Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets

Maud Fagny,[1,2,3] Etienne Patin,[1,2] David Enard,[4] Luis B. Barreiro,[5] Lluis Quintana-Murci,*[1,2] and Guillaume Laval*[1,2]

[1]Institut Pasteur, Human Evolutionary Genetics, Department of Genomes and Genetics, Paris, France

[2]Centre National de la Recherche Scientifique, URA3012, Paris, France

[3]Université Pierre et Marie Curie, Cellule Pasteur UPMC, Paris, France

[4]Department of Biology, Stanford University

[5]Department of Pediatrics, Sainte-Justine Hospital Research Center, University of Montreal, Montreal, Quebec, Canada

***Corresponding author:** E-mail: glaval@pasteur.fr; quintana@pasteur.fr.

Associate editor: Ryan Hernandez

Abstract

Genome-wide scans for selection have identified multiple regions of the human genome as being targeted by positive selection. However, only a small proportion has been replicated across studies, and the prevalence of positive selection as a mechanism of adaptive change in humans remains controversial. Here we explore the power of two haplotype-based statistics—the integrated haplotype score (iHS) and the Derived Intraallelic Nucleotide Diversity (DIND) test—in the context of next-generation sequencing data, and evaluate their robustness to demography and other selection modes. We show that these statistics are both powerful for the detection of recent positive selection, regardless of population history, and robust to variation in coverage, with DIND being insensitive to very low coverage. We apply these statistics to whole-genome sequence data sets from the 1000 Genomes Project and Complete Genomics. We found that putative targets of selection were highly significantly enriched in genic and nonsynonymous single nucleotide polymorphisms, and that DIND was more powerful than iHS in the context of small sample sizes, low-quality genotype calling, or poor coverage. As we excluded genomic confounders and alternative selection models, such as background selection, the observed enrichment attests to the action of recent, strong positive selection. Further support to the adaptive significance of these genomic regions came from their enrichment in functional variants detected by genome-wide association studies, informing the relationship between past selection and current benign and disease-related phenotypic variation. Our results indicate that hard sweeps targeting low-frequency standing variation have played a moderate, albeit significant, role in recent human evolution.

Key words: **positive selection, whole-genome sequence data, human populations, neutrality statistics.**

Introduction

The detection of genomic regions that have been targeted by recent positive selection has proved a powerful tool for delineating genes contributing to adaptation to environmental variables and for informing functions accounting for phenotypic diversity. Over the last decade, many genome-wide scans for selection have been reported in humans, fueled by the advent of whole-genome single nucleotide polymorphism (SNP) data sets. These studies have made use of various statistical methods based on the predictable effects of positive selection on patterns of genetic variation. These effects include a decrease in haplotype diversity (Voight et al. 2006; Frazer et al. 2007; Sabeti et al. 2007; Tang et al. 2007; Pickrell et al. 2009), high fraction of rare alleles (Carlson et al. 2005; Kelley et al. 2006), or major shifts of allele frequency between populations (Akey et al. 2002; Hinds et al. 2005; Weir et al. 2005; Frazer et al. 2007; Barreiro et al. 2008; Chen et al. 2010; Oleksyk et al. 2010; Jin et al. 2012). These approaches have led to the identification of several hundred genomic regions displaying selection signals, suggesting the presence in these regions of new beneficial mutations that have spread rapidly through the population.

The more recent advent of whole-genome sequence (WGS) data sets has provided unbiased information relating to the spectrum of allelic variation, overcoming the SNP ascertainment biases that characterize SNP genotyping data sets, with a power of ~99% to detect variants with a population frequency above 1%, for most of the genome (Abecasis et al. 2012). For example, the 1000 Genomes (1000G) project, both its Pilot and Phase 1 releases (1000 Genomes Project Consortium 2010; Abecasis et al. 2012), and the Complete Genomics (CG) data set (Drmanac et al. 2010) have provided with 12–38 million SNPs from various populations worldwide. This dramatic increase in the amount of sequence information available, corresponding to up to ten times that provided by the HapMap Consortium (Frazer et al. 2007; Altshuler et al. 2010), should provide increased power for evaluating the impact and prevalence of selection on the human genome. In this context, a recent study of the 1000G Pilot data set has defined a list of genes for which there was compelling evidence of positive selection (Grossman et al. 2013).

doi:10.1093/molbev/msu118 Advance Access publication April 1, 2014

Despite the considerable contribution of genome-wide scans to our understanding of the effects of natural selection on patterns of genome diversity, replication in different studies and functional support for adaptive significance have been demonstrated for only a handful of genes (Akey 2009). Furthermore, and more generally, the importance of positive selection in shaping human diversity remains an open question. Some studies have reported enrichment of certain functional SNP classes among selection signals, suggesting a nonnegligible prevalence of positive selection as a driving force of human adaptation (Voight et al. 2006; Frazer et al. 2007; Barreiro et al. 2008; 1000 Genomes Project Consortium 2010; Jin et al. 2012). However, others have suggested that these enrichment signals might actually result, at least in part, from the action of background selection (Coop et al. 2009; Pritchard et al. 2010; Hernandez et al. 2011). In addition, some studies indicate that selection following the "hard sweep" model, in which new advantageous mutations arise and spread rapidly to fixation, has occurred only rarely in recent human evolution (Hernandez et al. 2011; Granka et al. 2012). Indeed, it has been proposed that many adaptive events have occurred through other, largely undetected forms of positive selection, such as polygenic adaptation or selection on standing variation (Pritchard and Di Rienzo 2010; Pritchard et al. 2010).

The lack of agreement between these selection studies highlight the need to assess the power of statistical methods for detecting the effects of positive selection in the context of human demography and specifically of WGS (e.g., coverage, SNP calling, number of individuals). For example, simulations of populations of *Drosophila* and *Anopheles* mosquitoes (i.e., large populations with constant sizes of 10^6 individuals) have already shown that low coverage can potentially impact the power to detect selective sweeps (Crawford and Lazzaro 2012). It also remains unclear whether the evidence of positive selection—that is, enrichment of genic regions, as opposed to nongenic regions, among selection signals (Voight et al. 2006; Barreiro et al. 2008; Jin et al. 2012)—can be extended to WGS data sets and is robust to alternative selection scenarios, such as background selection. In light of the increasing amount of WGS data sets, there is a methodological need to address these issues as an indispensable prerequisite to explore the occurrence of selection in the genomes of humans and other species.

In this study, we aimed to explore the prevalence of recent, strong positive selection (i.e., the hard sweep model) in human adaptation, using WGS data sets. To do so, we first performed a simulation study based on realistic models of human demography and determined the power of relevant neutrality statistics for detecting recent population-specific positive selection, considering the features of current WGS data sets, such as differences in coverage and sample size. We next evaluated the sensitivity of these statistics to other selective regimes, such as polygenic adaptation, positive selection on standing variation and background selection. We then analyzed the 1000G and CG data sets and found enrichment of some functional SNP classes among selection signals, controlling explicitly for potential confounding factors. Lastly, we searched for functional support of the adaptive significance of genomic regions enriched in selection signals and found that these regions are indeed enriched for SNPs associated with phenotypic variation, both benign and disease related.

Results

Power to Detect Recent Hard Sweeps from Next-Generation Sequencing Data

We first evaluated the power to detect recent hard sweeps over a large range of allele frequencies, from next-generation sequencing data. We simulated autosomal regions under neutral and hard sweep assumptions, using for both the same calibrated model designed to match realistic scenarios of human demography (fig. 1*A*; supplementary text, table S1, and fig. S1, Supplementary Material online) (Voight et al. 2005; Laval et al. 2010; Gravel et al. 2011). Indeed, publicly available WGS data sets, such as the 1000G and CG data sets, include continental populations with different demographic histories, a feature known to affect the power of neutrality statistics (Pickrell et al. 2009; Li 2011). We focused on two haplotype-based statistics that are known to exhibit high power to detect positive selection over a large range of allele frequencies (Voight et al. 2006; Barreiro et al. 2009) and expected to be insensitive to background selection, that is, there is no prior reason that background selection differentially affects the haplotypes sharing the ancestral or derived allele. This contrast with statistics based on population differentiation, such as F_{ST}, where distinguishing the effects of positive and background selection is more challenging (Hernandez et al. 2011). We thus used the integrated haplotype score (iHS), which measures the difference in haplotype homozygosity associated with the ancestral and derived alleles (Voight et al. 2006), and the Derived Intraallelic Nucleotide Diversity (DIND) test, which measures the differences in nucleotide diversity associated with the ancestral and derived alleles (Barreiro et al. 2009). This choice was also based on the fact that iHS has been successfully used to detect strong signals of positive selection in genotyping data—that is, significant enrichment of functional sites among selection signals (Voight et al. 2006), and DIND was designed to make full use of resequencing data (Barreiro et al. 2009). Furthermore, in line with our aims, both iHS and DIND exhibit substantial power over a large range of allele frequencies of the selected mutation (Voight et al. 2006; Barreiro et al. 2008), in contrast with other statistics such as XP-EHH or composite likelihood ratio (CLR), which are known to detect almost-completed or recently completed sweeps (i.e., frequency of the selected allele > 0.8) (Nielsen et al. 2005; Sabeti et al. 2007; Williamson et al. 2007; Casto et al. 2010).

To validate our simulation process, we estimated the power of iHS and DIND, assuming genotypes and gametic phases to be known (i.e., "full sequence data," fig. 1*B*, see Materials and Methods). We set $2Ns$ (N being the effective size and s being the selection coefficient) to 100 and simulated 100-kb DNA regions. Consistent with simpler scenarios of populations of constant size that used similar parameters

