Signatures of positive selection apparent in a small sample of human exomes

Jacob A. Tennessen, Jennifer Madeoy and Joshua M. Akey

Genome Res. 2010 20: 1327-1334 originally published online August 6, 2010
Access the most recent version at doi:10.1101/gr.106161.110

Signatures of positive selection apparent in a small sample of human exomes

Jacob A. Tennessen,[1] Jennifer Madeoy, and Joshua M. Akey[1]

Department of Genome Sciences, University of Washington, Seattle, Washington 98195-5065, USA

Exome sequences, which comprise all protein-coding regions, are promising data sets for studies of natural selection because they offer unbiased genome-wide estimates of polymorphism while focusing on the portions of the genome that are most likely to be functionally important. We examine genomic patterns of polymorphism within 10 diploid autosomal exomes of European and African descent. Using coalescent simulations, we show how polymorphism, site frequency spectra, and intercontinental divergence in these samples would be influenced by different modes of positive selection. We examine putatively selected loci from four previous genome-wide scans of SNP genotypes and demonstrate that these regions indeed show unusual population genetic patterns in the exome data. Using a series of conservative criteria based on exome polymorphism, we are able to fine-scale map signatures of selection, in many cases pinpointing a single candidate SNP. We also identify and evaluate novel candidate selection genes that show unusual patterns of polymorphism. We sequence a portion of one novel candidate locus, *IVL*, in 74 individuals from multiple continents and examine global genetic diversity. Thus, we confirm, narrow, and supplement existing catalogs of putative targets of selection, and show that exome data sets, which are likely to soon become common, will be powerful tools for identifying adaptive genetic variation.

[Supplemental material is available online at http://www.genome.org.]

Regions of the human genome that have recently evolved via positive selection have been sought for decades, but often remain elusive or difficult to confirm (Bodmer and Cavalli-Sforza 1976; Akey 2009). Initial single locus approaches have now yielded to genome-wide analyses that extensively sample loci across all chromosomes, even if they do not sample the whole genome exhaustively (Hinds et al. 2005; The International HapMap Consortium 2005; Akey 2009). Multiple genomic scans have produced extensive, often poorly overlapping lists of candidate genes under positive selection in humans (Kelley et al. 2006; Voight et al. 2006; Williamson et al. 2007; Akey 2009). A limitation of most of these scans is that they have primarily focused on ascertained SNP markers (Hinds et al. 2005; The International HapMap Consortium 2005), complicating population genetics inferences. In order to extract well-supported regions of recent adaptation from existing catalogs of putatively selected loci, it is important to reevaluate and refine such lists using data that are free from ascertainment biases. Fortunately, more ideal genome-wide data sets are beginning to emerge. These include sets of all genomic exons, or "exomes," which are more practical to sequence at high coverage in multiple individuals than whole genomes. Although the sample sizes are still small, analysis of these genome-wide sequence data sets can be useful for evolutionary studies, as the unbiased estimates of polymorphism and divergence they provide can be used to assess previously identified candidate regions under selection and more precisely determine targets of selection.

Here, we analyze the autosomal exomes of four African and six European individuals (Ng et al. 2009). We first perform coalescent simulations with selection to evaluate whether selection could leave a signature in the exomes of a small number of individuals. We then test whether genomic regions previously identified as possible targets of positive selection show evidence of non-neutrality in the exome data, and we filter the candidate regions accordingly. We also identify and evaluate several novel regions of unusual polymorphism suggestive of positive selection, and we collect and analyze additional sequence data for one of the most interesting novel genes.

[1]**Corresponding authors.**
E-mail tennej@uw.edu; fax (206) 685-7301.
E-mail akeyj@uw.edu; fax (206) 685-7301.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.106161.110.

Results

Simulations

In order to test the hypothesis that polymorphism and divergence in a small exome data set of four African and six European individuals would reflect the action of positive selection if it had occurred, we performed extensive coalescent simulations for models with and without selection. We considered selection to be either local and recent (acting on a single continent since population splitting), or global and prolonged (acting constantly on all individuals for many generations prior to the split). We allowed the selection coefficient, s, to be weak (0.5%), moderate (1%), or strong (2%), and the scaled recombination rate to be 0 or 20. Details of the simulations are provided in the Methods.

Our results generally confirmed our expectations: Stronger or longer selection produced a greater effect, and recombination weakened its signature (Table 1; Fig. 1). On one or both continents, both nucleotide diversity (π) and Tajima's D were significantly reduced after prolonged weak selection or recent strong selection in the absence of recombination; with recombination, the effect was still significant after prolonged moderate or recent strong selection. Similarly, the proportion of fixed differences was significantly increased after recent moderate selection without recombination; for the African selection model, this was still true in the presence of recombination. These results suggest that polymorphism in the current exome data set has the power to reflect past selective events.

Genome-wide distribution of polymorphism

We divided the genome into nonoverlapping regions of 100 kb. We excluded 37 regions containing sites where all individuals in

Table 1. Summary of simulation results

Selection	R	Sites E	π_E (percent)	Tajima's D_E	Sites A	π_A (percent)	Tajima's D_A	$F_{ST} = 1$ (percent)
s = 0	0	1000	0.108	0.08	1000	0.130	−0.16	0.0002
s_E = 0.5%	0	783	0.122	0.19				0.0000
s_A = 0.5%	0				1098	0.137	−0.15	0.0009
s_{EA} = 0.5%	0	1925	0.090*	−0.06	2317	0.101***	−0.33*	0.0000
s_E = 1%	0	906	0.117	−0.01				0.0022**
s_A = 1%	0				1533	0.122	−0.31	0.0019***
s_{EA} = 1%	0	3467	0.068***	−0.23***	4184	0.074***	−0.56***	0.0001
s_E = 2%	0	1139	0.087*	−0.19**				0.0035***
s_A = 2%	0				2465	0.084***	−0.49***	0.0081***
s_{EA} = 2%	0	6639	0.046***	−0.38***	8002	0.053***	−0.70***	0.0004***
s = 0	20	1000	0.107	0.12	1000	0.129	−0.16	0.0001
s_E = 0.5%	20	1038	0.122	0.01				0.0009
s_A = 0.5%	20				1348	0.136	−0.14	0.0000
s_{EA} = 0.5%	20	2165	0.098	0.01	2554	0.116	−0.31	0.0000
s_E = 1%	20	1166	0.111	0.10				0.0000
s_A = 1%	20				1785	0.128	−0.14	0.0011***
s_{EA} = 1%	20	3704	0.082**	−0.13**	4419	0.100**	−0.37*	0.0004***
s_E = 2%	20	1390	0.095*	−0.17**				0.0000
s_A = 2%	20				2713	0.100**	−0.32	0.0026***
s_{EA} = 2%	20	6855	0.062***	−0.29***	8210	0.074***	−0.55***	0.0004***

