Limited Evidence for Classic Selective Sweeps in African Populations

Julie M. Granka*, Brenna M. Henn§, Christopher R. Gignoux**, Jeffrey M. Kidd§§,
Carlos D. Bustamante§, Marcus W. Feldman*

* Department of Biology, Stanford University, Stanford, CA, USA
§ Department of Genetics, Stanford University, Stanford, CA, USA
** University of California, San Francisco, CA, USA
§§ Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA
1. 385 Serra Mall, Stanford University, Stanford, CA, 94305

Copyright 2012.
Running Title: Selective Sweeps and African Populations

Keywords: positive selection, selective sweeps, humans, African populations, demography

Corresponding Author:
Julie M. Granka
385 Serra Mall
Stanford University
Stanford, CA, 94305
Email: jgranka@stanford.edu
Phone: 650-725-7727
Fax: 650-725-7745
Abstract

While hundreds of loci have been identified as reflecting strong positive selection in human populations, connections between candidate loci and specific selective pressures often remain obscure. This study investigates broader patterns of selection in African populations, which are underrepresented despite their potential to offer key insights into human adaptation. We scan for hard selective sweeps using several haplotype and allele frequency statistics with a dataset of nearly 500,000 genome-wide single nucleotide polymorphisms in 12 highly diverged African populations that span a range of environments and subsistence strategies. We find that positive selection does not appear to be a strong determinant of allele frequency differentiation among these African populations. Haplotype statistics do identify putatively-selected regions that are shared across African populations. However, as assessed by extensive simulations, patterns of haplotype sharing between African populations follow neutral expectations and suggest that tails of the empirical distributions contain false positive signals. After highlighting several genomic regions where positive selection can be inferred with higher confidence, we use a novel method to identify biological functions enriched among populations’ empirical tail genomic windows, such as immune response in agricultural groups. In general, however, it seems that current methods for selection scans are poorly suited to populations which, like the African populations in this study, are affected by ascertainment bias, have low levels of linkage disequilibrium, possibly old selective sweeps, and potentially reduced phasing accuracy. Additionally, population history can confound the interpretation of selection statistics, suggesting greater care is needed in attributing broad genetic patterns to human adaptation.

Introduction

Elucidating the selective pressures that human populations have encountered, as well as the means by which they have adapted to them, are central aims of evolutionary biology and anthropology. Recently, statistical methods in population genetics have been applied to genome-wide polymorphism data to identify genetic loci that may have experienced natural selection. Such inferences have primarily been made under the model of a hard selective sweep, where a new allele rapidly rises to fixation within a population due to positive selection (Sabeti et al., 2002, 2005, 2006; Voight et al., 2006; Sabeti et al., 2007; Williamson et al., 2007; Akey, 2009; Pickrell et al., 2009; Pritchard et al., 2010). While these genome-wide scans have detected hundreds of loci as focal sites of selective sweeps, most connections
between the loci and their selective pressures remain unknown. Lack of phenotypic or functional information, the relative effect of background selection, demography, and statistical noise have been suggested as some of the causes of the lack of agreement between studies (Teshima et al., 2006; Akey, 2009; Hermisson, 2009; Hernandez et al., 2011).

In light of these issues, it is crucial to understand broader patterns of adaptation in human populations before identifying individual putatively selected loci (Coop et al., 2009; Pickrell et al., 2009). One strategy is to compare across populations the regions of the genome that appear to be undergoing positive selection while incorporating knowledge of populations’ environments and lifestyles, providing insight into the selective forces acting on the same loci (Voight et al., 2006; Pickrell et al., 2009; Hancock et al., 2010b,a, 2011; Lachance et al., 2012). A second approach is to search for biological functions that could be under strong positive selection within populations, identifying functions that are more frequently associated with putatively-selected regions of the genome than expected by chance (see, for example, Voight et al. (2006) and Metspalu et al. (2011)).

Using both approaches, we sought a more complete picture of adaptation in human African populations. Previous studies (Voight et al., 2006; Sabeti et al., 2007; Jarvis et al., 2012; Lachance et al., 2012) have reported strong signals of positive selection in African populations. Local adaptation is expected to be common within Africa, as African populations practice a range of subsistence strategies and inhabit diverse environments to which adaptation may have occurred (DeMenocal, 2004; Campbell and Tishkoff, 2010; deMenocal, 2011). Selection may also have acted more efficiently in African populations, many of which have maintained large effective population sizes for the past 50,000 years, in contrast to populations outside of Africa (Tenesa et al., 2007; Lambert and Tishkoff, 2009; Campbell and Tishkoff, 2010; Henn et al., 2011). While African populations would appear ideal for understanding the extent of positive selection in humans, they have usually been underrepresented in such studies.

In a set of 12 highly diverged hunter-gatherer, pastoralist, and agriculturalist African populations, we search for genomic evidence of positive selection. We analyze genome-wide data of nearly a half a million single nucleotide polymorphisms (SNPs), and include populations from the Human Genome Diversity Project (HGDP) (Li et al., 2008), HapMap (Consortium, 2005; International HapMap Consortium et al., 2007; International HapMap 3 Consortium et al., 2010) and four hunter-gatherer populations for which thorough selection scans have not yet been conducted (Schuster et al., 2010;
As potential candidates for positive selection, we identify outliers in the empirical genomic distributions of several commonly-used selection statistics, including measures of population differentiation and hard selective sweeps.

**Methods**

**Samples**

We performed selection scans on genomic data from individuals sampled from 12 African populations using a set of nearly 500,000 single nucleotide polymorphisms (SNPs). Our sample of African populations includes agriculturalists, pastoralists, and hunter-gatherers, as well as individuals from rainforests, deserts, and other environments.

Populations included from the Human Genome Diversity Project (CEPH-HGDP) (sample sizes in parentheses) were the South African Bantu (8), Kenyan Bantu (11), Biaka Pygmies (22), Mbuti Pygmies (13), Mandenka (22), Namibian San (5), and Yoruba (21); Mozabites are excluded (Rosenberg et al., 2002; Li et al., 2008). We also included 12 Namibian San individuals from Schuster et al. (2010), who were combined with the HGDP Namibian San for a total of 17 Namibian San individuals; we also included additional individuals from the Hadza and Sandawe of Tanzania and the ≠Khomani Bushmen of South Africa, described by Henn et al. (2011). As in Henn et al. (2011), we removed closely related individuals, leaving 17 Hadza, 28 Sandawe, and 31 ≠Khomani Bushmen. (For \( F_{ST} \) calculations, we removed admixed individuals identified in Henn et al. (2011), leaving 11 Hadza and 23 ≠Khomani Bushmen). 

HapMap Maasai trio parents from Kinyawa, Kenya (MKK) (46 individuals) and Yoruba trio parents from Ibadan, Nigeria (YRI) (100 individuals) were also analyzed (Consortium, 2005; International HapMap Consortium et al., 2007; International HapMap 3 Consortium et al., 2010). (As Pemberton et al. (2010) found evidence of undocumented relatedness between some Maasai individuals, we removed a subset from the original dataset; see Supplemental Methods). For some analyses, we included HapMap CEU trio parents (88 individuals). HapMap Tuscans (88 individuals) were included for pairwise mean \( F_{ST} \) and population differentiation calculations (Henn et al., 2011). For a summary, see Table S1.

After merging the SNPs genotyped in the HapMap samples, in the CEPH-HGDP samples on the Illumina 650K platform, in the samples from Henn et al. (2011) on the Illumina 550K platform, and sequence data from Schuster et al. (2010), a total of 461,767 SNPs remained. For haplotype-based
selection statistics (see below), we pruned additional SNPs and phased genotype data from 461,154 SNPs using BEAGLE (Browning and Browning, 2007) (as in Henn et al. (2011), see Supplemental Methods). Known parent-offspring trios from the HapMap YRI and ≠Khomani Bushmen were used as seeds, with all default parameters, to increase phasing performance (Browning and Browning, 2011). All populations were phased in this manner except for the HapMap CEU, YRI, and M KK, for which phase information was available from the HapMap website. 457,682 autosomal SNPs were used for analyses of single SNP selection statistics (see below).

Ancestral alleles for all SNPs were determined by their state in chimpanzee; chimpanzee positions were obtained from the Ensembl database (www.biomart.org, accessed August 2011). If the ancestral state was unavailable or ambiguous, the SNP was removed. All datasets were merged to Human Genome Build hg18 (NCBI Build 36). When needed, we used the recombination map available through the HapMap (Consortium, 2005; International HapMap Consortium et al., 2007; International HapMap 3 Consortium et al., 2010).