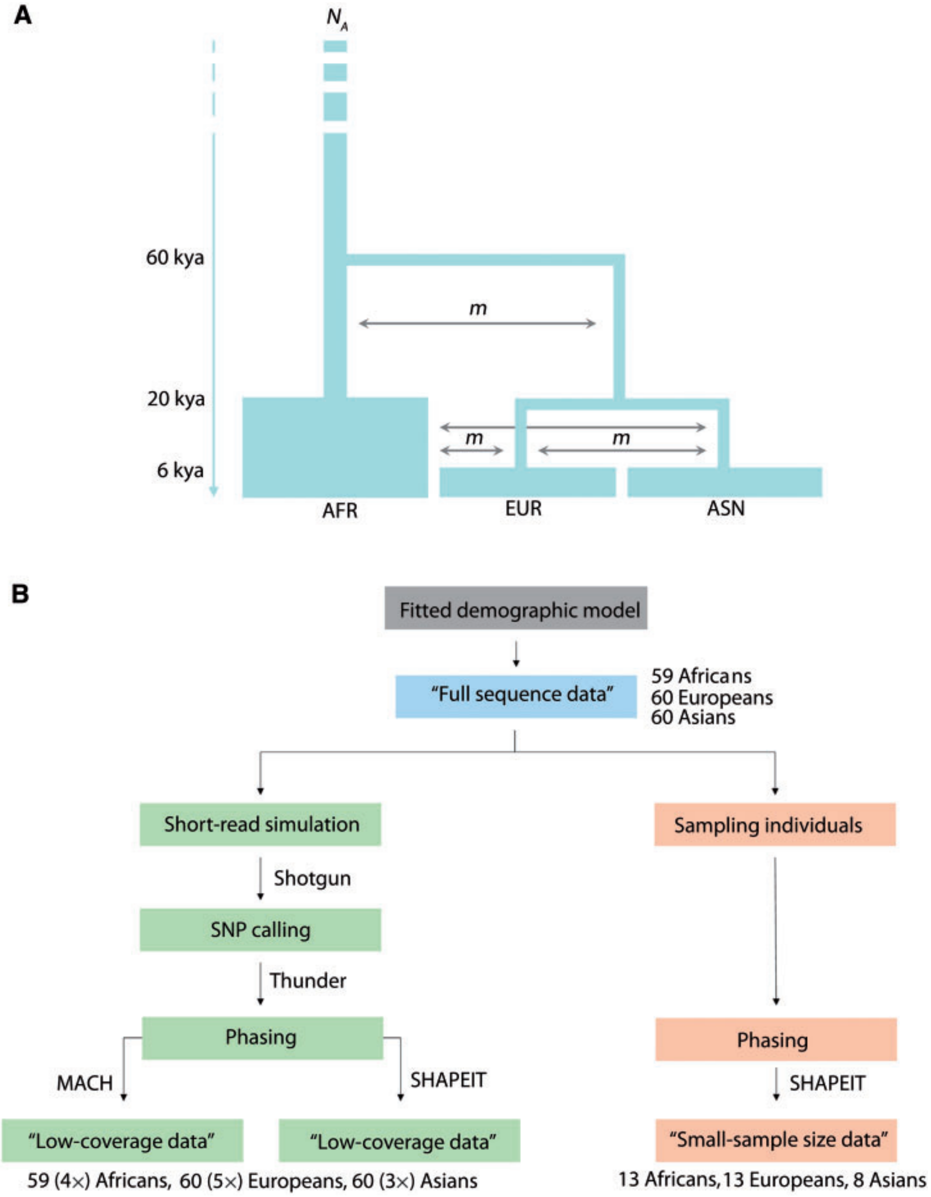

FIG. 1. Demographic model and flow chart used for the simulations. (*A*) Demographic model. The model used for the simulations considers that the ancestral Eurasian population split from the initial ancestral population (N_A) 60,000 years ago (60 kya) and went through a bottleneck reducing by half its effective population size. European and Asian populations diverged 20,000 years ago (20 kya) and went through recent expansions, corresponding to the Neolithic transition, increasing their effective size by 100. The expansion of the African population increased its effective size by 50. The migration parameter *m* is set to 1.3×10^{-5}. We used this calibrated demographic model to perform all subsequent simulations, that is, neutral simulations as well as those under various models of selection (recent selective sweep, background selection, and interaction between recent selective sweep and background selection, see Materials and Methods). (*B*) Flow chart for the simulations. To mimic 1000G Pilot data (green pipeline), we simulated low coverage from the "full sequence data" and inferred gametic phases. To mimic CG data (orange pipeline), we randomly sampled individuals from the "full sequence data" simulations and inferred gametic phases. Given the high coverage of the CG data set (read depth per site of 50× in average with 99% confidence interval ranging from 26× to 107× in Africa, 29× to 110× in Europe, and 17× to 75× in Asia), we did not simulate coverage, as this should not impact the power of DIND and iHS.

(Voight et al. 2006; Barreiro et al. 2009), iHS and DIND had a power of almost zero for selected allele frequencies (SAF) below 0.2, increasing rapidly to 80–100% for SAFs above 0.4 (fig. 2). In addition, the power computed as a function of SAF was similar to that found in a previous study specifying selection intensity on the basis of the age and the final frequency of the selected allele (Grossman et al. 2013). As expected, iHS and DIND clearly outperformed various neutrality statistics based on the allele frequency spectrum (AFS), even when we assumed realistic demographic models (supplementary table S2, Supplementary Material online). At the population level, the power of iHS and DIND was higher in African (78.90% and 76.06%, respectively) than in Eurasian populations (40.98% and 38.20%, respectively), highlighting the impact of demography on the power of these statistics. Furthermore, the power of both statistics was found to be similar after

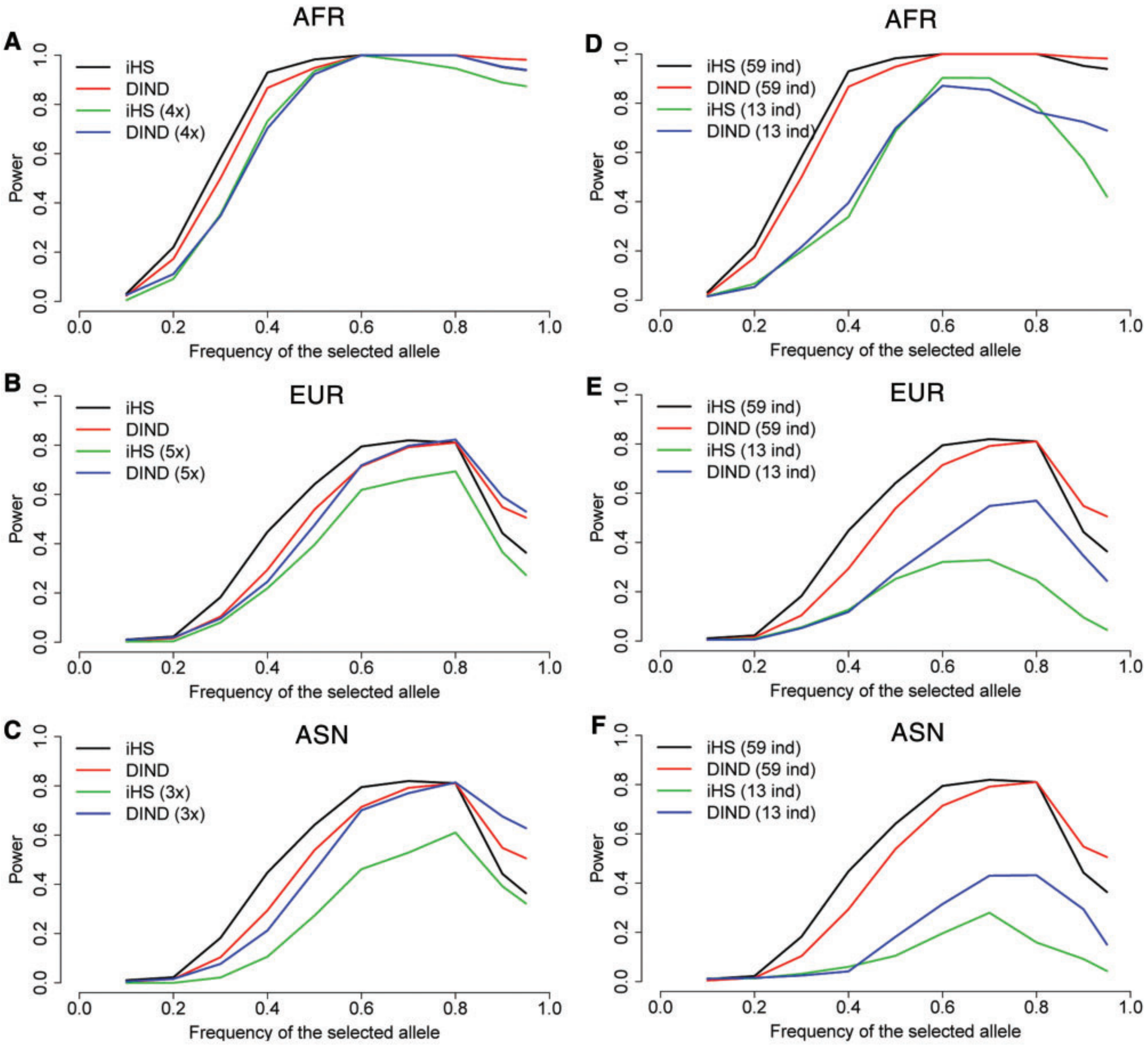

Fig. 2. Power of iHS and DIND to detect recent hard sweeps as a function of SAF. Critical values for both statistics, at FPR = 0.01, were obtained from 10^4 neutral simulations (2*Ns* = 0). For each simulation performed under recent positive selection (2*Ns* = 100), we used the proportion of extreme iHS and DIND values (see Materials and Methods). (*A*–*C*) Simulated "full sequence data" and "low-coverage data" (5× for Africans, 4× for Europeans, and 3× for Asians). (*D*–*F*) Simulated "full sequence data" and "small-sample size data" (13 individuals for Africans and Europeans, 8 individuals for Asians). In each case, we performed a total of about 2,000 simulations. (*A*, *D*) African population. (*B*, *E*) European population. (*C*, *F*) Asian population.

simulation of variation in recombination rate (i.e., presence of hotspots) and mutation rate (i.e., SNP density) (supplementary figs. S2 and S3, Supplementary Material online). The only exception to this trend was when SNP density was very low, where the power of iHS dropped dramatically as previously observed (Crisci et al. 2013), while that of DIND decreased only moderately. Overall, our simulations indicated that these haplotype-based statistics constituted powerful tests for detecting the effect of recent hard sweeps on a large range of allele frequencies in the context of full sequence data sets, regardless of the demographic history of the population considered.

We then investigated the effects of variation in coverage and sample size, characterizing WGS data sets, on the power of iHS and DIND. The 1000G Pilot data set is characterized by sample sizes of ~60 individuals per population sequenced at low coverage (3–5×), whereas the CG data set is characterized by small sample sizes (8–13 individuals per population) sequenced at high coverage (50×). We thus simulated data sets mimicking the 1000G Pilot ("low-coverage data") and CG ("small-sample size data") data sets and considered the uncertainty associated with haplotype phasing using MaCH and SHAPEIT (Li et al. 2010; Delaneau et al. 2012) (see Materials and Methods, fig. 1*B*). The power of both statistics varied with the frequency of the selected allele, as previously shown (fig. 2). Comparison of the full sequence and low-coverage simulated data sets demonstrated that low coverage had no impact on the power of DIND but slightly affected the

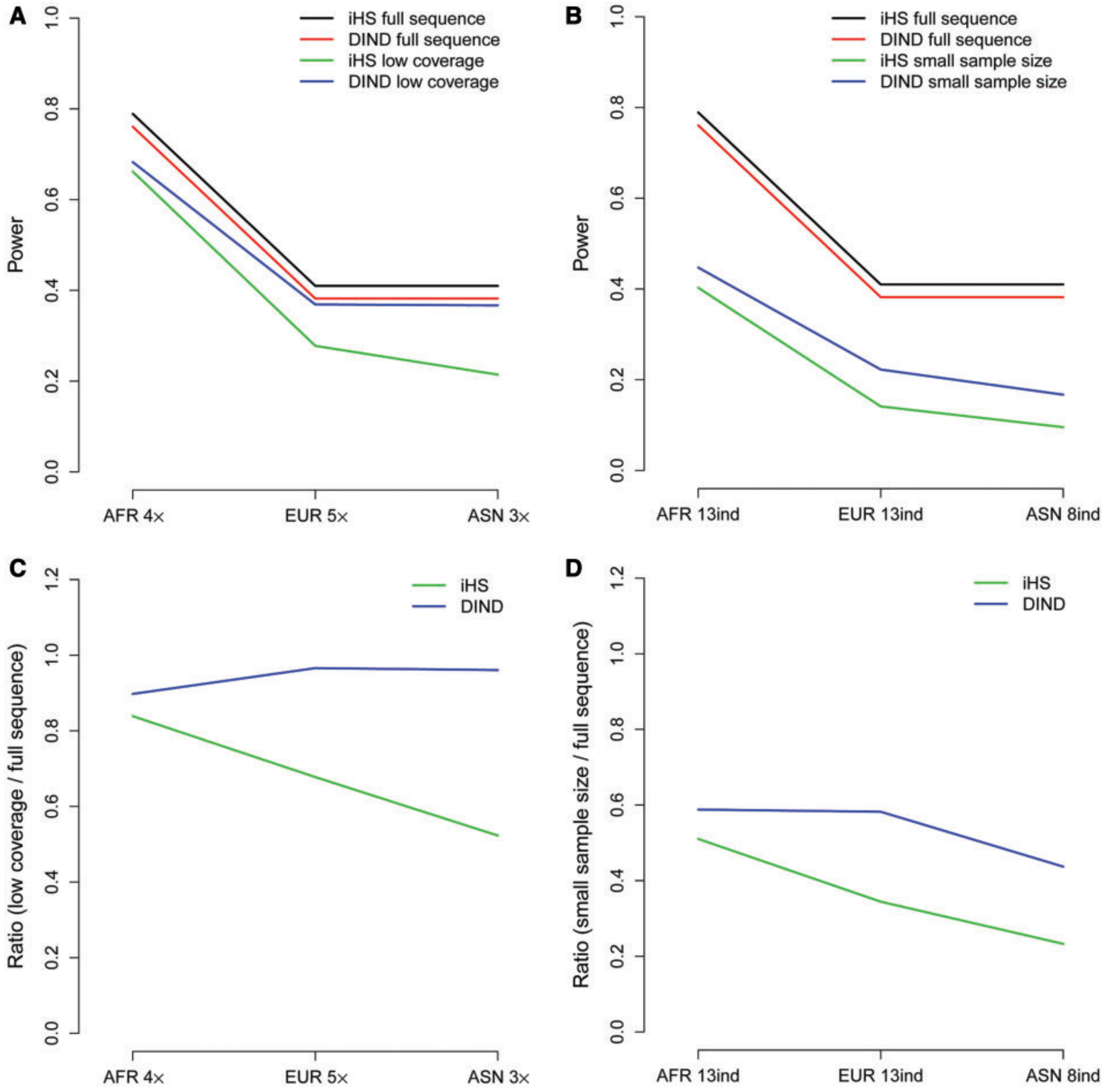

Fig. 3. Effect of low coverage and low sample size on the power to detect recent hard sweeps. (*A*) and (*B*) Power of iHS and DIND summed over a wide range of SAFs (simulations with SAF≥0.2 are considered together). (*C*) and (*D*) Power of iHS and DIND obtained within the context of next-generation sequencing data ("low-coverage data" or "small-sample size data") divided by the same power obtained with "full sequence data." For example, a ratio of 0.6 indicates that the power obtained with "low-coverage data" is 60% to that obtained with full sequence data. (*A*) and (*C*) "Low-coverage data" versus "full sequence data." The coverage is indicated for each population. (*B*) and (*D*) "Small-sample size data" versus "full sequence data" (60 individuals). The number of individuals is indicated for each population.

