doi: 10.1038/nature06250

Supplemental Methods

LRH test, single SNP and windowed

REHH

To compare EHH values for a region across different groups of chromosomes, the LRH test first calculates a relative EHH (REHH)[1]. REHH is the ratio of EHH in one such group *g* to the average of EHH values in all other groups, with each group weighted by the probability of two chromosomes chosen from the combined data set belonging to *g*. More explicitly, if there are *M* chromosome groups, each with C_i chromosomes and an EHH value of EHH_i, REHH can be calculated by the formula

$$\mathrm{REHH}_i = \mathrm{EHH}_i \Bigg/ \left[\frac{\sum_{\substack{j=1 \\ j\neq i}}^{M} \binom{C_j}{2} \times \mathrm{EHH}_j}{\sum_{\substack{i=1 \\ j\neq i}}^{M} \binom{C_i}{2}} \right]$$

When only two groups are considered, REHH of group *g* is the ratio of EHH of *g* to that in the other group.

LRH calculation

We define the single-SNP LRH test with respect to a given core SNP, a given population, and a given direction (centromere distal or proximal). We focus on an area from the core SNP up to 1MB away from it in the specified direction. We pick a SNP X in this region such that its EHH with respect to the whole population is as close as possible to 0.04; if there is no SNP with such an EHH of between 0.03 and 0.05, the LRH test is skipped. Otherwise, we split the members of the population according to the core SNP allele they carry. For each allele, we note the pair [allele frequency, REHH at SNP X].

We performed the single-SNP LRH test in both directions and for all SNPs in all populations in the HapMap Phase II dataset. However, we ignored SNPs whose minor allele had a frequency below 5%, because their low sample counts made their REHH scores unreliable. For alleles of comparable frequency, we found the resulting distribution of ln(REHH) scores (in both simulations and in the human genome) to be approximately normal. Thus, for each population, we split our results into 20 equally sized allele frequency bins, and normalized the associated ln(REHH) scores such that the ln(REHH) scores in every bin had zero mean and unit variance. We denoted these normalized ln(REHH) by "LRH scores". Outlying LRH scores are potentially indicative of selection[1].

As in Voight et al.[2], we can reduce our false positive rate (or, alternately, reduce our threshold for defining "outlying") by choosing to declare a region significant only when a cluster of nearby SNPs has outlying LRH scores. In this windowed LRH test, we divide the genome into 100kb windows, each overlapping the next one by 50kb, and identify candidate regions for selection as those in which more than 0.1 fraction of SNPs within them have an LRH score above 3.92.

iHS test, single SNP and windowed

Following Voight et al.[2], we define the single-SNP iHS test with respect to a given core SNP and a given population. We perform the test only for biallelic SNPs whose minor allele frequency is above 5%. We split the members of the population according to the core SNP allele they carry. Let A denote the ancestral allele and D, the derived allele. Considering only the chromosomes carrying A, we calculate EHH scores between the core SNP and every biallelic SNP within 2.5MB. By linearly interpolating between successive biallelic SNPs, we integrate EHH with respect to genetic distance (cM). The integral extends the two points (centromere distal and proximal) at which EHH drops to exactly 0.05. If, however, EHH doesn't drop in both directions below 0.05 within 2.5MB of the core SNP, we skip the iHS test for that SNP. Otherwise, we denote the value of the integral by iHH_A (integrated haplotype homozygosity, ancestral). We follow an analogous procedure on the chromosomes carrying D to determine iHH_D. The unstandardised integrated haplotype score, or iHS, is defined as $\ln(iHH_A / iHH_D)$.

We calculated unstandardised iHSs for every SNP and population in the HapMap Phase II dataset. For SNPs whose derived allele frequency is comparable, the resulting distribution of unstandardised iHSs (in both simulations and in the human genome) has been shown to be approximately normal. Thus, for each population, we split our results into 20 equally sized allele frequency bins, and normalized the scores such that the set of scores in every bin has zero mean and unit variance. Due to the different population structure of chromosome X, we normalized its iHSs separately from those of the other chromosomes. We denote these normalized scores by simply "iHSs" (integrated haplotype scores). Outlying iHSs are potentially indicative of selection.