We performed coalescent simulations of neutral sites linked to selected sites and tested whether various statistics summarizing polymorphism at these sites were affected by selection. Three modes of selection were considered (s_E = selection in Europe only after the Europe–Africa split, 3500 generations; s_A = selection in Africa only after the Europe–Africa split, 3500 generations; s_{EA} = prolonged selection both continents, 17,000 generations), at four selection coefficients (0, 0.5%, 1%, or 2%), with two levels of recombination (R = 0 or 20). "Sites" indicates the number of linked neutral sites analyzed for each simulation in Europe (E) and Africa (A). The polymorphism statistics π_E, π_A, Tajima's D_E, and Tajima's D_A are discussed in the text. "$F_{ST} = 1$ (percent)" indicates the percentage of linked neutral sites fixed for different variants between continents. All statistics were significantly different for at least some parameter combinations between sites linked to selected sites and sites linked to neutral sites.
*Wilcoxon rank-sum test, $P < 0.05$.
**Wilcoxon rank-sum test, $P < 0.001$.
***Wilcoxon rank-sum test, $P < 0.0001$.

both populations were heterozygous (Supplemental Table 1), as these may represent variation between undocumented paralogs (Doggett et al. 2006). Thus, our final data set consisted of 10,497 regions of 100 kb, each with over 500 bp of exon sequence, comprising 25,769 kb of exome sequence in each of 10 individuals. The mean length of the exon sequence in each region was 2455 bp (standard deviation = 2061; range = 501–19,362).

Overall mean π was 0.038% among Europeans and 0.049% among Africans, while mean π at synonymous sites (π_S) was 0.083% among Europeans and 0.108% among Africans. Overall mean Tajima's D was 0.01 in Europe and −0.20 in Africa, while mean Tajima's D at synonymous sites (Tajima's D_S) was 0.10 in Europe and −0.13 in Africa. Both π and Tajima's D were positively correlated between Europe and Africa (Supplemental Figs. 1, 2). Approximately 15% of the regions exhibited no polymorphism in exons (Supplemental Fig. 1); on each continent, the proportion of invariant regions was ~25% (Fig. 2). There were 35 highly divergent SNPs with an F_{ST} = 1 between Europeans and Africans, falling into 20 regions under 1 Mb (Supplemental Table 2).

Putatively selected regions identified in genome-wide scans of SNP data exhibit distinct patterns of polymorphism

We chose four catalogs of candidate regions under positive selection that were generated from distinct statistics: linkage disequilibrium (Voight et al. 2006), site-frequency spectra (Kelley et al. 2006), composite likelihood ratios (Williamson et al. 2007), and population differentiation (Akey 2009). These candidate lists were all generated from large-scale SNP data sets (Hinds et al. 2005; The International HapMap Consortium 2005). Although other lists of candidate genes exist (for review, see Akey 2009), we consider these four to be exemplary representatives of these four major methods for detecting selection. A total of 1043 of the 10,497 100-kb regions overlapped a region on at least one list, with each individual list comprising 26–425 regions (Table 2). The proportion of missing data was not significantly different between any candidate list and the remaining regions (Wilcoxon rank sum tests, $P > 0.05$ for all).

All four candidate lists are significantly different ($P < 0.05$) from the background distribution of polymorphism and divergence on both continents according to at least one statistic (Table 2). We tested three statistics on both continents for all three lists, and the majority of tests remained significant after a Bonferroni correction for multiple hypothesis testing. In contrast, only 4.7% of randomly generated lists of 100-kb regions were significant at one or more test statistics using this method. Correcting for differences in total exon length among the 100-kb regions produced nearly identical results, so we ignored exon length in subsequent analyses other than by continuing to exclude regions with 500 bp or fewer exon sequence. After merging the four lists, the overall list of candidates still had lower π and Tajima's D than non-candidate regions on both continents (Fig. 2). A parallel analysis that examined each gene individually, instead of binning them into 100-kb regions, gave similar results (Supplemental Table 3).

Filtered catalogs of selected loci

Because our sample size was small and ~15% of the regions were invariant in exons, a conventional outlier approach to detect regions under selection was not possible. Therefore, in order to develop a more refined catalog of selected loci, we used a series of

