UC Berkeley

Title

Localizing Recent Adaptive Evolution in the Human Genome

Permalink

https://escholarship.org/uc/item/5204q547

Journal

PLOS Genetics, 3(6)

ISSN

1553-7390

Authors

Williamson, Scott H

Hubisz, Melissa J

Clark, Andrew G

et al.

Publication Date

2007-06-01

DOI

10.1371/journal.pgen.0030090

Copyright Information

Peer reviewed

Localizing Recent Adaptive Evolution in the Human Genome

Scott H. Williamson[1*], Melissa J. Hubisz[1¤a], Andrew G. Clark[2], Bret A. Payseur[2¤b], Carlos D. Bustamante[1], Rasmus Nielsen[3]

1 Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America, 2 Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, United States of America, 3 Center for Bioinformatics and Department of Biology, University of Copenhagen, Copenhagen, Denmark

Identifying genomic locations that have experienced selective sweeps is an important first step toward understanding the molecular basis of adaptive evolution. Using statistical methods that account for the confounding effects of population demography, recombination rate variation, and single-nucleotide polymorphism ascertainment, while also providing fine-scale estimates of the position of the selected site, we analyzed a genomic dataset of 1.2 million human single-nucleotide polymorphisms genotyped in African-American, European-American, and Chinese samples. We identify 101 regions of the human genome with very strong evidence ($p < 10^{-5}$) of a recent selective sweep and where our estimate of the position of the selective sweep falls within 100 kb of a known gene. Within these regions, genes of biological interest include genes in pigmentation pathways, components of the dystrophin protein complex, clusters of olfactory receptors, genes involved in nervous system development and function, immune system genes, and heat shock genes. We also observe consistent evidence of selective sweeps in centromeric regions. In general, we find that recent adaptation is strikingly pervasive in the human genome, with as much as 10% of the genome affected by linkage to a selective sweep.

Introduction

Describing how natural selection shapes patterns of genetic variation within and between species is critical to a general understanding of evolution. With the advent of comparative genomic data, considerable progress has been made toward quantifying the effect of adaptive evolution on genome-wide patterns of variation between species [1–5], and the effect of weak negative selection against deleterious mutations on patterns of variation within species [1,5,6]. However, relatively little is known about the degree to which adaptive evolution affects DNA sequence polymorphism within species and what types of selection are most prevalent across the genome. Of particular interest is the effect of very recent adaptive evolution in humans. If one can localize adaptive events in the genome, then this information, along with functional knowledge of the region, speaks to the selective environment experienced by recent human populations. Another reason for the interest in genomic patterns of selection is that recent studies [3,5] have suggested a link between selected genes and factors causing inherited disease; furthermore, several established cases of recent adaptive evolution in the human genome involve mutations that confer resistance to infectious disease (e.g., [7,8]). Therefore, knowledge of the location of selected genes could aid in the effort to identify genetic variation underlying genetic diseases and infectious disease resistance. From a theoretical perspective, both the relative rate of adaptive evolution at the molecular level and the degree to which natural selection maintains polymorphism have been the subjects of intense debate in population genetics and molecular evolution [9–12]. With genome-scale polymorphism data becoming available, it is now possible to address these decades-old problems directly.

Adaptive events alter patterns of DNA polymorphism in the genomic region surrounding a beneficial allele, so population genetic methods can be used to infer selection by searching for their effects in genomic single-nucleotide polymorphism (SNP) data. Several recent studies [13–16] have taken this approach to scan the human genome for evidence of recent adaptation. These studies identify several regions of the genome that have recently experienced selection, and they suggest that adaptation is a surprisingly pervasive force in recent human evolution. However, the results of these analyses can only be considered preliminary. All of these studies have focused on the empirical distribution of a given test statistic, reasoning that loci with extreme values will be the most likely candidates for selective sweeps. This approach

Abbreviations: CLR, composite likelihood ratio; DPC, dystrophin protein complex; FDR, false discovery rate; OR, olfactory receptor gene; SFS, site-frequency spectrum; SNP, single-nucleotide polymorphism

* To whom correspondence should be addressed. E-mail: sw292@cornell.edu

¤a Current address: Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America,

¤b Current address: Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin, United States of America

Author Summary

A selective sweep is a single realization of adaptive evolution at the molecular level. When a selective sweep occurs, it leaves a characteristic signal in patterns of variation in genomic regions linked to the selected site; therefore, recently released population genomic datasets can be used to search for instances of molecular adaptation. Here, we present a comprehensive scan for complete selective sweeps in the human genome. Our analysis is complementary to several recent analyses that focused on partial selective sweeps, in which the adaptive mutation still segregates at intermediate frequency in the population. Consequently, our analysis identifies many genomic regions that were not previously known to have experienced natural selection, including consistent evidence of selection in centromeric regions, which is possibly the result of meiotic drive. Genes within selected regions include pigmentation candidate genes, genes of the dystrophin protein complex, and olfactory receptors. Extensive testing demonstrates that the method we use to detect selective sweeps is strikingly robust to both alternative demographic scenarios and recombination rate variation. Furthermore, the method we use provides precise estimates of the genomic position of the selected site, which greatly facilitates the fine-scale mapping of functionally significant variation in human populations.

provides a sensible way to rank loci according to their signal of recent adaptation, but because we do not know how common selection is in the genome, the "empirical p value" approach does not directly test the hypothesis of selection for any individual locus, and it provides no means for quantifying how common selection is across the genome [17,18]. For instance, the null hypothesis of selective neutrality could be true for the entire genome, in which case even the most extreme values would carry no information regarding selection. Also, there are no a priori criteria available for deciding how extreme a region needs to be in order to identify selection. In short, these previous studies do not estimate their uncertainty in identifying selection. Another concern is that the statistical properties of previous methods have only been explored under the very simplest evolutionary models. Complex factors such as demographic events in the history of the population, recombination rate variation, and the biasing effects of SNP ascertainment protocols all have the potential to systematically cause false signals of natural selection, yet previous methods for identifying recent adaptation have not been thoroughly tested for their robustness to these complicating factors.

In this paper, we present a full statistical analysis of evidence for selective sweeps in the human genome using a method for detecting sweeps that has been thoroughly tested for robustness to demography and recombination rate variation, and that explicitly incorporates SNP ascertainment protocols. We apply this approach to dense genomic polymorphism data [19] with uniform SNP discovery protocols. A recent selective sweep (a bout of adaptive evolution that fixes a beneficial mutation) alters patterns of allele frequency at linked sites, eliminating variation at tightly linked loci and creating a relative excess of alleles at very low and very high frequencies at more distant loci [20–22]. Because the effect of a selective sweep will depend on the genomic distance away from the beneficial mutation, we use a statistical method (test

2 in [22]) that searches for the unique spatial pattern of allele frequencies along a chromosome that is found after a selective sweep. Essentially, the test uses a composite likelihood ratio (CLR) to compare a neutral model for the evolution of a genomic window with a selective sweep model. In the neutral null model, allele frequency probabilities are drawn from the background pattern of variation in the rest of the genome. In the selective sweep model, allele frequency probabilities are calculated using a model of a selective sweep that conditions on the background pattern of variation. Allele frequency probabilities also depend on two parameters: the genomic position of the selective sweep (ψ), and a compound parameter (α) that measures the combined effects of the strength of selection and the recombination rate between a SNP and the selected site.

Extensive simulations under a variety of evolutionary models indicate that this CLR approach is not misled by demographic events in the population's history, such as population size changes, divergence, subdivision, or migration. Furthermore, simulations indicate that this is the only available method for detecting sweeps that is not highly sensitive to assumptions about the underlying recombination rate or recombination hotspots. This lack of dependence on demography and recombination allows us to calculate p values for individual loci that are consistent across a wide range of selectively neutral null models. Hence, we can reliably measure our uncertainty in identifying selective sweeps, and we can obtain rough estimates of the prevalence of recent adaptation across the genome. Also, the present analysis is one of the first to fully correct for the bias introduced by SNP discovery protocols, and we account for the effects of multiple hypothesis testing using a false discovery rate approach [23,24]. The method we use provides an accurate estimate of the genomic location of the selected allele, a feature that greatly facilitates mapping of the genomic targets of natural selection. A final important difference between our analysis and previous work is that the method we use searches for the signature of "complete" selective sweeps (i.e., adaptation where the beneficial mutation has recently attained a frequency of ~100% in the population). In contrast, methods based on extended haplotype length and high linkage disequilibrium [14–16] have the most power to detect "partial" selective sweeps [15] (i.e., where the beneficial mutation has not yet spread throughout the entire population). Therefore, the two approaches are complementary, and most loci where we discover evidence for recent adaptation were not detected by previous genome-wide scans for selection or targeted candidate gene approaches.

Results

Table 1 lists the 101 genomic locations that show very strong evidence for a recent, complete selective sweep (CLR $p < 10^{-5}$), excluding locations where the estimate of sweep position was greater than 100 kb from a known gene, and excluding centromeric regions. Genomic locations with very strong evidence for a selective sweep, but not within 100 kb of a known gene, are shown in Table S1, and application of the CLR test via sliding window analyses of all autosomes are given in Table S2. Under the model of a recent and strong selective sweep, the composite likelihood estimate of the

Table 1. The 101 Regions of the Human Genome with the Strongest Evidence ($p < 0.00001$, CLR Test) for a Recent Selective Sweep from a Sliding Window Analysis of the Combined, African-American, European-American, and Chinese Samples