**Single SNP Statistics**

We calculated genome-wide (over all SNPs) mean $F_{ST}$, a measure of population differentiation, between all pairs of African populations, and for each African population against the HapMap Tuscans, using the program GENEPOP (Raymond and Rousset, 1995; Rousset, 2008). GENEPOP was also used to calculate per-SNP $F_{ST}$ values between all pairs of African populations. We also calculated $F_{ST}$ over all populations for each SNP as in Akey et al. (2002), following Weir and Cockerham (1984). For these analyses, we included only the HGDP Yorubans (not the HapMap YRI due to their close relation to the HGDP population), for a total of 11 African populations.

We classified a SNP as genic if it was within 2 kb of a gene, and non-genic otherwise. Gene locations were obtained from UCSC refFlat mappings of RefSeq genes to hg18 (downloaded in May 2010, www.genome.ucsc.edu). We arranged values of $F_{ST}$ calculated over all populations for each SNP in bins with increments of 0.05. Within each bin, an enrichment value was calculated as the proportion of genic (or nongenic) SNPs in that bin divided by the genome-wide proportion of genic (or nongenic) SNPs. To assess significance, we resampled 200 kb segments of the genome and re-calculated enrichments in each $F_{ST}$ bin (Coop et al. (2009)). Upper and lower 2.5% values over 1,000 bootstrap samples were used to obtain 95% confidence intervals. We also calculated the absolute value of the derived allele frequency
difference ($|\delta|$) between all pairs of populations (55, or $\binom{11}{2}$, pairs). We averaged the $|\delta|$ values over all pairs for each SNP (to create one measure of differentiation) and calculated genic enrichments (in bin sizes of 0.03), assessing significance as above. We also calculated enrichments in bins of (non-absolute) $\delta$ values for individual pairs of African populations, as well as between the HGDP Yorubans and HapMap Tuscans, again assessing significance by bootstrapping (see Supplemental Methods).

For all population pairs, we evaluated the relationships between the top 99.99% tail values of $|\delta|$ (over all SNPs) and mean $F_{ST}$ (averaged over all SNPs), and fitted a curve using the lowess function in R ([www.r-project.org](http://www.r-project.org)) (see Supplemental Methods for additional analyses). We simulated the expected relationship under neutrality between the top 99.99% tail of $|\delta|$ values and mean $F_{ST}$ using the “beta-binomial” method of Balding (2003) (also used in Coop et al. (2009)), where allele frequencies in each population are drawn from a beta-binomial distribution whose variance is a function of the $F_{ST}$ between the two populations (see Supplemental Methods). For each pair of populations, we repeated this simulation independently for 457,682 loci (the number of autosomal SNPs), and determined the 99.99th percentile value of $|\delta|$ over all loci. For each population pair we plotted the mean and standard deviation of the 99.99% $|\delta|$ values over 25 simulations. A best fit curve through the simulated mean values was drawn using the lowess function in R.

Haplotype Statistics

We calculated several haplotype statistics widely used for detection of selective sweeps: the iHS (Integrated Haplotype Score) (Voight et al., 2006), which identifies ongoing and incomplete sweeps, and the Cross Population Extended Haplotype Homozygosity Test (XP-EHH) (Sabeti et al., 2007), which detects complete or nearly complete sweeps occurring in one population but not in another “reference” population (see Supplemental Methods for more details). Both statistics were calculated using scripts from the HGDP Selection Browser website (Pickrell et al., 2009) ([www.hgdp.uchicago.edu](http://www.hgdp.uchicago.edu)). Phased genotypes, ancestral SNP states, and recombination maps were determined as described previously.

Reference populations used for the XP-EHH calculations were the HapMap CEU (“XP-EHH CEU”), HapMap YRI (“XP-EHH YRI”), HapMap Maasai (“XP-EHH MKK”), and ≠Khomani Bushmen (“XP-EHH KHB”) (see Supplemental Methods for details). Using a more closely-related African reference population with a relatively more stable demographic history avoids artifacts caused by an out-of-Africa expansion, which could strongly affect the XP-EHH CEU statistic. iHS and XP-EHH CEU were calcu-
lated for all 12 populations, XP-EHH YRI for ten (HGDP Yorubans and HapMap YRI were excluded), and XP-EHH MKK and XP-EHH KHB for 11 (Maasai and #Khomani Bushmen, respectively, were excluded) (see Table S1). Per-SNP statistics were normalized in bins of SNPs with similar minor allele frequencies (in increments of 0.05) for iHS as in Voight et al. (2006), and over all SNPs for XP-EHH as in Sabeti et al. (2007).

**Windows and Haplotype Statistics**

We broke the genome into non-overlapping windows of 100 kb to identify regions of the genome under selection. Because of the lower linkage disequilibrium (LD) in African populations (Henn et al., 2011), we reduced the 200 kb window size used by Pickrell et al. (2009) (see Supplemental Methods; see also Conrad et al. (2006)).

As a summary statistic for each window, we used the maximum XP-EHH value of all SNPs within a window; for iHS, we used the proportion of SNPs in the window with \(|iHS| > 2\) (Pickrell et al., 2009). The variance of SNP statistics within 100 kb windows was significantly smaller than that among randomly selected SNPs (see Supplemental Methods and Results), confirming that the windows effectively grouped SNPs with more similar statistic values.

We binned genomic windows according to their numbers of SNPs in increments of ten SNPs (combining the few windows with \(\geq 50\) SNPs into one bin). We then obtained an empirical p-value for each window based on its ranking (according to its window statistic) in its SNP frequency bin. (See Supplemental Methods; Coop et al. (2009).)

**Overlap of Signals Between Populations**

As a summary, we first obtained the five genomic windows with the lowest empirical p-values in each population, and their p-values in all other populations, for iHS, XP-EHH CEU, XP-EHH YRI, XP-EHH MKK, and XP-EHH KHB (see Supplemental Methods for more details).

To quantify the similarity of selection signals between pairs of populations, we counted the number of 100 kb genomic regions present in the top empirical 1% (and 0.1%) of both populations of the pair using the `intersectBed` program in the program suite BEDTools (Quinlan and Hall, 2010). For each haplotype statistic, this resulted in a matrix of values over all population pairs. For each population pair we also computed a correlation between the two populations’ genome-wide window statistic values.
to construct another matrix for each haplotype statistic. We calculated a series of Mantel correlations of these matrices with a matrix of between-population $F_{ST}$ values; we obtained p-values by comparing observed correlations to those of 5,000 permutations, where rows and columns of the original matrices were randomly permuted, using the `mantel.rtest()` function in R. See Supplemental Materials for additional analyses.

We also tested the relationship of the matrix of overlaps of empirical tail windows with a matrix of average measures of Bantu ancestry for the populations of each pair (with a Mantel test as above). To calculate this ancestry measure, we first averaged the proportions of Bantu ancestry, inferred from the population structure analyses of Henn et al. (2011), over all individuals in each population separately. Then, for each population pair, we averaged the two population Bantu ancestry values.

**Coalescent Simulations**

We performed genome-wide simulations of XP-EHH and iHS using the coalescent simulation program *msms* (Ewing and Hermisson, 2010) under models that were neutral and models that included positive selection. We simulated a three population model: one population served as the reference population for XP-EHH, and the other two were the populations for which XP-EHH and iHS were calculated.

We assumed an equilibrium model and used the formula $(1 - F_{ST}) = (1 - \frac{1}{2N_e})^t$, where $t$ is the divergence time in generations, to translate $F_{ST}$ values to divergence times (Holsinger and Weir, 2009). To mimic the XP-EHH CEU statistic, we simulated an $F_{ST}$ of 0.15 between the reference population and the two daughter (African) populations (see $F_{ST}$ between African populations and Tuscans, Table S2). We simulated daughter population divergence times to correspond to $F_{ST}$ values of 0.01, 0.05, and 0.10, spanning the range of $F_{ST}$ values between the African populations in our dataset (Table S2). We assumed 12,000 as the effective size ($N_e$) of each population (Schaffner et al. (2005); see Supplemental Materials and Figure S12 for details.)

We did not rigorously model the ascertainment strategy of the Illumina genotyping platform, as we were not primarily interested in its precise effect (Eberle et al., 2007). As a proxy, we first removed all simulated SNPs with frequency less than 0.05 in the “European” reference population. We calculated allele frequencies for each SNP across all African populations of our dataset and grouped them by minor allele frequency (MAF) in bins of size 0.05; we then randomly thinned the simulated SNPs to match this MAF distribution and observed SNP density (Sabeti et al., 2007). Although this is only an approxima-
tion to the properties of the SNPs of our dataset, it should be sufficient for a window-based analysis of haplotype statistics (Conrad et al., 2006).

