rate/generation) and t (the time, in years measured from the present, at which the population began increasing in size). Values considered for α and t were: $\alpha = [0.0005, 0.001, \dots, 0.01]$ and $t = [10,000, 20,000, \dots, 100,000]$. The population structure under an island model is specified by the population migration rate between demes, $M = 4N_e m$, where N_e and m are the effective subpopulation size and fraction of migrants in each subpopulation per generation, respectively. Values of M considered were $M = [1, 2, \dots, 10]$. The structure model assuming population splitting with no subsequent migration is determined by the parameter t (the time in years since the populations diverged). Values of t considered were $t = [1,000, 2,000, \dots, 10,000]$. In all simulations we assumed an effective population size of 10,000 and a generation time of 25 y in order to facilitate comparisons to a previous study (Sabeti et al. 2002). The parameter space for each model included a full grid search, so we tested 160, 100, 10, and 10 parameter combinations for the bottleneck, expansion, structure (island), and structure (splitting) models, respectively. We performed 10^4 simulations for each parameter combination.

For each demographic model, we calculated the average value of Tajima's D, Fu and Li's D*, and Fu and Li's F* and compared the results to the observed values of these statistics. For the bottleneck and exponential expansion models, we identified the parameter values that most closely matched the observed data by identifying the parameter combination that minimized the function $S = \sum_{i=1}^3 |T_{O_i} - T_{S_i}|$, where T_{O_i} and T_{S_i} denote the observed and simulated averages of Tajima's D, Fu and Li's D*, and Fu and Li's F*. For the demographic models of population structure we selected parameter values that matched the observed F_{ST} . Finally, we reassessed the significance of the observed values of Tajima's D, Fu and Li's D*, Fu and Li's F*, and Fay and Wu's H by 10^4 coalescent simulations for each demographic model using the best-fit parameter values.

References

- Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 12: 1805–1814.
- Akey JM, Zhang K, Xiong M, Jin L (2003) The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol Biol Evol* 20: 232–242.

Supporting Information

Table S1. Summary Statistics of the 132 Genes

DOI: 10.1371/journal.pbio.0020286.st001 (266 KB DOC).

Table S2. Neutrality Test Statistics

DOI: 10.1371/journal.pbio.0020286.st002 (534 KB DOC).

Table S3. Overlap of Genes Analyzed by Clark et al. (2003)

DOI: 10.1371/journal.pbio.0020286.st003 (87 KB DOC).