Information on the ancestral state of SNPs was provided by the International Haplotype Map Consortium. The ancestral allele was taken to be the chimpanzee base, where available, or the macaque base otherwise. If neither base was available, no ancestral state was inferred. For the ~7% of SNPs whose ancestral alleles were unavailable, we did not perform an iHS test. The genetic distances with respect to which we integrated were also those determined by the HapMap Project. We also chose to implement Voight et al's ad-hoc procedure to correct for large inter-SNP gaps in the data, although its effect was negligible in the high SNP-density Phase II data.

The iHSs reported by Voight are slightly different than those that would be obtained following the above procedure. In particular, their iHH_A is actually calculated by integrating the quantity (EHH-0.05+1/N), with N being the number of chromosomes carrying A (personal communication), and similarly for iHH_D. We have chosen to reproduce this peculiarity to compare iHS, LRH and XPop as fairly as possibly, but found this correction to have a negligible effect on calculated iHSs.

Similarly to the windowed LRH test, we performed a windowed iHS test, where a 100kb window of the genome was identified if 0.3 fraction of iHSs had absolute value above 3.13.

XP-EHH methods

We define the XP-EHH test with respect to two populations, A and B, a given core SNP, and a given direction (centromere distal or proximal). First, we consider all the SNPs for which there is data for both A and B that are up to 1MB from the given core SNP in the given direction. We pick a SNP X in this region such that its EHH with respect to all chromosomes in *both* populations is as close as possible to 0.04; if there is no SNP with such an EHH of between 0.03 and 0.05, the XP-EHH test is skipped. Next, we restrict our attention to the chromosomes in population A: we calculate EHH at all SNPs between the core SNP and X, and, similarly to the iHS test, integrate it within these bounds with respect to genetic distance. We call the result I_A. We proceed analogously with respect to population B, and call the result I_B. We define an XP-EHH logratio as $\ln(I_A/I_B)$.

For each population pair, we performed the XP-EHH test in both directions and for all SNPs in the HapMap Phase II. Empirically, the resulting distribution of XP-EHH logratios (in both simulations and in the human genome) is approximately normal. We note, however, that, in general, there was a small skew towards one population; we neglect this asymmetry when calculating significance scores. We normalize the XP-EHH logratio such that the set of all such logratios has zero mean and unit variance. We denote these normalized XP-EHH logratios by " XP-EHH scores". Outlying XP-EHH scores are potentially indicative of selection in a particular population. An XP-EHH score is directional: a positive score suggests selection is likely to have happened in population A, whereas a negative score suggests the same about population B. We include the region as a candidate if XP-EHH in one population pairwise comparison is above 5.1 or if XP-EHH in 2 population pairwise comparisons is above 4.34. The distribution of scores in the HapMap Phase 2 dataset, and corresponding percentiles are given in Figure S9.

Simulations and Power Calculations

We simulated the evolution of a 1MB section (around 1.23 cM) of 120 chromosomes each of the three populations, European (CEU), Yoruba (YRI) and Chinese/Japanese (CHB+JPT), using a previously validated demographic model [3]. We simulated neutrally evolving loci and twenty scenarios of positive selection, in which a new allele experienced positive selective pressure starting 5ky, 10ky, 15ky, 20ky and 30ky, reaching in the present population 20%, 40%, 60%, 80% and 100% frequency. Positive selection was modelled separately in each of the three populations, using a deterministic allele frequency trajectory for the selected allele. The selected allele was omitted from the final data set (a conservative choice for calculating power), and the remaining SNPs were thinned randomly to match the HapMap Phase II data in density and allele frequency. For neutrality we produced 10,000 independent simulations. For the 10ky and 15ky scenario, we produced 1000 idependent simulations, and for the remaining scenarios, we produced 100 independent simulations.