Using these simulated and pruned SNP data, in each of the two daughter populations we calculated iHS and XP-EHH using the first population as a reference. We followed the same procedure as for observed data: we normalized statistic values across the genome, broke the genome into 100 kb windows, and calculated empirical p-values for windows based on their numbers of SNPs. We calculated the overlap in tail windows of the two daughter populations using the p-value cutoffs 0.05, 0.01, 0.005, and 0.001. We repeated this genome-wide simulation for 100 independent replicates for each daughter-population $F_{ST}$ value (0.01, 0.05, and 0.10).

We simulated 20 haplotypes from each population. For neutral simulations, we simulated segments of ten Mb and concatenated 300 of these segments to mimic an entire genome (3000 Mb in total). We also simulated data under a model of a selective sweep occurring in the first daughter population; the second daughter population and the reference population were assumed to evolve neutrally. Since we used an empirical outlier approach, we simulated only 50% of the genome under this model of selection (300 five Mb segments); we simulated another 150 ten Mb segments under neutrality for a total of 3000 Mb (see Supplemental Methods).

For simulations under selection, the position in the middle of each five Mb segment (at 2.5 Mb) was the selected site, although the site was not sampled. The positive selection strength of the selected allele was $s = 0.2$; while unrealistic, this allowed for efficient simulation of very strong selective sweeps that could be detected with a high true positive rate by both iHS and XP-EHH (see Supplemental Methods). Based on preliminary simulations, we set the time of selection as 62.4 generations before the present, again allowing both XP-EHH and iHS to have reasonable ability to detect the sweeps (see Supplemental Methods). We only included simulations where the frequency of the selected allele was non-zero at sampling time (i.e., it had not been lost), and where the selected allele had arisen only once (i.e., at a frequency $1/N$ for a hard sweep); we discarded all others. See Supplemental Methods for additional details.

For each simulation, we calculated the true positive rate as the percentage of genomic windows in each empirical tail that were simulated under selection (i.e., were genomic windows within a five Mb selection simulation segment). Because signals of selection may not extend across an entire five Mb, this may be an overestimate of the true positive rate, and an underestimate of the number of false positives.

Limitations of the msms program at the time of analysis prevented reliable simulation of parallel
selective sweeps at the same site in both daughter populations (Gregory Ewing, personal correspondence).

**Gene Annotation Enrichment for Haplotype Statistics**

To determine whether Gene Ontology (GO) biological process terms ([www.geneontology.org](http://www.geneontology.org)) were significantly over-represented within populations in genomic regions with the most extreme selection statistics, we implemented a permutation approach similar to that of Begun et al. (2007) (see Figure 1 and Supplemental Methods). To assess significance in the top \( x\% \) of genomic windows, we randomly sampled \( x\% \) of genomic windows from the genome, accounting for the number of SNPs within each window as when assigning p-values to observed windows (see Figure 1), and obtained the number of occurrences of each GO term. The distribution of the number of occurrences of a GO term over 20,000 of these randomizations was used as a null distribution to obtain a p-value for its number of occurrences in the observed top \( x\% \). In addition to examining all GO terms, we examined a subset of terms present in the PANTHER database (≈ 260 terms, associated with 28,173 genes), studied by Voight et al. (2006) ([www.pantherdb.org](http://www.pantherdb.org); downloaded in November 2010). To account for multiple hypothesis testing of many GO terms, for each population we controlled the false discovery rate (FDR) at the 5% level using the procedure of Benjamini and Hochberg (1995). Python and R scripts used to perform the analysis are available upon request; see Supplemental Methods for more details.

We focused on enrichments in the 0.1% empirical tail for each haplotype statistic for each population, based on results from between-population comparisons of empirical tail windows (see Results and Table S3). We merged adjacent empirical tail genomic windows and did not report GO terms that achieved significance only because a gene spanned multiple contiguous windows. Since significant terms associated with only one genomic window merely indicate that they are rare, we also only reported terms associated with more than one window. To investigate the difference in results obtained when using alternative methods, we replicated the GO enrichment method used by Voight et al. (2006) for iHS in the HapMap YRI (see the Supplemental Methods).
Results

Single SNP Statistics

We first examined population differentiation of single SNPs. We calculated per-SNP autosomal $F_{ST}$ over 11 African populations (see Methods, Tables S1 and S2), and plotted the enrichment of genic and nongenic SNPs, in comparison to their genome-wide frequencies, in bins of $F_{ST}$ values (Figure 2A). Frequent strong positive selection on genic regions would be expected to cause genic SNPs to display higher population differentiation on average than nongenic SNPs (Coop et al., 2009; Barreiro et al., 2008). However, SNPs with extreme $F_{ST}$ were not significantly more often genic than nongenic. At intermediate $F_{ST}$ values (≈ 0.2 - 0.3), a slight enrichment of genic SNPs was significant.

$F_{ST}$ calculated across all 11 populations could have poor power to identify loci under positive selection if populations have experienced mostly local, or independent, selection. For each SNP we also calculated the derived allele frequency difference ($\delta$) between all (55, or $\binom{11}{2}$) pairs of populations, following Coop et al. (2009) (Figure S1A and B). Less than half of the pairs of populations exhibited $|\delta|$ values greater than 0.8, though most highly differentiated SNPs were genic (Figure S1C). The most highly differentiated SNPs ($|\delta| > 0.8$) were found in comparisons with the Hadza and Mbuti Pygmies, where one allele was nearly completely absent (see also Figure S2D). There were also very few SNPs (often less than 5) in each population pair’s most extreme $\delta$ value bins. As a result, when assessing significance of enrichment for individual pairs, a majority of the bootstrap samples of genomic regions did not even contain SNPs with extreme $\delta$ values – invalidating the bootstrap confidence intervals (see Supplemental Results and Figure S2). Thus, for no pair did we find a significant enrichment of genic SNPs for the most extreme $\delta$ values. Genic enrichments in the next-most-extreme $\delta$ bins, with valid bootstrap confidence intervals, were not significant.

Upon averaging $|\delta|$ over all pairs of populations to obtain a measure of differentiation similar to $F_{ST}$ for each SNP, enrichment of genic SNPs at high values was also not significant (Figure 2B). There was a significant enrichment of genic SNPs at low averaged $|\delta|$ values, possibly due to purifying selection; however, this enrichment was not observed for $F_{ST}$ calculated over all 11 populations (Figure 2A; see also Figure S1C).

We also examined genic enrichments for $\delta$ between the HapMap Tuscans and HGDP Yorubans, as a comparison to the analysis of Coop et al. (2009) for Perlegen Type A SNPs between the HapMap CEU
and YRI (Hinds et al. (2005); Figure S3). Unlike among African populations, confidence intervals were valid, since all bootstrap samples contained SNPs in each δ bin. However, a genic enrichment among SNPs with high δ was not significant for either the Tuscans or the Yorubans.

For each African population pair, we examined the relationship between their mean $F_{ST}$ (calculated over all SNPs) and the upper 99.99% tail value (over all SNPs) of $|δ|$ (Figure 3). As mean $F_{ST}$ between African populations increased, the 99.99% tail value of $|δ|$ also increased (as in Coop et al. (2009); see also Figure S4). Under a neutral equilibrium model, we simulated extreme $|δ|$ values for all pairs of populations given their observed pairwise $F_{ST}$ values (following Coop et al. (2009); see Methods). The simulated data, as well as the lowess smoothed curve, corresponded closely to observed data (Figure 3).

**Haplotype Statistics**

Haplotype-based statistics iHS (Voight et al., 2006) and XP-EHH (Sabeti et al., 2007) calculated with phased genotype data may be less biased by SNP ascertainment than SNP-based measurements such as $F_{ST}$ (Conrad et al., 2006). For a total of four statistics, we used HapMap CEU, HapMap YRI, HapMap MKK, and the #Khomani Bushmen (KHB) as reference populations (see Methods). Candidate genomic windows for positive selection should have selection statistic values in the tails of an empirical distribution across all genomic windows, if such selection is relatively rare (Hernandez et al., 2011).