Accession Numbers

LocusLink ID numbers (<http://www.ncbi.nlm.nih.gov/LocusLink/>) for the genes discussed in this paper are *ABO* (28), *ACE2* (59272), *APOH* (350), *BDKRB2* (624), *BF* (629), *C2* (717), *CCR2* (1231), *CD36* (948), *CEBPB* (1051), *CRF* (10882), *CRP* (1401), *CSF2* (1437), *CSF3* (1440), *CSF3R* (1441), *CYP4A11* (1579), *CYP4F2* (8529), *DCN* (1634), *EPHB6* (2051), *F10* (2159), *F11* (2160), *F12* (2161), *F13A1* (2162), *F2* (2147), *F2R* (2149), *F2RL1* (2150), *F2RL2* (2151), *F2RL3* (9002), *F3* (2152), *F5* (2153), *F7* (2155), *F9* (2158), *FGA* (2243), *FGB* (2244), *FGG* (2266), *FGL2* (10875), *FSBP* (10646), *GPIBA* (2811), *ICAM1* (3383), *IFNG* (3458), *IGF2* (3481), *IGF2AS* (51214), *IL10* (3586), *IL10RA* (3587), *IL10RB* (3588), *IL11* (3589), *IL12A* (3592), *IL12B* (3593), *IL13* (3596), *IL15RA* (3601), *IL17B* (27190), *IL19* (29949), *IL1A* (3552), *IL1B* (3553), *IL1R1* (3554), *IL1R2* (7850), *IL1RN* (3557), *IL2* (3558), *IL20* (50640), *IL21R* (50615), *IL22* (50616), *IL24* (11009), *IL2RB* (3560), *IL3* (3562), *IL4* (3565), *IL4R* (3566), *IL5* (3567), *IL6* (3569), *IL8* (3576), *IL9* (3578), *IL9R* (3581), *IRAK4* (51135), *ITGA2* (3673), *ITGA8* (8516), *JAK3* (3718), *KEL* (3792), *KLK1* (3816), *KLKB1* (3818), *KNG* (3827), *LTA* (4049), *LTB* (4050), *MAP3K8* (1326), *MC1R* (4157), *MMP3* (4314), *MMP9* (4318), *NOS3* (4846), *PFC* (5199), *PLAT* (5327), *PLAU* (5328), *PLAUR* (5329), *PLG* (5340), *PON1* (5444), *PON2* (5445), *PPARA* (5465), *PPARG* (5468), *PROC* (5624), *PROCR* (10544), *PROS1* (5627), *PROZ* (8858), *PTGS2* (5743), *SCYA2* (6347), *SELE* (6401), *SELL* (6402), *SELP* (6403), *SELPLG* (6404), *SERPINA5* (5104), *SERPINC1* (462), *SERPINE1* (5054), *SFTPA1* (6435), *SFTPA2* (6436), *SFTPB* (6439), *SFTPC* (6440), *SFTPD* (6441), *SMPI* (23585), *STAT4* (6775), *STAT6* (6778), *TF* (7018), *TFPI* (7035), *TGFB3* (7043), *THBD* (7056), *TIRAP* (114609), *TNF* (7124), *TNFAIP1* (7126), *TNFAIP2* (7127), *TNFAIP3* (7126), *TNFRSF1A* (7132), *TNFRSF1B* (7133), *TRAF6* (7189), *TRPV5* (56302), *TRPV6* (55503), *VCAM1* (7412), *VEGF* (7422), and *VTN* (7448).

Coriell (<http://coriell.undm.edu/>) repository numbers for human genomic DNAs sequenced for this study are as follows. DNAs from African-Americans were NA17101–NA17116 and NA17133–NA17140. DNAs from European-Americans were NA06990, NA07019, NA07348, NA07349, NA10830, NA10831, NA10842–NA10845, NA10848, NA10850–NA10854, NA10857, NA10858, NA10860, NA10861, NA12547, NA12548, and NA12560.

Acknowledgments

We would like to thank members of the SeattleSNPs team (M. Ahearn, T. Armel, E. Calhoun, M. Chung, C. Hastings, P. Keyes, P. Lee, S. Kuldaneck, M. Montoya, C. Poel, E. Toth, and N. Rajkumar) for cataloging the variation data. We would also like to thank D. Akey and D. Crawford for critical reading of this manuscript and J. Fay for helpful discussions. This work was supported by a National Science Foundation Postdoctoral Research Fellowship in Interdisciplinary Informatics (JMA) and grants from the National Heart Lung and Blood Institute Program for Genomic Applications (HL66682 to DAN and MJR; HL66642 to LK), the National Institute of Mental Health (MH59520 to LK), and the National Institutes of Health Pharmacogenetics Research Network (U01 HL69757 to DAN). LK is a James S. McDonnell Centennial Fellow.

Conflicts of interest. The authors have declared that no conflicts of interest exist.

Author contributions. MJR, CSC, and DAN conceived and designed the experiments. JMA, MAE, MDS, and LK analyzed the data. JMA, DAN, and LK wrote the paper. ■

- Andolfatto P (2001) Adaptive hitchhiking effects on genome variability. *Curr Opin Genet Dev* 11: 635–641.
- Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, et al. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 74: 1111–1120.
- Cavalli-Sforza LL (1966) Population structure and human evolution. *Proc R Soc Lond B Biol Sci* 164: 362–379.

- Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, et al. (2003) Inferring non-neutral evolution from human-chimp-mouse orthologous gene trios. *Science* 12: 1960–1963.
- Combarros O, Sanchez-Guerra M, Infante J, Llorca J, Berciano J (2002) Gene dose-dependent association of interleukin-1A [-889] allele 2 polymorphism with Alzheimer's disease. *J Neurol* 249: 1242–1245.
- Crackower MA, Sarao R, Oudit GY, Yagil C, Kozieradzki I, et al. (2002) Angiotensin-converting enzyme 2 is an essential regulator of heart function. *Nature* 417: 822–828.
- Crawford ED (2003) Epidemiology of prostate cancer. *Urology* 62: 3–12.
- Currat M, Trabuchet G, Rees D, Perrin P, Harding RM, et al. (2002) Molecular analysis of the β -globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the β (S) Senegal mutation. *Am J Hum Genet* 70: 207–223.
- De Cosmo S, Tassi V, Thomas S, Piras GP, Trevisan R, et al. (2002) The Decorin gene 179 allelic variant is associated with a slower progression of renal disease in patients with type 1 diabetes. *Nephron* 92: 72–76.
- Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405–1413.
- Feldman MW, Cavalli-Sforza LL (1989) On the theory of evolution under genetic and cultural transmission with application to the lactose absorption problem. In: Feldman MW, editor. *Mathematical evolutionary theory*. Princeton: Princeton University Press. pp. 145–173.
- Flatz G, Rotthauwe HW (1973) Lactose nutrition and natural selection. *Lancet* 2: 76–77.
- Florez JC, Hirschhorn J, Altshuler D (2003) The inherited basis of diabetes mellitus: Implications for the genetic analysis of complex traits. *Annu Rev Genomics Hum Genet* 4: 257–291.
- Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133: 693–709.
- Fullerton SM, Bartoszewicz A, Ybazeta G, Horikawa Y, Bell GI, et al. (2002) Geographic and haplotype structure of candidate type 2 diabetes susceptibility variants at the calpain-10 locus. *Am J Hum Genet* 70: 1096–1106.
- Gilad Y, Rosenberg S, Przeworski M, Lancet D, Skorecki K (2002) Evidence for positive selection and population structure at the human MAO-A gene. *Proc Natl Acad Sci U S A* 99: 862–867.
- Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D (2003) Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: A multi-locus approach. *Genetics* 165: 1269–1278.
- Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet* 70: 369–383.
- Holden C, Mace R (1997) Phylogenetic analysis of the evolution of lactase digestion in adults. *Hum Biol* 69: 605–628.
- Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, et al. (2001) Lactase haplotype diversity in the Old World. *Am J Hum Genet* 68: 160–172.
- Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol* 23: 183–201.
- Hutson AM, Atmar RL, Graham DY, Estes MK (2002) Norwalk virus infection and disease is associated with ABO histo-blood group type. *J Infect Dis* 185: 1335–1337.
- Jorde LB, Watkins WS, Bamshad MJ (2001) Population genomics: A bridge from evolutionary history to genetic medicine. *Hum Mol Genet* 10: 2199–2207.
- Karjalainen J, Hukkoniemi J, Pessi T, Huhtala H, Nieminen MM, et al. (2002) The IL1A genotype associates with atopy in nonasthmatic adults. *J Allergy Clin Immunol* 110: 429–434.
- Karjalainen J, Joki-Erkkila VP, Hukkoniemi J, Pessi T, Nieminen MM, et al. (2003) The IL1A genotype is associated with nasal polyposis in asthmatic adults. *Allergy* 58: 393–396.
- Kawaguchi Y, Tochimoto A, Ichikawa N, Harigai M, Hara M, et al. (2003) Association of IL1A gene polymorphisms with susceptibility to and severity of systemic sclerosis in the Japanese population. *Arthritis Rheum* 48: 86–92.
- Kim Y, Stephan W. (2000) Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* 155: 1415–1427.
- Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217: 624–626.
- King JL, Jukes TH (1969) Non-Darwinian evolution. *Science* 164: 788–798.