We further studied the effect of bottlenecks on our tests by simulating recent bottlenecks with a range of intensity. For this purpose we employed a simplified version of the above demography: three populations, branched as before, but of constanst size (Ne = 10,000), with no migration or bottlenecks. A single bottleneck was then introduced into one population (the "European" population) 750 generations ago and 1 Mb segments were simulated. One thousand segments were generated for each of four intensities (as measured by the inbreeding coefficient): 0.0, 0.1, 0.2 and 0.3.

We analysed the simulation data using the LRH, iHS and XP-EHH tests described above When normalizing scores, we calibrated the neutral simulation scores to have zero mean and unit variance, and then used the same parameters to normalize scores in all other simulated scenarios.

To compare the effectiveness of our individual tests, we estimated two properties with simulation: power and false positive rate (FPR). Power can be estimated by observing the fraction of simulations in which selection is detected, and depends on the strength of the selective pressure and on population structure. Conversely, FPR can be estimated by observing the fraction of neutral simulations where selection is erroneously detected. We measured FPR with respect to 10000 simulations of neutral evolution, and averaged the results over all three populations.

For LRH and iHS, power of each test has been extensively studied in previous papers [1,2,4], although not directly compared. Comparing the tests on simulated data, we found that they have similar power to detect recent selection but with some differences. The iHS test has slightly lower power at low haplotype frequency, while the LRH test has slightly lower power at high frequency. This can be seen in applications to HapMap data (phase 1), where the iHS test misses the well-known cases of *HBB* and *CD36* and the LRH test misses the *SULT1C2* region[2,5]. While both tests are based on the concept of EHH, we observed that the false positives produced by the two tests in simulations tend not to overlap and thus that signals detected by both tests have a very low FPR.

Each XP-EHH test involves two populations, so we quantified its effectiveness slightly differently. When there are N simulations, there are actually 2N results for XP-EHH tests between population A and one of the other two populations. Thus, we estimated XPop's power for detecting selection in A by observing the fraction of these results that showed signals for selection in A. We estimated XP-EHH 's FPR by observing the fraction of pairwise XP-EHH tests between neutral simulations that showed signals of selection. When comparing multiple tests, we adjusted the test thresholds for claiming "significant results" so that all tests had equal FPRs.

F_{ST} – F_{ST} was calculated for each pair of populations using the unbiased estimator of Weir and Cockerham[6]. For this study, individual marker F_{ST}s were calculated.

Derived Allele Frequency – Information on the ancestral state of SNPs was provided by the International Haplotype Map Consortium. The ancestral allele was taken to be the chimpanzee base, where available, or the macaque base otherwise. If neither base was available, no ancestral state was inferred.

The error rate in assigning the derived state using the chimpanzee genome for outgroup comparison is low (0.5%)[7] . Moreover, the iHS and XP-EHH tests, are designed to allow for the possibility of incorrect assignment of derived state. For localizing the signal of selection to particular polymorphisms, we used the derived state only as a guide, and still delineate highly differentiated alleles associated with the long haplotype.

Sweep

We developed a Java program, Sweep, to perform EHH-like analyses (PV, BF, PCS, ESL unpubished, www.broad.mit.edu/mpg/sweep). Sweep can import genotyping data in various formats, run various selection tests on it, and then visualize and export the results. At its core, Sweep acts as an EHH calculator, atop which the different selection tests are layered. For haplotype-based LRH tests, Sweep can also automatically identify haplotype blocks according to Gabriel et al's method. Sweep can be used to draw haplotype bifurcation diagrams for each allele in a haplotype. Moreover, Sweep can infer ancestral trees from modern-day haplotypes, using any available ancestral gene data to improve this inference. All visualizations made by Sweep can be exported to many bitmap (e.g., GIF, JPG, PNG) or vector (e.g., PDF) image formats.