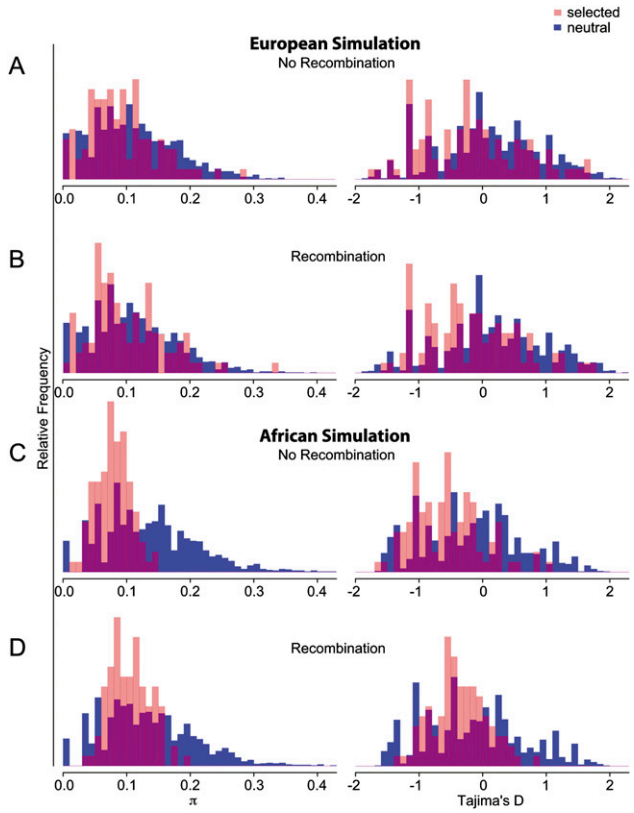

Figure 1. Distribution of summary statistics from coalescent simulations of neutral and positively selected loci. Blue and light red bars represent summary statistics from neutral and selected simulations, respectively. Areas where the two distributions overlap are shaded purple. *x*-Axes indicate values for π (% per nucleotide) or Tajima's *D*, while *y*-axes are unit-less relative frequencies. (*A*,*B*) Recent (3500 generations) strong selection (s_E = 2%) in Europe; (*C*,*D*) prolonged (17,000 generations) moderate selection (s_E = 1%) in Africa. For ease of visualization, very rare values beyond the ranges of these graphs are not shown, and regions from simulations with selection (light red) are scaled as if their frequency were identical to that of regions from neutral simulations (blue). Relative to blue bars, light-red bars are skewed toward the *left*, indicating that regions linked to sites under selection tend to have lower values of π and Tajima's *D*; recombination diminishes this effect only slightly.

conservative criteria to identify existing candidate loci with the strongest evidence of selection in the exome data. Instead of choosing outliers in the tail of a distribution, we instead excluded genes that possessed patterns of polymorphism trending in the opposite direction from what we would expect if they were genuine targets of selective sweeps; for example, if they showed a positive Tajima's *D* value or a π value greater than the median for the continent in question (see Methods). By repeating this conservative approach for several polymorphism-based statistics, we reduced the combined candidate list of 1043 regions down to 281 candidate regions. We classified these as Europe-centered (161 regions), Africa-centered (109 regions), or both (11 regions), based on the continent(s) where the selective sweep was originally hypothesized. Interestingly, only 36 of the 161 Europe-centered and 21 of the 109 Africa-centered regions show compelling evidence of continent-specific selection (Supplemental Tables 4, 5) based on high levels of intercontinental divergence. The patterns of selection for the remaining Europe- and Africa-centered regions (Supplemental Table 6) are more geographically ambiguous and exhibit reduced levels of genetic variation, but not strong intercontinental differentiation; a larger data set is needed to clarify the spatial heterogeneity, if any, of signatures of selection at these regions.

Our filtered lists of candidate selection regions retained many notable regions considered to be targets of selection, such as those containing *TRPV6* (Akey et al. 2006), *CYP3A5* (Thompson et al. 2004), *SLC24A5* (Lamason et al. 2005), and *KITLG* (Miller et al. 2007). For the continent-specific lists, few nonsynonymous sites in each region met the criterion of F_{ST} greater than 0.5, typically within a single gene; thus, we were usually able to identify a single candidate gene for the region, and often a single candidate SNP (Supplemental Tables 4, 5).

Novel candidate targets of selection

There were 13 100-kb regions that, despite not being on any of the four candidate lists, stood out as having extreme patterns of polymorphism (Table 3). All had at least one nonsynonymous site with F_{ST} greater than 0.78 (empirical P = 0.01), and all were invariant in the exome samples from at least one continent. For each region, there was a single gene containing the high F_{ST} nonsynonymous SNP(s). These represent novel candidates that may have been influenced by positive selection. The proportion of missing data was not significantly different between this novel candidate list and the remaining regions (Wilcoxon rank sum tests, $P > 0.1$).

One particularly interesting example, *IVL*, is a keratinization gene linked to hair and skin texture (Crish et al. 1993; Candi et al.

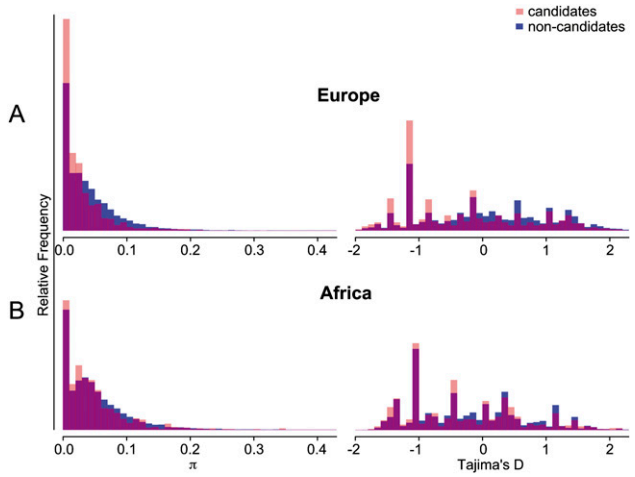

Figure 2. Polymorphism and site-frequency spectra for candidate and noncandidate regions. (*A*) Distribution of π and Tajima's *D* in the European sample; (*B*) distribution of π and Tajima's *D* in the African sample. Blue bars represent 100-kb regions that do not include candidate selection genes in any of the four predefined catalogs; light-red bars superimposed over these represent 100-kb regions that do include previously defined candidate selection genes; areas where the two distributions overlap are shaded purple. *x*-Axes are π (% per nucleotide) or Tajima's *D*, while *y*-axes are unit-less relative frequencies. For ease of visualization, very rare values beyond the ranges of these graphs are not shown, and candidate regions are scaled as if their frequency were identical to that of noncandidate regions. On each continent, ~25% of regions show no polymorphism (far *left* bars). Relative to putatively neutral regions (blue bars), candidate selection loci (light-red bars) are skewed toward the *left*, indicating that candidate regions tend to have lower values of π and Tajima's *D*. Light-red bars on the *right* halves of the histograms represent candidate regions for which the exomes provide no supporting evidence for positive selection; eliminating these was the first step in generating filtered candidate lists (see Supplemental Tables 4–6).