**Comparison of Empirical Tail Windows**

Genome-wide selection scans often compare candidate signals across multiple populations (Pickrell et al., 2009; Henn et al., 2011). An example of this is Figure 4, which shows the top five 100 kb genomic windows with the lowest empirical p-values for XP-EHH MKK in each population (see also Figures S5, S6, S7, and S8).

For an analysis of patterns of sharing of selective sweep statistics beyond what can be obtained from Figure 4, for every pair of populations we calculated the number of 100 kb windows in the 1% tail of the empirical distributions of both populations (approximately 250 tail windows). As a population pair’s mean $F_{ST}$ increased (i.e., as populations became more diverged), the number of windows appearing in the top 1% of both populations decreased predictably. Mantel correlations between matrices of mean $F_{ST}$ and number of overlapping 1% tail windows between pairs of populations were significant for iHS, XP-EHH CEU, XP-EHH MKK, and XP-EHH KHB (all except XP-EHH YRI) (Figures 5 and 6 and
Table S3). We observed a similar, though less pronounced, relationship between mean $F_{ST}$ and extent of sharing for the 0.1% tail (Figure S9 and Table S3). For most statistics, the extent of sharing of candidate sweep signals thus appeared to be difficult to separate from levels of population divergence (as measured by $F_{ST}$).

As overlap of tail genomic regions for the XP-EHH YRI statistic did not appear to be related to population divergence, we suspected that patterns might be more related to the Bantu expansion, which resulted in extensive admixture within Africa over the past several thousand years (Beleza et al., 2005; Berniell-Lee et al., 2009). We found a significant negative Mantel correlation of tail overlap (in the top 1%) with the average Bantu ancestry of the populations in the pair (p-value = 0.0006; see Table S3 and Figure S11). The more gene flow a pair of populations had with the Bantu (for which the YRI are considered a proxy for a source population), the fewer empirical tail genomic windows they shared.

We then assessed whether between-population patterns in the empirical tails, expected to reflect positive selection, were qualitatively different from between-population patterns across all genomic windows (summarized by the correlation of two populations' window statistic values across the whole genome). For all statistics, there was a significant positive Mantel correlation between 1% tail overlap and correlation of the two populations’ genome-wide window statistics (Table S3, Figure S10). Outlying genomic windows thus did not exhibit between-population patterns that were different from those of genome-wide windows. A more extreme tail may reveal more meaningful signatures of selection, as overlap in the top 0.1% was less related to the genome-wide correlation of two populations’ window statistics (Table S3; see Supplemental Results for related analyses).

Neutral Coalescent Simulations

To understand the relationship between $F_{ST}$ and the extent of overlap of empirical tail genomic windows under neutrality, we performed neutral coalescent simulations of genome-wide SNP data (see Methods). XP-EHH and iHS were calculated for two daughter populations with various divergence times, and a third outgroup “reference” population was simulated to mimic the XP-EHH CEU statistic (Figure S12).

As the two daughter populations became more highly diverged (i.e., had a larger $F_{ST}$ value), the number of windows overlapping in their 1% empirical tails for XP-EHH decreased predictably (Figure 6 and Figure S13). This trend was also apparent for iHS, although there was no significant difference in overlap for different $F_{ST}$ values (Figure 5 and Figure S14). This negative relationship between tail
overlap and $F_{ST}$ was even stronger for less extreme empirical tails (5% tail), and weaker for more extreme tails (0.1%), as in our observed data; this is likely because more extreme tails contained fewer windows (Figures S13 and S14). Thus, depending on their levels of divergence, populations can be expected to share some genomic windows in their empirical tails, even without positive selection acting in parallel on those windows.

In both observed and simulated neutral data, there was a similar relationship between population divergence and the extent of sharing of outlying genomic windows between populations. However, there were more overlapping tail windows in the observed data than in the neutral simulations for both XP-EHH and iHS (Figures 5 and 6). While the simulations were modeled after the XP-EHH CEU statistic, simulated overlap values were comparable to those observed for XP-EHH KHB.

**Selective Sweep Coalescent Simulations**

To understand the relationship between $F_{ST}$ and the extent of overlap of empirical tail genomic windows given selective sweeps, we performed coalescent simulations as above with hard sweeps occurring in one of the two daughter populations, and the other populations evolving neutrally (see Methods). We simulated sweeps with a relatively high selection coefficient ($s = 0.2$) to enable them to be detected by both XP-EHH and iHS and to allow more efficient simulations (see Supplemental Methods). We also simulated a relatively recent selection time ($\approx 62$ generations ago), preventing most selected alleles from reaching fixation so that they could be detected by iHS (Supplemental Methods, Figures S15 and S16; Voight et al. (2006) and Sabeti et al. (2007)).

For both XP-EHH and iHS, there was no relationship between $F_{ST}$ and the number of 100 kb windows overlapping in the tails of the daughter populations. The extent of overlap appeared consistent with that expected by chance under independence (Figures 5, 6, S15 and S16). For XP-EHH, the number of overlapping windows was considerably lower than that of observed data and neutral simulations (Figure 6); for iHS, the number of overlaps was only slightly lower (Figure 5).

In simulations of less recent positive selection (144 generations ago), the true positive rate of detecting sweeps for iHS was low – genomic windows simulated under neutrality, as well as selection, appeared in the empirical tails (Figure S18, see also Figure S17). Because of these false positive windows, the relationship of $F_{ST}$ with the extent of sharing of tail windows between populations was more apparent (Figure S18, see Supplemental Results). With the same simulation parameters for XP-EHH, the true positive rate
remained high, and there was no relationship between $F_{ST}$ and overlap of tail windows.

**Enrichment of Biological Functions**

Another approach used to integrate results of genome-wide selection scans asks whether any biological functions (Gene Ontology biological process terms, or GO terms) appear to be more frequent targets of positive selection than expected by chance (www.geneontology.org). Within each population, we used a novel permutation procedure (see Methods and Figure 1) to search for biological functions significantly associated with genes in outlying genomic windows (according to values of iHS, XP-EHH CEU, YRI, M KK and KHB). We focused on the 0.1% tail, motivated by the lower correlation of between-population patterns of this more extreme tail with those of the whole genome (see previous section and Table S3).

Many GO terms appeared to be under strong selection within populations (Tables 1 and 2 and Table S4). For example, the Biaka Pygmies showed an enrichment of genes related to histone methylation, and the Namibian San showed an enrichment for terms related to apoptosis (Table 1). In the Mandenka, several GO terms related to antigen processing were enriched (Table 2). Several terms appeared to be under selection in multiple populations. In the Biaka Pygmies, Mbuti Pygmies, ≠Khomani Bushmen, Mandenka, and HapMap YRI, terms related to transcription were enriched primarily in the tails for XP-EHH CEU. In the Mandenka and Yoruba, immune response was significantly enriched in top windows for XP-EHH YRI and XP-EHH M KK, respectively. In the Namibian San and the Yoruba, the term brain development was enriched, while the ≠Khomani Bushmen and HapMap YRI both showed enrichment for ubiquitin cycle. The term sodium ion transport was enriched in both the Kenyan Bantu and HapMap YRI, which could motivate future studies in light of the salt retention hypothesis (Thompson et al., 2004). No GO terms were significant in the Hadza for any statistic, and terms were only significant for the Kenyan Bantu at less extreme p-value cutoffs (see Supplemental Results).

We found no significantly enriched terms in the empirical tails of iHS for the HapMap YRI at any significance level (Table 2). In contrast, in their iHS analysis of the same population, Voight et al. (2006) found enrichment of several terms including olfaction and MHC-1 mediated immunity. There could be several reasons for this lack of agreement. While we searched for enrichment among genomic windows, Voight et al. (2006) searched for enrichments among genes using statistics from 50 contiguous SNPs (see Supplemental Methods); they also analyzed a larger set of SNPs. Since we were able to replicate the results of Voight et al. (2006) using our SNP dataset and their enrichment procedure (see Supplemental
Methods and Results), the primary cause of the difference in results appeared to be the statistical method for assessing enrichment.

**Discussion**

We computed statistics for hard selective sweeps and patterns of SNP variation in a diverse set of human African populations following the approaches of many recent genome-wide scans for positive selection (Sabeti et al., 2002, 2005, 2006; Voight et al., 2006; Sabeti et al., 2007; Akey, 2009; Pickrell et al., 2009). Patterns of allele frequency differentiation and sharing of haplotype selection signals among African populations were consistent with expectations under neutrality, suggesting the presence of false positive signals in empirical tails. We confirmed that the effects of positive selection do contribute to some empirical tail loci by identifying biological functions significantly associated with them using a novel permutation procedure. However, population history can confound the interpretation of selection statistics, suggesting that greater care is needed in attributing broad genetic patterns to human adaptation.