power of iHS (fig. 3*A* and *C*). By contrast, small sample sizes similar to those of the CG data set had a strong impact on the power of both statistics, this effect being most pronounced for iHS (fig. 3*B* and *D*). Note that the phasing process did not alter the power of iHS and DIND, as it was found to be similar with either inferred or known gametic phases (fig. 3*A*). In addition, the power was found to be similar when individual gametic phases were inferred either with MaCH or SHAPEIT (data not shown).

Overall, we found that sample size had a stronger effect on the power of these tests than coverage, which had little impact on power, with the DIND test being insensitive to even very low depth of coverage (~3×).

Robustness of iHS and DIND to Alternative Selective Regimes

Selective processes such as background selection, that is, the reduction in variability at neutral or nearly neutral sites due to selection against linked deleterious alleles (Charlesworth et al. 1997; Charlesworth 2012), can mimic the patterns left by positive selection, generating spurious "positive selection" signals in some cases (Hernandez et al. 2011). We determined the extent to which iHS and DIND were sensitive to background selection. Given that 30–40% of human nonsynonymous mutations have been suggested to be highly deleterious or lethal ($|s| > 1\%$ i.e., 2*Ns* lower than −200 in humans, see Boyko et al. [2008]), we simulated genomic regions with 20% of sites under negative selection, mimicking the selective features that can be observed in coding regions. We used various values of the population genetic selection parameter 2*Ns*, ranging from −1 to −500 (supplementary table S3, Supplementary Material online). The proportion of simulated sequences under background selection detected with iHS and DIND at a false positive rate (FPR) of 1% ranged from 0% to 2.5% (average ~1%), indicating that neither of these tests could detect this selective regime. Because the patterns of genetic variation can be the target of multiple modes of selection (Hernandez et al. 2011), we next explored whether background selection can alter the signal of a hard sweep. We tested whether negatively selected mutations segregating near positively selected variants affect the power of iHS and DIND, by simulating 100-kb regions in which a new advantageous mutation (2*Ns* = 100) was inserted in a genetic background where 20% of sites were negatively selected (see Materials and Methods). We found that background selection does not alter the power to detect selection following a hard sweep model (supplementary fig. S4, Supplementary Material online).

Alternative models of positive selection, such as polygenic adaptation or selection on standing variation, can also play an important role in adaptation, but their effects are more difficult to detect (Pritchard and Di Rienzo 2010; Pritchard et al. 2010). We evaluated the power of iHS and DIND to detect polygenic adaptation, which was modeled here as weak positive selection acting on many independent loci. This model of polygenic adaption has been proposed as an alternative model to rapid genetic adaptation, in light of the highly polygenic architecture of many traits in humans (Turchin et al. 2012). We thus simulated positive selection models with a low 2*Ns* (2*Ns* = 5, supplementary table S4, Supplementary Material online), keeping unchanged all the other parameters used to simulate hard sweeps (see Material and Methods). Neither DIND nor iHS detected a selection signal at low values of 2*Ns*, as low 2*Ns* values lead to small shifts in the frequency of the selected alleles, as predicted under a model of polygenic adaptation acting through weak selection (Pritchard and Di Rienzo 2010; Pritchard et al. 2010). Consequently, our results support the notion that conventional methods have little power to detect signatures of polygenic adaptation (Chevin and Hospital 2008; Pritchard and Di Rienzo 2010; Pritchard et al. 2010).

Finally, we performed simulations of positive selection on standing variation, that is, a neutral or mildly deleterious allele that is already segregating in the population at a frequency greater than 1/2*N* suddenly becomes positively selected and increases in frequency (Przeworski et al. 2005; Pritchard and Di Rienzo 2010; Pritchard et al. 2010). We evaluated the power of iHS and DIND for an initial frequency of the selected allele from 0.01 to 0.5 and used values of 2*Ns* ranging from 100 to 1,000 (supplementary table S5, Supplementary Material online). To do so, we used mpop software (Pickrell et al. 2009), which allows simulations only in a constant-size population model (see Materials and Methods). The power of iHS and DIND was found to decrease with increasing initial frequency of the selected allele, as high initial frequencies reduce the signature of the sweep around the selected site (Przeworski et al. 2005). For example, the power of both statistics was lower, by a factor of 4, for initial frequencies of the selected allele ≥0.2 and a 2*Ns* = 100. The application of such a decrease in power to the results of iHS and DIND obtained considering appropriate demographic histories and mimicking WGS data (fig. 3*A* and *B*) would yield a power of less than 10% for non-African samples. Moreover, no signals of positive selection on standing variation were detected (data not shown) when simulations were performed with low values of 2*Ns* (2Ns < 10), because the frequency shifts of the selected alleles were, as for polygenic adaptation by weak selection, too small to be detected.

Our simulation results demonstrate that iHS and DIND are insensitive to background selection and underpowered for the detection of polygenic adaptation or recent positive selection on standing variation when the selected allele has an initial frequency of 0.2 or above. Thus, the signals of positive selection detected by DIND and iHS in WGS data sets should reflect the effects of recent, strong positive selection targeting either a newly arisen allele (i.e., hard sweep *stricto sensu*) or standing mutations with a preselection frequency lower than 0.2 (nearly hard sweep).

Assessment of the Genome-Wide Extent of Selection Using Functional SNP Classes

To assess the extent of positive selection at the genome-wide level and to evaluate whether iHS and DIND are able to detect enrichment in selection signals in particular SNP functional classes from WGS data sets, we analyzed the 1000G and CG data sets (supplementary fig. S5 and table S6, Supplementary Material online). In classical outlier approaches, which identify SNPs presenting extreme values for a given statistic as displaying evidence of selection, the proportion of false positives remains unknown and can be high (Kelley et al. 2006; Teshima et al. 2006). Here, we overcome this caveat by applying the following rationale: if positive selection has preferentially targeted functionally important loci, then we would expect an enrichment of certain functional SNP classes among extreme values for a particular statistic (Voight et al. 2006; Barreiro et al. 2008; Jin et al. 2012). For example, it has been shown that positive selection can create strong clustering of extreme iHS values yielding strong enrichments of such extreme values within genes (Voight et al. 2006). We therefore investigated whether iHS and DIND outliers (the top 1% of values for each statistic) were more strongly enriched in putatively functional SNPs (i.e., genic or nonsynonymous SNPs) than in nongenic SNPs (supplementary table S7, Supplementary Material online). This approach, which allows quantifying the proportions of false-positive signals, should make it possible to deduce the proportion of outliers genuinely targeted by positive selection.

We thus calculated iHS and DIND for the phased data of each population of the 1000G Pilot and the CG data sets, using windows of 100 kb centered on each SNP (i.e., the core SNP) and retaining only those for which the derived state of the core SNP was unambiguously determined. We minimized the FPR by excluding windows in which the core SNP had a derived allele frequency (DAF) below 0.2, given that these tests had a power close to zero in such conditions (fig. 2 and supplementary table S2, Supplementary Material online). We assessed enrichment of SNP classes among outliers by logistic regression, generating an odds ratio (OR) for the effect of recent positive selection. If selection has occurred in genic regions, an OR $>$ 1 would be expected, reflecting the enrichment of genic SNPs among outliers (e.g., OR = 1.25 when there are 20% true and 80% false-positive SNPs among genic outliers). Otherwise (i.e., 100% of false positives among genic outliers), we would expect an OR $\leq$ 1, indicating that the proportion of genic SNPs among outliers is no greater than the proportion of genic SNPs among all SNPs (~38% for the 1000G and CG data sets, supplementary table S7, Supplementary Material online). We also controlled for various potential confounding factors, such as genomic variation in coverage, recombination rate, and the number of SNPs per window, and calculated corrected ORs (OR_C, see Materials and Methods).

With DIND, significant enrichment in genic SNPs was observed for both the 1000G Pilot and CG data sets (table 1). These enrichments were found to be robust to the confounding factors tested ($\mathrm{OR}_C > 1$) and were highly significant when compared with several genomic resamplings (table 1, see Materials and Methods). Likewise, DIND outliers displayed a greater enrichment in nonsynonymous SNPs, with respect to nongenic SNPs, although the statistical significance of this enrichment was lower due to the small number of nonsynonymous SNPs tested (table 1). In contrast with the results obtained for Africans and Europeans, almost no significant enrichment was observed for Asians from the 1000G Pilot and CG data sets ($\mathrm{OR}_C = 0.97$ and $\mathrm{OR}_C = 0.98$, table 1). In the CG data set, sample size was the smallest for the Asian population, confirming the critical nature of this experimental specification (fig. 3*B* and *D*). In the 1000G Pilot data set, the low coverage of Asians (~3×) would not be expected to mask the enrichments resulting from selection, given the results of our simulations (fig. 3*A* and *C*). We thus reasoned that another aspect of the data, such as genotype calling errors, might have decreased the power to detect selection in the pilot data. We tested this hypothesis using the 1000G Phase 1 data set, in which genotype quality and coverage were improved for African (AFR) and Asian (ASN) samples (Abecasis et al. 2012). Using this data set, we retrieved a signal in the Asian sample, which displayed highly significant enrichment ($\mathrm{OR}_C = 1.49$, table 1). This finding clearly indicates that DIND is insensitive to low coverage (e.g., 4.3× for the Phase 1 ASN sample) but highlights the ability of the genotype calling errors inherent to low-coverage data sets to wipe out the selection signal.

By contrast, no significant enrichment of genic SNPs was observed among iHS outliers in any of the various data sets studied (table 1). This finding contrasts with previous results for iHS and the HapMap genotyping data set, reporting highly significant enrichments (Voight et al. 2006). We first investigated whether our methodology (i.e., resampling scheme for significance calculation and window size) could account for such an absence of enrichment (supplementary text, Supplementary Material online). We found that the enrichment was replicated when our method was applied to the HapMap data and that the results obtained were independent of the window size used (supplementary fig. S6, tables S8 and S9, Supplementary Material online). Our results may, therefore, simply reflect an inadequacy of iHS to detect an enrichment in putative targets of selection in the specific context of the WGS data set used (i.e., low coverage and poor genotype calling quality, or small sample size). When we used the 1000G Phase 1 data set, for which the genotype calling quality was higher, we observed a slightly significant enrichment in Europeans, despite the similar mean coverage between the two 1000G data sets (4.5× for Phase 1 vs. 5.1× for Pilot) (table 1). These results suggest that iHS, which we

Table 1. Enrichment of Genic and Nonsynonymous SNPs, as Opposed to Nongenic SNPs, among iHS and DIND Outliers Calculated on 100-kb Windows.