As of this writing, we've coded the LRH, iHS and XP-EHH tests into Sweep, as well as provided an interactive way to adjust the tests' parameters and visualize the effects of these changes. To facilitate more automated analyses of larger data sets, like HapMap2, most of Sweep's functionality can be invoked through the command line.

Sweep has limited support for other recent selection tests (like F_{ST} and derived allele frequency), but has been designed to be easily extended. We hope Sweep may serve as a platform that allows other researchers to run existing selection tests on fertile new datasets, as well as a base on which to develop new tests for selection.

Recombination rate variation between populations

Studies of the fine-scale recombination rate in the human genome have indicated that population variation in recombination rate may exist in some regions of the genome. These differences could affect our signals, as a long-haplotype might be generated by reduced recombination in one population rather than by a single chromosome rising in frrequency. For our top regions detected by LRH and iHS, where the selected allele is still polymorphic in the population. We therefore use the other haplotypes in the population for comparison and to control for local or population variation in recombination rate.

For our top regions identified by the XP-EHH, we carefully examined the region surrounding each candidate, to rule out variation in recombination rate as a source of the

XP-EHH signal. For candidate regions where markers are still polymorphic in the population (like *LCT*) we use other non-selected alleles in the population assess possible population variation in recombination rate. In these XP-EHH candidate regions, the non-selected allele shows similar EHH decay as those alleles in other populations suggesting that recombination rate differences is not the source of the signal. For many of the top XP-EHH signals (like *SLC24A5*) the selected allele has gone to 100% frequency, as have the nearby alleles. Recombination differences between populations is not a significant issue because the signal is driven by the lack of polymorphism, and only enhanced by the occurrence of long-haplotypes from other polymorphisms further away within the region.

Copy Number Variation (CNV)

Several of the selected regions overlap with reported copy-number- variant (CNV) regions; while CNVs make appealing candidate loci for selection, current reports of CNV have insufficient spatial resolution for a true assessment of whether CNVs lie within the selected regions, and have generally lacked accurate sample-by-sample genotypes that could be used to assess whether copy number variants segregate on selected haplotypes or merely appear in the same general regions as selected loci. To assess whether signatures of selection at our strongest 22 candidates might be due to reported copy number variants in those regions, we reviewed underlying hybridization data for the HapMap samples from both a BAC arrayCGH platform[8] and a high-resolution oligonucleotide platform (Affymetrix GenomeWide 6.0, unpublished data). We report their coordinates in Table S11 along with the corresponding reference.

We further developed assays to assess whether CNV's could account for signatures of selection in the regions containing *EDAR* and *SLC24A5* (Figure S10). Copy number variation was previously reported in the *EDAR* region on BAC probes spanning 108.392-108.536 Mb [9]and 107.908-108.682 Mb[8]. Analysis of oligonucleotide array data showed that these observations were due to a 600-kb duplication variant spanning the 600kb region between segmentally duplicated sequence at 107.951-107.977 and 108.568-108.594 Mb (and therefore likely to have resulted from non-allelic homologous recombination between those sequences). The duplication allele was observed in two related YRI individuals (NA18870 and NA18872) but in no other HapMap samples, and is therefore unlikely to explain the signature of selection in this region. Overlapping the *SLC24A5* region, copy number variation has been reported by a single BAC probe spanning 46.296-46.451 Mb[8]; however, despite the fact that this region contained 60 probes on the oligonucleotide array, we observed no evidence for a CNV in any of the HapMap samples in this region, and suggest that the earlier report is a false discovery.

Fraction of SNPs estimated to be genotyped in the HapMap and to be identified in dbSNP

We estimated these numbers using full sequence data from the ENCODE project, assuming it is representative of the true genome, and applied a correction for those SNPs likely missed by ENCODE (only important for very low frequency SNPs, < 5%).