Sample	Chr.	CMLE Position[a]	CLR	Genes (Distance in kb)[b]	Notes
African-American	1	13427120	29.024	PRDM2 (0)	
	1	195876600	41.904	PTPRC (19 kb), ATP6V1G3 (78 kb)	PTPRC encodes a leukocyte cell-surface molecule and contains suceptibility alleles for multiple sclerosis
	4	177391500	29.622	GPM6A (0)	GPM6A is a neuronal membrane glycoprotein
	5	29062440	59.662	LOC340211 (0)	
	6	66157130	59.88	EGFL11 (0)	
	8	4886706	40.618	CSMD1 (47 kb)	
	10	38121540	42.777	ZNF248 (0)	
	11	55171790	48.233	OR4P4 (9 kb)	Position estimate is within a cluster of olfactory receptor genes; six OR genes within 100 kb
	15	89572970	35.422	SV2B (4 kb)	SV2B is synaptic vesicle glycoprotein 2B, which is expressed primarily in the cerebral cortex
	20	20149280	43.999	C20orf26 (0)	
European-American	1	52897800	42.055	SCP2 (11 kb)	SCP2 plays a role in the intracellular movement of cholesterol
	2	158371000	41.014	KIAA1189 (0), PSCDBP (100 kb)	
	3	144901300	44.16	SLC9A9 (13 kb)	SLC9A9 is a sodium/hydrogen exchanger with a suggestive association with ADHD
	3	189987700	33.127	LPP (70 kb)	
	5	110427700	37.645	TSLP (55 kb), WDR36 (76 kb)	TSLP is part of a family of B cell–stimulating factors
	5	133570600	37.535	SKP1A (0), TCF7 (11 kb)	SKP1A is a transcription regulator with a suggested involvement with nervous/sensory development, especially the inner ear
	6	105777300	32.39	PREP (0)	
	7	136657300	35.646	DGKI (0)	Mutations in Drosophila DGKI causes degeneration of photoreceptor cells
	8	35614900	38.744	UNC5D (0)	
	10	21268430	32.164	NEBL (0)	NEBL encodes an actin-binding protein, and mutations in NEBL have been shown to cause nemaline myopathy, which causes several problems including decreased muscle density and problems with reflexes
	10	22739870	44.449	SPAG6 (29 kb), PIP5K2A (90 kb)	Mutations in mouse SPAG6 are known to cause sperm motility problems
	10	74357920	37.558	TTC18 (0), MRPS16 (1 kb)	
	11	36601700	33.082	LOC119710 (0), RAG2 (18 kb), RAG1 (37 kb)	
	12	42894650	47.363	DKFZp434K2435 (0)	
	12	99399670	37.529	NR1H4 (0), GAS2L3 (70 kb), SLC17A8 (82 kb)	NR1H4 is a nuclear hormone receptor relating to phenotypes of serum cholesterol, bile acid, lipoprotein, and triglycerides
	15	26994330	39.48	APBA2 (0)	The APBA2 protein binds the amyloid-beta (A4) precursor, and is a candidate gene for Alzheimer disease
	15	27655440	32.385	TJP1 (53 kb)	The tight-junction protein 1 (TJP1) associates with a protein (CagA) injected into gastric epithelial cells by H. pylori
	15	86739850	35.154	MRPS11 (0), MRPL46 (0), DET1 (45 kb)	
	17	59013260	32.4	APPBP2 (0)	The APPPBP2 protein binds the amyloid (beta-A4) precursor, and is a candidate gene for Alzheimer disease
	17	59681810	39.782	BCAS3 (0)	
	18	28723870	50.461	C18orf34 (46 kb)	
	18	30398320	51.283	DTNA (0)	DTNA is dystrobrevin-alpha, a component of the dystrophin protein complex
	18	44260350	39.481	KIAA0427 (57 kb)	
	18	64896900	44.055	C18orf14 (26 kb)	
	18	65739330	37.6	CD226 (0)	The CD226 protein is involved in T cell and natural killer cell cytotoxicity
	19	47672850	32.195	CEACAM1 (30 kb), UNQ473 (34 kb), LIPE (50 kb), CNFN (87 kb), SBP1 (98 kb)	
Chinese	1	57813740	33.199	DAB1 (0)	DAB1 plays a role in establishing the laminar organization of the cerebral cortex
	1	66817090	30.064	MI-ER1 (0), SLC35D1 (23 kb), FLJ23129 (56 kb)	
	1	103041700	40.208	COL11A1 (5 kb)	COL11A1 is a collagen associated with two disorders: (1) Stickler syndrome, which is characterized by progressive myopia and retinal detachment; and (2) Marshall's syndrome, which causes abnormalities in facial development
	1	158541900	33.019	SDHC (0), LOC257177 (9 kb), MPZ (45 kb)	SDHC is associated with hereditary paragangliomas, which involves nonmalignant tumors in vascular tissue
	2	109198300	40.035	EDAR (0)	EDAR is associated with ectodermal dysplasia, and it is involved in hair follicle, sweat gland, and tooth development

Table 1. Continued.

Sample	Chr.	CMLE Position[a]	CLR	Genes (Distance in kb)[b]	Notes
	2	189810100	54.195	DIRC1 (0)	
	2	216482300	29.141	FN1 (0), ATIC (65 kb)	
	3	17387700	43.978	TBC1D5 (0)	
	3	115642400	31.113	ZBTB20 (0)	
	3	144899200	38.179	SLC9A9 (11 kb)	See entry for SLC9A9 in the European-American sample
	4	6024760	47.629	FLJ46481 (0), CRMP1 (66 kb), MARLIN1 (95 kb)	
	4	13404330	31.993	FAM44A (23 kb)	
	4	41912200	57.385	SLC30A9 (0), TMEM33 (39 kb)	
	4	106988000	36.517	FLJ20184 (0), LOC57117 (74 kb)	
	5	42060400	33.69	FBXO4 (73 kb)	
	6	12902840	30.008	PHACTR1 (0)	
	6	26350950	44.027	HIST1H4F (2 kb)	Position estimate is in a large cluster of histone-1 genes, 20 of which are within 100 kb
	6	54864430	40.376	C6orf143 (10 kb),	
	6	158234200	36.574	SNX9 (0), SYNJ2 (78 kb)	SNX9 is an intracellular trafficking protein that regulates the degradation of ectodermal growth factor receptor
	7	100731700	51.119	EMID2 (0), MYLC2PL (85 kb)	EMID2 is a collagen expressed in the testis and ovary, and the protein is found in the extracellular matrix
	7	136674800	31.625	DGKI (0)	See entry for DGKI in the European-American sample
	8	50815690	38.22	SNTG1 (58 kb)	SNTG1 is a subunit of the dystrophin protein complex
	8	66983090	29.969	DNAJC5B (1 kb)	
	8	98234550	29.599	TSPYL5 (5 kb)	
	8	106772400	38.378	ZFPM2 (0)	ZFPM2 is a transcription factor with an important role in heart development
	8	136395000	36.856	KHDRBS3 (45 kb)	
	9	74370350	37.428	RFK (87 kb)	RFK plays a role in metabolizing riboflavin
	9	102273200	37.709	SMC2L1 (0)	SMC2L1 is involved in the maintenance and segregation of chromosomes during cell division
	10	22732610	37.798	SPAG6 (22 kb), PIP5K2A (97 kb)	See entry for SPAG6 in the European-American sample
	10	45409270	42.153	ANUBL1 (0), MARCH8 (35 kb), FAM21C (98 kb)	
	10	55292980	41.377	PCDH15 (0)	PCDH15 is involved in morphogeneisis of stereocilia in the inner ear
	10	81881400	42.407	TSPAN14 (0), C10orf58 (24 kb)	
	11	36610870	33.832	LOC119710 (0), RAG2 (27 kb), RAG1 (46 kb)	
	11	60688890	29.627	VPS37C (0), CD5 (18 kb), PGA5 (95 kb)	VPS37C is part of the endosomal sorting complex, which is recruited for viral budding
	12	24305690	29.76	SOX5 (0)	
	12	34031300	39.953	ALG10 (35 kb)	ALG10 is a regulator of potassium channels
	12	53770680	41.792	OR9K2 (38 kb)*, NEUROD4 (64 kb)	*Estimate is at the edge of a cluster of OR genes
	12	84651660	54.887	PAMCI (49 kb)	
	12	91589690	37.412	FLJ46688 (42 kb)	
	13	18052490	38.147	PSPC1 (0), HSMPP8 (47 kb)	
	14	21862100	29.019	MYH6 (0), MYH7 (10 kb), CKLFSF5 (23 kb), IL17E (27 kb), EFS (37 kb), SLC22A17 (51 kb), PABPN1 (77 kb), BCL2L2 (91 kb)	Both MYH6 and MYH7 have been associated with cardiac myopathy
	14	43313740	36.514	C14orf28 (42 kb), BTBD5 (74 kb)	
	14	75923480	33.061	AHSA1 (0), THSD3 (7 kb)	AHSA1 activates the heat shock protein hsp90, and is involved in stress response
	15	29051590	29.039	TRPM1 (0), MTMR10 (53 kb)	
	15	61878600	42.35	DAPK2 (36 kb), HERC1 (37 kb)	
	15	86742750	40.079	MRPS11 (0), MRPL46 (3 kb), DET1 (42 kb)	
	17	44236980	45.792	FLJ25168 (42 kb)	
	17	44710400	29.86	LOC284058 (0)	
	17	59681810	29.413	BCAS3 (0)	
	17	64527940	39.821	MGC33887 (0)	
	18	14001290	32.131	ZNF519 (93 kb)	
	18	28715730	45.289	C18orf34 (53 kb)	
	18	30406890	62.627	DTNA (0)	See entry for DTNA in the European-American sample
	18	44351560	32.265	KIAA0427 (0)	
	20	3532485	33.116	ATRN (0)	ATRN is homologous to the mouse mohogany gene, and it plays a role in several processes in mouse, including pigmentation, adaptive immunity, and obesity
	20	31004100	31.003	BCL2L1 (0), COX4I2 (26 kb), ID1 (65 kb), TPX2 (68 kb)	
	21	16307440	29.563	C21orf34 (57 kb)	
Combined	1	113016400	23.963	LRIG2 (50 kb)	
	1	154941000	24.15	FCRL2 (0), FCRL1 (40 kb), FCRL3 (54 kb)	CMLE for position in the middle of a cluster of FCRL genes, which are thought to play a role in B cell development
	1	211644800	47.46	PTPN14 (0)	

Table 1. Continued.