Population	OR	Genic vs. Nongenic						Nonsynonymous vs. Nongenic					
		1000G Pilot		CG		1000G Phase1		1000G Pilot		CG		1000G Phase 1	
		DIND	iHS	DIND	iHS	DIND	iHS	DIND	iHS	DIND	iHS	DIND	iHS
AFR	OR	1.34***	0.76	1.27***	0.96	1.50***	0.87	1.23**	0.68	1.40***	0.98	1.53***	0.68
	OR_C	1.27***	0.81	1.09**	0.98	1.28***	0.92	1.18	0.72	1.17*	1.00	1.47***	0.67
EUR	OR	1.22***	0.82	1.13***	1.01	1.34***	0.96	1.25**	1.02	1.04	1.04	1.30**	1.10
	OR_C	1.22***	1.02	1.28***	1.07	1.29***	1.11*	1.22*	1.21	1.13	1.09	1.24**	1.17
ASN	OR	0.95	0.79	0.86	1.02	1.38***	0.90	1.01	0.63	0.95	0.65	1.33***	0.91
	OR_C	0.97	0.96	0.98	1.07	1.49***	1.02	1.16*	0.86	1.05	0.69	1.62***	0.98

NOTE.—OR_C indicates that the logistic regression used to calculate the OR controlled for the following confounding factors: mean recombination rate, mean coverage, and number of SNPs per window.

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

found to be slightly more sensitive to coverage than DIND (figs. 2 and 3), is also sensitive to the genotype calling quality of the data.

When several potentially confounding factors were taken into account, our analyses showed that DIND was more powerful than iHS for detecting an enrichment of certain SNP functional classes among putative targets of selection, in the context of both small sample sizes (CG data set) and low genotype calling quality/low coverage (1000G Pilot and Phase 1 data sets). Given that our simulations showed that DIND was mostly insensitive to other selective regimes, these enrichments probably reflect the action of recent, strong positive selection.

Genomic Regions Enriched in Selection Signals Present Robust Evidence of Positive Selection

To assess whether genomic regions enriched in selection signals are biologically meaningful, we considered the distribution of SNPs presenting outlier iHS and DIND values along the genome, to detect regions potentially targeted by selection, given that positive selection tends to create local clustering of outliers (Voight et al. 2006). We evaluated the extent to which these regions overlap with genes displaying identified, robust signals of positive selection. Specifically, we searched for regions presenting the greatest clustering of extreme values of iHS and DIND (i.e., the 1% of 100-kb sliding windows presenting the highest proportion of iHS or DIND outliers) in each data set, 1000G Pilot, 1000G Phase 1, and CG (supplementary tables S10–S12, Supplementary Material online). Because the number of SNPs varies across sliding windows, we grouped windows into bins presenting similar numbers of SNPs and determined the 1% highest proportion of iHS or DIND outliers separately for each bin (Voight et al. 2006). As iHS and DIND were found to have maximum power for the detection of positive selection with high 2*Ns*, our set of genes with extreme outlier clustering should overlap, to a large extent, the regions of the genome previously found to present robust signatures of strong positive selection. We thus compared it to 1) the list of genes presenting signals consistent with the hard sweep model (2*Ns* = 200–800) using iHS (Voight et al. 2006) and 2) the top list of candidate genes obtained with various linkage disequilibrium (LD)-based statistics, for example, LRH and XP-EHH (Sabeti et al. 2007). The genes with extreme DIND outlier clustering in the 1000G Pilot, 1000G Phase 1, and CG data sets included 42%, 50%, and 42%, respectively, of the top genes detected by these previous studies, whereas iHS detected only 19%, 34%, and 19%, respectively (table 2).

DIND retrieved well-known signals of positive selection, some of which were strongly supported by functional data (fig. 4*A* and *B*; supplementary figs. S7–S11, Supplementary Material online). For example, DIND consistently identified, across the three WGS data sets, the emblematic case of the rs4988235 mutation in the lactase (*LCT*) gene region, which is known to be associated with persistence of lactase activity in adulthood (Enattah et al. 2002; Bersaglieri et al. 2004; Kelley and Swanson 2008). This mutation is the core SNP of a 100-kb window located in the peak of the DIND signal and containing the second highest proportion of SNP outliers of the *LCT* region (fig. 4*A*; supplementary figs. S7–S11, Supplementary Material online). Likewise, DIND retrieved the well-known cases of the *ADH* cluster and the *EDAR* gene (table 2) (Osier et al. 2002; Carlson et al. 2005; Sabeti et al. 2007; Barreiro et al. 2008). For *EDAR*, the signal retrieved from the 1000G data set encompassed the nonsynonymous V370A mutation (rs3827760), which has been associated with hair thickness, tooth morphology, and the number of eccrine sweat glands (Fujimoto et al. 2008; Kamberov et al. 2013) (fig. 4*B*; supplementary figs. S10–S11, Supplementary Material online). Conversely, none of these emblematic cases of selection was detected by iHS in the CG and 1000G Pilot data sets (table 2, fig. 4*A* and *B*; supplementary figs. S7–S9, Supplementary Material online), with the exception of the *LCT* region in the CG data, but only if 1-Mb windows were used (supplementary fig. S9*A*, Supplementary Material online). The strong selection signals of both *EDAR* and *LCT* were restored when iHS was applied to the 1000G Phase 1 data set (supplementary figs. S10*B* and S11*A* and *B*, Supplementary Material online). These examples highlight again the sensitivity of iHS to both low coverage and the quality of genotyping calls. Consistent with the results of enrichment analyses among functional SNP classes, DIND did replicate signals of strong, recent positive selection more effectively than iHS.

Functional and Medical Relevance of Regions Enriched in Signals of Selection

To provide additional support to the adaptive significance of the genomic regions enriched in selection signals, we next investigated the extent to which these regions were enriched in SNPs that are likely to have functional consequences, that is SNPs associated with phenotype traits or disease by genome-wide association studies (here termed as GWAS-SNPs, see Materials and Methods) (Hindorff et al. 2009). Given the higher performance of DIND, with respect to iHS, with WGS data sets, we restricted our analysis to DIND outlier regions. We found a genome-wide enrichment of GWAS-SNPs in Africans and, even more so, in Europeans, as expected, given that most GWAS have been performed in populations of European descent (fig. 5*A* and *B*; supplementary table S13, Supplementary Material online). Likewise, when focusing on particular traits or diseases overrepresented among DIND outliers in each population, Europeans displayed the highest number of enriched categories (supplementary table S14, Supplementary Material online). For example, various outlier SNPs were found to be associated with skin pigmentation, such as rs1667394 A/G in *OCA2* or rs916977 A/G in *HERC2*, for which the selected allele is associated with fairer skin, hair, and eye color (supplementary table S15, Supplementary Material online). This observation is consistent with a Gene Ontology (GO) analysis in which subcategories relating to pigmentation are enriched in genes with extreme DIND outlier clustering in Europeans (e.g., melanocyte differentiation, pigment cell differentiation, supplementary table S16, Supplementary

Table 2. Overlap of Extreme iHS and DIND Outlier Clustering, Calculated from the 1000G Pilot and Phase 1 and CG Data Sets, with Regions Previously Found to Present Robust Signatures of Positive Selection.

Position	Population	Gene	Voight et al. (2006)[a]	Sabeti et al. (2007)[a]	1000G Pilot[b]		CG[b]		1000G Phase 1[b]	
					DIND e-value[c]	iHS e-value	DIND e-value	iHS e-value	DIND e-value	iHS e-value
1q23.3-q24	ASN	*BLZF1, SLC19A2*		LRH, iHS	1.0000	0.0262	0.0005	1.0000	0.0108	0.2367
1p31.3	ASN	*SLC44A5*	iHS		0.0074	0.0441	0.0037	0.0932	0.0153	0.0898
1p34.3	EUR	*NCDN, TEKT2*	iHS		0.1058	1.0000	1.0000	1.0000	0.0178	1.0000
2q13	ASN	*EDAR*		LRH, iHS, XP-EHH	0.0006	0.1026	0.0422	1.0000	0.0011	0.0013
2q21.3	ASN	*SULT1C* cluster	iHS		0.0019	0.0108	0.0011	0.1469	0.0006	0.0054
2q21.3-q22.1	EUR	*LCT*	iHS	LRH, iHS, XP-EHH	0.0002	0.1205	0.0003	0.0172	0.0002	0.0046
2p23.3	AFR	*NCOA1, ADCY3*	iHS		0.0203	0.0069	0.0169	0.0872	0.0054	0.0286
2q31.2	ASN	*PDE11A*		LRH, iHS, XP-EHH	0.0219	0.0714	0.0241	0.0009	0.0084	0.0291
	EUR				0.0106	1.0000	0.0031	0.1755	0.0197	0.0806
4p13	ASN	*SLC30A9*		LRH, iHS, XP-EHH	0.0052	0.0593	1.0000	1.0000	0.0000	0.0326
4q21-23	ASN	*ADH* cluster	iHS		0.0058	0.0735	0.0008	0.0859	0.0085	0.0395
8q11.21-23	AFR	*SNTG1*	iHS		0.0021	0.0340	0.0291	0.0222	0.0011	0.0621
	ASN		iHS		0.0021	0.0003	0.0011	0.0421	0.0014	0.0308
	EUR		iHS		0.0011	0.0764	0.0199	0.0307	0.0026	0.0089
9p22.3	ASN	*C9orf93*	iHS		1.0000	0.2567	1.0000	0.197	0.0572	0.0436
10q21.1	ASN	*PCDH15*		LRH, iHS, XP-EHH	0.0171	0.0004	0.0163	0.0024	0.0024	0.0021
12q21.2	AFR	*SYT1*	iHS		0.0008	0.0003	0.0036	0.0003	0.0017	0.0003
15q21.1	EUR	*SLC24A5*[d]		XP-EHH	NA	NA	NA	NA	NA	NA
15q22	ASN	*HERC1*		XP-EHH	9e-05	0.1508	1.0000	1.0000	0.0001	0.0233
16q22.3-q23.1	AFR	*CHST5, ADAT1, KARS*		LRH,iHS	0.0535	1.0000	0.0159	0.0085	0.0278	0.0742
	ASN				0.0997	1.0000	0.0284	0.0868	0.0815	0.0252
17q23	EUR	*BCAS3*		XP-EHH	0.0057	0.0421	0.0032	0.0891	0.0267	0.0007
20cen	AFR	*SPAG4*	iHS		0.0137	0.0698	0.0444	0.2307	0.0131	0.0374
	EUR		iHS		0.0144	1.0000	0.0010	0.011	0.0069	0.0129
20cen	ASN	*ITGB4BP, CEP2*	iHS		1.0000	0.0792	1.0000	1.0000	1.0000	0.0385
22q12.3	AFR	*LARGE*		LRH	1.0000	1.0000	0.1153	0.0284	0.0395	0.0059

[a] Empty cells correspond to regions not present in the list of the top selection targets in these studies.

[b] The genes with at least one window showing extreme proportion of outliers are underlined (the windows are grouped into bins with similar numbers of SNPs, and the 1% most extreme proportion of outliers are determined separately for each bin).

[c] The e-value is based on the calculation of the proportion of outliers within sliding windows of 100 kb, centered on each SNP (outlier clustering). The e-value is the genome-wide proportion of windows, with an outlier clustering greater than the maximum clustering value observed for the gene.

[d] For *SLC24A5*, NA indicates that no SNP with a DAF over 0.2 was found in this gene. Note that the contiguous genes *SLC12A1* and *FBN1*, which are located 80 kb and 250 kb away from *SLC24A5*, respectively, were detected using the 1000G Pilot and Phase 1 data sets.