Table 2. Overlap between previously identified candidate regions and extreme values in the exome data

List of regions	No. of regions[a]	No. of polymorphic regions[b]	π (percent)[c]	Tajima's D[d]	$\log(\pi_A/\pi_E)$[e]
All E	10,497	7799	0.038 ±0.049	0.01 ±1.01	0.24 ±1.00
All A	10,497	8151	0.049 ±0.056	−0.20 ±0.92	0.24 ±1.00
Akey E	425	292	0.025*** ±0.033	−0.19** ±1.09	0.42** ±1.13
Kelley E	164	112	0.017*** ±0.028	−0.65*** ±0.88	0.69*** ±1.05
Williamson E	68	47	0.031 ±0.043	−0.25* ±1.03	0.48* ±1.19
Voight E	153	103	0.029** ±0.054	−0.22* ±1.05	0.19 ±1.13
Akey A	425	323	0.041* ±0.044	−0.20 ±0.95	0.42** ±1.13
Kelley A	129	107	0.055 ±0.061	−0.52* ±0.76	0.03 ±0.90
Williamson A	26	19	0.052 ±0.050	−0.52 ±0.95	−0.29* ±0.93
Voight A	197	146	0.039* ±0.045	−0.27 ±1.05	0.16 ±1.05

"E" and "A" refer to European and African samples, respectively. For each set of candidate regions, we tested whether each of three summary statistics was significantly different from the equivalent statistic calculated from all remaining regions.
[a]Number of 100-kb regions analyzed, excluding all regions with 500 or fewer base-pair coding sequence.
[b]Number of 100-kb regions analyzed that harbored polymorphisms, allowing Tajima's D to be calculated.
[c]Mean percentage value of π for each region, ±SD.
[d]Mean value of Tajima's D for each polymorphic region, ±SD.
[e]The log of π in Africa divided by π in Europe, ±SD.
*Wilcoxon rank-sum test, $P < 0.05$, Bonferroni corrected for three tests.
**Wilcoxon rank-sum test, $P < 0.001$, Bonferroni corrected for three tests.
***Wilcoxon rank-sum test, $P < 0.0001$, Bonferroni corrected for three tests.

2005). This locus has three nonsynonymous sites with F_{ST} values of 0.89 in the exome data, all three with the derived allele at high frequency in Europe. The 100-kb region encompassing *IVL* is fixed in our European exome samples, but is in the top 5% of most polymorphic regions in Africa ($\pi_A = 0.15\%$) (Fig. 3A). *IVL* is part of the highly polymorphic epidermal differentiation complex (EDC) (mean $\pi = 0.112\%$) that spans approximately 2 Mb on chromosome 1 (Mischke et al. 1996), and its protein product, involucrin, cross-links to itself and to other structural proteins encoded by this chromosomal region to form the cornified envelope (Fig. 3A; Supplemental Fig. 3; Steinert and Marekov 1997, 1999; Steinert et al. 2003). One of the high-F_{ST} SNPs results in a glutamic acid to glutamine substitution (E237Q) that has been previously noted as divergent between Africans and Europeans (Urquhart and Gill 1993). Because glutamine residues cross-link to lysines all along the length of the protein, this polymorphism may affect the degree of cross-linking.

In order to investigate the evolutionary history of *IVL* in more detail, we sequenced an ~800-bp region in 74 individuals from seven populations (Fig. 3B,C; Supplemental Data 1). This region encompassed two of the high-F_{ST} nonsynonymous SNPs, including the glutamine/glutamic acid polymorphism, as well as a variable 30-bp indel ("B^S") (Urquhart and Gill 1993). Unlike in the exome samples, European (CEU) π at *IVL* in the new samples (0.055%) was higher than the genome-wide average from the exomes (0.038%), although it was lower than *IVL* π in most other populations. Allelic structure at this locus is still unusual and suggestive of selection. The two nonsynonymous SNPs were in perfect linkage disequilibrium with each other, and with two synonymous SNPs, thus defining the two divergent haplotypes that differed substantially in frequency among populations. One haplotype was common in Africa but rare in Europe and did not happen to be sampled in the European exomes. Between the CEU samples and the Southern African samples, F_{ST} between these haplotypes was 0.75. Populations with intermediate frequencies of both haplotypes were characterized by remarkably high π and Tajima's D; for example, among Han Chinese, $\pi = 0.25\%$ and Tajima's $D = 2.60$ ($P < 0.01$). In contrast, Tajima's D was highly negative for populations with skewed allele frequencies, such as CEU (Tajima's $D = -1.33$; $P = 0.1$). Coalescent simulations suggest that the observation of only two haplotypes given four segregating sites, as seen in several populations, is significantly unusual ($P < 0.05$). The B^S indel was not in perfect linkage disequilibrium with the SNPs and did not vary as much between populations, but it was at higher frequency in Europe, Eastern Asia, and South America than elsewhere. Additional tandem 10-residue repeat polymorphisms, which we did not sequence, also vary among human populations (Urquhart and Gill 1993); these indels may have been missed in the exome data set, since repeats can be difficult to assemble accurately using next-generation sequencing.

Discussion

A small sample of exome sequences like this one can offer useful insights into evolutionary history. Exome sequence data and SNP data are mutually reinforcing with respect to identifying outlier regions that may be evolving under positive selection. All SNP-based lists of candidate regions, generated using distinct methods, showed significant departures from background levels of polymorphism for at least one sequence-based statistic (Table 2; Fig. 2). In addition, our filtered lists of candidate genes retain many regions that are well characterized as being under selection (Supplemental Tables 4–6). Our simulations suggest that π, Tajima's D, and F_{ST} in small population samples can detect positive selection provided that the selection coefficient is strong enough (Table 1; Fig. 1). Thus, small population samples of sequence data are sufficient to show unusual regions as unusual, and large population samples of SNP data, despite gaps and ascertainment biases, do pick up genomic regions with low polymorphism or skewed site-frequency spectra.