Sample	Chr.	CMLE Position[a]	CLR	Genes (Distance in kb)[b]	Notes
	2	141425500	44.172	*LRP1B* (0)	
	2	202042300	26.795	*MGC39518* (3 kb), *ORC2L* (12 kb), *NIF3L1* (72 kb), *PPIL3* (86 kb), *NDUFB3* (96 kb)	
	3	29922840	25.623	*RBMS3* (0)	
	3	43323910	27.861	*SNRK* (0), *FLJ10375* (44 kb)	
	3	144913600	23.908	*SLC9A9* (26 kb), *MGC33365* (93 kb)	See entry for *SLC9A9* in the European-American sample
	4	71991670	27.388	*IGJ* (0), *ENAM* (13 kb), *SAS10* (28 kb), *RIPX* (62 kb)	*IGJ* is an immunoglobulin with two known functions: linking immunoglobulin monomers and binding these immunoglobulins to secretory component
	4	169845700	24.098	*FLJ20035* (0)	
	5	15527500	41.987	*FBXL7* (26 kb)	
	6	128601800	33.418	*PTPRK* (0)	
	8	57052930	25.735	*RPS20* (16 kb), *MOS* (22 kb), *PLAG1* (70 kb), *LYN* (80 kb)	
	10	45462260	26.114	*ANUBL1* (10 kb), *FAM21C* (44 kb)	See entry for *ANUBL1* in the Chinese sample
	12	81503770	24.547	*DKFZp762A217* (79 kb)	
	13	36706830	29.695	*UFM1* (15 kb)	
	15	37567860	35.829	*THBS1* (21 kb), *FSIP1* (40 kb)	
	15	89573760	39.016	*SV2B* (5 kb)	See entry for *SV2B* in the African-American sample
	16	81827590	24.175	*HSPC105* (3 kb), *HSD17B2* (20 kb)	
	18	30386860	42.249	*DTNA* (0)	See entry for *DTNA* in the European-American sample
	18	44272270	25.806	*KIAA0427* (45 kb)	

Also shown are all known genes within 100 kb of the estimate of the position of the selective sweep. The 65 genomic regions which exhibited very strong evidence for a recent selective sweep that is more than 100 kb from a known gene are not shown.
[a]Physical map estimate of the location of the sweep for the window with the highest local test statistic.
[b]Lists all refseq genes within 100 kb of the estimate of sweep position.
doi:10.1371/journal.pgen.0030090.t001

position of the selective sweep is very accurate (to within ~20 kb in regions with typical recombination rates; see [22]), so the gene nearest the estimate of sweep position is generally the best candidate as the target of selection. However, we cannot rule out the possibility that unknown functional elements or, in very gene-dense or low-recombination regions, another nearby gene might be the true target of selection.

The genomic region with the strongest evidence for a recent selective sweep is in the *DTNA* gene on Chromosome 18; this location shows very strong evidence for selection in the Chinese, European-American, and combined samples. In the Chinese sample, the observed CLR statistic in this region is 62.63. In contrast, the highest CLR statistic for the Chinese population over 100,000 selectively neutral simulations is 24.34, and the 95th percentile of the simulated neutral datasets is 9.57. These simulations were performed with population bottleneck parameters that have been fit to human data [25] and with a recombination rate that is slightly less than that of the *DTNA* region. *DTNA* encodes the dystrobrevin protein, a component of the dystrophin protein complex (DPC). Aside from *DTNA*, several other genes that contribute to the DPC show evidence for recent selective sweeps (Table S3), including several syntrophin and sarcoglycan genes. The DPC primarily functions as a key structural component in the architecture of muscle tissue [26], suggesting that the selective sweeps at DPC genes may involve a muscle-related phenotype. Furthermore, several other muscle-related genes show very strong evidence for recent

selective sweeps, including *NEBL* and two tightly linked, cardiac-specific myosin heavy-chain genes (*MYH6* and *MYH7*).

One of the most conspicuous features of our genomic scan is that several centromeric regions have extreme spatial patterns of allele frequency consistent with recent selective sweeps. For instance, the region spanning the centromere of Chromosome 16 shows strong evidence of recent selection. The size of the affected area is remarkable: the combined, European-American, and Chinese samples exhibit skewed frequency spectra and very low p values by the CLR test over 16 Mb. Of the 17 autosomes for which we have data spanning the centromere, we observe evidence of selective sweeps in centromeric regions of Chromosomes 1, 3, 8, 11, 12, 16, 18, and 20 (Figure 1). Because the CLR test is not very sensitive to the underlying recombination rate [22], it is unlikely that this signal is an artifact of reduced recombination rates in centromeric regions. The large genomic distance over which the signature of selection extends in many of these regions complicates the identification of the selected target. However, the consistent signal of selective sweeps and the paucity of known genes in centromeric regions suggest the hypothesis that the centromeres themselves may be the functional genomic elements targeted by selection. One interesting possibility in this regard is that selection in centromeric regions may be the result of meiotic drive [27–29] (e.g., during female meiosis, any variant which even slightly decreases the probability that a chromosome segregates to a polar body will carry a huge selective advantage [30]). Also, centromeres are strong candidates for regions affecting chromosomal segregation.
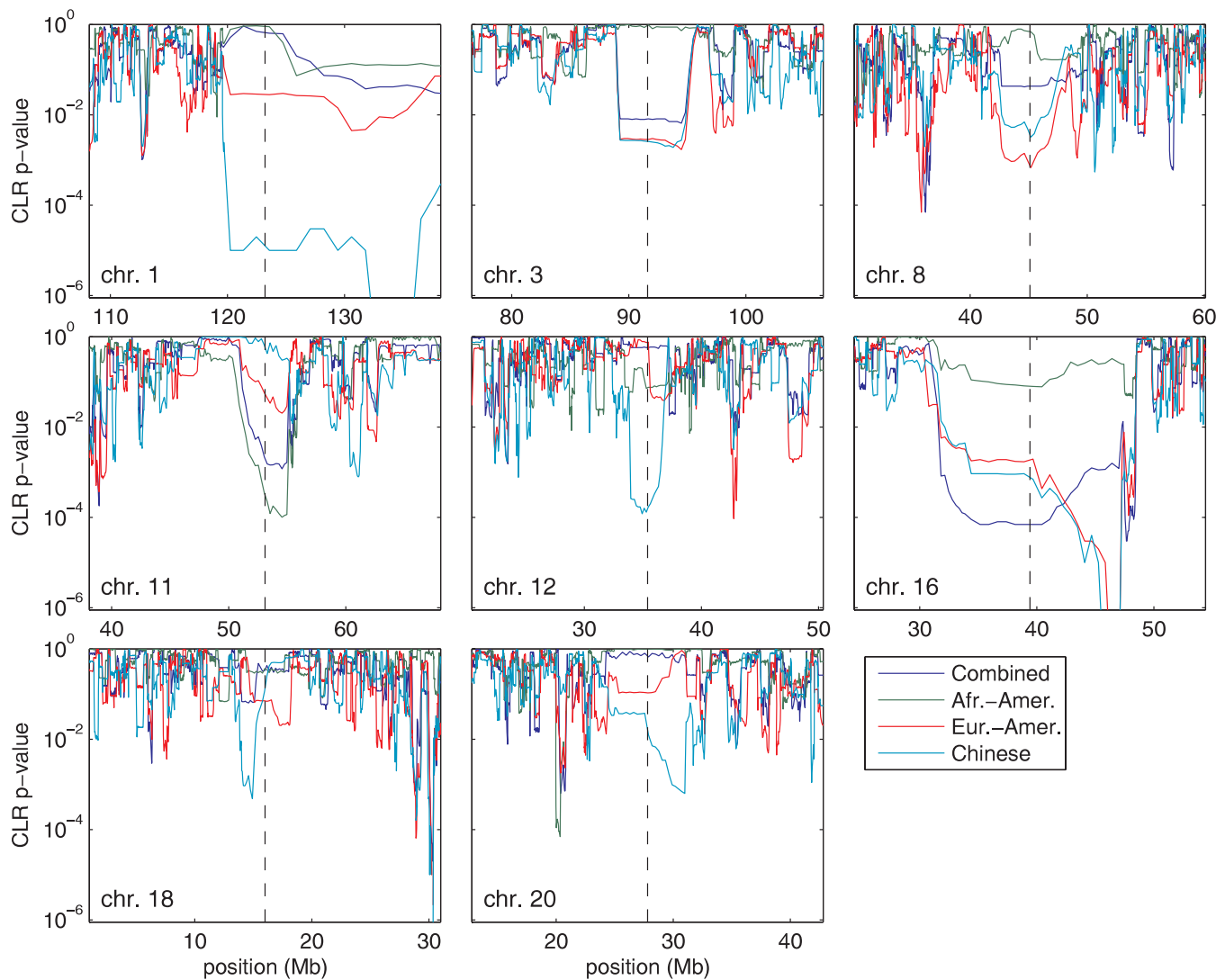
Figure 1. Evidence for Selective Sweeps in Centromeric Regions of Several Chromosomes, as Measured by the *p* Value of the CLR Test in Three Human Populations

Vertical dashed lines indicate the positions of the centromere, and *p* values are plotted on a log scale.

doi:10.1371/journal.pgen.0030090.g001

Because of the time scale in which the CLR test has power to detect a selective sweep (within the last ~200,000 y), it is useful for identifying selected changes that occurred in one or more populations since the time of population divergence (the continental populations represented by the samples probably diverged within the last 100,000 years). Such population-specific selective sweeps should be evident in our analysis as a high CLR statistic and low CLR *p* value in only one of the continental groups that was sampled. Along these lines, Jablonski and Chaplin [31] suggested that global variation in skin pigmentation is due to adaptation to local environments, noting that skin pigmentation in indigenous human populations correlates very strongly with the local average intensity of UV radiation. To investigate the role of local adaptation in shaping global patterns of human skin pigmentation, we interrogate pigmentation candidate genes (Table 2) for evidence of population-specific selective sweeps. *KITLG*, which encodes a signaling molecule that stimulates melanocyte proliferation, growth, and dendricity [32], shows

strong evidence for selective sweeps in the European-American and Chinese samples (Figure 2). Notably, the coding sequence of *KITLG* is 218 kb away from our estimate of the sweep position, whereas the next-nearest gene is 550 kb away, indicating that *KITLG* is the likely target of selection. Furthermore, the distance between our estimate of the sweep position and the *KITLG* coding sequence suggests the hypothesis that the selected mutation may be regulatory in nature. The presence of a selective sweep or sweeps at *KITLG*, along with experimental phenotypic effects of the gene, suggests that *KITLG* may be an important quantitative trait locus underlying variation in human skin pigmentation.

