Natural selection has driven population differentiation in modern humans

Article *in* Nature Genetics · April 2008

DOI: 10.1038/ng.78 · Source: PubMed

5 authors, including:

Guillaume Laval
Institut Pasteur International Network
120 PUBLICATIONS **25,271** CITATIONS

SEE PROFILE

Helene Quach
Institut Pasteur
90 PUBLICATIONS **3,953** CITATIONS

SEE PROFILE

Etienne Patin
French National Centre for Scientific Research
175 PUBLICATIONS **3,599** CITATIONS

SEE PROFILE

Lluis Quintana-Murci
Institut Pasteur
375 PUBLICATIONS **11,662** CITATIONS

SEE PROFILE

nature genetics

Natural selection has driven population differentiation in modern humans

Luis B Barreiro[1,2], Guillaume Laval[1,2], Hélène Quach[1], Etienne Patin[1] & Lluís Quintana-Murci[1]

The considerable range of observed phenotypic variation in human populations may reflect, in part, distinctive processes of natural selection and adaptation to variable environmental conditions. Although recent genome-wide studies have identified candidate regions under selection[1–5], it is not yet clear how natural selection has shaped population differentiation. Here, we have analyzed the degree of population differentiation at 2.8 million Phase II HapMap single-nucleotide polymorphisms[6]. We find that negative selection has globally reduced population differentiation at amino acid–altering mutations, particularly in disease-related genes. Conversely, positive selection has ensured the regional adaptation of human populations by increasing population differentiation in gene regions, primarily at nonsynonymous and 5′-UTR variants. Our analyses identify a fraction of loci that have contributed, and probably still contribute, to the morphological and disease-related phenotypic diversity of current human populations.

Natural selection can act at the level of genes, if particular genotypes allow for increased fitness in specific environments. For example, there is evidence that the population prevalence of some human phenotypes, such as resistance to malaria or lactose tolerance in adulthood, results from natural selection in response to idiosyncratic conditions[7,8]. In this study, we aimed to evaluate, at the genome-wide scale, the impact of natural selection on worldwide population differentiation and to identify the type of genetic variants preferentially targeted by selection. We applied a statistical approach that considers the degree of population differentiation (F_{ST})[9,10] (**Supplementary Note** online) at single nucleotide polymorphisms (SNPs) throughout the genome, with respect to the physical location and functional impact of these SNPs. Under an assumption of neutrality, F_{ST} is determined by demographic history (that is, genetic drift and gene flow), which affect all loci similarly. By contrast, natural selection acts in a locus-specific manner: negative or balancing selection tends to decrease F_{ST}[11] (**Supplementary Fig. 1** online), whereas local positive selection tends to increase F_{ST}[11]. We hypothesized that selection preferentially targets genic over nongenic regions. We also reasoned that variants leading to amino-acid changes (nonsynonymous mutations) or located in *cis*-regulatory regions (5′ UTR and 3′ UTR) would be under stronger selective pressure than 'silent' genic mutations (synonymous and intronic variants). We estimated F_{ST} for more than 2.8 million Phase II HapMap SNPs[6]. The entire dataset was divided into the following SNP classes: nongenic, genic, intronic, 5′ UTR, 3′ UTR, synonymous and nonsynonymous (**Supplementary Note**). This genome-wide approach is novel in that it compares different SNP classes that are equally influenced by demography. Therefore, any deviation in the degree of population differentiation between SNP classes should be attributable to selection.

The estimated mean F_{ST} values for the different SNP classes were similar ($\sim$0.11) and concordant with genome-wide estimates[12,13] (**Supplementary Note**). However, we detected significant differences in the fraction of SNPs presenting low F_{ST} values among different SNP classes. Overall, genic SNPs presented a significant excess of low F_{ST} values ($F_{ST} < 0.05$) with respect to nongenic SNPs (χ^2 test, $P = 3.1 \times 10^{-11}$; **Fig. 1a**,**b**). Notably, this excess was particularly marked for nonsynonymous SNPs (χ^2 test, $P = 2.0 \times 10^{-67}$). However, heterogeneous ascertainment bias between different SNP classes, particularly for nonsynonymous SNPs, can complicate inferences of natural selection[14]. To test whether this ascertainment bias could explain the observed excess of low F_{ST} among nonsynonymous SNPs, we restricted our analyses to those SNPs that were discovered using a genome-wide homogeneous resequencing scheme and that were genotyped without regard to gene location, spacing or frequency—the 'class A' SNPs from Perlegen[15] (**Supplementary Note**). Using this homogeneously biased dataset, we observed a consistent excess of low F_{ST} values among nonsynonymous SNPs (χ^2 test, $P = 8.7 \times 10^{-8}$, **Fig. 1c**). Thus, the lower degree of population differentiation observed among nonsynonymous SNPs, which cannot be accounted for solely by ascertainment bias, can be explained by negative and/or balancing selection. We thus sought to determine the range of allele frequencies associated with the excess of low F_{ST} values by comparing nongenic and nonsynonymous SNPs matched for bins of global minor allele frequency (MAF). We observed that, for both datasets, the excess of low-F_{ST} nonsynonymous SNPs was restricted to low-frequency bins (**Fig. 2**); excess of low-F_{ST} nonsynonymous SNPs was not apparent in intermediate-frequency bins, as would have been expected under balancing selection. This excess seems to be primarily

[1]Human Evolutionary Genetics Unit, Centre National de la Recherche Scientifique-Unité de Recherche Associée (CNRS-URA3012), Institut Pasteur, 25 rue Dr. Roux, Paris 75015, France. [2]These authors contributed equally to this work. Correspondence should be addressed to L.Q.-M. (quintana@pasteur.fr).