The average number of SNPs in of our 26 strongest candidates of selection was 809 with 195 – 1951 for the 95% confidence interval (CI). Given that 46% of SNPs with MAF > 5% are in HapMap, we thus estimated the typical region (95% CI) to have 424 – 4240 SNPs.

Targeting sites of transcription factors and microRNAs

We predicted potential binding sites for all mammalian transcription factors deposited in the Transfac database (version 7.4). To identify matching instances of each factor, we calculate a log-odds score to evaluate how well a sequence matches the positional weight matrix (p_{ij}) of the factor. The log-odds score is defined as $LO = \sum_i \log_2(p_{i,\, j(i)} / b_{j(i)})$ where $j(i)$ is the nucleotide at position i of the sequence, and b_j is the background frequency of the nucleotide j. We calculated the mean (μ) and variance (σ^2) of the log-odds score over a set of control sequences. An instance is called a matching instance if its log-odds score is above the threshold: μ+4.5σ. Upon identifying a matching motif instance in human, we determined if the instance is conserved in orthologous regions of other mammals. We proceeded by first extracting aligned sequences in the whole-genome alignment of 12 mammals (from UCSC Genome website). We then determined those species in which the corresponding aligned sequence also contains a matching instance. We defined an instance as *conserved* if the evolutionary tree connecting all species with a matching instance has a total branch length (measured in rate of mutations per nucleotide) greater than 0.85. (For reference, the total branch length connecting human, mouse, rat, and dog is 0.76).

We also predicted the targeting sites of microRNAs in 3'-UTRs of genes using the method as described in Xie et al[10].

Expression analysis

We obtained expression intensities of 44,000 probes representing the majority of the human gene complement for the HapMap individuals from the Wellcome Trust Sanger Institute's website[11]. We normalized the four sets of data from each individual by fitting a nested linear model to account for the two levels of technical duplication (in vitro transcription and chip hybridization) performed by the data generators, using the provided detection probability as a weighting. Of 289 UCSC known genes (H. sapiens build 17) in our regions, 109 were represented on the expression platform and had median detection probabilities >0.95, giving us high confidence that they were reliably detected. We used intensities from each of these genes as quantitative traits in a standard association test to the SNPs within the gene's region of residence, estimating significance by permutation (Purcell S, Neale B, Daly MJ, Sham PC, submitted)

Alignment of human SLC24 amino acid sequences

Amino acid sequences of the six SLC24 proteins were obtained from the Uniprot protein database [12], aligned with ClustalW[13] and the alignment formatted with Boxshade (http://www.ch.embnet.org/software/BOX_form.html). Transmembrane Domain predictions were performed using TMHMM[14].

Species alignment

We aligned the amino acid sequences of our top candidates to orthologous sequences from 17 mammals and annotated features of interest including: fixed differences, exon numbers, conserved regions, nonsynonymous SNPs, and functional domains. We obtained human amino acid sequences and exon positions from the UCSC Genome Browser (http://genome.ucsc.edu/). We then used another genome browser, Alpheus (http://www.broad.mit.edu/~mclamp/alpheus/), to align the human amino acid sequences to their orthologs. We designated amino acids as being encoded by conserved genomic regions based on the phastConsElements17way dataset (hg17) from the UCSC Genome Browser. We acquired a table of SNPs (snp125 hg17) for each gene using the UCSC Genome Browser. Finally, we annotated functional domains and other protein features based on designations in the UniProt protein database (http://www.pir.uniprot.org/).

Conservation Graphs

We made graphs of conservation versus nucleotide position for 10kb regions surrounding our candidate SNPs in each of the following genes: EDAR, EDA2R, SLC24A5, and SLC45A2. We also marked interesting genomic features in these regions including: exons, SNPs, and protein domains. We obtained the conservation scores from the PhastCons17way dataset (hg17) from the UCSC Genome Browser (http://genome.ucsc.edu/). We also obtained exon and SNP positions from the UCSC Genome Browser. Finally, we marked nucleotides coding for protein domains and other protein features based on amino acid designations in the UniProt protein database[12].