- Kreitman M (2000) Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet* 1: 539–559.
- McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304: 581–584.
- Moore JH (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 56: 73–82.
- Neel JVA (1962) “Thrifty” genotype rendered detrimental by “progress.” *Am J Hum Genet* 14: 353–362.
- Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity* 86: 641–647.
- Nijenhuis T, Hoenderop JGJ, Nilius B, Bindels RJM (2003) (Patho)physiological implications of the novel epithelial Ca^{2+} channels TRPV5 and TRPV6. *Pflugers Arch* 446: 401–409.
- Otto SP (2000) Detecting the form of selection from DNA sequence data. *Trends Genet* 16: 526–529.
- Paiss T, Worner S, Kurtz F, Haeussler J, Hautmann RE, et al. (2003) Linkage of aggressive prostate cancer to chromosome 7q31–33 in German prostate cancer families. *Eur J Hum Genet* 11: 17–22.
- Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, et al. (1998) Estimating African American admixture proportions by use of population-specific alleles. *Am J Hum Genet* 63: 1839–1851.
- Pluzhnikov A, Di Rienzo A, Hudson RR. (2002) Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* 161: 1209–1218.
- Przeworski M (2002) The signature of positive selection at randomly chosen loci. *Genetics* 160: 1179–1189.
- Przeworski M, Hudson RR, Di Rienzo A (2000) Adjusting the focus on human variation. *Trends Genet* 16: 296–302.
- Ramensky V, Bork P, Sunyaev S (2002) Human nonsynonymous SNPs: Server and survey. *Nucleic Acids Res* 30: 3894–3900.
- Rana BK, Hewett-Emmett D, Jin L, Chang BH, Sambuughin N, et al. (1999) High polymorphism at the human melanocortin 1 receptor locus. *Genetics* 151: 1547–1557.
- Rockman MV, Hahn MW, Soranzo N, Goldstein DB, Wray GA (2003) Positive selection on a human-specific transcription factor binding site regulating IL4 expression. *Curr Biol* 13: 2118–2123.
- Ronchetti F, Villa MP, Ronchetti R, Bonci E, Latini L, et al. (2001) ABO/Secretor genetic complex and susceptibility to asthma in childhood. *Eur Respir J* 17: 1236–1238.
- Rozas J, Gulland M, Blandin G, Aguade M (2001) DNA variation at the rp49 Gene region of *Drosophila simulans*: Evolutionary inferences from an unusual haplotype structure. *Genetics* 158: 1147–1155.
- Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832–837.
- Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141: 413–429.
- Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, et al. (1998) Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *Am J Hum Genet* 62: 1507–1515.
- Storey JD (2002) A direct approach to false discovery rates. *J R Stat Soc B* 64: 479–498.
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
- 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.
- Tishkoff SA, Verrelli BC (2003) Patterns of human genetic diversity: Implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet* 4: 293–340.
- Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, et al. (2001) Haplotype diversity and linkage disequilibrium at human G6PD: Recent origin of alleles that confer malarial resistance. *Science* 293: 455–462.
- van de Graaf SF, Hoenderop JG, Gkika D, Lamers D, Prenen J, et al. (2003) Functional expression of the epithelial Ca^{2+} channels (TRPV5 and TRPV6) requires association of the S100A10-annexin 2 complex. *EMBO J* 22: 1478–1487.
- Wall JD (1999) Recombination and the power of statistical tests of neutrality. *Genet Res* 74: 65–79.
- Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7: 256–275.
- Wissenbach U, Niemeyer BA, Fixemer T, Schneidewind A, Trost C, et al. (2001) Expression of CaT-like, a novel calcium-selective channel, correlates with the malignancy of prostate cancer. *J Biol Chem* 276: 19461–19468.
- Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, et al. (2001) Comparison of human genetic and sequence-based physical maps. *Nature* 409: 951–953.