Received 25 April 2007; accepted 11 December 2007; published online 3 February 2008; doi:10.1038/ng.78

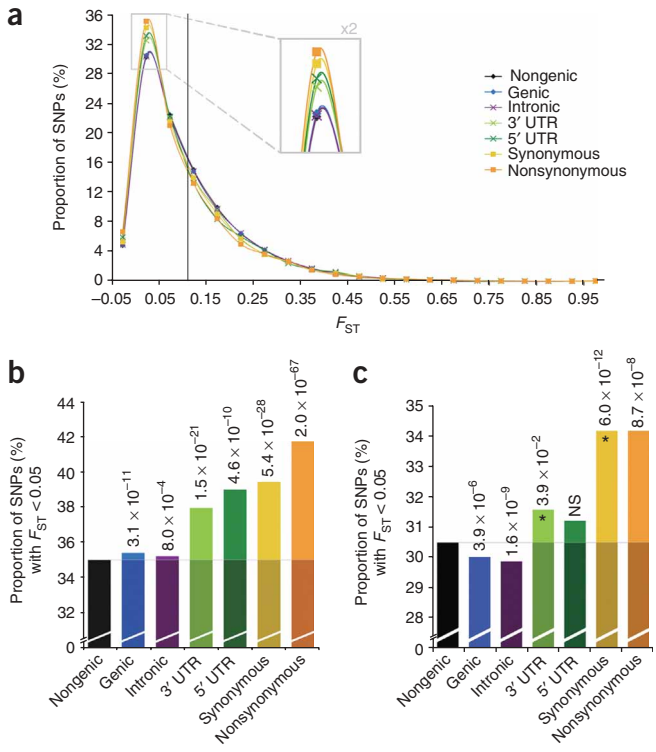

Figure 1 Consistent enrichment of nonsynonymous SNPs showing low degrees of population differentiation (F_{ST}). (**a**) Global F_{ST} distribution among the four HapMap populations for each SNP class. The vertical line indicates the genome-wide mean F_{ST} value ($F_{ST} \sim 0.11$). (**b**) Observed excess of low F_{ST} values for the different SNP classes, with respect to nongenic regions, using the global Phase II HapMap dataset. (**c**) Observed excess or deficit of low F_{ST} values for the different SNP classes, with respect to nongenic regions, when we restricted the analyses of the HapMap dataset to the Perlegen 'class A' SNPs ('restricted HapMap dataset'). Asterisks (*) indicate that the observed significant increases of low F_{ST} values for these two SNP classes were not replicated when we analyzed the Perlegen dataset *per se*.

due to an excess of rare variants among nonsynonymous SNPs (**Supplementary Note**). Altogether, the most plausible explanation for the lower levels of population differentiation observed among nonsynonymous mutations is that negative selection acts to maintain the status quo of essential proteins.

We subsequently predicted the effects of the 15,259 HapMap nonsynonymous SNPs[6] on fitness (benign, possibly damaging, or probably damaging) using the Polyphen algorithm[16]. Consistent with negative selection, mutations identified as possibly or probably damaging were significantly more heavily represented among low-F_{ST} SNPs (χ^2 test, $P \leq 6.0 \times 10^{-4}$, **Fig. 3a**). This result is attributable primarily to the observed lower population frequencies of 'damaging' mutations in the human genome (*t*-test, $P \leq 4.6 \times 10^{-20}$, **Fig. 3b,c**). Thus, by retaining damaging variants at low population frequencies, negative selection has not allowed them to differentiate as much as they could under neutral conditions (**Supplementary Note**). Our genome-wide results further support previous studies that, on the basis of the site-frequency spectrum of 106 and 301 human genes[17,18], proposed that negative selection acts on deleterious mutations. We then evaluated the direct impact of low-F_{ST} nonsynonymous variants on human health by retrieving the Online Mendelian Inheritance of Man (OMIM) morbidity status of the corresponding genes for each nonsynonymous SNP. Low-F_{ST} nonsynonymous SNPs were significantly more frequent in genes known to modulate disease (χ^2 test, $P = 6.4 \times 10^{-7}$, **Supplementary Fig. 2** online). Thus, low-F_{ST} nonsynonymous SNPs—particularly those predicted to be 'damaging'—are probably deleterious and may be of special interest in medical research.