Using a series of conservative criteria, we have produced filtered lists of putative candidate regions under selection. The lists of continent-specific selection are especially short (Supplemental Tables 4, 5), and these loci warrant further attention. Notably prevalent are genes encoding mitochondrial proteins, metal ion-binding proteins, bitter taste receptors, and factors involved in

Table 3. Genes showing extreme patterns of polymorphism that are not on the four predefined candidate lists

Chromosome	High-F_{ST} sites	F_{ST}	Gene	π_E (percent)	π_A (percent)	Lists	Most specific GO term
1	151,149,234; 151,149,606; 151,150,335	0.89	*IVL*	0.000	0.152[a]	Tang et al. (2007)	isopeptide cross-linking via N6-(L-isoglutamyl)-L-lysine
1	159,285,664	0.78	*ARHGAP30*	0.023	0.000	The International HapMap Consortium (2007)	GTPase activator activity
2	20,687,979	0.78	*HS1BP3*	0.243[b]	0.000	na	phosphoinositide binding
2	219,587,000	0.89	*CCDC108*	0.000	0.042	Wang et al. (2006); Kimura et al. (2007)	structural molecule activity
3	113,667,715	0.89	*BTLA*	0.000	0.032	Kimura et al. (2007)	negative regulation of alpha-beta T cell proliferation
4	48,758,629	0.89	*FLJ21511*	0.000	0.039	Wang et al. (2006)	GPI anchor biosynthetic process
5	33,987,450	1.00	*SLC45A2*	0.000	0.040	Kimura et al. (2007)	melanosome membrane
6	36,554,953	0.89	*KCTD20*	0.014	0.000	na	voltage-gated potassium channel complex
6	110,214,210	0.78	*FIG4*	0.081	0.000	na	polyphosphoinositide phosphatase activity
9	5,547,708	0.89	*PDCD1LG2*	0.000	0.030	na	negative regulation of T cell proliferation
13	102,247,203	0.89	*KDELC1*	0.000	0.014	na	endoplasmic reticulum lumen
20	45,301,259	1.00	*ZMYND8*	0.000	0.000	na	zinc ion binding
22	17,803,250	0.89	*MRPL40*	0.000	0.028	Voight et al. (2006) (for Africa)	mitochondrial ribosome

All genes are within 100-kb regions that are fixed in at least one continent and all have at least one nonsynonymous site with $F_{ST} > 0.78$ ("High-F_{ST} sites;" position in the hg18 alignment is indicated). "π_E" and "π_A" refer to the respective 100-kb region, not just the gene. "Lists" indicates other genome scans for selection, not the four we examine in this study, which have highlighted these genes as possible targets of selection; "na" indicates that no previous lists include it. The most specific Gene Ontology (GO) term, defined as that which is shared with the fewest other genes, is indicated.
[a]Top 5% π for that continent.
[b]Top 0.5% π for that continent.

DNA damage response. Five candidate loci (*SLC24A5*, *RTTN*, *ANKRD45*, *TMED5*, and *SLC30A9*) reside within exonically invariant regions and have a nonsynonymous SNP with F_{ST} greater than 0.78, thus meeting the same criteria we used to identify novel candidates. Interestingly, the 200-kb span encompassing the developmental symmetry gene *RTTN* includes two candidate 100-kb regions for local adaptation, each on a different continent; for both high-F_{ST} SNPs in this gene, the derived allele is at high frequency on the continent exhibiting low π in that region, suggesting that selection may have acted on both continents, but targeted different regions of the gene in each case (Supplemental Fig. 4).

Although the present exome data set is underpowered for detecting non-neutral evolution given the relatively small sample size, several regions that were not identified in the four previous studies considered here do show unusual patterns and are worthy of further investigation (Table 3). These novel candidates are characterized by a high-F_{ST} nonsynonymous SNP within a region with locally invariant exons, often with relatively high π on the other continent, the combination of which is unusual. They include a motor neutron gene (Supplemental Fig. 5), two T-lymphocyte inhibitors, a mitochondrial ribosomal protein, a pigmentation gene, and others. Some, but not all of these genes have been highlighted by previous selection scans other than the four studies we chose to evaluate (Wang et al. 2006; The International HapMap Consortium 2007; Kimura et al. 2007; Tang et al. 2007). The most striking novel candidate gene was *IVL*, showing both unusually high F_{ST} across multiple nonsynonymous SNPs and an unusually high ratio of π_A to π_E (Fig. 3). By sequencing this locus in samples from seven human populations, we confirmed that the atypical patterns at *IVL* were not merely an artifact of the small exome data set. Although *IVL* is not actually invariant in Europe, F_{ST} is still quite high among continents, and the locus possesses a unique haplotype structure with at least four SNPs in perfect linkage disequilibrium, leading to both high and low values of Tajima's *D* in different populations, based largely on the frequencies of these haplotypes. Given its role in the epidermis and its remarkable geographically structured polymorphism, we hypothesize that *IVL* alleles may contribute to the pronounced differences in hair or skin morphology observed among populations (Wesley and Maibach 2003; Loussouarn et al. 2007).

Despite these promising results, the power to infer natural selection from an exome data set such as this one is limited in certain ways. Although sequencing coverage of this data set was relatively high (51X), some data are inevitably missing, which increases the noise in our estimates of polymorphism. Our simulations suggest that instances of weaker selection ($s < 1\%$) may not have a noticeable effect on all polymorphism statistics, especially in the presence of recombination. In addition, it's meaningless to identify "outliers" of low polymorphism or extreme Tajima's *D* values in this data set, as 15% of the regions are invariant in exons. At the opposite extreme, these data are of limited utility for detecting balancing selection or other evolutionary processes that would increase π, because the regions of highest polymorphism had to be discarded prior to analysis as possible undocumented paralogs. Furthermore, spatially or temporally complex modes of selection may not leave a clear signal. One well-documented example of positive selection in humans, *LCT* (Bersaglieri et al. 2004), is excluded as a selection candidate by our methods. The unknown specific ancestry of our European samples might include populations where *LCT* was not strongly selected, and the strongest signatures of selection may occur in noncoding regions.