**Population Differentiation**

Positive selection does not appear to have substantially shaped present-day allele frequency differences among the African populations in our dataset (Figures 2 and 3 and Figures S1 and S2). SNP density may partially explain this finding. While Coop et al. (2009) found a significant enrichment of genic SNPs among those highly differentiated between HapMap CEU and YRI for Perlegen Type A SNPs (which have a uniform ascertainment based on a multi-ethnic discovery set; Hinds et al. (2005)), our analysis of just the Illumina 550K genotype data for HGDP Yorubans and HapMap Tuscons showed no significant enrichment (Figure S3). We were also unable to identify a significant genic enrichment for SNPs with extreme \( \delta \) values for any pair of African populations (Figures S1A, S1B, S2 and Supplemental Results). This suggests that in addition to SNP density, ascertainment bias could contribute to our inability to identify significant genic enrichments.

In another analysis we discovered phenomena that may be unique to African populations. Extreme values of allele frequency differences between pairs of African populations were closely related to a pair’s overall divergence (as measured by mean \( F_{ST} \)), in near perfect concordance with data simulated under a neutral equilibrium model (Figure 3 and Figure S4). (Since selective sweeps are believed to have been
relatively rare in recent human evolution, mean $F_{ST}$ is likely a valid measure of divergence (Hernandez et al., 2011.) In contrast, a study of worldwide populations using a comparable set of SNPs (Coop et al., 2009) suggested a stronger role of positive selection in increasing levels of allele frequency differentiation beyond that predicted by population history. In addition, most strong signals of differentiation in our dataset resulted from extreme allele frequencies in the Hadza and Mbuti Pygmies, which have each undergone recent extreme bottlenecks (see Henn et al. (2011), Supplemental Results, and Figure S2D).

Our findings from SNP array data do not support positive selection as a more important determinant of allele frequency differentiation among African populations than population history. This cautions that population differentiation measures used with an empirical outlier approach could lead to false positive selection signals in this dataset.

**Comparison of Empirical Tail Windows**

With the haplotype statistics iHS and XP-EHH, we examined empirically-outlying genomic windows of 100 kb as candidates for selection. A common strategy used to interpret the generated lists of putatively selected loci is to compare them between populations (see, for example, Figure 4). Such comparisons have been used to generate hypotheses, such as shared selective pressures, to explain why particular regions of the genome might be under selection in multiple populations (Coop et al., 2009; Pickrell et al., 2009; Johnson et al., 2011). Analyzing coalescent simulations (performed with msms; Ewing and Hermisson (2010)) under models of both neutrality and local adaptation, we find that conclusions from comparative methods can be misleading if there are a large number of false positive signals.

Under neutrality, empirical tail genomic windows are expected to be shared between populations, with the amount of sharing determined by a population pair’s level of divergence (Figures 5 and 6). In simulations under a model of local adaptation, with selective sweeps occurring in only one population of a pair, there was no relationship between divergence and the extent of sharing of outlier haplotype statistics if the statistics had a high ability to detect sweeps (i.e, a high true positive rate and a low false negative rate). Under this scenario, the amount of sharing of haplotype signals between populations was that expected by chance under independence regardless of $F_{ST}$ value (Figures 5, 6, S15, and S16). In contrast, if haplotype statistics had insufficient ability to detect selection (due in our simulations to the timing of selection), a slight negative relationship was expected (Figure S18). This occurred because false positive neutral genomic windows, rather than only those under selection, appeared in empirical
tails. Lower power and a higher false positive rate may indeed be expected for sweeps with more realistic parameter values than those we simulated (chosen for efficiency and to maximize the ability of iHS and XP-EHH to detect selection, see Methods) (Voight et al., 2006; Sabeti et al., 2006, 2007).

Due to limitations of msms at the time of analysis, we were unable to simulate selection under a model of parallel, convergent selective sweeps at the same locus in multiple populations (see Methods). Given the above results, if statistics have a high true positive rate, empirical tail signals shared between populations should correctly reflect populations’ true amount of convergent adaptation and appear to have no relationship with population history.

In African populations, we observed patterns of between-population divergence and overlap of outlier loci that were very similar to those simulated under neutrality (Figures 5 and 6; while simulated data did not precisely match observed data, differences were likely due to the simulations’ simplicity and non-uniform SNP density of observed data). The observed strong negative relationship between $F_{ST}$ and extent of overlap of tail windows suggests that many signals in the 1% empirical tails may not be driven primarily by local (or convergent) adaptation – under which only a weak relationship would be expected (Figures 5 and 6). (In the more extreme 0.1% tail, this negative relationship was less pronounced.) Furthermore, between-population empirical tail overlap for the top 1% seemed to follow genome-wide correlations (Figure S10 and Table S3). In light of our simulations, false positive outlier loci seem to be a plausible explanation for the close relationship of population history with selective sweep patterns in African populations.

Our results agree with Coop et al. (2009) and Pickrell et al. (2009), who found that selective sweep signals tend to cluster by broad geographic and continental regions, and who emphasize that demography can influence haplotype selection statistics (see also Conrad et al. (2006); Jakobsson et al. (2008)). When methods do not fully or accurately identify signals of selection, sharing of selection signals between populations can have an important contribution from population history.

Between-population patterns of observed data and data simulated under both selection and neutrality can thus illuminate the prevalence of false positives in observed data. Previous studies may have spuriously attributed a common sweep signal among populations to selection because they did not examine these broader patterns. In light of this, summaries such as Figure 4 can be misleading (see Pickrell et al. (2009)). Determining significance of selection with empirical p-values has also elicited criticism; since the true proportion of the genome that is under natural selection is unknown, so are the most appro-
priate empirical cutoffs (see Results; Teshima et al. (2006); Thornton and Jensen (2007); Thornton et al. (2007); Akey (2009); Hernandez et al. (2011); Lohmueller et al. (2011)).

Finally, it is not usual to calculate XP-EHH using multiple, diverse reference populations, as we do with XP-EHH versus Europeans, Yorubans, Maasai and ≠Khomani San. Previous studies of XP-EHH in Africans have primarily used an out-of-Africa reference population (i.e., XP-EHH CEU), which identifies different signals of selection from when using a more closely-related reference population (see Supplemental Results). Divergence time of the reference population also affects the amount of sharing between populations of outlying genomic windows: fewer tail windows are shared for XP-EHH using as references the Maasai and Yorubans, which are more recently diverged compared to Europeans and ≠Khomani Bushmen (see Figure 6 and Figures S5, S6, S7, and S8). As seen with the Yorubans as a reference, gene flow may also play a role (Figure S11). Thus, divergence time and demography of the reference population should be considered when interpreting comparisons of XP-EHH between populations.

**Enrichment of Biological Functions**

As an alternative approach to identifying important signals of selection, we searched for biological functions over-represented in the tail of genomic window distributions for the iHS and XP-EHH. Motivated by the haplotype statistic analyses and simulations, which suggested a high rate of false positives in the 1% empirical tail, we selected the top 0.1% as a more appropriate cutoff (see Results). Using a statistically rigorous permutation procedure (see Methods), we detected enrichment of many Gene Ontology (GO) biological process terms in outlying genomic windows of nearly all studied populations (except for the Hadza, likely because of their recent severe bottleneck, Henn et al. (2011); Tables 1 and 2). Many of these enrichments suggest previously-unreported selective pressures within individual populations (see Results): examples include the terms blood coagulation in the Sandawe of Tanzania and response to virus in the Maasai.

Such findings increase our confidence that some signals inferred using selective sweep statistics do indeed identify areas of the genome that are under selection in African populations. Immune response, for instance, was significantly enriched in empirical tail genomic windows of both the Mandenka and Yoruba, for which other terms such as B-cell differentiation and antigen processing and presentation were also enriched. Since the Mandenka and Yoruba both inhabit western Africa, this may indicate strong pressure from infectious or parasite-related diseases in this region, or may reflect recent selective pressure
associated with adoption of agriculture only 5,000 years ago (Gignoux et al., 2011). These enrichments agree with recent studies, which have highlighted pathogens as a primary driver of recent human evolution (Fumagalli et al. (2011); Novembre and Han (2012)), and which have presented evidence that HLA and other disease-related genes are under selection in West Africans (Bhatia et al., 2011).