Other pigmentation candidate genes with strong evidence of population-specific selective sweeps include *RAB27A*, *MATP*, *MC2R*, *ATRN*, *TRPM1*, and *SLC24A5*. *SILV* and *OCA2* show marginally significant evidence for population-specific sweeps. Mouse orthologs of most of these genes carry coat color phenotypes, and *SLC24A5* was recently shown to contain a common mutation affecting skin pigmentation in

Table 2. Candidate Genes for Variation in Human Skin Pigmentation and Evidence of Population-Specific Selective Sweeps

| Gene	Chr	Position (Mb)	CLR p Value, African-American	CLR p Value, European-American	CLR p Value, Chinese
POMC	2	25.36	0.654 (0.433)	0.295	0.150
MITF	3	69.83	0.181	0.254 (0.182)	0.658 (0.627)
KIT	4	55.48	0.828 (0.813)	0.618	0.301
F2r11	5	76.21	0.808	0.870	0.933
MATP	5	34.01	0.976	**0.00014**	0.658
DTNBP1[a]	6	15.70	0.913 (0.416)	0.644 (0.599)	0.037
TYRP1[a]	9	12.69	0.652	0.326	0.421
TYR	11	88.66	0.746 (0.725)	0.145 (0.117)	0.221 (0.209)
SILV	12	54.64	0.092	0.050	**0.007**
KITLG	12	87.44	**0.014**	**0.000007**	**0.00002**
DCT	13	92.81	0.812 (0.796)	0.335	0.305
OCA2 [a]	15	25.77	0.400 **(0.046)**	0.140 (0.055)	**0.020 (0.0023)**
TRPM1	15	29.04	0.992	0.707 (0.689)	**0.00004 (0.00002)**
SLC24A5[a]	15	46.14	0.287	**0.0008**	0.868
MYO5A[a]	15	50.43	0.382	0.492 (0.454)	0.398
RAB27A	15	53.23	0.885 (0.814)	**0.0025**	**0.00020**
MC1R	16	89.73	0.274	0.556	0.405
MC2R	18	13.88	0.839	0.125	**0.0005**
ATRN	20	35.19	0.613	0.608 (0.582)	**0.00020 (0.00006)**
ASIP	20	33.57	0.518	0.749	0.375

Reported p values are from the genomic window with a midpoint nearest the midpoint of the gene.
Values in parentheses indicate the minimum p value of windows with a center between the start and stop codon of the gene, which is reported only if it is different from the midpoint p value. Bold typeface indicates p values with nominal significance below 5%.
[a]Genes previously identified as experiencing partial selective sweeps in the European population [15].
doi:10.1371/journal.pgen.0030090.t002

humans [33]. Considered as a whole, pigmentation candidate genes are enriched for significant CLR tests. For instance, in the genome scan of the Chinese sample, pigmentation genes contain more than twice as many significant CLR tests (at the $p < 0.01$ level) compared with the expectation from the rest of the genome; this enrichment is marginally significant ($\chi^2_{(1)} = 6.04$, $p = 0.007$). Using a more stringent significance level for the CLR test, the enrichment of pigmentation genes becomes more pronounced (i.e., at the $p < 0.001$ level), and pigmentation genes are more than 5-fold enriched for significant tests, compared with the genomic expectation ($\chi^2_{(1)} = 17.3948$, $p = 1.5 \times 10^{-5}$). A similar pattern emerges in the European-American sample: at the CLR $p < 0.01$ level, we observe twice as many significant pigmentation genes as expected ($\chi^2_{(1)} = 2.6297$, $p = 0.052$), and at the $p < 0.001$ level, we observe a nearly 5-fold enrichment ($\chi^2_{(1)} = 9.057$, $p = 0.0013$). In a similar analysis, Voight and coworkers [15] identified a signal of partial selective sweeps in the European population for *OCA2, MYO5A, DTNBP1, TYRP1,* and *SLC24A5,* all of which are pigmentation candidate genes. Likewise, Izagirre and coworkers [34] found evidence of a partial selective sweep at *TP53B1* and *RAD50* in African populations, and at *TYRP1* and *SLC24A5* in European populations. A partial sweep occurs when the beneficial mutation has not spread throughout the entire population, whereas the CLR test is designed to detect beneficial mutations that have recently reached a frequency of 100% (complete sweeps). Thus, the two analyses should be complementary, and there is little overlap between the analyses in terms of which pigmentation genes are identified as selected in which populations. Taken together, these results indicate that population-specific selective sweeps, both partial and complete, have been common in genes in skin pigmentation

pathways, suggesting that adaptation to local environments has driven the evolution of human skin pigmentation.

Several other gene categories and pathways show a striking pattern of recent adaptation. For instance, we observe evidence for a selective sweep mainly in the African-American sample in a region surrounding a cluster of olfactory receptor *(OR)* genes on Chromosome 11. Recent adaptive evolution appears to be a pervasive force among *OR* genes. Among 29 autosomal clusters of *OR* genes, 16 clusters show evidence of a selective sweep (CLR $p < 0.05$) in at least one of the populations. These findings corroborate work on adaptation in *OR* genes [35], and suggest that many changes in the human olfactory repertoire may have occurred very recently. Similarly, candidate genes for hair morphology show consistent signals of recent adaptation. Keratin-associated proteins (KRTAPs) are thought to play an important role in the shape of hair follicles, and we observe evidence for recent adaptation at four out of five clusters of *KRTAP* genes, mostly in the European-American sample. Perhaps the most surprising category of genes that show consistent evidence of recent adaptation is heat shock proteins (Table S4). Among 56 unlinked heat shock genes, 28 showed evidence of a recent selective sweep in at least one population at the $p < 0.05$ level. Several genes with functional roles in the development and function of the nervous system show very strong evidence (CLR $p < 10^{-5}$) for a recent selective sweep. For example, *SV2B*, a gene encoding a synaptic vesicle protein with highest expression during brain development [36], exhibits strong evidence for a selective sweep in the African-American sample. Likewise, the protein encoded by *DAB1* plays a developmental role in the layering of neurons in the cerebral cortex and cerebellum [37], and exhibits strong evidence for a selective sweep in the Asian sample. Other nervous system genes with strong evidence for a selective sweep include two
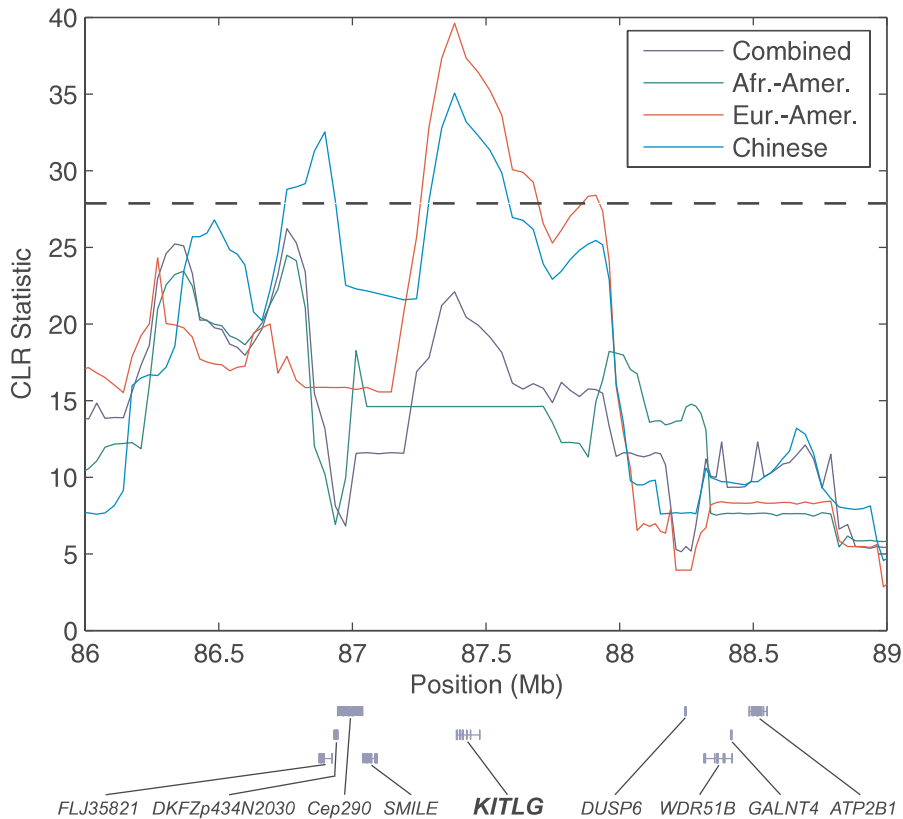
Figure 2. Sliding Window Analysis of the *KITLG* Region of Chromosome 12, Along with Gene Models of All refseq Genes in the Region
The horizontal dashed line represents the *p* < 0.001 critical value of the population-specific CLR tests generated using a conservative estimate of the average recombination rate in the region.
doi:10.1371/journal.pgen.0030090.g002

candidate genes for Alzheimer disease (*APPBP2* and *APBA2*) that bind the amyloid-beta precursor protein, two genes (*SKP1A* and *PCDH15*) with a role in sensory development, and several others with various roles in nervous system development and function (*PHACTR1, ALG10, PREP, GPM6A,* and *DGKI*).

Several analyses (e.g., [3–5]) suggest genes that play a role in immunity and pathogen response are among the most common targets of adaptive evolution. Consistent with these results, we observe very strong evidence of recent adaptation (CLR $p < 10^{-5}$) within or very close to several immune system genes. These include: (1) two genes thought to play a role B-cell development (*FCRL2* and *TSLP*); (2) two somatic recombination-activating genes (*RAG1* and *RAG2),* which help generate the diversity of immunoglobulins and T cell receptors; (3) *CD226,* a *trans*-membrane protein involved in the cytotoxicity of natural killer cells and T cells; and (4) *IGJ,* an immunoglobulin responsible for linking other immunoglobulins to each other and to the secretory component. In addition, two genes that are not part of the immune system, but which might play an important role in pathogen interactions, also show very strong evidence of a recent sweep; these are *TJP1* and *VPS37C.* The TJP1 protein associates with the CagA protein [38], which is translocated into gastric epithelial cells by the human pathogen *Helicobacter pylori.* The *TJP1*–CagA interaction is thought to play a role in the pathogenicity of *H. pylori,* and the selective sweep in the *TJP1* region suggests the hypothesis that the selected variation may have affected the pathogenic effects of *H. pylori* infection. The VPS37C protein is a subunit of the endosomal sorting

complex, which is recruited by HIV and other viruses to promote viral budding from infected cells [39].