Material online). Similarly, seven SNPs associated with height were among those presenting the strongest signals of positive selection in Europeans, five of which have been associated with increased height. Likewise, four SNPs associated with height were found to be under positive selection among Africans and four among Asians (supplementary table S17, Supplementary Material online). Finally, one SNP associated with age at menarche was among the strongest signals of positive selection in Europeans and three in Africans, all the selected alleles being associated with an older age at menarche onset (supplementary table S18, Supplementary Material online).

For disease-associated SNPs, several categories, such as immune-related diseases and cancers, were overrepresented in DIND outliers. Interestingly, for some of the GWAS-SNPs associated with immune-related disorders, we observed a clear directionality (e.g., risk or protection) of the selective pressure, with most of the selected alleles increasing disease risk (~70%, table 3; supplementary table S19, Supplementary Material online). By contrast, no clear directionality of the selection pressure was observed for GWAS-SNPs associated with other human diseases, including cancers (table 3; supplementary table S19, Supplementary Material online) or diet-related traits, such as fat metabolism and cholesterol levels (supplementary table S20, Supplementary Material online).

Discussion

The aim of this study was not to perform a hypothesis-generating genome-wide scan of selection using classical outlier approaches. The overlap of outlier loci among existing studies remains limited (Akey 2009), owing to the heterogeneity of statistics used, threshold definitions of "outlier," time frames of the selective events recovered, and high false discovery rates (FDRs) (Kelley et al. 2006; Teshima et al. 2006), emphasizing the need for studies that consider demography and other selective models. Our aim here was instead to provide a global assessment of the genome-wide prevalence of recent,

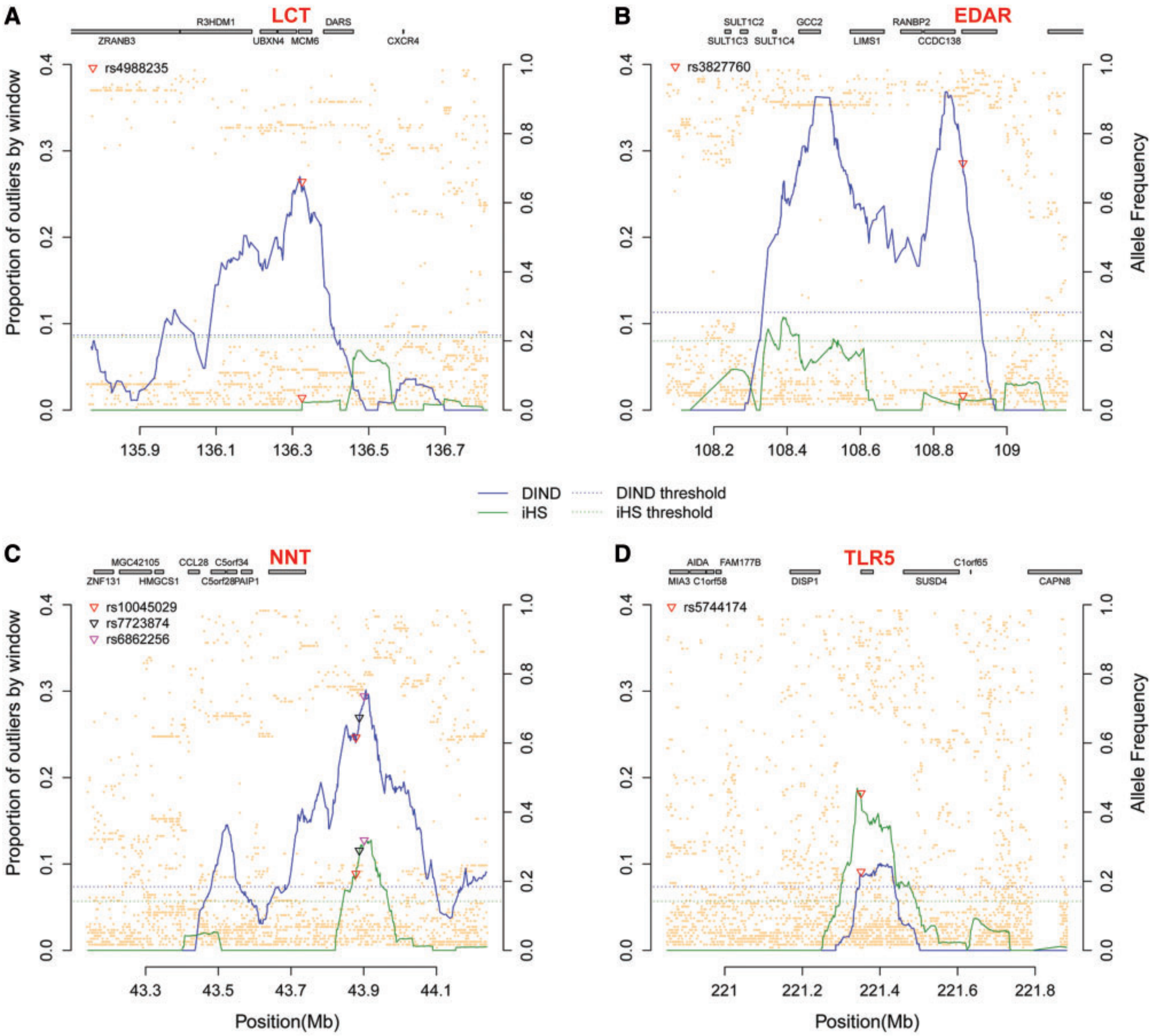

Fig. 4. Examples of candidate genomic regions under positive selection in the 1000G Pilot data set. iHS and DIND were calculated for 100-kb windows. Lines show the proportion of iHS (in green) and DIND (in blue) outliers by window. The dotted lines represent, for iHS and DIND, the threshold defining the 1% most extreme proportions of outliers by window (100 kb). The orange dots are the DAFs. The gray rectangles show the position of the genes. (*A*) *LCT*. Evidence of positive selection in the EUR population at locus 2q21, centered on SNP rs4988235, responsible for lactase persistence in adulthood (red triangle). (*B*) *EDAR*. Evidence of positive selection in the ASN population at locus 2q13, around the SNP rs3827760, associated with hair morphology (red triangle). (*C*) *NNT*. Evidence of positive selection in the AFR population at locus 5p12, implicated in familial glucocorticoid and cortisol deficiency, and particularly around the SNPs rs10045029, rs7723874, and rs6862256, associated with *NNT* expression (red, dark blue, and magenta triangles, respectively). (*D*) *TLR5* region. Evidence of positive selection in the AFR population at locus 10q24, involved in the recognition of bacterial flagellin, and, in particular, around SNP rs5744174, a nonsynonymous mutation (L616F) associated with lower levels of NF-kB signaling in response to flagellin (red triangle).

strong positive selection as a mechanism of adaptive change in humans. We found that the haplotype-based iHS and DIND statistics are both powerful to detect hard sweeps, in the context of WGS data sets and human demography, and insensitive to background selection and other modes of selection. By applying these statistics to WGS data sets, we provide evidence of positive selection targeting specific functional SNP classes, that is, enrichments of genic and nonsynonymous SNPs among selection signals, and that such selection signals are enriched in SNPs associated with phenotypic variation.

First, our simulation study showed that the haplotype-based statistics iHS and DIND are powerful to detect selection over a large range of allele frequencies. We also found that the power of these statistics remained virtually unchanged when simulating variation of mutation and recombination rate, and that it is not affected by the reconstruction of gametic phases. Notably, our simulation study demonstrated the almost total

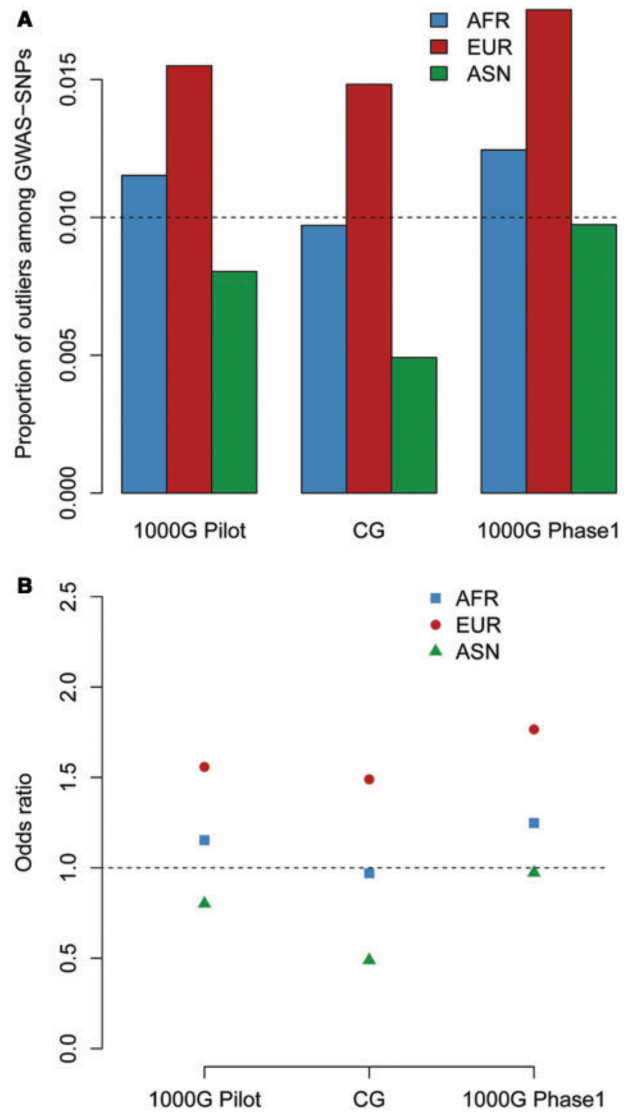

Fig. 5. Enrichment in GWAS-SNPs among DIND outliers. DIND was calculated for 100-kb windows (results for DIND calculated for 1-Mb windows are available in supplementary table S13, Supplementary Material online). GWAS-SNPs were filtered for *P* value lower than 10^{-7}. A single entry was retained for each SNP-trait association, and LD was accounted for (see Materials and Methods). (*A*) Proportion of GWAS-SNPs that are DIND outliers. Bar plots show the proportion of outliers among GWAS-SNPs for each data set and each population (from left to right, AFR, EUR, ASN). The black dotted line indicates the proportion of outliers among all the SNPs of the genome. (*B*) Enrichments of GWAS-SNPs among DIND outliers. Relative enrichment of GWAS-SNPs among DIND was measured using OR. The black dotted line corresponds to an OR equal to 1. An OR equal to or smaller than 1 indicates no enrichment. An OR greater than 1 indicates enrichment of GWAS-SNPs among DIND outliers (EUR in all data sets, and AFR in 1000G Pilot and Phase 1 data sets).

Table 3. Numbers of Selected Risk Alleles, Protection Alleles, and Nonreported Effect Alleles for Diseases for Which DIND Outliers Displayed Enrichment.

Disease categories	Population		Risk[a]	Protection[b]	NR[c]
Immune-related diseases	ALL	Count[d]	13	6	0
		Percent	68.42%	31.58%	0.00%
	AFR	Count[d]	1	2	0
		Percent	33.33%	66.67%	0.00%
	EUR	Count[d]	9	3	0
		Percent	75.00%	25.00%	0.00%
	ASN	Count[d]	3	1	0
		Percent	75.00%	25.00%	0.00%
Cancer	ALL	Count[d]	7	7	1
		Percent	46.67%	46.67%	6.67%
	AFR	Count[d]	0	4	0
		Percent	0.00%	100.00%	0.00%
	EUR	Count[d]	3	2	1
		Percent	50.00%	33.33%	16.67%
	ASN	Count[d]	4	1	0
		Percent	80.00%	20.00%	0.00%
Other diseases	ALL	Count[d]	6	8	6
		Percent	30.00%	40.00%	30.00%
	AFR	Count[d]	2	2	1
		Percent	40.00%	40.00%	20.00%
	EUR	Count[d]	2	4	3
		Percent	22.22%	44.44%	33.33%
	ASN	Count[d]	2	2	2
		Percent	33.33%	33.33%	33.33%

[a] The selected allele (derived allele) is associated with a higher risk of developing disease.