The overlap between exome data and candidate lists generated by different methods shows that these methods successfully identify unusual patterns of genetic diversity, but our candidates

could just represent the tail end of a neutral distribution. As with all tests for selection, it is possible that demographic and stochastic processes could produce population genetic patterns that appear non-neutral; however, it is generally assumed that these processes will affect the entire genome, while selection will only affect a small subset of regions (Lewontin and Krakauer 1973; Akey 2009). Furthermore, our simulations explicitly modeled the major demographic processes that characterize the history of these populations. Thus, assuming that there exists a real set of positively selected loci, our candidate lists are likely to overlap it substantially, but further understanding and confirmation of selection will require additional neutrality tests with larger sample sizes, including analyses of noncoding sequences in these regions in case the true targets of selection reside there, followed by phenotypic and functional assays.

The data examined here are but a harbinger of the approaching tide of genomic data. Although more extensive data sets representing European and African humans are forthcoming, our results are promising for studies of human populations at finer scales of sampling, and nonhuman, nonmodel species, for which large genome-scale sample sizes will rarely be available. The ultimate goal of genome scans for selection is to pinpoint advantageous mutations with high confidence, laying the groundwork for studies of their phenotypic effects (Akey 2009; Grossman et al. 2010). Our analysis is still preliminary, and with additional data we anticipate increased power and precision to identify adaptive polymorphisms.

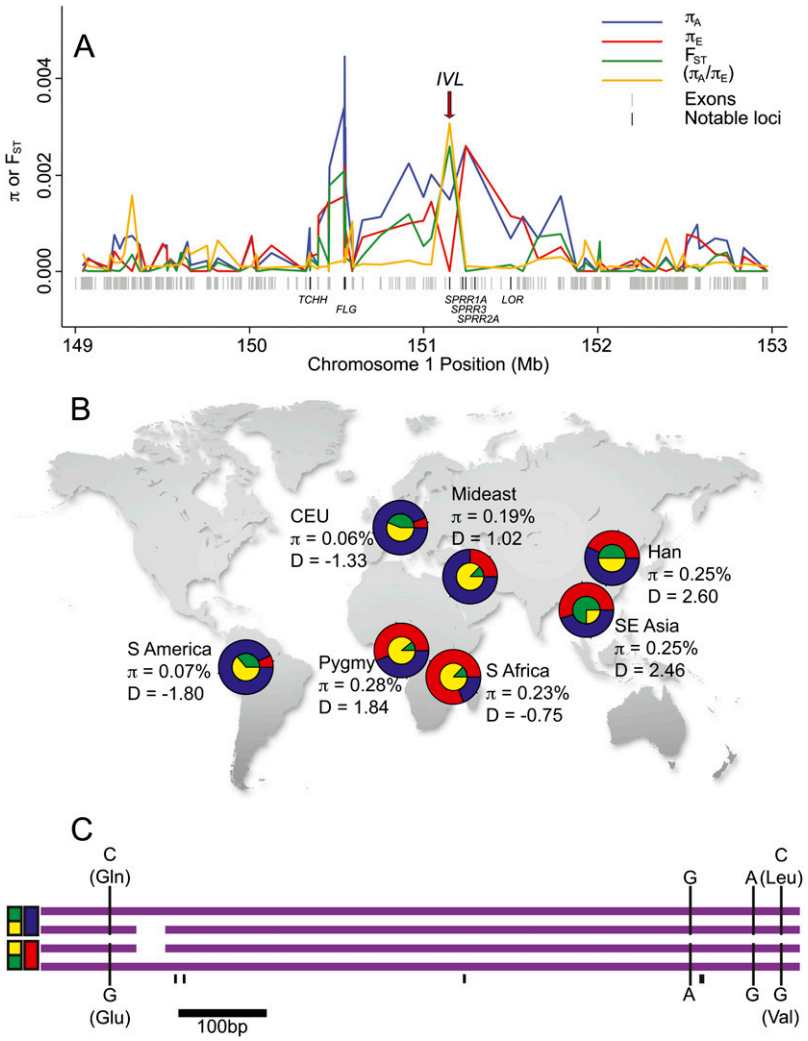

Figure 3. Population genetic patterns at the *IVL* locus. (*A*) Sliding window analysis of exome polymorphism and divergence in the EDC. While high π is common throughout the EDC, *IVL* shows very high F_{ST}, low π_E, and high π_A. The ratio of π_A to π_E has been scaled by dividing by 10,000. The positions of all exons and several notable genes that interact with *IVL* (Supplemental Fig. 3) are indicated. (*B*) Global *IVL* allele frequencies for SNP-based divergent haplotypes (red vs. blue) and a length polymorphism (green vs. yellow). We sequenced and analyzed a portion of *IVL* in seven populations: CEU (n = number of chromosomes = 34), S Africa (n = 16), Han (n = 30), Pygmy (n = 18), Mideast (n = 16), SE Asia (n = 20), and S America (n = 14). The glutamic acid carrying haplotype (red) was highest in Africa, while the glutamine carrying haplotype (blue) was highest in Europe. Tajima's D and π values varied remarkably among populations. Population differences for the green/yellow length polymorphism were less striking but still substantial between Africa/Mideast and elsewhere. (*C*) The four most common alleles in the 828–858-bp portion of *IVL* that we sequenced are shown. Four SNPs, two synonymous and two nonsynonymous, in perfect linkage disequilibrium defined two haplotypes (red and blue). An extra copy of a 30-bp repetitive sequence was present (green) or absent (yellow) on some copies of both haplotypes. All additional polymorphisms, indicated by short unmarked black lines, were rare and endemic to Africa.