We next investigated the impact of local positive selection on population differentiation by testing for an excess of high F_{ST} values among different SNP classes. We measured the deviation (λ) between the expected and observed proportions of each SNP class in the various F_{ST} bins (**Supplementary Note**). High-F_{ST} bins were significantly enriched in genic SNPs: the proportion of genic SNPs with $F_{ST} > 0.65$ was 1.36-fold higher than expected under neutrality (χ^2 test, $P = 9.0 \times 10^{-24}$; **Fig. 4**). However, a higher gene density surrounding high-F_{ST} genic SNPs could have contributed to the observed excess of high F_{ST} among this SNP class, as a result of genetic hitchhiking. In this case, a single event of selection extending into neighboring genes would increase the overall proportion of genic SNPs presenting high F_{ST}. We compared the gene density around high-F_{ST} genic SNPs with respect to that around average-F_{ST} genic SNPs. No significant correlation was observed between gene density and F_{ST} values (**Supplementary Fig. 3** online), reinforcing a genuine excess of selective events among genic SNPs with high F_{ST}. This excess was accounted for primarily by a disproportionate number of nonsynonymous and 5′-UTR SNPs, which present a 2.61-fold increase for nonsynonymous SNPs (χ^2 test, $P = 1.0 \times 10^{-13}$) and a 2.42-fold increase for 5′-UTR SNPs (χ^2 test, $P = 1.1 \times 10^{-4}$) in the proportion of SNPs presenting $F_{ST} > 0.65$ (**Fig. 4c**). We controlled again for potentially varying ascertainment bias associated with different HapMap SNP classes by restricting our analyses to the 'class A' SNPs from Perlegen[15]. We observed a consistent 3.9-fold increase for nonsynonymous SNPs (χ^2 test, $P = 4.3 \times 10^{-12}$) and a 1.9-fold increase for

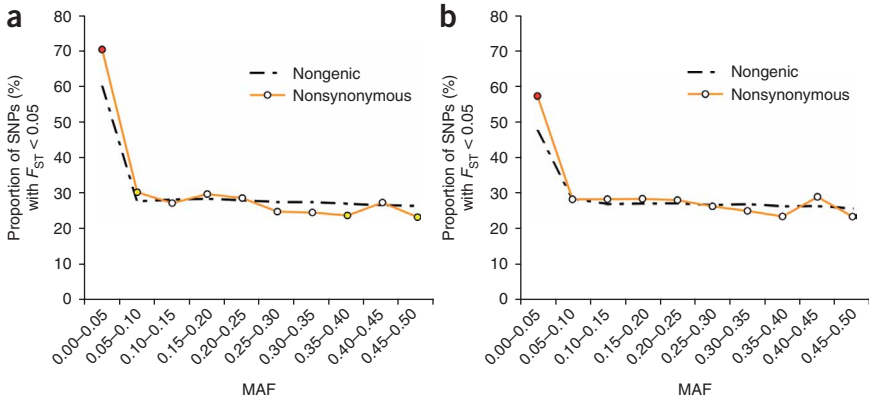

Figure 2 Enrichment of nonsynonymous SNPs presenting low F_{ST} among low-frequency variants. (**a**,**b**) Observed excess of low F_{ST} values for nonsynonymous SNPs with respect to nongenic SNPs when constraining the analyses to SNPs presenting the same global MAF estimated over the four HapMap populations, for the entire Phase II HapMap dataset (**a**) and the restricted HapMap dataset (**b**). The colors of the circles indicate statistical significance: white (not significant), yellow ($P < 0.05$), green ($P < 1 \times 10^{-3}$), and red ($P < 1 \times 10^{-10}$).

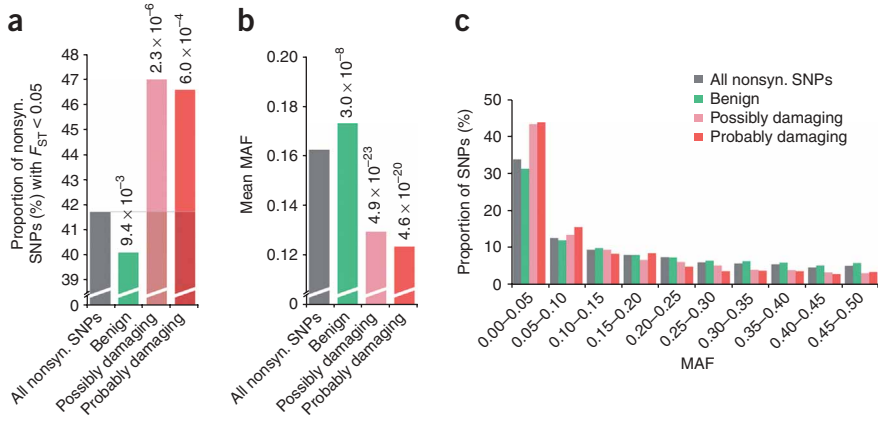

Figure 3 Imprints of negative selection in the human genome. (**a**) Observed excess of low F_{ST} values for the different SNP fitness categories predicted by Polyphen, with respect to all nonsynonymous SNPs. (**b**) Mean MAF among all populations, for the different SNP fitness categories, with respect to all nonsynonymous SNPs. (**c**) Global distribution of MAFs for the different SNP fitness categories. The observed genome-wide excess of low-frequency variants—particularly those with MAF lower than 0.05—among damaging mutations is also observed when considering single populations separately (data not shown).

5′-UTR SNPs (χ^2 test, $P = 0.18$) in the proportion of SNPs presenting $F_{ST} > 0.65$ (**Fig. 4d** and **Supplementary Fig. 4** online). The nonsignificance of the excess of 5′-UTR SNPs among high F_{ST} values is explained by the limited number of 5′-UTR SNPs (1,612 SNPs) in this replication process. Finally, the finding of excess of genic SNPs, and particularly nonsynonymous SNPs, with high F_{ST} was replicated when we constrained the analyses for both datasets to SNPs presenting similar global allele frequencies (**Fig. 5**). These observations are consistent with the recent Phase II HapMap data, which reported an excess of high F_{ST} (>0.5) among nonsynonymous SNPs with respect to synonymous SNPs when matching for similar derived allele frequencies[6].