[b] The selected allele (derived allele) is not the risk allele defined in the NHGRI GWAS database.

[c] The risk allele was not reported in the NHGRI GWAS database.

[d] Counts were obtained taking LD into account.

insensitivity of DIND to low coverage (as low as 3×). Indeed, because the nucleotide diversity π is not particularly sensitive to low-frequency variants DIND is particularly insensitive to low coverage mainly affecting these low-frequency variants. Furthermore, we showed that the power to detect hard sweeps was greatest for the 1000G data set, because sample size appeared to have a stronger effect on the power of iHS and DIND than coverage variation. Indeed, the small sample size of the CG data set was not compensated by its deep coverage (~40×) for the detection of signals of strong ongoing selective sweeps.

Second, echoing the simulation results, the analysis of the WGS data sets showed that DIND performed better than iHS in the context of small sample sizes, as shown for the CG data set, and low coverage, as shown for the 1000G data set. In this context, iHS failed to replicate the enrichment of genic SNPs among outliers previously obtained with HapMap ((Voight et al. 2006) and this study) and to detect well-known signals of positive selection (Voight et al. 2006; Sabeti et al. 2007; Pickrell et al. 2009). iHS is sensitive not only to low coverage but also to genotype calling errors. Following the improvement of data quality in the 1000G Phase 1 release, the OR_C of iHS increased and reached significance in Europeans (table 1), and some of the strongest signals of selection, including *LCT* and *EDAR*, were restored. In addition, the iHS signal-to-noise ratio is lower in WGS data sets than in genotyping data sets, because extended haplotypes are more rapidly broken in the presence of low-frequency variants (Grossman et al. 2013). That low-frequency variants are more common in WGS (1000 Genomes Project Consortium 2010) could explain the absence (or weakness) of enrichment in genic SNPs within iHS outliers, while such enrichment has been observed in genome-wide SNP data sets. This would not affect DIND, as π is not particularly sensitive to low-frequency variants, consistent with significant enrichments of genic SNPs among DIND outliers only. Likewise, when simulating DNA sequence data under selection, we observed a lower clustering of outliers for iHS with respect to DIND. In addition, the breakdown of extended haplotypes by low-frequency variants is exacerbated in genic regions, where a higher proportion of low-frequency variants is observed (Abecasis et al. 2012), because moderately deleterious mutations are maintained at low frequency by negative selection (30–42% of human nonsynonymous mutations are moderately deleterious, $0.01\% < |s| < 1\%$, see (Boyko et al. 2008)). Accordingly, we obtained lower OR_C for Africans than for non-Africans with iHS, for all data sets, as expected, given that negative selection is more efficient in populations with large effective sizes.

The enrichments of genic and nonsynonymous SNPs among DIND outliers, and its substantial power to detect well-known signals of selection, provide an important proof-of-concept of the detection of genuine positive selection events in WGS data sets. For example, we identified the functionally validated selection signal at *TLR5* in Africans (Grossman et al. 2013) (fig. 4*D*; supplementary figs. S7–S11, Supplementary Material online). *TLR5* is an innate immunity receptor involved in the recognition of bacterial flagellin, highlighting the importance of the selective pressures imposed by pathogens during human evolution (Barreiro and Quintana-Murci 2010; Quintana-Murci and Clark 2013). In addition, several new signals are of particular interest because they involved SNPs predicted to be functional by other studies (supplementary tables S10–S12, Supplementary Material online). For example, we detected a strong signal on chromosome 5p12 in the African population, for all WGS data sets (fig. 4*C*; supplementary figs. S7–S11, Supplementary Material online). The peak signal was located 150 kb downstream from the *NNT* gene, and the SNPs with the highest DIND scores have been associated with *NNT* expression (i.e., expression quantitative trait loci, eQTLs) in Africans ($P = 6.4 \times 10^{-8}$; [Pickrell et al. 2010]). *NNT* has recently been implicated in familial glucocorticoid deficiency, which triggers low cortisol levels, hypoglycemia, and hyperpigmentation (Meimaridou et al. 2012). Glucocorticoids are steroid hormones that mediate homeostatic responses to environmental stressors, and these responses are known to vary among human populations (Maranville et al. 2011). Finally, among the strongest selection signals in east Asians, three gene regions have been linked to breast cancer. These include *RAD51L1* and the *ECHDC1-RNF146* region identified by GWAS (Hoggart et al. 2007; Gold et al. 2008), and *HERC1*, which has previously been reported as a selection target and is mutated in breast cancer (Grossman et al. 2013). These observations highlight the need for further studies to better understand the extent to which cancer, which is generally a rather late-onset disease, has been a selective factor by itself or a by-product of other selective forces exerting pressure on pleiotropic genes.

Further support to the adaptive significance of the genomic regions enriched in selective signals came from the overlap with GWAS, providing new insight into the relationship between past selection and benign and disease-related phenotypic variation. We found global enrichments in GWAS-SNPs among DIND outliers, supporting again the notion that we may detect true selective events from WGS data. Importantly, we were able to infer the phenotypic directionality of selective events in some cases. For example, although it has been suggested that height-associated SNPs are subject to polygenic adaptation by weak selection (Turchin et al. 2012), we detected five SNPs associated with this polygenic trait that displayed signatures of strong selection favoring high stature in European populations (supplementary table S17, Supplementary Material online). Likewise, we detected four SNPs in African and European samples for which positive selection has favored a later onset of menarche (supplementary table S18, Supplementary Material online). It has been suggested that the increasing complexity of human societies (e.g., the emergence of farming) has delayed psychosocial maturity (Gluckman and Hanson 2006) and that the occurrence of sexual maturity in psychosocially immature females is detrimental. Our analyses suggest that selection has acted to compensate for this trend by shifting sexual maturity to older ages. Importantly, we also observed a strong skew in selection, targeting alleles associated with a higher risk of immune-related diseases. Our results further support the hypothesis that the incidence of immune-related disorders in modern societies may at least partly reflect the consequences of past selection for stronger immune responses to combat infection (Barreiro and Quintana-Murci 2010; Raj et al. 2013).

More generally, our results must be seen in the context of recent debates as to the prevalence of hard sweeps in the human genome. Two recent studies have suggested that classic selective sweeps have been relatively rare during human evolution (Hernandez et al. 2011; Granka et al. 2012) and that most of these "sweeps" could be explained by the widespread action of background selection (Hernandez et al. 2011). Here we show that the two haplotype-based statistics used are robust to background selection and underpowered for the detection of positive selection events other than hard, or nearly hard, sweeps. Our results should, therefore, highlight only the occurrence of recent, strong positive selection. However, although clearly significant, the ORs obtained (OR_C) in the enrichment analyses for functional SNP classes were generally modest (1.2–1.5), supporting the notion that the prevalence of such sweeps is moderate. For example, an OR_C of 1.25 indicates that no more than 20% of candidate genes are true targets of positive selection. This observation may explain the limited overlap of outlier loci among other studies (Akey 2009), as well as between DIND and iHS in this study. However, using the OR_C of the 1000G Phase 1 data set, we can roughly estimate the number of genes under selection at approximately 70–100 in each of the different population groups. However, these numbers may represent the lower bound of genes under selection, given that the actual power to detect selection is lower than 100% and that we neglect the occurrence of selection in nongenic regions, for example, overlap of iHS and eQTL signals (Kudaravalli et al. 2009). Taken together, our results indicate that recent, hard sweeps have played a moderate, but significant, role over the last ~60,000 years of human evolution. Given that positive selection regimes other than the hard sweep model, such as polygenic adaptation by weak selection and selection on standing variation, cannot be detected by our approach, the degree of positive selection *lato sensu* acting on the human genome is undoubtedly higher than suggested here and in previous studies.

We conclude that low-coverage WGS data can be efficiently used for the detection of selective sweeps, revealing genes and functions accounting for adaptive phenotypic variation in humans or other species. The development of methods that can be safely used in the context of low-coverage data is of particular importance for the design of population genetic studies, as the sequencing of many individuals at high coverage remains costly. It is now time to refine analyses by focusing on populations living in extreme environmental conditions—high altitude, Artic climate, forest- or savannah-based populations—or with different modes of subsistence. Whole-genome sequencing of individuals from these populations, even at low coverage, should improve our understanding of the genetic basis of human adaptation to specific environments.

Materials and Methods

Data

The low-coverage part of the 1000G Pilot Project consists of data for samples from four populations: 59 unrelated Yoruba from Ibadan, Nigeria (AFR), 60 unrelated Utah residents with Western and Northern European ancestry (EUR), and 60 unrelated Asians (ASN), 30 Han Chinese from Beijing and 30 Japanese from Tokyo. All these samples were sequenced at low mean coverage: 3.7× for the AFR panel, 5.1× for EUR panel, and 2.8× for the ASN panel. In total, we analyzed 9,760,562 SNPs for AFR, 6,858,242 SNPs for EUR, and 5,674,252 SNPs for ASN. The ancestral state of each SNP was retrieved from the 1000G Project website (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_03/pilot1, last accessed April 9, 2014).

We evaluated the influence of genotype call quality by incorporating into our analyses a subset of the Phase 1 data of the 1000G Project, in which the quality of genotype calls was significantly improved. For comparison purposes, we included only individuals for whom data were already present in the Pilot release. Our subset of the 1000G Phase 1 data set consisted of 52 AFR, 45 EUR, and 58 ASN individuals (supplementary table S6, Supplementary Material online). These samples were sequenced at low mean coverage: 4.4× for the AFR panel, 4.5× for EUR panel, and 4.3× for the ASN panel. In total, we analyzed 12,848,493 SNPs for AFR, 7,577,087 SNPs for EUR, and 7,161,377 SNPs for ASN. We rendered the Pilot and Phase 1 data sets comparable, by removing the singletons from the 1000G Phase 1 data set (supplementary fig. S5, Supplementary Material online). The ancestral state of each SNP was retrieved from the 1000G Project website (ftp://ftp.ncbi.nih.gov/1000genomes/ftp/technical/working/20120316_phase1_integrated_release_version2/, last accessed April 9, 2014).

We also studied the high-coverage data of the CG public data set (software version 1.10.0.26). We selected samples from nonadmixed populations only and pooled together populations presenting close genetic affinities, to increase sample size (supplementary table S6, Supplementary Material online). We pooled together nine Yoruba from Ibadan, Nigeria, and four Luhya from Webuye, Kenya, to form a single panel of 13 unrelated Africans (AFR). We pooled together nine Utah residents with northern and western European ancestry from the centre d'etude du polymorphisme humain (CEPH) collection and four individuals from Tuscany, Italy, to form a single panel of 13 unrelated Europeans (EUR). We pooled together four Han Chinese from Beijing, China, and four Japanese from Tokyo, Japan, to form a single panel of eight unrelated Asians (ASN). All these samples were sequenced with a high mean coverage of over 50×. We removed from the analysis all SNPs presenting 5% or more low-quality Illumina calls (i.e., calls with a mapping and assembly with qualities [MAQ] mapping quality of 0). We also removed from the analysis the 11q region in which we found an accumulation of Mendelian errors. In total, we analyzed 10,070,271 SNPs for AFR, 6,281,785 SNPs for EUR, and 5,065,417 SNPs for ASN. The ancestral states of the SNPs were determined from the ancestral sequence provided by the 1000G Project and the genomes of five primates: gorilla, chimpanzee, orangutan, macaque, and marmoset (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/, last accessed April 9, 2014).