Methods

Simulations

We used the ancestral selection graph (Neuhauser and Krone 1997) to evaluate the effect of various evolutionary scenarios on patterns of polymorphism and divergence. Our implementation (Ronald and Akey 2007; Skelly et al. 2009) allows for arbitrarily complex demographic history with recombination and applies to evolution in which selection acts additively. We simulated the genealogy of a single neutral site linked to a single site under selection. We followed the cosi demographic model (Schaffner et al. 2005) and set the following parameters: ancestral nucleotide = A; scaled population mutation rate = 0.008; selection for C, G, or T (s) = 0.005, 0.01, or 0.02; and population scaled recombination rate = 0 or 20. To match the empirical data, we simulated eight African haploid samples and 12 European haploid samples. We allowed selection to either act only on Europeans after the populations split 3500 generations ago (recent, s_E), only on Africans after the populations split (recent, s_A), or in both populations since the initial expansion of modern humans 17,000 generations ago (prolonged, s_{EA}). For each combination of parameters, we simulated 200,000 sites and concatenated all neutral sites for which a beneficial nucleotide appeared at the linked selected site, merging these independent sites into a single sequence. For this reason, it was impractical to simulate a lower mutation rate, as that would result in too few

beneficial mutations, and thus uninformatively short merged sequences of linked neutral sites. We performed 100 replicates each of these selection simulations. As neutral controls, we ran simulations as described above, except that selection was set to zero and only 1000 sites were simulated, all of which were analyzed; we performed 2000 replicates of each of these control simulations. For each combination of parameters, we calculated the polymorphism statistic π_E and/or π_A (for Europeans and Africans, respectively), Tajima's D_E and/or Tajima's D_A (for selection in Europe or Africa, respectively [Tajima 1989]), and the proportion of sites for which F_{ST} (Weir 1996) was 1. Due to non-normality of some parameters, we tested whether these values were significantly different from the corresponding neutral control values using nonparametric Wilcoxon rank-sum tests.

Genome-wide distribution of polymorphism

We analyzed exome data from four Yoruba HapMap individuals ("Africans") and six individuals of European ancestry ("Europeans"), two of whom are CEPH HapMap individuals and four of whom are non-HapMap individuals of European American ancestry (Ng et al. 2009). Although Ng et al. (2009) also examined two individuals of Asian descent, we excluded these from analysis because of the small sample size.

We divided the human genome into 100-kb regions and analyzed all autosomal sections of the exome data set with over 500 bp of exon sequence. Variation in lengths over 500 bp did not substantially affect polymorphism statistics (Supplemental Fig. 6). To avoid biasing our estimates of polymorphism, we excluded all regions in which any SNP had all individuals heterozygous and fewer than three individuals missing data, as these could represent undocumented paralogs scored as single loci. We calculated π, π_E, π_A, Tajima's D, Tajima's D_E, Tajima's D_A, the equivalent of these parameters at synonymous sites (π_S, π_{ES}, π_{AS}, Tajima's D_S, Tajima's D_{ES}, Tajima's D_{AS}), F_{ST} at individual sites and averaged across sites (where F_{ST} at invariant sites is set to zero), and various combinations of these parameters such as π_A/π_E (where 0.00005 is added to both numerator and denominator to avoid dividing by zero). Throughout this study, π is calculated per nucleotide and reported as a percentage. We calculated human-specific divergence from the common ancestor of human and chimpanzee (d_H), with macaque as an outgroup, using the panTro2/hg18 and rheMac2/hg18 alignments on the UCSC Genome Browser. As we did not have access to the proportion of missing data at invariant sites, we calculated the proportion of missing data at polymorphic sites for each region as a proxy for missing data overall. Analyses were performed using Perl, including the Perlymorphism package (Stajich and Hahn 2005), and R (R Development Core Team 2009).

Patterns of polymorphism in putatively selected regions identified in genome-wide scans of SNP data

Using the exome data, we evaluated four lists of candidate genes under positive selection (Kelley et al. 2006; Voight et al. 2006; Williamson et al. 2007; Akey 2009). For the list of Akey (2009), we included only the F_{ST}-based list, not the meta-analysis list of regions supported by multiple other studies. For each continent and candidate list, we calculated π, Tajima's D (Tajima 1989), and the ratio of π_A to π_E. Due to non-normality of some parameters, we used Wilcoxon rank-sum tests to evaluate whether these statistics were significantly different at the candidate genes than for the exome overall. Because we calculated three test statistics for each continent-list combination, we used a Bonferroni correction and divided our α value by three. In order to test the reliability of our method, we randomly generated 1000 lists of 425 false "candidate regions," corresponding to the size of the largest real candidate list, and tested how often they were significantly different from the remaining regions at these test statistics. In order to test whether variation in total exon length among the 100-kb regions affected our conclusions, we used weighted multiple regression. We calculated the residuals from the regression of each statistic of interest versus exon length, weighted according to the correlation between the log-squared values of these residuals and exon length. We tested all candidate lists both with and without the use of these weighted residuals; given that they had no major effect on our results, all presented results are based on tests without them.

Filtering catalogs of selected loci

We combined the lists of candidate regions and used a set of conservative criteria to filter this list based on the exome data. For each region, we sought evidence of selection on the continent(s) where the sweep was originally hypothesized. To detect continent-specific selection, we expected high intercontinental divergence, and thus sought the following criteria: local π and π_S less than the median (where "local" refers to the continent in question); local nonpositive Tajima's D and Tajima's D_S; π_A/π_E less (Africa) or greater (Europe) than the median; at least one nonsynonymous SNP with $F_{ST} > 0.5$. If selection did not generate high intercontinental divergence, the current data set may be insufficient to distinguish persistent global selection from more complex spatiotemporal models. To detect such geographically ambiguous selection, we sought the following criteria: overall and local π, π_S, and π/d_H less than their medians, overall and local nonpositive Tajima's D and Tajima's D_S, and appearing on a candidate list other than the F_{ST}-based list (Akey 2009), as that method is designed to identify continent-specific selection.

Novel candidate targets of selection

In order to identify regions that may have recently evolved via continent-specific positive selection, but which were not previously identified as such, we chose genes that met the following three criteria: within a 100-kb region showing no polymorphism on at least one continent; at least one nonsynonymous SNP with $F_{ST} > 0.78$ (corresponding to a maximum of a single shared allele in Africa or two shared alleles in Europe); not on any of the four candidate lists (Kelley et al. 2006; Voight et al. 2006; Williamson et al. 2007; Akey 2009). We further examined these regions with sliding window graphs (window size = 2000 cumulative base pairs of exon sequence).