1 P. C. Sabeti, D. E. Reich, J. M. Higgins et al., *Nature* **419** (6909), 832 (2002).
2 B. F. Voight, S. Kudaravalli, X. Wen et al., *PLoS Biol* **4** (3), e72 (2006).
3 S. F. Schaffner, C. Foo, S. Gabriel et al., *Genome Res* **15** (11), 1576 (2005).
4 P. C. Sabeti, S. F. Schaffner, B. Fry et al., *Science* **312** (5780), 1614 (2006).
5 *Nature* **437** (7063), 1299 (2005).
6 B. S. Weir and C. C. Cockerham, *Evolution* **38**, 1358 (1984).
7 *Nature* **437** (7055), 69 (2005).
8 R. Redon, S. Ishikawa, K. R. Fitch et al., *Nature* **444** (7118), 444 (2006).
9 D. P. Locke, A. J. Sharp, S. A. McCarroll et al., *Am J Hum Genet* **79** (2), 275 (2006).
10 X. Xie, J. Lu, E. J. Kulbokas et al., *Nature* **434** (7031), 338 (2005).
11 B. E. Stranger, M. S. Forrest, M. Dunning et al., *Science* **315** (5813), 848 (2007).
12 *Nucleic acids research* **35** (Database issue), D193 (2007).
13 R. Chenna, H. Sugawara, T. Koike et al., *Nucleic acids research* **31** (13), 3497 (2003).
14 A. Krogh, B. Larsson, G. von Heijne et al., *Journal of molecular biology* **305** (3), 567 (2001).