All things considered, and after excluding a number of potentially confounding factors, we conclude that the observed excess of strong population differentiation in genic SNPs, particularly in nonsynonymous and 5′-UTR variants, must therefore result from the action of local positive selection. Notably, the signature of positive selection observed at these SNP classes was not restricted to a single population or a broad geographic area; instead, it was observed in all study populations, as attested by the similar results obtained when using population-pairwise F_{ST} estimates (**Supplementary Fig. 5** online). Additional support for our conclusions comes from the observation that genic SNPs, and particularly nonsynonymous variants, are significantly enriched for long-range haplotypes with respect to nongenic SNPs (data not shown). In parallel, we observed a significant excess of long-range haplotypes among genic and nonsynonymous SNPs presenting high F_{ST} with respect to all genic and nonsynonymous SNPs considered together (**Supplementary Fig. 6** online).

Classical outlier approaches to detect natural selection across the genome are limited in that they cannot quantify the proportion of genomic regions presenting extreme values for a given statistic that are real targets of selection[19–22]. Our approach—comparing whole-genome F_{ST} distributions between different functional classes of SNPs—showed that at least 60% (lighter color, **Fig. 4c**) of the genes presenting extreme levels of population differentiation for nonsynonymous and 5′-UTR variants (**Table 1**) are indeed under positive selection. Notably, an appreciable fraction of the genes identified by our analyses as being under positive selection has been shown to be associated with long-range haplotypes[3], on the basis of the LRH[23], the

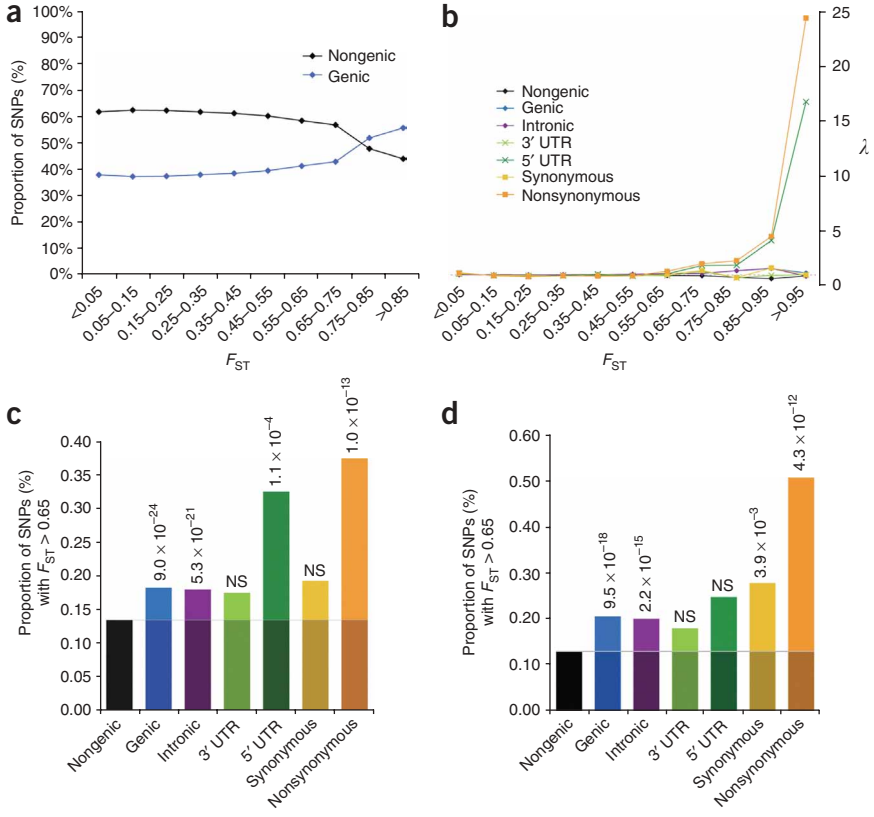

Figure 4 Imprints of positive selection in the human genome. (**a**) Enrichment of genic SNPs among high-F_{ST} bins. (**b**) Deviation (λ) between the expected and observed proportions of each SNP class per F_{ST} bin. Under neutral conditions, we expect the proportion of each SNP class to be maintained in each bin of the global F_{ST} distribution. For example, if nonsynonymous SNPs account for 0.54% of the 2.8 million SNPs analyzed, this proportion should be constant for all F_{ST} bins ($\lambda = 1$). A significant distortion of λ ($\lambda > 1$ or $\lambda < 1$) indicates natural selection. (**c**,**d**) Observed excess of high F_{ST} values for the different SNP classes, with respect to nongenic regions, using the entire Phase II HapMap dataset (**c**) and the restricted HapMap dataset (**d**).