Several loci in the human genome have been previously identified as targets of recent adaptive evolution. Because these loci were identified using independent data and different statistical methods, they are to some extent positive controls (i.e., if selection is truly operating in these regions and if the CLR test has sufficient power, then we should observe evidence for selective sweeps at many of these loci using our approach). One such locus is the *LCT* gene on Chromosome 2. Numerous studies have identified evidence for one or more functional polymorphsims in *LCT* that affect lactose metabolism in adults [40,41], and Bersaglieri and coworkers [42] found that very recent positive selection in European populations has strongly affected the frequency of this polymorphism. Concordantly, we observe evidence for a selective sweep in the European-American sample (CLR p = 0.012), but not the other samples. Notably, the proposed beneficial mutation in *LCT,* the lactase persistence allele, is not completely fixed in European populations; rather, its frequency is 77% [42]. Even though the CLR test considers a model of a complete selective sweep in which the beneficial allele reaches a frequency of 100%, the significant result at *LCT* suggests that the CLR test has at least some power to detect recent adaptive events that deviate from the assumptions of the complete sweep model. The *HFE* gene on Chromosome 6 is another locus for which previous work suggests a selective sweep [43]. For the genomic window centered on *HFE,* we find significant evidence for a selective sweep in the vicinity of *HFE* in the Chinese (p = 0.00006),

European-American ($p = 0.002$), and combined ($p = 0.0006$) samples. *HFE* contains a relatively high-frequency recessive mutation, C282Y, which causes hereditary hemochromatosis [44], an iron-overload disorder. Although positive selection is thought to operate somewhere in the vicinity of *HFE*, it is unknown whether the C282Y mutation attained high frequency through selection directly (positive selection on C282Y itself) or indirectly (positive selection on a nearby beneficial mutation associated with C282Y). Our composite likelihood estimate of the position of the selective sweep is within a cluster of histone genes, 150 kb away from *HFE*, suggesting that C282Y may have attained high frequency through association with a nearby beneficial allele. If this hypothesis of C282Y rising to high frequency indirectly is correct, then it carries the interesting implication that populations experiencing selective sweeps may sometimes incur indirect costs: occasionally, selective sweeps may carry tightly linked, initially rare, deleterious, and potentially disease-causing variation to relatively high frequencies [45]. Essentially, a recent selective sweep may have a localized effect in the genome similar to a population bottleneck (i.e., a sweep is somewhat analogous to a genomically localized reduction in effective population size), and deleterious disease alleles in these regions may obtain observable frequency by chance in this situation. Other regions where previous research has suggested positive selection, and the signal is confirmed by our analysis, include the cluster of *ADH* genes on Chromosome 4 [46], which show evidence for a recent sweep only in the Chinese sample (CLR $p = 0.00015$), and the opioid receptor *PDYN* [47], which also shows evidence of a selective sweep only in the Chinese sample (CLR $p = 0.002$). Loci that have been previously identified as targets of recent or ongoing selective sweeps, but do not show evidence for a selective sweep in the present analysis, include *MMP3* [48], *CD40LG* [8], *CCR5* [7], *ASPM* [49], and *MCPH* [50]. Like *LCT*, previous work indicates a partial selective sweep at these loci, and in all of the above cases, the frequency of the putatively beneficial allele is relatively low (between 10% and 70%). Because these loci are thought to deviate more strongly from the complete sweep model, the CLR test probably does not have adequate power to detect selection at these loci.

Another means of validation for our genomic scan is to compare the spatial distribution of evidence for selection along chromosomes with the distribution of known functional elements in the genome (i.e., if a large proportion of positive tests are false positives, then one would not expect positive tests to be associated with functional elements). For example, Voight and coworkers [15] found that genic regions of chromosomes are strongly enriched for extreme values of the integrated extended haplotype homozygosity statistic, an observation that is not readily explainable by factors that can cause a false signal of selection, such as demography or ascertainment bias. Using a similar approach, we tested regions surrounding known genes for an enrichment of significant CLR tests. We used a contingency table approach to test for enrichment (i.e., we compared the proportion of significant tests in windows nearest the center of known genes to the proportion of significant tests in the remainder of the genome). The results of these analyses are given in Table S5. Notably, in the European-American and Chinese samples, we observe a strong excess of significant tests in genic regions, and this signal becomes stronger as the significance level

applied to the CLR test becomes more stringent. For example, in the European-American sample at a significance level of $p < 0.001$, we observe 40% more significant tests than expected at gene centers, based on the total number of significant tests and the total number of windows at gene centers. Because centromeric regions have strong evidence of selection and low gene density, this signal becomes even stronger if centromeric regions are excluded. We conclude, therefore, that extreme values of the CLR statistic are strongly associated with genic regions of chromosomes, and this association has two important implications. First, it further corroborates the results of our genomic scan for selective sweeps, as this association is not predicted if a high proportion of significant tests are false positives. Second, the association between genes and selection in this paper and in the Voight et al. [15] study suggests that the empirical follow-up to genomic scans for selection will be at least somewhat experimentally tractable. Identifying beneficial mutations and determining their phenotypic effects will be much easier if the beneficial mutation is within a known gene.

Another interesting comparison is the contrast between our analysis and previously published genomic scans for selective sweeps. This comparison does not necessarily provide a means of validating ours or previous analyses, as the statistics used in the different genomic scans may be correlated even under selective neutrality, and the statistics have power to detect different types of selective sweeps. However, the comparison does provide a general sense of the consistency of population genetic methods for identifying selective sweeps from genomic variation data. Table S6 gives the CLR statistics and p values for the most extreme regions of the genome identified in [16] using two different approaches: population differentiation (Table 9 in [16]) and extended haplotype homozygosity [8] (Table 10 in [16]). In the Chinese sample, genes containing nonsynonymous SNPs that exhibit high levels of population differentiation in the Hapmap data [16] are enriched for CLR tests significant at the $p < 0.01$ level ($\chi^2_{(1)} = 10.6$; $p = 0.0011$). Similarly, genomic regions with the most extreme patterns of extended haplotype homozygosity in the Hapmap data [16] also have more significant CLR tests than would be expected if the two statistics were statistically independent. However, even among the most extreme regions of the genome in the Hapmap analysis, the CLR analysis does not always show evidence of a selective sweep. This inconsistency is likely the result of differential power of the alternative approaches in detecting different types of selection. For example, considering that extended haplotype approaches [8] have the most power to detect partial selective sweeps [15], it would not be surprising if the most extreme regions of the genome by these approaches were the result of a partial sweep. Furthermore, the CLR approach probably has limited power to detect this type of selection because it does not leave a population genetic signature similar to that of a complete sweep. In conclusion, it is encouraging that the CLR test is not independent of other statistics, which suggests some consistency among genomic scans for selective sweeps. However, it is also encouraging that the CLR test is not completely correlated with other approaches; if it were, then we would not have uncovered any previously unknown selective sweeps in this analysis.

In addition to the statistical exploration of the CLR test by
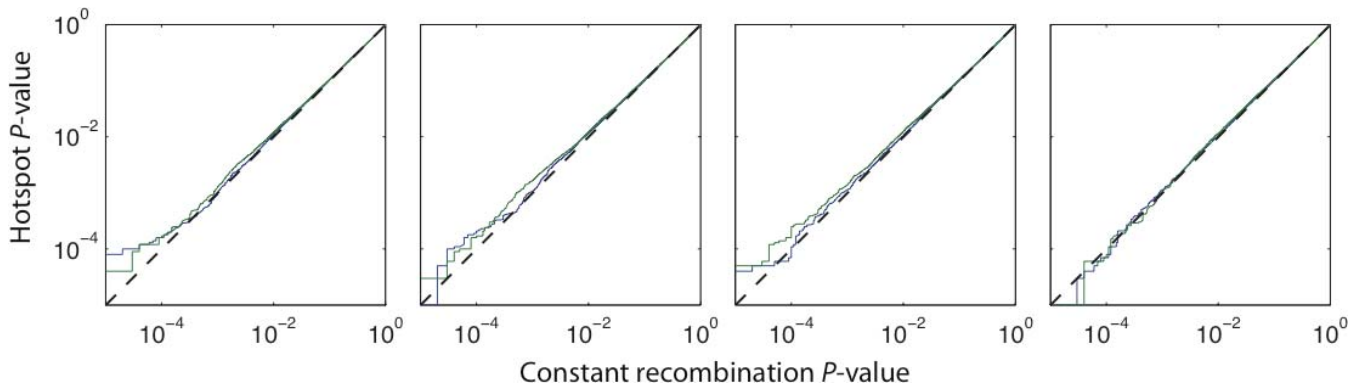
Figure 3. A Comparison of *p* Values of the CLR Test, Calculated from Simulations of Models Assuming a Constant Recombination Rate and Models That Include Recombination Hotspots

(A) The combined sample.
(B) The African-American sample.
(C) The European-American sample.
(D) The Chinese sample.
p Values are highly consistent between constant recombination and hotspot models, indicating that the CLR test is robust to recombination rate variation. Note that both axes are on a log scale.
doi:10.1371/journal.pgen.0030090.g003

Nielsen et al. [22], we performed extensive neutral simulations to determine how robust the CLR approach is to both recombination rate variation and complex demography. Recent work suggests that recombination rate variation is a pervasive feature of the human genome, and most recombination events occur in recombination hotspots [51,52]. To investigate how recombination rate variation might affect our analysis, we performed coalescent simulations with recombination hotspots, as well as SNP ascertainment, missing data, and different demographic scenarios. Recombination hotspots were represented as randomly spaced 5 kb fragments with an average distance between hotspots of 50 kb, and within the hotspot, the recombination rate was assumed to be 8-fold higher than the background rate. Figure 3 shows a comparison of *p* values calculated from a constant recombination model and a hotspot model with an equal average recombination rate. Recombination rate variation appears to have no effect on the null distribution of the CLR statistic, and *p* values calculated under the hotspot and constant recombination models are strikingly consistent. We observe some minor differences in *p* values calculated for very extreme test statistics ($p < 10^{-4}$), but these differences are readily explainable by Monte Carlo error in the estimation of *p* values via simulation.

We also performed simulations under a variety of demographic models beyond those considered by Nielsen et al. [22] in order to more fully explore the robustness of the CLR test to complex population demography. In particular, we investigated how the strength of the population bottlenecks experienced by non-African populations affects the null distribution of the CLR statistic. We simulated data under population bottlenecks with a constant duration and varying severity, with the temporary reduction in population size ranging from 50% to 99% only for non-African populations. Surprisingly, the null distribution of the CLR statistic is shifted toward lower values under the strong bottleneck model (99% reduction) compared with the equilibrium model (Figure 4), and the variance in the CLR statistic is much lower. This result indicates that, if the strong bottleneck model accurately reflects history, but we use the equilibrium model (random mating, constant population size)

to obtain *p* values of the CLR test, our results will be strongly conservative. These surprising results for the strong bottleneck model can be explained by a coalescent argument: with a strong and recent bottleneck, the vast majority of the coalescences and the most recent common ancestor of the sample typically occur during the bottleneck, which reduces the stochasticity due to the ancestral process. This reduced stochasticity results in less variation in the site-frequency spectrum (SFS) across the genome and, consequently, less extreme CLR statistics. Under a weak bottleneck (50% reduction), the null distribution of the CLR statistic is nearly unaffected. Intermediate-strength bottlenecks (90%–95% reduction) cause the most problems: compared with the equilibrium model, the CLR statistic shows slightly more variation under intermediate bottlenecks, and the upper tail of the null distribution is slightly heavier. Similar to the case of an intermediate bottleneck model, the complex model approximated by Schaffner et al. [53] results in slightly more variation in the CLR statistic with a heavier upper tail. Therefore, the equilibrium neutral model will be somewhat anticonservative when applied to a population that has experienced an intermediate bottleneck or multiple weak bottlenecks, as in the case of the Schaffner et al. [52] model. However, compared with the effect of demography on standard methods for detecting selection, the CLR approach is very robust to even the most extreme demographic effects. The robustness of the CLR approach to demographic effects is reflected in the general consistency of *p* values obtained under alternative demographic models (Figure S1).