Supplemental Figures

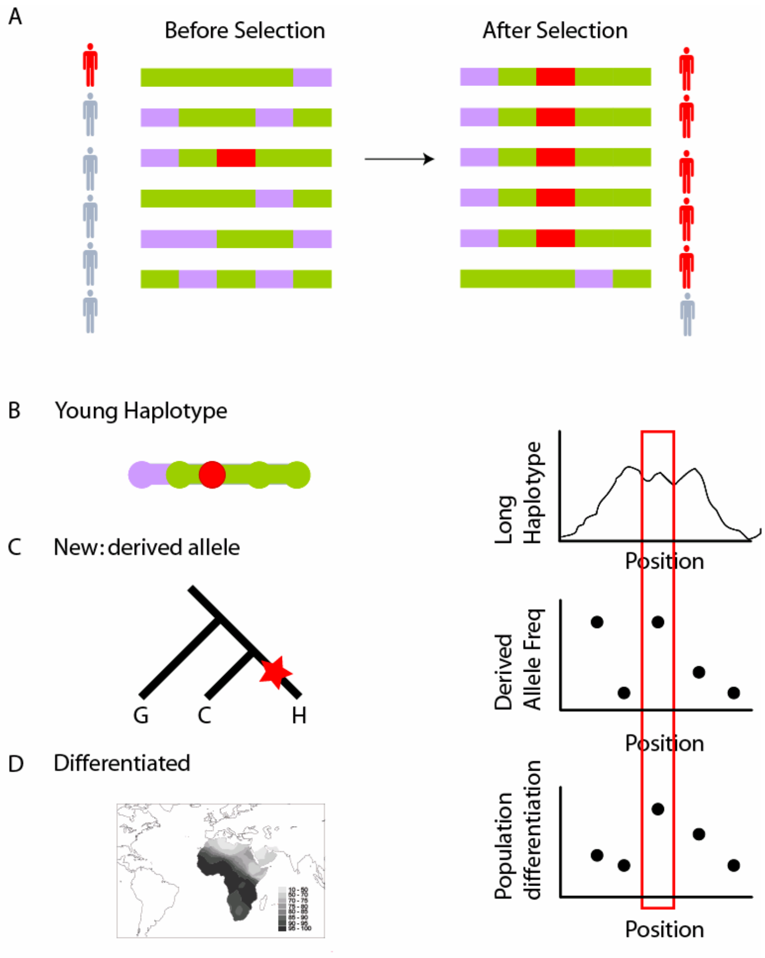

Figure S1 Localizing signal in a candidate region for natural selection, identified by long haplotypes. A). We show a cartoon of 5 polymorphisms in a candidate region rising to high frequency along with a positively selected (red) allele. Derived alleles are shown in purple. B) Long-haplotpe tests identify regions where variants have risen to high frequency so rapidly that recombination has not had time to break down links between nearby variants. Many variants within the region, will thus share the signal of long-haplotype, as they are all reciprically linked to each other. C) Given that long-haplotype methods are designed to identify young alleles, we expect the selected allele to be a derived allele on the long-haplotype identified. D) Given that recent selection is often a local phenomenon, we expect the selected allele to be differentiated between populations with and without signals of selection. Only a subset SNPs in a candidate region will share these characteristics, and an even smaller subset will be functional.

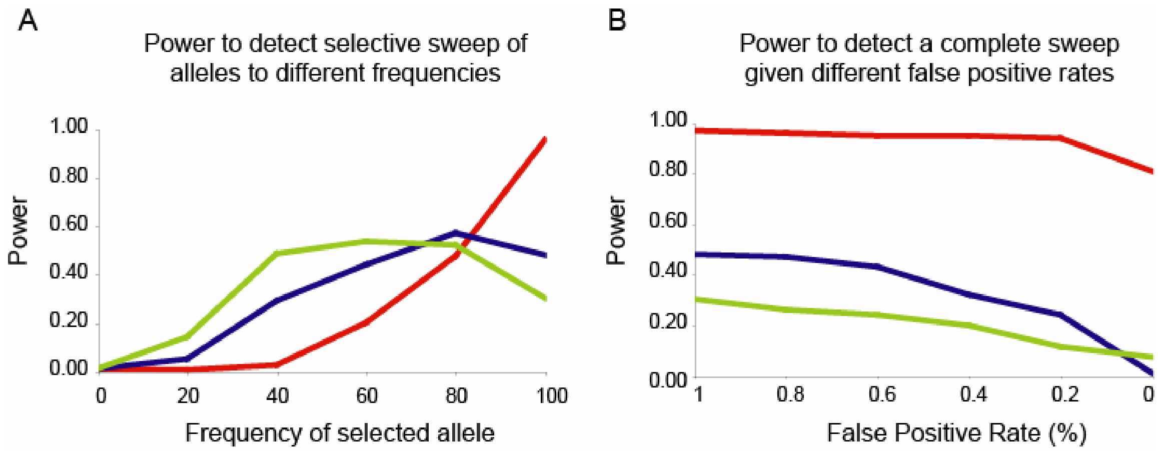

Figure S2 Power Calculations. (A) The estimated power, based on simulations, for the LRH test (green), the IHS test (blue), and the XP-EHH test (red). The power for simulations is given where the selected allele originated 15kya in Europe, given a 1% false positive rate (FPR) in neutral simulations (Methods). (B) The estimated power to detect a complete sweep (100% frequency of selected allele) for the 3 tests as the FPR for neutral simulation is decreased from 1% to 0%. See Table S1-6 for results for other time frames, populations, demographies, and thresholds.