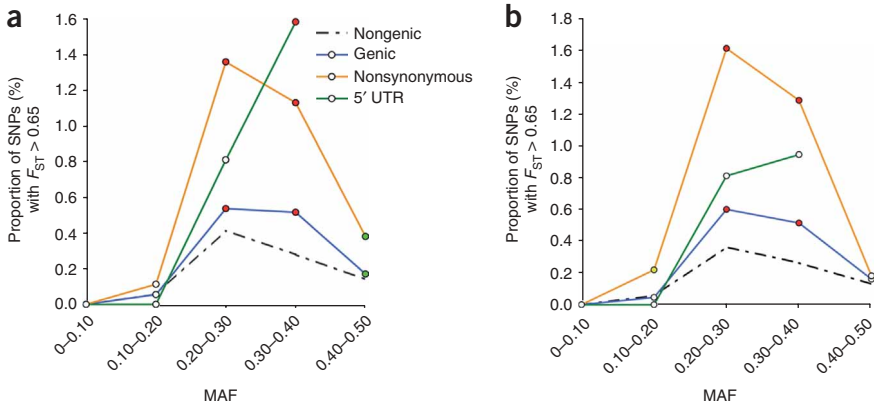

Figure 5 Enrichment of genic SNPs presenting high F_{ST} when matching for different allele frequency bins. (**a**,**b**) Observed excess of high F_{ST} values among genic SNPs, particularly nonsynonymous and 5′-UTR variants, with respect to nongenic SNPs when constraining the analyses to SNPs presenting the same global MAF estimated over the four HapMap populations for the entire Phase II HapMap dataset (**a**) and the restricted HapMap dataset (**b**). The colors of the circles indicate statistical significance: white (not significant), yellow ($P < 0.05$), green ($P < 1 \times 10^{-3}$), and red ($P < 1 \times 10^{-10}$).

iHS[4] and/or the newly developed XP-EHH tests[3] (**Table 1** and **Supplementary Table 1** online). Because long-range haplotypes persist for relatively short time periods (<30,000 years)[21], genes presenting high F_{ST} together with significant long-range haplotypes should correspond to those genes that have been hit by more recent positive selection, but that present a selective coefficient strong enough to explain the high levels of population differentiation we observed.

Of note, among the highly differentiated genes with known functions, several control variable morphological traits in humans (**Table 1**). Furthermore, most of these genes are pleiotropic: that is, they are individually involved in several different traits. For example, *EDAR* regulates hair follicle density and the development of sweat glands and teeth in humans and mice[24,25]. In humans, selective pressures on *EDAR* favoring changes in body temperature regulation and hair follicle density in response to colder climates may have influenced tooth shape, although this trait probably does not affect population fitness. This anecdotal example shows how 'phenotypic hitchhiking' in genes under positive selection may have substantially increased the observed number of physiological and morphological traits differentiating modern human populations.

Table 1 Genes showing the strongest signatures of positive selection

Phenotype category	Genes
Morphological traits (for example, skin pigmentation and hair development)	*ABCC11*, ***EDAR***, *SLC45A2*, *PKP1*, *PLEKHA4*, ***SLC24A5***
Immune response to pathogens	*CEACAM1*, *CR1*, ***DUOX2***, *VAV2*
DNA repair and replication	*MPG*, *POLG2*, *TDP1*
Sensory functions (for example, olfaction and eye development)	*COL18A1*, *OR52K2*, *RP1L1*
Insulin regulation, metabolic syndrome (obesity, diabetes, hypertension)	*ALMS1*, *CEACAM1*, *ENPP1*
Various metabolic pathways (for example, ethanol, intestinal zinc and citrulline)	***ADH1B***, *ASS1*, *SLC39A4*
Miscellaneous	*FBXO31*, *RTTN*, ***SPAG6***
Unknown	*ABCC12*, ***ADAT1***, *AK127117*[a], C17orf46, C8orf14, COLEC11, ***CPSF3L***, *DNAJC5B*, *DNHD1*, *ETFDH*, *EXOC5*, ***FAIM***, ***CCDC142***[b], *FLJ37464*[a], *FXR1*, *GCN5L2*, *KIAA0984*[a], *LAMB4*, *LOC648511*[a], ***LIMCH1***, ***PCGF1***[b], *PLEKHG4*, *POL3S*[a,c], ***RNF135***, ***SLC30A9***, *SYTL3*, *TEX15*, ***TTC31***[b], *VPS33B*, *ZNF646*[c]

These genes contain at least one nonsynonymous or 5′-UTR mutation with $F_{ST} > 0.65$. An exhaustive list of 582 genes containing other classes of genic SNPs with $F_{ST} > 0.65$ is provided in **Supplementary Table 1**. Genes in bold correspond to those also presenting significant long-range haplotypes, as measured by the iHS statistic[4], or defined as top candidates for recent selective sweeps[3].
[a]These genes have not yet been attributed a HUGO-approved symbol. [b]These three genes are located in a linkage-disequilibrium block in chromosome 2. [c]These two genes are located in a linkage-disequilibrium block in chromosome 16.