By preserving the genomic features of our dataset, our permutation method accounts for important factors such as non-uniform SNP density, different gene lengths, the existence of regions of the genome where genes with similar functions are clustered, and the method used to assign empirical p-values to genomic windows (Figure 1). While our method follows that of Begun et al. (2007), permutation methods are not often implemented for GO enrichment analyses in human genetics (Metspalu et al., 2011). Those commonly used are designed for microarray studies, which present different statistical challenges from analyses of genomic windows, within which several genes can exist (Huang et al., 2009a,b; Metspalu et al., 2011). Fumagalli et al. (2011), for example, used the method GONOME (Stanley et al., 2006); although this method accounts for gene size, it does not account for gene clustering within genomic windows (see also Subramanian et al. (2005); Wang et al. (2007); Holmans et al. (2009); Hancock et al. (2010b); Fournier-Level et al. (2011); Hancock et al. (2011); Lehne et al. (2011)). While McLean et al. (2010) developed a method for enrichments among genomic segments, it is not appropriate for larger genomic windows containing multiple genes.

Studies have acknowledged that these biases could result in false positive signals of enrichment. Voight et al. (2006), for example, suggested that gene clusters like the HLA region may have contributed to their significant results (see also Metspalu et al. (2011)). In fact, our results for outlying genomic regions for iHS in the HapMap YRI differed from those of Voight et al. (2006), likely due to our use of different statistical methodology.

Complicating Factors for Studies of Selection

It has been suggested that selection may be more common outside of Africa than within Africa (Kayser et al., 2003; Storz et al., 2004; Williamson et al., 2007; Johansson and Gyllensten, 2008; Myles et al., 2008; Coop et al., 2009; Hofer et al., 2009). While we have identified plausible signals of selection in our African dataset, demographic and technical factors complicate the application of selection statistics. The absence of observed genic enrichment for highly differentiated polymorphisms may be due to ascertainment bias, as the SNPs analyzed here were ascertained primarily in European populations and
are unlikely to be in LD with unique alleles that may be under selection in Africans (Voight et al., 2006; Eberle et al., 2007; Nielsen et al., 2007; Henn et al., 2011). A similar explanation has been invoked to explain the failure of genome-wide association studies (GWAS) in some African populations, e.g. for studies of malaria (Need and Goldstein, 2009). The lower LD of African populations may also make it difficult to identify SNPs undergoing allele frequency differentiation due to selection, to detect signatures of increased LD due to a sweep, and to infer chromosomal phasing accurately (Henn et al., 2011). To address some of these concerns related to LD, we analyzed genomic windows of 100 kb, in contrast to previous studies that have used windows of twice this size (Pickrell et al., 2009; Henn et al., 2011; Metspalu et al., 2011). Our analyses suggest that SNP density can also affect statistical significance, even with a relatively large dataset of nearly 500,000 SNPs and in populations where ascertainment bias is minimal (Figure S3).

Additionally, small sample sizes in some populations may have reduced our ability to accurately estimate allele frequency differentiation or detect hard selective sweeps. The power of the iHS and XP-EHH statistics to detect selective sweeps in African populations may also be limited if the selective events have occurred too far in the past (greater than $\approx 30,000$ years ago, see Sabeti et al. (2006)). One such example could be the transition to a desert environment in the $\neq$Khomani Bushmen, which may have exerted selective pressure too long ago to be detected.

All of these factors could contribute to the difficulty of unequivocally inferring positive selection. While SNP density and ascertainment bias on existing genotyping arrays are influential, whole-genome sequencing may not be a panacea for the issues we raise (preliminary results; see also Lachance et al. (2012)). The sensitivity of widely-used selective sweep statistics to population history, selection timing, phasing error, lower LD, and sample size will likely remain problematic.

Although we focus exclusively on hard selective sweeps, other models of selection, such as polygenic adaptation or adaptation based on standing genetic variation, also have the potential to uncover signals of positive selection (Coop et al., 2010; Hancock et al., 2010a,b; Pritchard et al., 2010; Pritchard and Di Rienzo, 2010; Fumagalli et al., 2011; Hancock et al., 2011; Hernandez et al., 2011; Tennessen and Akey, 2011). Awareness of the issues raised here should also improve analyses of these alternative models of selection.
Data Access

All genotype data has been published in previous work and is publicly available. For details, see Henn et al. (2011), Li et al. (2008), Consortium (2005), International HapMap Consortium et al. (2007), International HapMap 3 Consortium et al. (2010), and Schuster et al. (2010).

Acknowledgments

We would like to thank Nandita Garud, Laura Rodríguez-Botigué, Philip Greenspoon, Philipp Messer, David Enard, Martin Sikora, and Jonathan Pritchard for helpful comments and discussions. J.M.G is supported by the National Science Foundation (NSF) Graduate Research Fellowship, grant number DGE-0645962. B.M.H. and C.D.B. are supported by NIH grant R01HG003229. C.R.G. is supported in part by NIH Training Grant T32 GM007175 and the UCSF Chancellor’s Research Fellowship. J.M.K. is supported by NIH Grant T32HG000044. This research is supported in part by NSF grant GM28016 to M.W.F.
Figures

List of Figures

1 Permutation procedure used to assess significance of enrichment of Gene Ontology (GO) terms in the top $x\%$ of 100 kb genomic windows. Procedure is performed for a particular statistic (iHS, XP-EHH CEU, XP-EHH YRI, XP-EHH MKK, or XP-EHH KHB) for a particular population. Note that the third and fourth boxes demonstrate the method used to assign p-values to genomic windows based on the number of SNPs per window; see Methods for details.

2 Enrichment of genic and non-genic SNPs for values of $F_{ST}$ and for average values of derived allele frequency difference ($|\delta|$) between pairs of populations. (A) Enrichments of genic and non-genic SNPs for values of $F_{ST}$ (calculated across all 11 populations for each SNP), divided into bins of width 0.05. Dashed lines indicate upper and lower 95% confidence intervals for each of the genic and non-genic categories, assessed by bootstrap re-sampling (see Methods). (B) Enrichments for average $|\delta|$ values ($|\delta|$ is calculated between all pairs of populations for each SNP; average is taken over the pairs), divided into bins of width 0.03. Dashed lines are as in (A).

3 Extreme $|\delta|$ values versus mean $F_{ST}$ for pairs of populations. Each point is a different population pair (55 pairs in total). Y-axis is the upper 99.99% tail value (over all autosomal SNPs) of derived allele frequency difference ($|\delta|$) between the pair of populations; X-axis is mean $F_{ST}$ (over all SNPs) between the two populations. Black points are observed data; red points are the mean of data points simulated using the beta-binomial method (described in Methods) over 25 independent simulations (error bars indicate +/- 2 standard deviations over the simulations). Best-fit lowess curves are drawn through the points of the same color.

4 P-values for the 5 most extreme 100 kb genomic windows according to the XP-EHH MKK statistic in each population. Red indicates $p < 0.01$, orange indicates $p < 0.05$, yellow indicates $p < 0.10$, and white indicates $p > 0.10$. “Genes” column lists the genes located within the indicated windows.
Number of shared 100 kb genomic windows in the top empirical 1% vs. mean $F_{ST}$ for pairs of populations for iHS. X-axis is mean $F_{ST}$, Y-axis indicates the number of shared windows in the top empirical 1%. (A) Each point is a different population pair (some points are labeled, see key). A best-fit lowess curve is drawn through the points. Significance of the p-value for the Mantel correlation between the $x$ and $y$ variables is indicated in the upper right corner: “NS”: not significant, “*”: < 0.05, “**:”: < 0.01, “***”: < 0.001. (B) Boxplot of overlap in the top empirical 1% of iHS between populations with the indicated simulated $F_{ST}$ values, for 100 independent neutral simulations. (C) Means +/- standard errors of overlap, for the neutral simulations of part (B). (D) Boxplot of overlap in the top empirical 1% of iHS between populations with the indicated simulated $F_{ST}$ values, for 100 independent simulations of a selective sweep in only one population with selection strength $s = 20\%$ and time of selection 62 generations ago. (E) Means +/- standard errors of overlap, for the simulations in part (D).