A Realistic Human Demographic History

We aimed to simulate data under a realistic demographic scenario that was tractable with the forward-in-time simulation software SFS-CODE (Hernandez 2008). We determined realistic values of key demographic parameters by comparing observed data (i.e., 20 independent 1.33-kb noncoding regions previously resequenced in 95 Africans, 60 Europeans, and 60 Asians, see Laval et al. [2010]), with simulations of various demographic scenarios obtained with SFS-CODE, summarizing each data set in terms of the mean and standard deviation of several statistics (Tajima's *D*, the number of segregating sites *S* and F_{ST}) (supplementary text and table S1, Supplementary Material online). We simulated a 1.33-kb DNA fragment with θ ($\theta = 4N\mu$, μ is the per generation per site mutation rate) and ρ ($\rho = 4Nr$, *r* is the per generation rate of recombination between adjacent loci) equal to 0.001. We simulated three populations mimicking the African, European and Asian populations and tested several scenarios by varying the age and strength of bottlenecks and expansions. For each simulation, 95 individuals were sampled from the African population, and 60 were sampled from each non-African population (10,000 simulations for each scenario). We used an ABC approach (Beaumont et al. 2002) to estimate the posterior probability of each demographic model, as previously described (Laval et al. 2010) (supplementary text, Supplementary Material online). We retained the demographic scenario with the highest posterior probability in order to perform all subsequent simulations, that is, recent selective sweep, background selection, interaction of recent selective sweep, and background selection as well as the neutral simulations that were used to determine the thresholds applied to detect selection.

Consistent with the general model of human evolution (Voight et al. 2005; Laval et al. 2010; Gravel et al. 2011), the retained scenario consisted of an ancestral African population of constant size ($N = 10{,}000$) that split into two populations (African and non-African) 60,000 years ago (fig. 1A). An expansion resulted in an instantaneous 50 times increase in the African population, 20,000 years ago. This time frame corresponds to the mean of the times corresponding to the Bantu expansions (Diamond and Bellwood 2003) and a more ancient expansion that may have occurred in Africa (e.g., ~30,000 years ago [Voight et al. 2005; Laval et al. 2010]). The bottleneck accompanying the out-of-Africa exodus caused an instantaneous decrease in the ancestral non-African population, which was halved. This population then split again into two populations (European and Asian) 20,000 years ago. Finally, both these populations underwent an instantaneous 100-fold expansion 6,000 years ago, corresponding to the Neolithic expansion (Laval et al. 2010). The migration rate (m) was set to 1.3×10^{-5} and was fixed according to what is commonly admitted concerning modern human evolution. We minimized computation time by using an ancestral effective size $N = 100$, although the effective population size for humans is generally considered to be $N = 10{,}000$. Indeed, if it is desired to simulate over t generations a population with parameter values N, μ, ρ, and s, then a simulation using instead N/λ, $\lambda\mu$, $\lambda\rho$, and λs, evolved for t/λ generations, for some $\lambda > 1$, will generate approximately the desired AFS and patterns of LD (Hoggart et al. 2007). Consequently, the AFS simulated using SFS-CODE are not affected by the simulated population size (Hernandez 2008) (see also the SFS-CODE documentation). In addition, we tested the effect of this scaling on the power of statistics based on the levels of LD surrounding a positively selected allele such as iHS and found no effect (data not shown).

Simulating Full Sequence Data

We used SFS-CODE to simulate DNA regions according to the demographic model, mutation, and recombination rates described above. We used this calibrated demographic model to perform all subsequent simulations, that is, all neutral simulations as well as those under the various models of selection investigated. For each simulation, 59 individuals were sampled from the African population and 60 from one of the two non-African populations, for matching with the 1000G Pilot samples (largest number of sampled individuals for the data sets analyzed). We first simulated neutrally evolving DNA regions and positive selection models, assuming the hard sweep model (Pritchard et al. 2010). A new advantageous mutation with a population genetics selection parameter $2Ns$ was inserted into the middle of the sequence, at a frequency of $1/2N$, in a specific population, at a specific time t. We simulated 100-kb DNA regions with $2Ns$ equal to 100, a combination of parameters, that is, length of DNA region and strength of selection, which was previously used to consistently estimate the power to detect recent positive selection in humans (Voight et al. 2006; Barreiro et al. 2009). The time t was drawn from a range of recent values (7,500; 10,000; 15,000; 20,000; 25,000; 37,500; 50,000) to obtain a large range of SAF ($0 \leq \text{SAF} \leq 1$) values, covering the frequency spectrum from 0 to 1.

We then simulated background selection. We assumed that 20% of the mutations of each 100-kb region were negatively selected, and we explored a wide range of $2Ns$ ranging from -500 to -1. We also simulated models of interaction between positive and background selection. To do so, a new advantageous mutation ($2Ns = 100$) was inserted (frequency of $1/2N$), in a specific population at a specific time t (same range of recent values as above), into the middle of a sequence, where 20% of sites were set as negatively selected with identical $2Ns$ values. We explored various $2Ns$ values including $2Ns = -1$, $2Ns = -100$, and $2Ns = -500$.

We also aimed to simulate scenarios of positive selection on standing variation. Unfortunately, to our knowledge, it is not possible to simulate positive selection on standing variation with SFS-CODE. We therefore used mpop (Pickrell et al. 2009) for forward simulations, assuming a population of constant effective size ($2N = 1{,}000$ chromosomes). Indeed, mpop can simulate positive selection on standing variation only for populations of constant size. We set the per locus mutation rate ($\theta = 4N\mu$) and the rate of recombination between adjacent loci ($\rho = 4Nr$) to 0.001, as in previous studies. We simulated standing variation scenarios by adding a selective advantage of $s = 0.1$ ($2Ns = 100$) to a previously neutral allele of frequency 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, or 0.5.

Simulating Next-Generation Sequencing Data

To simulate low-coverage data, we used the short-read simulator ShotGun (Kang et al. 2013). It allowed us to simulate 100-bp reads, with realistic read depth distribution following a negative binomial distribution, which is a Gamma mixture of Poisson. Indeed, the read depth distribution is known to follow a Poisson distribution, but stochastic and experimental limitations result in overdispersed read depths across bases. The level of overdispersion is controlled by a shape parameter set to be equal to 4 (Kang et al. 2013). The sequencing error rate specified was set to 0.001 (Shendure and Ji 2008). In order to simulate the SNP calling step, we used Thunder (Li et al. 2011), which takes into account the LD information to call genotypes. Thunder is an extension of MaCH, the genotype imputation and phase reconstruction software (Li et al. 2010). We simulated an average coverage of 4× for the African individuals, 5× for the European individuals, and 3× for the Asian individuals by using negative binomial distributions with means of 4, 5, and 3. These values correspond to the per individual average coverage calculated for the AFR, EUR, and ASN samples of the 1000G Pilot data set. The lower and upper bounds of the 99% confidence intervals are equal to 0–14 in Africa, 1–17 in Europe, and 0–11 in Asia. After simulating coverage and SNP calling steps, we then removed every singleton, as observed in the 1000G data set. In addition, for each of these simulated data sets, we reconstructed the gametic phase of each individual using Thunder/MaCH and SHAPEIT (Delaneau et al. 2012), without the use of genealogical information.

We simulated small sample sizes by randomly drawing individuals from the simulations under both neutrality and positive selection described above. We randomly drew 13 individuals from the African and one of the non-African populations and eight from the other non-African population, for matching to CG data.

Statistics

To detect mutations targeted by recent positive selection, we used the haplotype-based statistics iHS and DIND (Voight et al. 2006; Barreiro et al. 2009), which are population- and SNP-specific. They were designed to directly detect mutations targeted by recent positive selection, in contrast with other approaches (e.g., AFS-based statistics, such as Tajima's *D*) that cannot identify the local target of selection because they are calculated over a given region. In addition, iHS was designed to determine whether the ancestral or derived allelic state of each mutation has been targeted by recent positive selection, whereas DIND detects positive selection on the derived allele only. Both methods are based on the same principle: the comparison of haplotypes carrying the ancestral allele with haplotypes carrying the derived allele of a given SNP.

The iHS statistic is therefore calculated when the ancestral and derived allelic states are known unambiguously. This statistic is based on extended haplotype homozygosity (EHH) (Sabeti et al. 2002), a statistic assessing the identity of haplotypes carrying the ancestral or derived alleles of a given SNP over increasing distances. It is based on the rationale that an allele targeted by strong positive selection increases in frequency much more rapidly than a neutral allele, therefore, displaying high levels of haplotype homozygosity over much greater distances than would be expected under neutrality (indeed, the neighboring region accumulates much less recombination). More specifically, the iHS is based on the integral of the observed decay of EHH (summed in both directions away from the core SNP until EHH reaches 0.05) denoted iHH. The iHS statistic is calculated as follows:

$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left(\ln\left(\frac{iHH_A}{iHH_D}\right)\right)}{SD_p\left(\ln\left(\frac{iHH_A}{iHH_D}\right)\right)}$$

with $\mathrm{iHH_A}$ and $\mathrm{iHH_D}$ being the iHH calculated with haplotypes carrying the ancestral and derived alleles, respectively; E_p and $\mathrm{SD_p}$ are the expectation and standard deviation estimated from the empirical distribution for SNPs with a DAF p matching the frequency at the core SNP. Consequently, an extremely negative value of iHS denotes positive selection on the derived allele ($\mathrm{iHH_D} > \mathrm{iHH_A}$), whereas a highly positive iHS indicates positive selection on the ancestral allele ($\mathrm{iHH_A} > \mathrm{iHH_D}$).

The DIND is also calculated for unambiguously known ancestral and derived allelic states. This statistic is based on nucleotide diversity (π), which is used to measure the genetic diversity of haplotypes carrying the ancestral or derived allele of a given SNP. It is based on the rationale that alleles targeted by strong positive selection increase in frequency more rapidly than neutral alleles and therefore tend to have a lower nucleotide diversity than would be expected under a hypothesis of neutrality (indeed, the neighboring region accumulates fewer mutations). More specifically, DIND is the ratio $\pi_\mathrm{A}/\pi_\mathrm{D}$, with π_A and π_D being the haplotype diversity calculated with haplotypes carrying the ancestral and derived alleles, respectively. The DIND statistic is calculated as follows:

$$DIND = \frac{\pi_A}{\pi_D} = \frac{\dfrac{\sum_{i=1}^{n_A-1}\sum_{j=i+1}^{n_A} d_{ij}}{C_{n_A}^2}}{\dfrac{\sum_{k=1}^{n_D-1}\sum_{l=k+1}^{n_D} d_{kl}}{C_{n_D}^2}}$$

with n_A and n_D being the number of ancestral and derived alleles, respectively, d_{ij} being the number of differences between two haplotypes *i* and *j* carrying the ancestral allele, and d_{kl} being the number of differences between two haplotypes *k* and *l* carrying the derived allele. Consequently, very high values of DIND indicate the occurrence of positive selection on the derived allele: that is, $\pi_\mathrm{D} << \pi_\mathrm{A}$. Note that DIND was initially designed to capture selection targeting the derived allele but can easily be extended to detect positive selection targeting ancestral alleles. These two statistics require individual gametic phases (the effect on the power of these statistics of the phasing procedure used was evaluated, see Results).

Phasing the Data

As described above, the iHS and DIND statistics are based on haplotypic information and must therefore be calculated for individual gametic phases. For the low-coverage part of the 1000G data set (Pilot and Phase 1 releases), phased data were obtained from the MaCH website (Center for Statistical Genetics, University of Michigan, http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G-2010-06.html (last accessed April 9, 2014) for the Pilot, and http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G.2012-02-14.html (last accessed April 9, 2014) for the Phase 1 release). The phasing procedure imputed all missing genotypes, which were found at SNPs presenting 20% or more low-quality Illumina calls (i.e., calls with MAQ mapping quality of 0). For CG public data, the phased data were inferred with SHAPEIT, by merging populations (Delaneau et al. 2012). The phasing process was improved by the use of the Yoruba from Ibadan, Nigeria (YRI) trio to phase the AFR, and of 13 members of the same family (pedigree) to phase the EUR. Only the founders of AFR and EUR families were retained for positive selection analyses.