Genes under positive selection are thought to have an important role in human survival and to affect complex phenotypes of medical relevance. Indeed, as reported for negative selection, nonsynonymous SNPs showing signs of positive selection are observed in genes involved in disease more frequently than expected (χ^2 test, $P = 1.0 \times 10^{-9}$, **Supplementary Fig. 2**). For example, we observed a missense mutation in the *CR1* gene, the derived state of which has a frequency of 85% in Africans, but which is absent elsewhere (rs17047661; $F_{ST} = 0.85$, **Supplementary Note**). As this gene modulates the severity of malarial attacks in Papua New Guineans[26], our analysis strongly suggests that this particular *CR1* mutation has been positively selected for in Africans because it modifies host susceptibility to malaria. Another important selective pressure that has confronted modern humans is adaptation to variable nutritional resources. Several genes involved in the regulation of insulin and in metabolic syndrome seem to have undergone positive selection (**Table 1**). For example, *ENPP1* harbors a mutation with a derived state known to protect against obesity and type II diabetes[27] that is present in ~90% of non-Africans but virtually absent in Africans (rs1044498; $F_{ST} = 0.77$, **Supplementary Note**). *ENPP1* and several other examples of derived protective alleles[28] indicate that, in contrast to the situation with mendelian diseases, alleles that increase complex disease risk are not necessarily new mutations, but rather ancestral alleles that have become disadvantageous after changes of environment and lifestyle.

In conclusion, we have identified a fraction of loci that have influenced the morphological and disease-related phenotypic diversity characterizing modern human populations. These results open multiple avenues for future research, as they may facilitate genetic explorations of medical conditions by identifying strong candidate genes for diseases in which prevalence depends on ethnic background. The next step will be to determine how genetic variation in loci found to be under selection, particularly in those genes of unknown function, modulates susceptibility to or the pathogenesis of human disease.

METHODS

HapMap data. We analyzed genome-wide data from release 20 of the International HapMap Project Phase II[6]. For our analysis, we considered only unrelated individuals. The population panel consisted of 60 Yoruba from Ibadan (Nigeria), 60 individuals of northwestern European ancestry, 45 Han Chinese from Beijing and 45 Japanese from Tokyo. We retained only SNPs that successfully genotyped in all four populations and that were polymorphic in at least one of the study populations. When considering the global Phase II HapMap dataset, we analyzed a total of 2,841,354 autosomal polymorphic SNPs (**Supplementary Note**). When restricting the analyses of HapMap data to the Perlegen 'class A' SNPs[15] (the so-called 'restricted HapMap dataset'), we analyzed a total of 851,846 SNPs (**Supplementary Note**).

SNP classes and annotation. We partitioned the global Phase II HapMap SNP dataset[6] according to the physical location and functional impact of SNPs. We assigned SNPs to two major classes: genic and nongenic SNPs. For genic SNPs, we further classified the mutations as intronic, 5′ UTR, 3′ UTR, synonymous or nonsynonymous. We determined function-class annotations for each SNP using the ENSEMBL gene model, and systematically verified them using the dbSNP classification. The results from ENSEMBL and dbSNP classification were highly concordant for all SNP classes, except for the class of UTR SNPs, where the concordance rate was 69%. To test whether this lower concordance would influence our conclusions regarding UTR SNPs, we replicated our analyses for these SNP classes by considering only UTR SNPs overlapping between the ENSEMBL and dbSNP classifications. All our conclusions remained unaltered (data not shown).

Estimates of F_{ST}. As all measures of population genetic distances are known to be highly correlated[12], we decided to use the F_{ST} estimate derived from ANOVA[10]. This estimate is equivalent to the unbiased estimates of F_{ST} described by Weir and Cockerham[9], when considering individual SNPs, as in our study. We calculated the F_{ST} for each single SNP among the four HapMap populations by considering three hierarchical levels: population, individuals within the population, and genotypes within individuals. F_{ST} is estimated as the proportion of genetic variance explained by population level. Considering S populations, F_{ST} can be estimated as follows:

$$F_{ST} = \frac{\sigma_A^2}{\sigma_T^2}$$

with

$$\sigma_A^2 = (MSD_{AP} - MSD_{AI/WP})/n_C$$

and

$$\sigma_T^2 = (MSD_{AP} - MSD_{AI/WP})/n_C + (MSD_{AI/WP} - MSD_{WI})/2 + MSD_{WI}$$

where

$$n_C = \left(\sum_i n_i - \frac{\sum_i n_i^2}{\sum_i n_i} \right) / (S-1)$$

Here, MSD_{AP} denotes the observed mean square deviation among populations, $MSD_{AI/WP}$ denotes the observed mean square deviation among individuals within the population, and MSD_{WI} denotes the observed mean square deviation within individuals. In the above formula, n_i denotes the sample size in the i^{th} subpopulation and n_c denotes the average sample size across the S samples, also incorporating and correcting for variation in sample size between subpopulations.

As originally defined, the range of F_{ST} lies between 0 and 1. However, the above unbiased method for estimating F_{ST} can produce negative values. This observation, which has no biological interpretation, simply reflects the consequences of sampling error when population subdivision is weak. However, sampling error affects all F_{ST} estimates in a similar fashion and, therefore, negative values were included in our analyses to prevent bias in the estimated F_{ST} distributions. This decision affects only the estimated mean F_{ST} values, and in no case affects our conclusions.

Genotyping errors on high-F_{ST} SNPs. Genotyping errors, like allele flipping or false monomorphisms, can theoretically be a source of aberrant high F_{ST} values. Although genotyping and annotation errors are a reality in large public SNP databases, their presence is not expected to be more accentuated in any particular SNP class; therefore, they should not influence our conclusions, which are based on the comparison of F_{ST} distributions between different SNP classes. However, we checked for potential genotyping errors on high-F_{ST} genic SNPs by comparing the HapMap population genotype frequencies with those retrieved from independent datasets (for example, Perlegen, Affymetrix and CEPH; **Supplementary Note**). In addition, we experimentally verified the genotype frequencies for the nonsynonymous and 5′-UTR high-F_{ST} SNPs presented in **Table 1** as well as for a random set of nongenic high-F_{ST} SNPs. Genotyping errors were not more heavily represented among genic SNPs with respect to nongenic SNPs (**Supplementary Note**), and the few genic SNPs found to present discordant genotype frequencies were excluded from all analyses. Because genotyping errors among nongenic SNPs also exist, the exclusion of genotyping errors only for genic SNPs renders our analyses extremely conservative.