False discovery rate (FDR) methods [23,24] use the distribution of *p* values among tests to correct for multiple hypothesis testing, providing an estimate of the probability that the null hypothesis is true for any particular test (the *q* value). The distribution of *p* values for the different windows is shown in Figure 5. In the Chinese and European-American samples, the distribution shows a strong excess of tests with very low *p* values from the CLR test, suggesting that the null hypothesis is false for many of these windows. In addition to correcting for multiple testing, FDR methods estimate the number of tests in which the null hypothesis is false (m_1). In the case of genomic scans for natural selection, m_1 is itself a
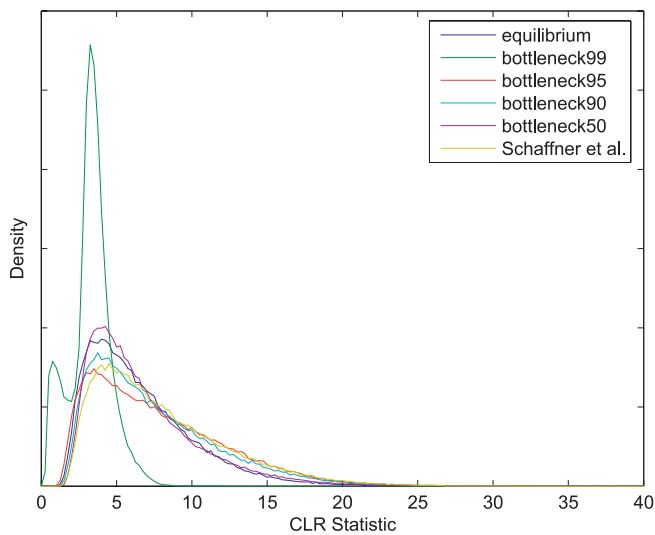
Figure 4. The Null Distribution of the CLR Statistic in a Non-African Population for Non-African Bottleneck Models of Varying Strength, As Well As the Complex Schaffner Model

doi:10.1371/journal.pgen.0030090.g004

parameter of interest, because it provides a rough indication of what proportion of the genome is affected by selective sweeps at linked sites. FDR estimates of the proportion of tests where the null hypothesis is false (m_1/m) is shown in Figure 6, using several alternative demographic models to obtain p values. All alternative models indicate that recent selective sweeps have been a pervasive force in the human genome, with ~10% of the genome affected by selective sweeps in the European-American and Chinese samples, ~1% in the African-American sample, and ~5% in the combined sample.

The FDR estimates of m_1 suggest that recent adaptation has had a strong effect on genome-wide patterns of nucleotide variation, to the point that a considerable fraction of the genome is evolving nonneutrally. However, this conclusion should be considered preliminary: m_1 is a very rough measure of the pervasiveness of selective sweeps, and estimates of the proportion of the genome affected by a sweep will of course depend strongly on what is meant by "affected." In our case, this means that selection has altered patterns of variation in the window sufficiently to drive the p value of the CLR test below ~0.05. The ability of selection to alter variation in a window will depend very much on the strength of selection, the genomic distance away from the beneficial mutation, the age of the selective event, and the type of selection. Fully describing the genomic effects of linked selection and estimating the number of selective events will require fitting a model of multiple selective events to the entire genome (perhaps including complete selective sweeps of varying age, different types of balancing selection, partial selective sweeps, and "soft" sweeps starting from standing variation), rather than fitting a model of a single selective sweep to a small window of the genome for a number of different windows. The primary utility of the present analysis lies in the fine-scale identification of individual loci that have experienced selection, which greatly facilitates the investigation of what human phenotypes have been affected by adaptation, and

what forces in the environment have driven recent human evolution.

Discussion

Here we have presented a comprehensive scan for selective sweeps across the human genome. Several general patterns emerge from the analysis. We find much more evidence for selective sweeps in Chinese and European-American populations than in the African-American population. This result is consistent with the hypothesis that, as anatomically modern humans migrated out of sub-Saharan Africa, the novel environments they encountered imposed new selective pressures, which in turn led to an increased rate of population-specific selective sweeps [54–56]. However, a caveat should be considered when interpreting the differences between African-American and non-African populations: the statistical power to detect selective sweeps is likely to be much lower in the African-American sample. Because the CLR test is based on a complete sweep model, the recent admixture of African and European lineages in the African-American population probably weakens the signal of Africa-specific selective sweeps. If a complete selective sweep occurred in African populations after the divergence of European populations, then the beneficial allele, and corresponding haplotypes, would not be fixed in the African-American sample. In other words, admixture is expected to fundamentally alter the molecular signature of a selective sweep, and it is therefore unsurprising that our results for the African-American sample are distinctly different from those of the European-American and Chinese samples. Another factor to consider is the extensive subdivision among African populations [57]. Subdivision within Africa may have allowed, or may have been driven by, adaptation to local environments within Africa. This sort of selection may not be evident in the African-American sample, which represents a nonrandom, continent-wide sampling of African lineages with some admixture of European lineages [58]. Subdivision within Africa may add further complications to the effect of admixture on the power of the CLR test (i.e., perhaps the proper demographic history of the African-American population includes the admixture of several diverged African populations), followed by large-scale (20%, from [59]) admixture with European populations. For example, in this demographic scenario, if a selective sweep occurred within Africa in a source population for the African-American population, the molecular signature of this sweep would be obscured by the admixture among African populations during the founding of the African-American population, and the signature would further be eroded by subsequent admixture with the European population. Considering that numerous factors suggest that selective sweeps will be much more difficult to detect in the African-American sample, compared with the non-African populations, it is premature to conclude that the rate of adaptation has increased in non-African populations.

Another general pattern that emerges from our analysis is that we observe more evidence for selective sweeps within subpopulations, compared with the cosmopolitan sample. This result suggests that adaptation to local environments has been an important force in recent human evolution. The relevance of local adaptation might be predicted considering
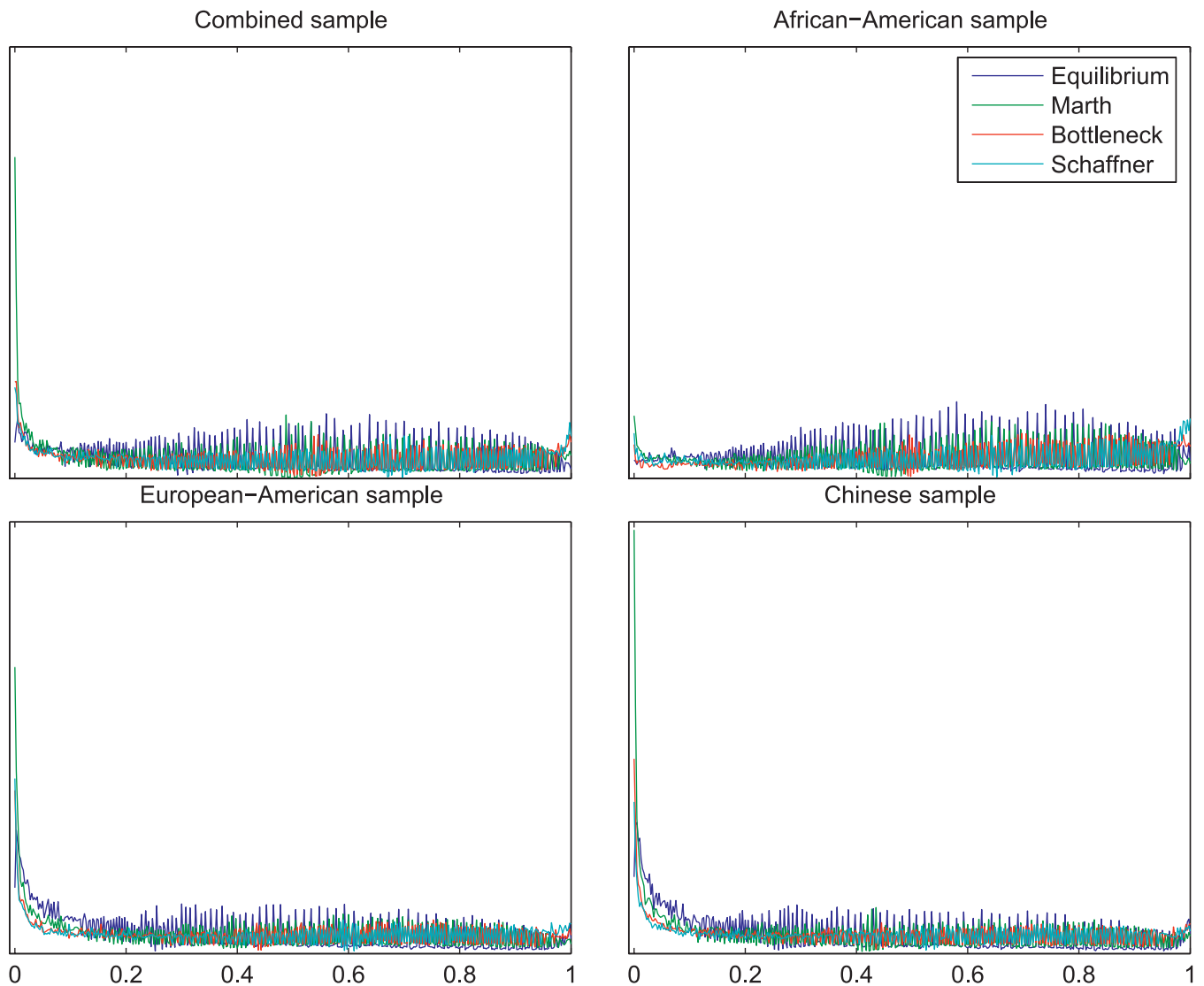
Figure 5. The Distribution of *p* Values for the CLR Test of a Selective Sweep
doi:10.1371/journal.pgen.0030090.g005

the extensive range expansions in recent human history, and the tremendous diversity of environments inhabited by indigenous human populations. However, the notable discrepancy between local and cosmopolitan sweeps is also difficult to interpret due to potential differences in the statistical power to detect different types of selective events. For example, if the power to detect sweeps were much greater in the local samples compared with the cosmopolitan sample, then one would expect to observe results similar to ours, even if the true number of local and cosmopolitan sweeps were equal. Fully evaluating the relative importance of localized and worldwide selective sweeps will require a detailed study of the statistical power to detect these types of sweeps under reasonable models of human demographic history.