Number of shared 100 kb genomic windows in the top empirical 1% vs. mean $F_{ST}$ for pairs of populations for XP-EHH. X-axis is mean $F_{ST}$, Y-axis indicates the number of shared windows in the top empirical 1%. (A-D) Each point is a different population pair (some points are labeled, see key and legend of Figure 5.) (E) Boxplot of overlap in the top empirical 1% of XP-EHH between populations with the indicated simulated $F_{ST}$ values, for 100 independent neutral simulations. (F) Boxplot of overlap in the top empirical 1% of XP-EHH between populations with the indicated simulated $F_{ST}$ values, for 100 independent simulations of a selective sweep only in one population with selection strength $s = 20\%$ and time of selection 62 generations ago.
Figure 1. Permutation procedure used to assess significance of enrichment of Gene Ontology (GO) terms in the top $x\%$ of 100 kb genomic windows. Procedure is performed for a particular statistic (iHS, XP-EHH CEU, XP-EHH YRI, XP-EHH MKK, or XP-EHH KHB) for a particular population. Note that the third and fourth boxes demonstrate the method used to assign p-values to genomic windows based on the number of SNPs per window; see Methods for details.
Figure 2. Enrichment of genic and non-genic SNPs for values of $F_{ST}$ and for average values of derived allele frequency difference ($|\delta|$) between pairs of populations. (A) Enrichments of genic and non-genic SNPs for values of $F_{ST}$ (calculated across all 11 populations for each SNP), divided into bins of width 0.05. Dashed lines indicate upper and lower 95% confidence intervals for each of the genic and non-genic categories, assessed by bootstrap re-sampling (see Methods). (B) Enrichments for average $|\delta|$ values ($|\delta|$ is calculated between all pairs of populations for each SNP; average is taken over the pairs), divided into bins of width 0.03. Dashed lines are as in (A).
Figure 3. Extreme $|\delta|$ values versus mean $F_{ST}$ for pairs of populations. Each point is a different population pair (55 pairs in total). Y-axis is the upper 99.99% tail value (over all autosomal SNPs) of derived allele frequency difference ($|\delta|$) between the pair of populations; X-axis is mean $F_{ST}$ (over all SNPs) between the two populations. Black points are observed data; red points are the mean of data points simulated using the beta-binomial method (described in Methods) over 25 independent simulations (error bars indicate +/- 2 standard deviations over the simulations). Best-fit lowess curves are drawn through the points of the same color.
Figure 4. P-values for the 5 most extreme 100 kb genomic windows according to the XP-DHH MKK statistic in each population. Red indicates $p < 0.01$, orange indicates $0.01 < p < 0.05$, yellow indicates $p < 0.1$, and white indicates $p > 0.1$. "Genes" column lists the genes located within the indicated windows.
Figure 5. Number of shared 100 kb genomic windows in the top empirical 1% vs. mean $F_{ST}$ for pairs of populations for iHS. X-axis is mean $F_{ST}$, Y-axis indicates the number of shared windows in the top empirical 1%. (A) Each point is a different population pair (some points are labeled, see key). A best-fit lowess curve is drawn through the points. Significance of the p-value for the Mantel correlation between the $x$ and $y$ variables is indicated in the upper right corner: “NS”: not significant, “*”: $< 0.05$, “**”: $< 0.01$, “***”: $< 0.001$. (B) Boxplot of overlap in the top empirical 1% of iHS between populations with the indicated simulated $F_{ST}$ values, for 100 independent neutral simulations. (C) Means +/- standard errors of overlap, for the neutral simulations of part (B). (D) Boxplot of overlap in the top empirical 1% of iHS between populations with the indicated simulated $F_{ST}$ values, for 100 independent simulations of a selective sweep in only one population with selection strength $s = 20\%$ and time of selection 62 generations ago. (E) Means +/- standard errors of overlap, for the simulations in part (D).
Figure 6. Number of shared 100 kb genomic windows in the top empirical 1% vs. mean $F_{ST}$ for pairs of populations for XP-EHH. X-axis is mean $F_{ST}$. Y-axis indicates the number of shared windows in the top empirical 1%. (A-D) Each point is a different population pair (some points are labeled, see key and legend of Figure 5.) (E) Boxplot of overlap in the top empirical 1% of XP-EHH between populations with the indicated simulated $F_{ST}$ values, for 100 independent neutral simulations. (F) Boxplot of overlap in the top empirical 1% of XP-EHH between populations with the indicated simulated $F_{ST}$ values, for 100 independent simulations of a selective sweep only in one population with selection strength $s = 20\%$ and time of selection 62 generations ago.
Tables

List of Tables

1. Significantly enriched Gene Ontology (GO) terms in the top empirical 0.1% of windows in each population. ................................................................. 33
2. Significantly enriched GO terms in the top empirical 0.1% of windows in each population, continued. ................................................................. 34
Table 1. Significantly enriched Gene Ontology (GO) terms in the top empirical 0.1% of windows in each population.

<table>
<thead>
<tr>
<th>Population</th>
<th>Statistic</th>
<th>GO term</th>
<th>Name of GO term</th>
<th>p value</th>
<th>Windows</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biaka Pygmy</td>
<td>XP-EHH CEU</td>
<td>GO:0045893</td>
<td>positive regulation of transcription, DNA-dependent</td>
<td>0.0002</td>
<td>8</td>
<td>TGFB11, TCF3, PCBD2, NSD1, TFE6, FOXXK1, CREB5, SMARC0D3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH KHB</td>
<td>GO:0006412 translation</td>
<td>0.01035</td>
<td>3</td>
<td>AC107956.8, EINF4,1, AL121593.10</td>
</tr>
<tr>
<td>Mbuti Pygmy</td>
<td>XP-EHH CEU</td>
<td>GO:0006366 (P)</td>
<td>transcription from RNA polymerase II promoter</td>
<td>0.00085</td>
<td>3</td>
<td>HLP, NFX, NFKB1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:0007186 G-protein coupled receptor protein signaling pathway</td>
<td>0.00215</td>
<td>4</td>
<td>CEN1DS, NFX, GPR35, NFKB1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:000981 phosphate transport</td>
<td>0.0067</td>
<td>2</td>
<td>COL5A3, COL6A1</td>
</tr>
<tr>
<td>Sandawe</td>
<td>XP-EHH YRI</td>
<td>GO:0007596</td>
<td>blood coagulation</td>
<td>0.01185</td>
<td>2</td>
<td>ADORA2A, IL2RB2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH KHB</td>
<td>GO:0046058 cAMP metabolic process</td>
<td>0.0001</td>
<td>2</td>
<td>PTPI, PTPIH</td>
</tr>
<tr>
<td>Namibian San</td>
<td>XP-EHH CEU</td>
<td>GO:0042981</td>
<td>regulation of apoptosis</td>
<td>0.0075</td>
<td>2</td>
<td>TNFSF5P1, MALT1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0006916 anti-apoptosis</td>
<td>0.0035</td>
<td>2</td>
<td>MALT1, YWHAAZ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0006754 ATP biosynthetic process</td>
<td>0.0002</td>
<td>2</td>
<td>ATP5B, SLC25A13, ATP6V1B2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0001764 neuron migration</td>
<td>0.00525</td>
<td>2</td>
<td>ITGAM, NTN1, MYH10, PAFAH1B1, YWHAE, CDK5R1, RELN, TWIST1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0003036 actin cytoskeleton organization</td>
<td>0.01555</td>
<td>2</td>
<td>CAP1, CRK</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0007420 brain development</td>
<td>0.00065</td>
<td>2</td>
<td>PTPTI, RELN</td>
</tr>
<tr>
<td>#Khomani Bushmen</td>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0004657 post-translational protein modification</td>
<td>0.00135</td>
<td>2</td>
<td>UBE2L3, UBE2E2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0051246 regulation of protein metabolic process</td>
<td>0.00135</td>
<td>2</td>
<td>UBE2L3, UBE2E2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0006457 protein folding</td>
<td>0.0043</td>
<td>2</td>
<td>DNAJC1, PPTL2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0006512 ubiquitin cycle</td>
<td>0.0193</td>
<td>3</td>
<td>HECW2, UBE2L3, UBE2E2</td>
</tr>
<tr>
<td>Bantu SA</td>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:00006814 sodium ion transport</td>
<td>0.01905</td>
<td>2</td>
<td>SLC4A5, ACCN5</td>
</tr>
<tr>
<td>Maasai</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>iHS</td>
<td>GO:0006945</td>
<td>response to virus</td>
<td>0.00225</td>
<td>2</td>
<td>CREBF2, IFNA5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0006584</td>
<td>response to nutrient</td>
<td>0.0014</td>
<td>2</td>
<td>LCT, UOT1A3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:004293</td>
<td>response to drug</td>
<td>0.0035</td>
<td>2</td>
<td>LCT, UOT1A3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:000648</td>
<td>protein N-linked glycosylation</td>
<td>0.00045</td>
<td>2</td>
<td>MGA5, KEL</td>
</tr>
</tbody>
</table>