Assessment of statistical significance. For each functional class, we used 2×2 contingency tables to compare the observed numbers of low F_{ST} ($F_{ST} < 0.05$) and high F_{ST} ($F_{ST} > 0.65$) SNPs of each genic class with the numbers of low and high F_{ST} SNPs observed among nongenic SNPs. Significance was assessed using a χ^2 test with 1 degree of freedom. Under a hypothesis of strict neutrality, the proportion of SNPs presenting high or low F_{ST} values should be similar in genic and nongenic SNPs. The magnitude of disparity between the observed and expected distributions for each SNP class indicates the extent to which natural selection has influenced population differentiation (altering the proportion of a given SNP class in a given F_{ST} bin). In our analyses, we used nongenic SNPs as the baseline above which natural selection can be considered irrefutable. However, it is now widely accepted that natural selection may also affect nongenic regions, suggesting that these genomic regions may be of functional relevance[29]. Thus, the use of nongenic SNPs as the baseline of 'neutral diversity', even if natural selection has affected some of these nongenic regions, makes our comparisons highly conservative. Our approach to detecting signs of natural selection thus identifies the lower limit from which selective pressures have influenced recent human evolution.

Calculation and statistical test of λ. We measured the deviation (λ) between the expected and observed proportions of SNPs of each SNP class in each F_{ST} bin. Here, $\lambda = p_{O,i}/p_E$, where $p_{O,i}$ is the observed proportion of SNPs of a given class in the i^{th} bin of the distribution and p_E is the expected proportion of SNPs of a given class in that same F_{ST} bin. For example, if nonsynonymous SNPs account for 0.54% of the 2.8 million SNPs analyzed, 0.54% is the expected proportion (p_E) of nonsynonymous SNPs in all F_{ST} bins (λ will be equal to 1). By contrast, if nonsynonymous SNPs are overrepresented or underrepresented in particular F_{ST} bins, λ will be higher or lower than 1, respectively. For example, when considering SNPs presenting F_{ST} values higher than 0.95, we observed that 13% ($p_{O,i}$) of the total number of such high-F_{ST} SNPs were nonsynonymous. This corresponds to a 24-fold increase ($\lambda = 24$) in the expected proportion of nonsynonymous SNPs. We tested the significance of the λ value obtained for each SNP class (intronic, 5′ UTR, 3′ UTR, synonymous and nonsynonymous), using a χ^2 test with 1 degree of freedom. As only small numbers of SNPs were observed in the tails of the distributions, particularly in those corresponding to high F_{ST} values, we also evaluated whether the estimated χ^2-test P values were reliable in these conditions, by means of the Z-test (**Supplementary Note**). Finally, the F_{ST} distributions of each SNP class (nongenic, genic, intronic, 5′ UTR, 3′ UTR, synonymous and nonsynonymous) were tested against the entire genome-wide F_{ST} distribution (that is, the entire Phase II HapMap dataset, including the particular SNP class tested) giving highly conservative P values in the χ^2 and Z-tests.

Long haplotype test. The iHS statistic for each Phase II HapMap SNP was downloaded from the Haplotter[4] website (see URLs section below). For nongenic SNPs, we analyzed 1,335,664 SNPs for Africans, 1,176,074 for Europeans and 1,062,190 for Asians. For genic SNPs, we analyzed 796,598 SNPs for Africans, 699,521 for Europeans and 638,017 for Asians. For nonsynonymous SNPs, we analyzed 9,520 for Africans, 8,877 for Europeans and 8,335 for Asians. We could not test for an enrichment of significant iHS values among high F_{ST} 5′-UTR SNPs, because of the very limited effective number of SNPs falling into this category ($\leq$13 SNPs).

Population genetic simulations of negative selection. We carried out simulations using the forward population genetics (FPG) simulation program, provided by J. Hey (State University of New Jersey). Specifically, we simulated two populations of 25 chromosomes each, with a diploid effective population size of 250 (ref. 30), presenting average levels of population differentiation for neutral sites similar to those observed in human populations ($F_{ST} \sim 0.11$). To simulate the effects of negative selection on F_{ST} estimates, we then incorporated a deleterious population selection coefficient (S) varying from 1 to a maximum of 15 (ref. 30). An additive fitness scheme was used in the simulations performed, although the use of other fitness schemes (for example, multiplicative or epistatic) seemed not to affect our conclusions (data not shown). We ran stochastic simulations until obtaining, for each value of S, a minimum of 1,000 independent deleterious and neutral mutations. We then estimated the F_{ST} values, on a single-SNP basis, for all the simulated variants (**Supplementary Fig. 1**). The precise command lines used in our simulation process are available upon request.