Power of iHS and DIND

The power was evaluated on 100-kb regions by simulations, assuming the demographic model described above. For each statistics, critical values were determined separately for each population, by neutral simulations (no selected site included in the 100-kb regions), to obtain an FPR of 1%. We calculated iHS and DIND for each mutation of the 100-kb region. As the variance and mean of the DIND statistic depend on the DAF, the values of DIND can only be compared for SNPs with similar DAF values. We therefore determined the extreme values of DIND from bins of DAFs. We grouped mutations by DAF bin (from 0 to 1, in increments of 0.025) and extracted the top 1% of DIND values for each bin. We normalized iHS by DAF bin (see equation above). For the sake of comparison, we used exactly the same procedure as for DIND. We grouped mutations by DAF bin (from 0 to 1, in increments of 0.025) and extracted the top 1% of absolute iHS values for each bin. In accordance with a previous study (Voight et al. 2006), we evaluated power on the basis of the proportion of extreme iHS or DIND values in each window. We determined the critical values defining 1% of the 100-kb regions with the highest proportion of extreme iHS or DIND values in 10^4 neutral simulations (equivalent to an FPR of 1%) for each simulated population. The power of each test to detect selection (i.e., either background selection or various regimes of positive selection) was then calculated as the proportion of simulations under selection effectively detected by this procedure (i.e., the percentage of simulations presenting proportions of extreme iHS or DIND values above the threshold defined for an FDR of 1%).

Genome-Wide Calculation of iHS and DIND and Identification of Outliers

To calculate iHS and DIND for each SNP of the WGS data sets analyzed, we first determined the ancestral and derived state of each mutation (see above). However, as these statistics are extremely sensitive to the misspecification of derived states, we calculated iHS and DIND only when the derived state was determined unambiguously. If the regions in which iHS and DIND were calculated overlapped with long gaps (>200 kb), the resulting statistics were excluded from the analysis. We carried out these calculations for 86.67%, 87.8%, and 90.71% of the mutations of the 1000G Pilot, 1000G Phase 1, and the CG data sets, respectively. Because the power of the iHS and DIND was estimated from sets of simulations over 100-kb regions, we calculated iHS and DIND over the same genomic regions of 100 kb surrounding each mutation. This ensures that we obtain values of the two statistics on strictly equivalent regions, in terms of the recombination rate, coverage, and AFS of mutations for each core SNP. We also calculated iHS and DIND over genomic regions of 1 Mb surrounding each mutation, to assess the sensitivity of our results to window size. Indeed, DIND uses information from the haplotype diversity over the entire window considered, while iHS may use information from only a part of the region concerned (i.e., iHH is computed only when EHH > 0.05, over a region whose length is mainly dependent on the intensity of selection). As previously described (Voight et al. 2006), we then extracted the 1% most extreme iHS and DIND values by using bins of DAFs (from 0 to 1, in increments of 0.025) and considered these extreme values (outliers) as potential targets of positive selection. For the identification of regions under positive selection, we focused on the degree of clustering of outliers (Voight et al. 2006). We quantified signal strength by determining the proportion of outliers recorded per 100-kb window. We binned the windows by SNP density and considered the 1% of windows with the highest proportion of outliers in each bin to be potentially under positive selection.

Enrichment in SNP Functional Classes and Resampling Method

We calculated the enrichment of genic and nonsynonymous SNPs among iHS and DIND outliers, by logistic regression, controlling for the genomic variation of certain confounding factors (Kudaravalli et al. 2009). These potential confounding factors include the coverage observed in the region surrounding an SNP (e.g., the power to detect positive selection is lower in regions with low coverage, see Results), recombination rate (Voight et al. 2006), and SNP density. We therefore retrieved these items of information for each window. The recombination rate was determined from HapMap recombination maps build 36 for the 1000G Pilot data set and HapMap recombination maps build 37 for the CG and 1000G Phase 1 data sets. We calculated the enrichment in genic and nonsynonymous SNPs from the logistic model as follows:

$$\begin{aligned} Logit[I(genic=1)] = {} & \beta_1 I(TEST_o=1) \\ & + \big[\beta_2 Cov + \beta_3 Rec + \beta_4 NbSNP \\ & + \beta_5 Cov * Rec + \beta_6 Rec * NbSNP \\ & + \beta_7 NbSNP * Cov\big] + \varepsilon, \end{aligned}$$

with I(genic = 1) being an indicator function equal to 1 if the SNP is located in a genic (nonsynonymous) region and equal to 0 otherwise, $I(\text{TEST}_o = 1)$ being an indicator function equal to 1 if the SNP shows a signal of selection (i.e., is an outlier) and equal to 0 otherwise, Rec being the mean recombination rate calculated in cM/bp, Cov being the mean coverage, and nbSNP being the number of SNPs in the window. The OR, which measures the relative enrichment of genic (nonsynonymous) SNPs among SNPs with selection signals (outliers), was estimated by exp(β_1), defined as follows:

$$OR = \left[\frac{P(genic\,|\,SEL)}{P(nongenic\,|\,SEL)}\right]\left[\frac{P(nongenic\,|\,not\ SEL)}{P(genic\,|\,not\ SEL)}\right]$$

with SEL being “with selection signal,” that is with a significant result in tests for selection ($\text{TEST}_o = 1$). The OR estimated from a logistic regression model incorporating all confounding factors and the interaction terms (see equation above) is denoted by OR_C. The odds ratio is denoted OR for logistic regression models not taking the confounding factors into account.

The P values associated with enrichment were obtained from 10,000 independent resamplings, taking into account the LD between SNPs, a source of noise that can increase the frequency of outliers in a given window. For each resampling, we drew nonoverlapping regions of 500 consecutive SNPs and arbitrarily assigned them to the genic class until we reached the number of genic SNPs observed in each population. We considered the remaining SNPs to be nongenic and calculated the OR for each resampling. To resample nonsynonymous SNPs, we first determined the distribution of the number of nonsynonymous SNPs per windows of 500 SNPs. We next drew nonoverlapping regions of 500 consecutive SNPs, and randomly assigned a number of SNPs to the nonsynonymous class so as to fit the real distribution, until we reached the number of nonsynonymous SNPs in each population. Considering the remaining SNPs to be nongenic, we calculated the OR for each resampling. For the calculation of the *P* values for OR_C, we first applied a linear regression to the iHS and DIND values, taking into account the same confounding factors.

$$STAT = C + \alpha_1 Cov + \alpha_2 Rec + \alpha_3 NbSNP + \alpha_4 Cov * Rec + \alpha_5 Rec * NbSNP + \alpha_6 NbSNP * Cov + \varepsilon$$

We then used the residual values (ε) to extract the outliers, before applying the resampling method.

GeneTrail and GWAS Analysis

We used the GeneTrail online tool (http://genetrail.bioinf.uni-sb.de, last accessed April 9, 2014) to analyze the enrichment of some GO biological functions (Ashburner et al. 2000) among DIND outliers. This made it possible to analyze the overrepresentation of each GO category among the outliers by comparing our sets of genes under positive selection with the human reference gene set. An FDR adjustment was applied to correct for multiple testing, and the significance threshold was fixed at 0.05.

The National Human Genome Research Institute (NHGRI) database (http://www.genome.gov/gwastudies/, last accessed April 9, 2014) summarizes results from all published genome-wide association (GWA) analyses for which the *P* values are below 1.0×10^{-5} (Hindorff et al. 2009). We first filtered the database to remove associated SNPs for which *P* values were greater than 1.0×10^{-7}, retaining a single entry for each SNP-trait association. We then calculated the proportion of GWAS-SNPs among DIND outliers by accounting for LD. To this end, all the SNPs associated with the same trait or disease and with the same outlier/nonoutlier status in a genic region were counted as one association. These genic regions were determined with the "mapped gene" field of the database. We then compared these results with those expected under neutrality (0.01 vs. 0.99). ORs were calculated for all associated SNPs together and by trait and disease category.

Supplementary Material

Supplementary text, supplementary figures S1–S11, and tables S1–S20 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Pierre Boutillier for assistance with data analyses, and Katie Siddle and Eddie Loh for discussions. This work was supported by the Institut Pasteur, the Centre Nationale de la Recherche Scientifique (CNRS), the Ecole Normale Supérieure de Lyon, and by the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement No. 281297.

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* 19:711–722.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12:1805–1814.

Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25:25–29.

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B, et al. 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* 5:e1000562.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet.* 40:340–345.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet.* 11:17–30.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

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.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4:e1000083.

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15: 1553–1565.

Casto AM, Li JZ, Absher D, Myers R, Ramachandran S, Feldman MW. 2010. Characterization of X-linked SNP genotypic variation in globally distributed human populations. *Genome Biol.* 11:R10.

Charlesworth B. 2012. The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the *Drosophila* X chromosome. *Genetics* 191: 233–246.

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res.* 70:155–174.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res.* 20:393–402.

Chevin LM, Hospital F. 2008. Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics* 180: 1645–1660.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009. The role of geography in human adaptation. *PLoS Genet.* 5:e1000500.

Crawford JE, Lazzaro BP. 2012. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Front Genet.* 3:66.

Crisci JL, Poh YP, Mahajan S, Jensen JD. 2013. The impact of equilibrium assumptions on tests of selection. *Front Genet.* 4:235.

Delaneau O, Marchini J, Zagury JF. 2012. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 9:179–181.

Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. *Science* 300:597–603.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78–81.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat Genet.* 30:233–237.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–861.

Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, et al. 2008. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet.* 17: 835–843.

Gluckman PD, Hanson MA. 2006. Evolution, development and timing of puberty. *Trends Endocrinol Metab.* 17:7–12.

Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, Friedman E, Narod S, Olshen AB, Gregersen P, et al. 2008. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc Natl Acad Sci U S A.* 105: 4340–4345.

Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. 2012. Limited evidence for classic selective sweeps in African populations. *Genetics* 192:1049–1064.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A.* 108:11983–11988.

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, et al. 2013. Identifying recent adaptations in large-scale genomic data. *Cell* 152:703–713.

Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* 24: 2786–2787.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331:920–924.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* 106:9362–9367.

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.

Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ. 2007. Sequence-level population simulations over large genomic regions. *Genetics* 177:1725–1731.

Jin W, Xu S, Wang H, Yu Y, Shen Y, Wu B, Jin L. 2012. Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Res.* 22:519–527.

Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, et al. 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* 152:691–702.

Kang J, Huang KC, Xu Z, Wang Y, Abecasis GR, Li Y. 2013. AbCD: arbitrary coverage design for sequencing-based genetic studies. *Bioinformatics* 29:799–801.

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.

Kelley JL, Swanson WJ. 2008. Positive selection in the human genome: from genome scans to biological significance. *Annu Rev Genomics Hum Genet.* 9:143–160.

Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. 2009. Gene expression levels are a target of recent natural selection in the human genome. *Mol Biol Evol.* 26:649–658.

Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 5: e10284.

Li H. 2011. A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Mol Biol Evol.* 28: 365–375.

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. 2011. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* 21:940–951.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol.* 34:816–834.

Maranville JC, Baxter SS, Torres JM, Di Rienzo A. 2011. Inter-ethnic differences in lymphocyte sensitivity to glucocorticoids reflect variation in transcriptional response. *Pharmacogenomics J.* 13: 121–129.

Meimaridou E, Kowalczyk J, Guasti L, Hughes CR, Wagner F, Frommolt P, Nurnberg P, Mann NP, Banerjee R, Saka HN, et al. 2012. Mutations in *NNT* encoding nicotinamide nucleotide transhydrogenase cause familial glucocorticoid deficiency. *Nat Genet.* 44: 740–742.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15:1566–1575.

Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci.* 365: 185–205.

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

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19:826–837.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464:768–772.

Pritchard JK, Di Rienzo A. 2010. Adaptation—not by sweeps alone. *Nat Rev Genet.* 11:665–667.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20:R208–R215.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59:2312–2323.

Quintana-Murci L, Clark AG. 2013. Population genetic tools for dissecting innate immunity in humans. *Nat Rev Immunol.* 13: 280–293.

Raj T, Kuchroo M, Replogle JM, Raychaudhuri S, Stranger BE, De Jager PL. 2013. Common risk alleles for inflammatory diseases are targets of recent positive selection. *Am J Hum Genet.* 92:517–529.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol.* 26:1135–1145.

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.

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

Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Hirschhorn JN. 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet.* 44: 1015–1019.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A.* 102:18508–18513.

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

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. 2005. Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* 15:1468–1476.

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.