In order to correct for the confounding effects of demographic history, we use a test [22] that compares allele frequencies in regions of the genome to the background pattern of variation. Simulations of a number of demographic models indicate that the methods are fairly robust to a wide variety of demographic histories; therefore, complex

demography should not increase the rate of false positives, but we cannot rule out the possibility that some complicated demographic scenarios could lead to an aberrant signal of selection. Even so, if selective sweeps have affected some regions of the human genome, we feel that the regions that we have identified with extreme frequency spectra are the best candidates for future studies. Another alternative explanation of the results of the CLR test is that weak negative selection operating on the SNPs themselves could locally skew allele frequencies toward rare alleles in a manner that could mimic a selective sweep. Although we cannot rule out this explanation, several factors suggest that localized weak selection does not have a systematic effect on our results. First, the vast majority of SNPs are in genomic regions with no known function (99.2% are noncoding). Second, in most of the regions where we identify selective sweeps, the sweep is population-specific, an observation that is difficult to explain with weak negative selection. And third, we observe greater evidence for selective sweeps in non-African populations than in the African-American sample. If weak
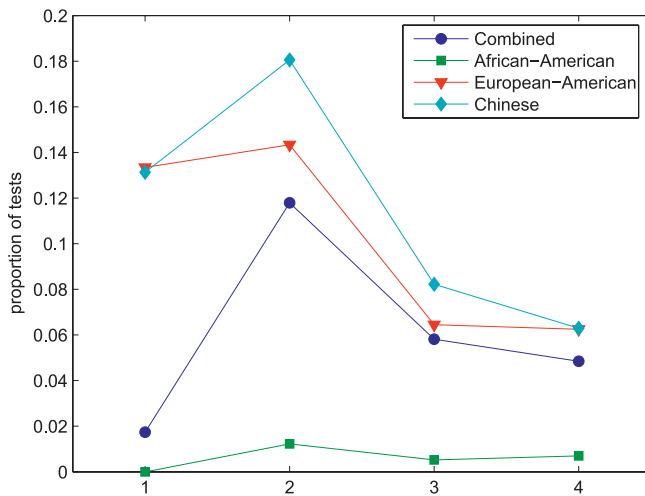
Figure 6. The Fraction of Tests for Which the Null Hypothesis Is False, Estimated Using a FDR Procedure and Shown for Four Alternative Evolutionary Models

(1) The equilibrium, random mating, neutral model. (2) The Marth et al. [25] bottleneck and growth model. (3) The most conservative non-African bottleneck model. (4) The complex demographic and recombination model calibrated by Schaffner et al. [50].
doi:10.1371/journal.pgen.0030090.g006

negative selection were the root cause for these deviations from neutrality, then one would expect a greater signal in the African-American sample because of the larger effective population size in African populations.

The approach we have taken here—detecting complete selective sweeps by their effects on variation at linked sites—is complementary to previous divergence-based approaches [1–5] characterizing adaptive evolution across the human genome. For instance, divergence-based approaches have been limited to detecting adaptive changes that have occurred via recurrent amino acid substitutions within a gene, whereas the present approach is capable of detecting adaptive changes at all functional genomic categories. The two approaches also differ in the time scale over which selection is detectable. Divergence-based approaches detect molecular adaptation that has occurred at any time on the lineage separating humans and chimps. Linked selection approaches, in contrast, are time-specific, detecting ongoing or very recent (within the last ~200,000 years) selection. Linked selection approaches are also much more amenable to investigating the adaptation of subpopulations to local environments at the molecular level. Given the complementary nature of divergence-based and linked selection methods, the present analysis fills in some of the gaps in our knowledge of human adaptive evolution. The challenge now is to use information about the genomic location of selective sweeps, in combination with the tools of functional genomics and knowledge of human ecology, to identify the traits that have been affected by recent adaptation and the selective forces that have shaped human populations.

Materials and Methods

Statistics. To correct for the confounding effect of demography, the CLR test of a selective sweep compares the SFS of a small region of the genome (a "window") to the SFS of the rest of the genome. The CLR test calculates the composite likelihood of the data in a window for two models: (1) a model which predicts the probability of SNP

frequencies using the genomic background SFS; and (2) a model of a very recent selective sweep. The composite likelihood in the sweep model is independent of demography because the SNP frequencies among lineages that were present before the sweep are predicted using the genomic background SFS. In essence, the CLR test works by considering the spatial pattern of allele frequencies along the genomic sequence, as predicted by a selective sweep model given the background pattern of variation. In an investigation of the statistical properties of methods for detecting selective sweeps, Nielsen et al. [22] demonstrate that, among several statistical tests for detecting selective sweeps, the CLR test is the most powerful and is the most robust to demography and the underlying recombination rate. The CLR test can be applied to either the SFS of the entire sample or to population-specific subsets of the data, enabling the detection of geographically restricted selective sweeps and balancing selection. For population-specific tests, we incorporate SNPs that are variable in the combined sample, but invariable within the subpopulation (i.e., the SFS describes the number of SNPs with minor allele counts of $I = 0, 1, 2 \ldots n/2$). The inclusion of invariable SNPs may significantly increase power to detect selective sweeps because, if a population-specific sweep has occurred recently, then one expects a strong excess of invariable SNPs within the population. By using SNPs that are invariable within a subpopulation, but variable in the combined sample, our methods should be robust to mutation rate heterogeneity across the genome, which would not be true if we included all invariable sites. A full description of the tests and an exploration of their statistical properties can be found in Nielsen et al. [22].

Because allele frequencies of linked SNPs are not statistically independent, we determine the null (selectively neutral) distributions of all test statistics using coalescent simulations [60]. For data analysis, we define genomic windows based on the number of SNPs in the window; therefore, we condition on an equal number of SNPs being present in our simulated datasets. Defining windows based on the number of SNPs makes the procedure robust to both mutation rate heterogeneity and the increased variance in regional nucleotide diversity caused by nonstandard demographies such as bottlenecks (K. Thornton, personal communication). To address the effect of SNP ascertainment, we incorporate the ascertainment scheme into our simulations by simulating the genealogy of both the genotyping sample and the sample in which the SNP was discovered, and keeping only those SNPs that are variable in the discovery sample. For each SNP, the discovery sample size was determined by a random draw from the empirical distribution of discovery sample sizes, which was provided by Perlegen Sciences (http://www.perlegen.com). We incorporate ascertainment into the simulations, rather than applying an explicit ascertainment correction [61,62], because the cosmopolitan discovery sample is computationally expensive to correct for in population-specific genotyping samples. The Monte Carlo approach to correcting for SNP ascertainment is greatly simplified by the uniform SNP discovery protocol used by Perlegen; for datasets with variable SNP ascertainment, such as the hapmap SNPs [16], it would be necessary to also model the autocorrelation of ascertainment along the chromosomes. Each iteration consisted of simulating a sample with a fixed number of ascertained SNPs, dividing the sample into African-American, European-American, and Chinese samples, then calculating the combined and population-specific CLR statistics. This procedure was repeated 10^5 times. Nielsen et al. [22] found that, among a variety of demographic models that have been fitted to human data, the equilibrium neutral model (random mating, constant population size) provides the most conservative critical values for the CLR test; therefore, all reported p values are from simulations of the standard neutral model. Finally, we incorporate SNPs with missing data by calculating the tests using SNP allele frequencies from a subsample of the data, summing over all possible allele frequencies in the subsample [25,62]. For the population-specific tests, the subsample size was set to $n = 44$ chromosomes, and for the combined test, it was set to $n = 132$. SNPs that did not have at least 44 chromosomes successfully genotyped in the African-American, European-American, and Chinese samples were excluded from further analysis. The correction for missing data was incorporated into the simulations of the CLR null distribution, and data was missing in the simulated data sets by randomly drawing the sample size for each SNP according to the empirical distribution of sample sizes.

The CLR statistic is weakly dependent on the underlying recombination rate: the test becomes somewhat more conservative if the assumed recombination rate is less than the true rate, and slightly anticonservative if the assumed rate is greater than the true rate. It is necessary to account for this weak dependence because: (1)

recombination rates are known to vary considerably across the genome [63]; and (2) we base the size of our genomic windows on a fixed number of contiguous SNPs, so that the size of the window in base pairs will vary with SNP density. To address these issues, we estimate the recombination rate for each window of the genome based on the size of the window and genetic map estimates [63] of the local recombination rate. Then, to make the tests more conservative, we downwardly bias our estimates by a factor of five. We have simulated the null distributions of all test statistics for regional recombination rates of $r = 0$, 10^{-5}, 3×10^{-5}, 10^{-4}, 3×10^{-4}, and 10^{-3}. To estimate the p value for each genomic window, we use our downwardly biased estimates of r to interpolate between p values calculated from the simulated null distributions with different r.

To account for multiple hypothesis testing, we apply FDR methods [23] that are specifically designed for genomic analyses [24]. FDR methods use the distribution of p values to estimate the number of tests in which the null hypothesis is false (m_1), and the probability that the null hypothesis is true for any particular test (the q value). One modification to the approach outlined by Storey and Tishirani [24] is the method we use for selecting the tuning parameter, λ. First, we represent the distribution of p values using a histogram of 500 bins. Next, we smooth the distribution by calculating the average density of the distribution in a window surrounding a particular p value. Let b be the number of bins in the window, $a(P)$ be the average density around P, and w be the width of the bins. Then we select the tuning parameter λ as the minimum P for which the following relation holds: $[a(P) - a(P + wb)] \,/\, a(P + wb) \leq \varepsilon$. For the CLR test, b was set to 12, and ε was set to 0.1. In essence, we use this procedure to estimate the point at which the distribution of p values flattens out. The procedure was used because the CLR test was designed to be conservative; therefore, one expects the distribution of p values to be skewed somewhat toward $p = 1$. Standard methods, such as splines [26], assume the distribution of p values is flat near $p = 1$.