Polyphen and OMIM analysis. We investigated whether the excess of nonsynonymous SNPs presenting low F_{ST} values resulted from negative selection by comparing the proportion of nonsynonymous variants with $F_{ST} < 0.05$ in the various predicted 'SNP fitness categories'. We predicted the fitness status of all nonsynonymous mutations using the Polyphen algorithm[16]. This method, which considers protein structure and/or sequence conservation information for each gene, has been shown to be the best predictor of the fitness effects of nonsynonymous mutations[18]. Using Polyphen analysis, we classified all 15,259 HapMap nonsynonymous SNPs into one of three fitness categories: 'benign', 'possibly damaging' or 'probably damaging'. We assessed the statistical significance of the observed differences in the proportion of low F_{ST} values between fitness categories using a χ^2 test with 1 degree of freedom. We also checked for significant differences in mean MAF between the different SNP fitness categories using Student's t-test.

We investigated whether SNPs presenting low and high F_{ST} values were more commonly observed than expected in genes known to modulate human disease by retrieving, for all HapMap nonsynonymous SNPs, the OMIM morbidity status of the corresponding genes. If a given SNP was located in a gene with a morbidity status entry, the SNP was labeled '1'. Conversely, if a given SNP was located in a gene with no morbidity status entry, the SNP was labeled '0'. We then used the χ^2 test to test for an association of low and high F_{ST} values with nonsynonymous SNPs located in genes known to modulate disease (labeled '1').

URLs. Haplotter, http://hg-wen.uchicago.edu/selection/haplotter.htm; HGDP-CEPH Human Genome Diversity Cell Line Panel, http://www.cephb.fr/HGDP-CEPH-Panel/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We acknowledge the International HapMap Consortium and Perlegen Sciences for making available their datasets to the scientific community; J. Hey for providing the forward population genetics (FPG) simulation program; S. Sunyaev for help with Polyphen analyses; M. Przeworski, R. Nielsen and E. Heyer for helpful suggestions and discussion; and L. Abel, T. Bourgeron, J.L. Casanova, S. Jamain, K. McElreavey and O. Neyrolles for critical reading of the manuscript. Financial support was provided by Institut Pasteur, by the Centre National de la Recherche Scientifique (CNRS) and by an Agence Nationale de la Recherche (ANR) research grant (ANR-05-JCJC-0124-01). L.B.B. is supported by a "Fundação para a Ciência e a Tecnologia" fellowship (SFRH/BD/18580/2004), and E.P. by the Fondation pour la Recherche Médicale (FRM).

AUTHOR CONTRIBUTIONS

L.B.B., G.L., E.P. and L.Q.-M. conceived the study. The data analyses were primarily performed by L.B.B and G.L., with contributions from E.P. H.Q. performed the genotyping experiments. The paper was written primarily by L.B.B. and L.Q.-M., with contributions from G.L. and E.P.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. The International Haplotype Map Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
2. Carlson, C.S. *et al.* Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* **15**, 1553–1565 (2005).
3. Sabeti, P.C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
4. Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
5. Williamson, S.H. *et al.* Localizing recent adaptive evolution in the human genome. *PLoS Genet.* **3**, e90 (2007).
6. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
7. Tishkoff, S.A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31–40 (2007).
8. Hamblin, M.T. & Di Rienzo, A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* **66**, 1669–1679 (2000).
9. Weir, C.L. & Cockerham, C.C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
10. Excoffier, L., Smouse, P.E. & Quattro, J.M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491 (1992).
11. Nielsen, R. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005).
12. Akey, J.M., Zhang, G., Zhang, K., Jin, L. & Shriver, M.D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002).
13. Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M. & Hill, W.G. Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* **15**, 1468–1476 (2005).
14. Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502 (2005).
15. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
16. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900 (2002).
17. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
18. Williamson, S.H. *et al.* Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA* **102**, 7882–7887 (2005).
19. Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W. & Akey, J.M. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* **16**, 980–989 (2006).
20. McVean, G. & Spencer, C.C. Scanning the human genome for signals of selection. *Curr. Opin. Genet. Dev.* **16**, 624–629 (2006).
21. Sabeti, P.C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
22. Teshima, K.M., Coop, G. & Przeworski, M. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* **16**, 702–712 (2006).
23. Sabeti, P.C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
24. Monreal, A.W. *et al.* Mutations in the human homologue of mouse dl cause autosomal recessive and dominant hypohidrotic ectodermal dysplasia. *Nat. Genet.* **22**, 366–369 (1999).
25. Mou, C., Jackson, B., Schneider, P., Overbeek, P.A. & Headon, D.J. Generation of the primary hair follicle pattern. *Proc. Natl. Acad. Sci. USA* **103**, 9075–9080 (2006).
26. Cockburn, I.A. *et al.* A human complement receptor 1 polymorphism that reduces Plasmodium falciparum rosetting confers protection against severe malaria. *Proc. Natl. Acad. Sci. USA* **101**, 272–277 (2004).
27. Meyre, D. *et al.* Variants of ENPP1 are associated with childhood and adult obesity and increase the risk of glucose intolerance and type 2 diabetes. *Nat. Genet.* **37**, 863–867 (2005).
28. Di Rienzo, A. & Hudson, R.R. An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet.* **21**, 596–601 (2005).
29. Drake, J.A. *et al.* Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.* **38**, 223–227 (2006).
30. Williamson, S. & Orive, M.E. The genealogy of a sequence subject to purifying selection at multiple sites. *Mol. Biol. Evol.* **19**, 1376–1384 (2002).