Given an unusual population genetic pattern at the *IVL* locus in the exome data, we sequenced a portion of this locus (sites 151,149,529–151,150,356 in the hg18 assembly) in 74 individuals from seven populations: CEU (CEPH Utah residents of European descent), Han (Los Angeles residents of Han Chinese descent), Mideast, Pygmy (Biaka and Mbuti), S Africa (southern sub-Saharan Africa), S America (South American Andes), and SE Asia (southeast Asia excluding Japanese and Chinese). We used DNA samples from Coriell Cell Repositories with the following repository numbers: (S America: NA17301, NA17302, NA17304, NA17307–NA17310; Han: NA17733–NA17739, NA17741, NA17747, NA17749, NA17752–NA17756; Middle East: NA17042, NA17044–NA17050; Pygmy: NA10469–NA10473, NA10492–NA10495; Southeast Asia: NA17081–NA17090; CEU: NA06990, NA07019, NA07349, NA10830, NA10831, NA10842–NA10845, NA10848, NA10850, NA10851, NA10857, NA10858, NA10860, NA10861, NA17201; Southern Africa: NA17341, NA17342, NA17344–17349). We calculated basic statistics such as π, Tajima's D, and F_{ST}, as described above.

Acknowledgments

We thank all members of the Akey laboratoary for helpful comments. This work was supported by the NIH grant 1R01GM076036-01 (J.M.A.).

References

Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res* **19:** 711–722.
Akey JM, Swanson WJ, Madeoy J, Eberle M, Shriver MD. 2006. *TRPV6* exhibits unusual patterns of polymorphism and divergence in worldwide populations. *Hum Mol Genet* **15:** 2106–2113.
Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74:** 1111–1120.
Bodmer WF, Cavalli-Sforza LL. 1976. *Genetics, evolution, and man*. W.H. Freeman, San Francisco, CA.
Candi E, Schmidt R, Melino G. 2005. The cornified envelope: A model of cell death in the skin. *Nat Rev Mol Cell Biol* **6:** 328–340.
Crish JF, Howard JM, Zaim TM, Murthy S, Eckert RL. 1993. Tissue-specific and differentiation-appropriate expression of the human involucrin gene in transgenic mice: An abnormal epidermal phenotype. *Differentiation* **53:** 191–200.
Doggett NA, Xie G, Meincke LJ, Sutherland RD, Mundt MO, Berbari NS, Davy BE, Robinson ML, Rudd MK, Weber JL, et al. 2006. A 360-kb interchromosomal duplication of the human HYDIN locus. *Genomics* **88:** 762–771.
Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O, et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327:** 883–886.
Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072–1079.
The International HapMap Consortium 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.
The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.
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.
Kimura R, Fujimoto A, Tokunaga K, Ohashi J. 2007. A practical genome scan for population-specific strong selective sweeps that have reached fixation. *PLoS ONE* **2:** e286. doi: 10.1371/journal.pone. 0000286.
Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, et al. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310:** 1782–1786.
Lewontin RC, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74:** 175–195.
Loussouarn G, Garcel AL, Lozano I, Collaudin C, Porter C, Panhard S, Saint-Léger D, de La Mettrie R. 2007. Worldwide diversity of hair curliness: A new method of assessment. *Int J Dermatol* **46:** 2–6.
Miller CT, Beleza S, Pollen AA, Schluter D, Kittles RA, Shriver MD, Kingsley DM. 2007. *cis*-Regulatory changes in kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell* **131:** 1179–1189.
Mischke D, Korge BP, Marenholz I, Volz A, Ziegler A. 1996. Genes encoding structural proteins of epidermal cornification and S100 calcium-binding proteins form a gene complex ('epidermal differentiation complex') on human chromosome 1q21. *J Invest Dermatol* **106:** 989–992.
Neuhauser C, Krone SM. 1997. The genealogy of samples in models with selection. *Genetics* **145:** 519–534.
Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461:** 272–276.
R Development Core Team. 2009. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
Ronald J, Akey JM. 2007. The evolution of gene expression QTL in *Saccharomyces cerevisiae*. *PLoS ONE* **2:** e678. doi: 10.1371/journal.pone.0000678.
Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* **15:** 1576–1583.
Skelly DA, Ronald J, Akey JM. 2009. Inherited variation in gene expression. *Annu Rev Genomics Hum Genet* **10:** 313–332.
Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. *Mol Biol Evol* **22:** 63–73.
Steinert PM, Marekov LN. 1997. Direct evidence that involucrin is a major early isopeptide cross-linked component of the keratinocyte cornified cell envelope. *J Biol Chem* **272:** 2021–2030.
Steinert PM, Marekov LN. 1999. Initiation of assembly of the cell envelope barrier structure of stratified squamous epithelia. *Mol Biol Cell* **10:** 4247–4261.
Steinert PM, Parry DAD, Marekov LN. 2003. Trichohyalin mechanically strengthens the hair follicle: Multiple cross-bridging roles in the inner root sheath. *J Biol Chem* **278:** 41409–41419.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.
Tang K, Thornton KR, Stoneking M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol* **5:** e171. doi: 10.1137/journal.pbio.0050171.
Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA, Di Rienzo A. 2004. *CYP3A* variation and the evolution of salt-sensitivity variants. *Am J Hum Genet* **75:** 1059–1069.
Urquhart A, Gill P. 1993. Tandem-repeat internal mapping (TRIM) of the involucrin gene: Repeat number and repeat-pattern polymorphism within a coding region in human populations. *Am J Hum Genet* **53:** 279–286.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4:** e72. doi: 10.1371/journal.pbio.0040072.
Wang ET, Kodama G, Baldi P, Moyzis RK. 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc Natl Acad Sci* **103:** 135–140.
Weir BS. 1996. *Genetic data analysis II: Methods for discrete population genetic data*. Sinauer Associates, Sunderland, MA.
Wesley NO, Maibach HI. 2003. Racial (ethnic) differences in skin properties: The objective data. *Am J Clin Dermatol* **4:** 843–860.
Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet* **3:** e90. doi: 10.1371/journal.pgen.0030090.

Received February 5, 2010; accepted in revised form August 2, 2010.