Data. We obtained allele frequency data for the Perlegen SNPs [19] from the Perlegen genotype browser website (http://genome.perlegen.com/browser/download.html), and ascertainment information was obtained directly from Perlegen Sciences. We limited the analysis to those SNPs that were discovered by Perlegen's chip-based resequencing in a worldwide sample of 24 individuals [64], including African-Americans, European-Americans, Native-Americans, and Asian-Americans. For analysis, we take a sliding window approach to scan the entire genome for evidence of selective sweeps and balancing selection. For a genomic window of 200 contiguous SNPs (on average ~500 kb), we perform the CLR test on the SFS of the combined sample (African-American + European-American + Chinese) and on the SFS of each of the individual populations. The values of all test statistics, corresponding significance levels, maximum likelihood estimates of the position of the sweep, and an estimate of the composite parameter α are then recorded along with the genomic position of the center of the window. We repeat this procedure for every tenth window of 200 SNPs across all autosomes. Chromosomal positions of genes and genetic map estimates of local recombination rates were retrieved using the July 2003 build of the human genome on the University of California Santa Cruz (UCSC) table browser [65]. A list of refseq genes mapped on to the same genomic build as the Perlegen SNPs is available either from the UCSC table browser or by request from the corresponding author.

Supporting Information

Figure S1. A Comparison of p Values Calculated from the Equilibrium Neutral Model with p Values Calculated from Alternative Neutral Null Models

Curves above the diagonal dashed lines indicate that the equilibrium model is anticonservative relative to the alternative null, and curves below the dashed line indicate that the equilibrium model conservatively identifies selection. The close correspondence between the curves and the diagonal dashed lines indicates that p values are largely consistent across alternative neutral null models, and demographic history does not systematically mislead the CLR approach.

Found at doi:10.1371/journal.pgen.0030090.sg001 (47 KB PDF).

Table S1. The 63 Genomic Regions with Strong Evidence for a Recent Selective Sweep ($p < 0.00001$, CLR test), but where the Estimate of the Position of the Beneficial Allele Is Not within 100 kb of the Coding Sequence of a Known Gene

Found at doi:10.1371/journal.pgen.0030090.st001 (111 KB DOC).

Table S2. A Genomic Scan for Selective Sweeps Using the CLR Test and a Sliding Window Approach

Each row contains the results of the CLR test for a 200 SNP window of the genome. Columns represent (1) chromosome; (2) position of the center of the window; (3) CLR statistic for the combined sample; (4) maximum composite likelihood estimate of sweep position in the combined sample; (5) CLR p value for the combined sample; (6) CLR statistic for the African-American sample; (7) maximum composite likelihood estimate of sweep position in the African-American sample; (8) CLR p value for the African-American sample; (9) CLR statistic for the European-American sample; (10) maximum composite likelihood estimate of sweep position in the European-American sample; (11) CLR p value for the European-American sample; (12) CLR statistic for the Chinese sample; (13) maximum composite likelihood estimate of sweep position in the Chinese sample; (14) CLR p value for the Chinese sample.

Found at doi:10.1371/journal.pgen.0030090.st002 (12 MB TXT).

Table S3. Evidence of Selective Sweeps at Genes Involved in the Dystrophin Protein Complex

p values are from the test of the genomic window nearest the midpoint of the gene, and values in parentheses represent the minimum p value for all windows within the gene, which is reported if different from the midpoint p value.

Found at doi:10.1371/journal.pgen.0030090.st003 (71 KB DOC).

Table S4. Evidence of Selective Sweeps at Heat Shock Genes

p values are from the test of the genomic window nearest the midpoint of the gene.

Found at doi:10.1371/journal.pgen.0030090.st004 (147 KB DOC).

Table S5. Contingency Table Analyses for Enrichment of Significant Results in Windows Nearest the Midpoint of Known Genes, Compared with the Remainder of the Genome

Different rows repeat the analysis for different CLR test significance levels (indicated in parentheses) and for different population samples. For the CLR test in the European-American and Chinese samples, we observe a highly significant enrichment of CLR tests that reject the null at gene centers, and this signal becomes stronger with more stringent significance levels.

Found at doi:10.1371/journal.pgen.0030090.st005 (74 KB DOC).

Table S6. Evidence of a Selective Sweep by the CLR Test in the Most Extreme Genomic Regions Identified by Other Methods in the Hapmap Analysis

Values in parentheses indicate p values of the CLR statistic.

Found at doi:10.1371/journal.pgen.0030090.st006 (99 KB DOC).

References

1. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227–1234.

2. Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in *Drosophila*. Nature 415: 1022–1024.

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

4. 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.

5. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al.

(2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153–1157.

6. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, et al. (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A 102: 7882–7887.

7. Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, et al. (1998) Dating the origin of the CCR5-Δ32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62: 1507–1515.

8. 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.

9. Dobzhansky T (1955) A review of some fundamental concepts and problems of population genetics. Cold Spring Harb Symp Quant Biol 20: 1.

10. Lewontin RC (1974) The genetic basis of evolutionary change. New York: Columbia University Press. 346 p.

11. Kimura M (1983) The neutral theory of molecular evolution. New York: Cambridge University Press. 367 p.

12. Gillespie JH (1991) The causes of molecular evolution. New York: Oxford University Press. 336 p.

13. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, et al. (2005) Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15: 1553–1565.

14. Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc Natl Acad Sci U S A 103: 135–140.

15. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4 (3): e72.

16. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.

17. Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res 16: 702–712.

18. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. 16: 980–989.

19. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, et al. (2005) Whole-genome patterns of common DNA variation in three human populations. Science 18: 1072–1079.

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

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

22. Nielsen R, Williamson SH, Hubisz MT, Kim Y, Clark AG, et al. (2005) Genomic scans for natural selection using ascertained SNP data. Genome Res 15: 1566–1575.

23. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B 85: 289–300.

24. Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. Proc Natl Acad Sci U S A 100: 9440–9445.

25. 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.

26. Ehmsen J, Poon E, Davies K (2002) The dystrophin-associated protein complex. J Cell Sci 115: 2801–2803.

27. Malik HS, Henikoff S (2002) Conflict begets complexity: The evolution of centromeres. Curr Opin Genet Dev 12: 711–718.

28. Malik HS, Bayes JJ (2006) Genetic conflicts during meiosis and the evolutionary origins of centromere complexity. Biochem Soc Trans 34: 569–573.

29. Pardo-Manuel de Villena F, Sapienza C (2001) Female meiosis drives karyotypic evolution in mammals. Genetics 159: 1179–1189.

30. Chevin LM, Hospital F (2006) The hitchiking effect of an autosomal meiotic drive gene. Genetics 173: 1829–1832.

31. Jablonski NG, Chaplin G (2000) The evolution of human skin coloration. J Hum Evol 39: 57–106.

32. Grichnik JM, Burch JA, Burchette J, Shea CR (1998) The SCF/KIT pathway plays a critical role in the control of normal human melanocyte homeostasis. J Invest Dermatol 111: 233–238.

33. Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, et al. (2006) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310: 1782–1786.

34. Izagirre N, Garcia I, Junquera C, de la Rua C, Alonso S (2006) A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. Mol Biol Evol 23: 1697–1706.

35. Gilad Y, Bustamante CD, Lancet D, Pääbo S (2003) Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am J Hum Genet 73: 489–501.

36. Bajjalieh SM, Peterson K, Linial M, Scheller RH (1994) Brain contains two forms of synaptic vesicle protein 2. Proc Nat Acad Sci U S A 90: 2150–2154.

37. Howell BW, Hawkes R, Soriano P, Cooper JA (1997) Neuronal position in the developing brain is regulated by mouse disabled-1. Nature 389: 733–737.

38. Amieva MR, Vogelmann R, Covacci A, Tompkins LS, Nelson WJ, et al. (2003) Disruption of the epithelial apical-junctional complex by *Helicobacter pylori* CagA. Science 300: 1430–1434.

39. Stuchell MD, Garrus JE, Muller B, Stray KM, Ghaffarian S, et al. (2004) The human endosomal sorting complex required for transport (ESCRT-I) and its role in HIV-1 budding. J Biol Chem 279: 36059–36071.

40. Cavalli-Sforza L (1973) Analytic review: Some current problems of population genetics. Am J Hum Genet 25: 82–104.

41. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, et al. (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30: 233–237.

42. 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.

43. Toomajian C, Ajioka RS, Jorde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165: 287–297.

44. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, et al. (1996) A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 13: 399–408.

45. Peck JR (1994) A ruby in the rubbish: Beneficial mutations, deleterious mutations and the evolution of sex. Genetics 137: 597–606.

46. Osier MV, Pakstis AJ, Soodyall H, Comas D, Goldman D, et al. (2002) A global perspective on genetic variation at the *ADH* genes reveals unusual patterns of linkage disequilibrium and diversity. Am J Hum Genet 71: 84–99.

47. Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, et al. (2005) Ancient and recent positive selection transformed opioid *cis*-regulation in humans. PLoS Biol. 3 (12): e387.

48. Rockman MV, Hahn MW, Soranzo N, Loisel DA, Goldstein DB, et al. (2004) Positive selection on MMP3 regulation has shaped heart disease risk. Curr Biol 14: 1531–1539.

49. Mekel-Bobrov N, Gilbert SL, Evans PD, Vallender EJ, Anderson JR, et al. (2005) Ongoing adaptive evolution of *ASPM*, a brain size determinant in *Homo sapiens*. Science 309: 1720–1722.

50. Evans PD, Gilbert SL, Mekel-Bobrov N, Vallender EJ, Anderson JR, et al. (2005) *Microcephalin,* a gene regulating brain size, continues to evolve adaptively in humans. Science 309: 1717–1720.

51. 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.

52. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321–324.

53. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, et al. (2005) Calibrating a coalescent simulation. Genome Res 15: 1576–1583.

54. Kayser M, Brauer S, Stoneking M (2003) A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol Biol Evol 20: 893–900.

55. Storz JF, Payseur BA, Nachman MW (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol Biol Evol 21: 1800–1811.

56. Stajich JE, Hahn MW (2005) Disentangling the effects of demography and selection in human history. Mol Biol Evol 22: 63–73.

57. Tishkoff SA, Williams SM (2002) Genetic analysis of African populations: Human evolution and complex disease. Nat Rev Genet 3: 611–621.

58. Dodson H, Diouf S (2004) In motion: The African-American migration experience. Washington (D. C.): National Geographic. 224 p.

59. 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.

60. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

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

62. Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168: 2373–2382.

63. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, et al. (2002) A high-resolution recombination map of the human genome. Nat Genet 31: 241–247.

64. Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. 8: 1229–1231.

65. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. Nuc Acids Res 32 (Suppl 1): D493–D496.