Significance determined as described in Methods and Figure 1, using an FDR cutoff of 0.05, within each population. A population is not shown if no terms were significantly enriched (i.e., the Hadza and Bantu Kenya). The “Statistic” column indicates the statistic resulting in significance. Next to a GO term, “(P)” indicates that the term was only significant when examining the PANTHER subset of terms; “(also P)” indicates that the term is significant both when examining PANTHER and when examining all GO terms; if nothing is next to the term, it indicates significance only when all GO terms were examined. We count the number of windows in the 0.1% empirical tail to which each term was associated (“Windows” column); rather than reporting those windows, we report the genes within those windows that were associated with the given GO term (in the “Genes” column).
<table>
<thead>
<tr>
<th>Population</th>
<th>Statistic</th>
<th>GO term</th>
<th>Name of GO term</th>
<th>p value</th>
<th>Windows</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mendenka</td>
<td>XP-EHH CEU</td>
<td>GO:0007588 excretion</td>
<td></td>
<td>0.00125</td>
<td>2</td>
<td>HMox1, KCNK5</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:0006813 potassium ion transport</td>
<td></td>
<td>0.00855</td>
<td>3</td>
<td>ABC8S, KCN11J, KCN11J, KCNK5</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:0045893 positive regulation of transcription, DNA-dependent</td>
<td></td>
<td>0.00805</td>
<td>2</td>
<td>MAP2K5, TCFS</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:0016311 dephosphorylation</td>
<td></td>
<td>0.00345</td>
<td>2</td>
<td>PTPRB, PTPTN2</td>
</tr>
<tr>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0002504 (also P) antigen processing and presentation of peptide or polysaccharide antigen via MHC Class II</td>
<td></td>
<td>0.0</td>
<td>2</td>
<td>HLA-DRB5, HLA-DRB9, HLA-DQB1, HLA-DQA1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH YRI</td>
<td>GO:0006814 sodium ion transport</td>
<td></td>
<td>0.02065</td>
<td>2</td>
<td>SCNN1B, ACCN5</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0006355 regulation of transcription, DNA-dependent</td>
<td></td>
<td>0.00845</td>
<td>7</td>
<td>CAMTA1, SCMH1, TCF3, IMD128, MCM5, AC093376.5, DLX5, DLX6</td>
</tr>
<tr>
<td>Yoruba</td>
<td>XP-EHH MKK</td>
<td>GO:0007588 excretion</td>
<td></td>
<td>0.00130</td>
<td>2</td>
<td>SCNN1B, HMox1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:0001666 response to hypoxia</td>
<td></td>
<td>0.00395</td>
<td>2</td>
<td>SCNN1B, HMox1</td>
</tr>
<tr>
<td>HapMap YRI</td>
<td>XP-EHH CEU</td>
<td>GO:0009684 sodium ion transport</td>
<td></td>
<td>0.00950</td>
<td>2</td>
<td>SCNN1B, ACCN5</td>
</tr>
<tr>
<td></td>
<td>XP-EHH CEU</td>
<td>GO:0006355 regulation of transcription, DNA-dependent</td>
<td></td>
<td>0.00845</td>
<td>7</td>
<td>CAMTA1, SCMH1, TCF3, IMD128, MCM5, AC093376.5, DLX5, DLX6</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0004593 camera-type eye morphogenesis</td>
<td></td>
<td>0.00015</td>
<td>2</td>
<td>NFI, RING1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0030199 collagen fibril organization</td>
<td></td>
<td>0.0003</td>
<td>2</td>
<td>NFI, COL1IA2</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:00030183 B cell differentiation</td>
<td></td>
<td>0.00095</td>
<td>2</td>
<td>CD79A, Pou1F1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0006836 neurotransmitter transport</td>
<td></td>
<td>0.00185</td>
<td>2</td>
<td>SLC32A1, SLC6A30</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0006612 ubiquitin cycle</td>
<td></td>
<td>0.0021</td>
<td>4</td>
<td>USP9I, RING1, TULP4, HECW1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0006809 (also P) lipid transport</td>
<td></td>
<td>0.00385</td>
<td>2</td>
<td>USP9I, ATG10</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0006955 (also P) immune response</td>
<td></td>
<td>0.00525</td>
<td>3</td>
<td>CD79A, CCR9, AL645940.4, AL645931.7, HLA-DPB1, HLA-DPA1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH KHB</td>
<td>GO:00050850 positive regulation of calcium-mediated signaling</td>
<td></td>
<td>0.0002</td>
<td>2</td>
<td>FCER1A, P2RX5</td>
</tr>
<tr>
<td></td>
<td>XP-EHH KHB</td>
<td>GO:0009684 sodium ion transport</td>
<td></td>
<td>0.00950</td>
<td>2</td>
<td>SCNN1B, ACCN5</td>
</tr>
<tr>
<td></td>
<td>XP-EHH KHB</td>
<td>GO:0006355 regulation of transcription, DNA-dependent</td>
<td></td>
<td>0.00845</td>
<td>7</td>
<td>CAMTA1, SCMH1, TCF3, IMD128, MCM5, AC093376.5, DLX5, DLX6</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0004593 camera-type eye morphogenesis</td>
<td></td>
<td>0.00015</td>
<td>2</td>
<td>NFI, RING1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0030199 collagen fibril organization</td>
<td></td>
<td>0.0003</td>
<td>2</td>
<td>NFI, COL1IA2</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:00030183 B cell differentiation</td>
<td></td>
<td>0.00095</td>
<td>2</td>
<td>CD79A, Pou1F1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0006836 neurotransmitter transport</td>
<td></td>
<td>0.00185</td>
<td>2</td>
<td>SLC32A1, SLC6A30</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0006612 ubiquitin cycle</td>
<td></td>
<td>0.0021</td>
<td>4</td>
<td>USP9I, RING1, TULP4, HECW1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0006809 (also P) lipid transport</td>
<td></td>
<td>0.00385</td>
<td>2</td>
<td>USP9I, ATG10</td>
</tr>
<tr>
<td></td>
<td>XP-EHH MKK</td>
<td>GO:0006955 (also P) immune response</td>
<td></td>
<td>0.00525</td>
<td>3</td>
<td>CD79A, CCR9, AL645940.4, AL645931.7, HLA-DPB1, HLA-DPA1</td>
</tr>
<tr>
<td></td>
<td>XP-EHH KHB</td>
<td>GO:00050850 positive regulation of calcium-mediated signaling</td>
<td></td>
<td>0.0001</td>
<td>2</td>
<td>FCER1A, P2RX5</td>
</tr>
<tr>
<td></td>
<td>XP-EHH KHB</td>
<td>GO:00050850 positive regulation of calcium-mediated signaling</td>
<td></td>
<td>0.0001</td>
<td>2</td>
<td>FCER1A, P2RX5</td>
</tr>
</tbody>
</table>
References


HOFER, T., N. RAY, D. WEGMANN and L. EXCOFFIER, 2009 Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet 73: 95–108.


ROSENBERG, N. A., J. K. PRITCHARD, J. L. WEBER, H. M. CANN, K. K. KIDD et al., 2002 Genetic

ROUSSET, F., 2008 GENEPOP’007: A complete re-implementation of the GENEPOP software for Win-


SABETI, P. C., S. F. SCHAFFNER, B. FRY, J. LOHMUELLER, P. VARILLY et al., 2006 Positive natural

SABETI, P. C., P. VARILLY, B. FRY, J. LOHMUELLER, E. HOSTETTER et al., 2007 Genome-wide

SABETI, P. C., E. WALSH, S. F. SCHAFFNER, P. VARILLY, B. FRY et al., 2005 The case for selection

SCHAFFNER, S. F., C. FOO, S. GABRIEL, D. REICH, M. J. DALY et al., 2005 Calibrating a coalescent

SCHUSTER, S. C., W. MILLER, A. RATAN, L. P. TOMSHO, B. GIARDINE et al., 2010 Complete Khoisan

STANLEY, S. M., T. L. BAILEY and J. S. MATTICK, 2006 GONOME: measuring correlations between
GO terms and genomic positions. BMC Bioinformatics 7: 94.

STORZ, J. F., B. A. PAYSEUR and M. W. NACHMAN, 2004 Genome scans of DNA variability in humans

SUBRAMANIAN, A., P. TAMAYO, V. K. MOOTHA, S. MUKHERJEE, B. L. EBERT et al., 2005 Gene
set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.