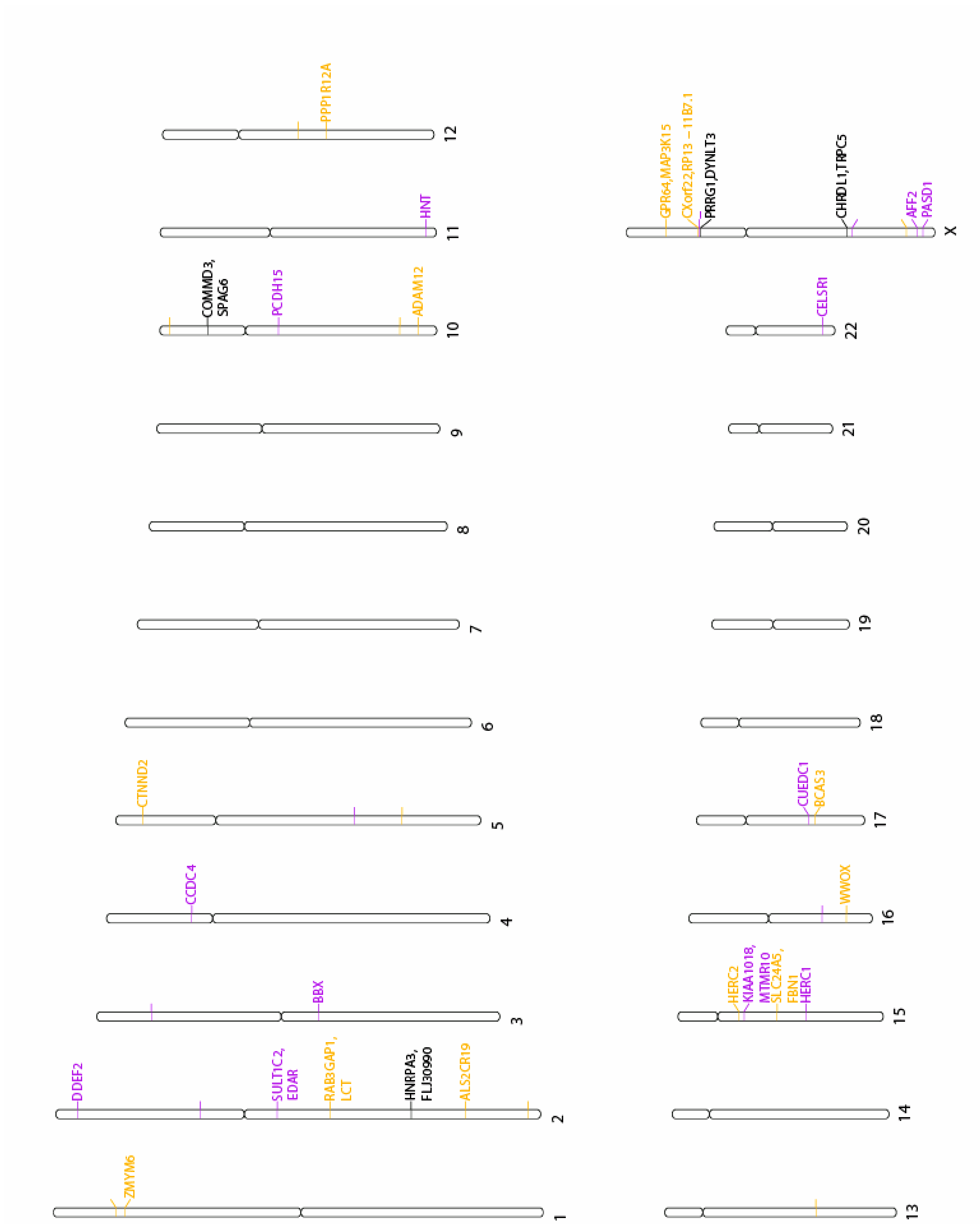

Figure S3: Top 43 XP-EHH candidates. The regions were identified given an FPR in simulations of 0.4% per 1MB region. The regions are given along with example genes in the region. The color indicates the population where selection was identified (orange-CEU, purple-JPT+CHB, black-CEU&JPT+CHB). Top candidates for the LRH and iHS tests are given in the HapMap Phase 2 paper, a companion paper in this issue.

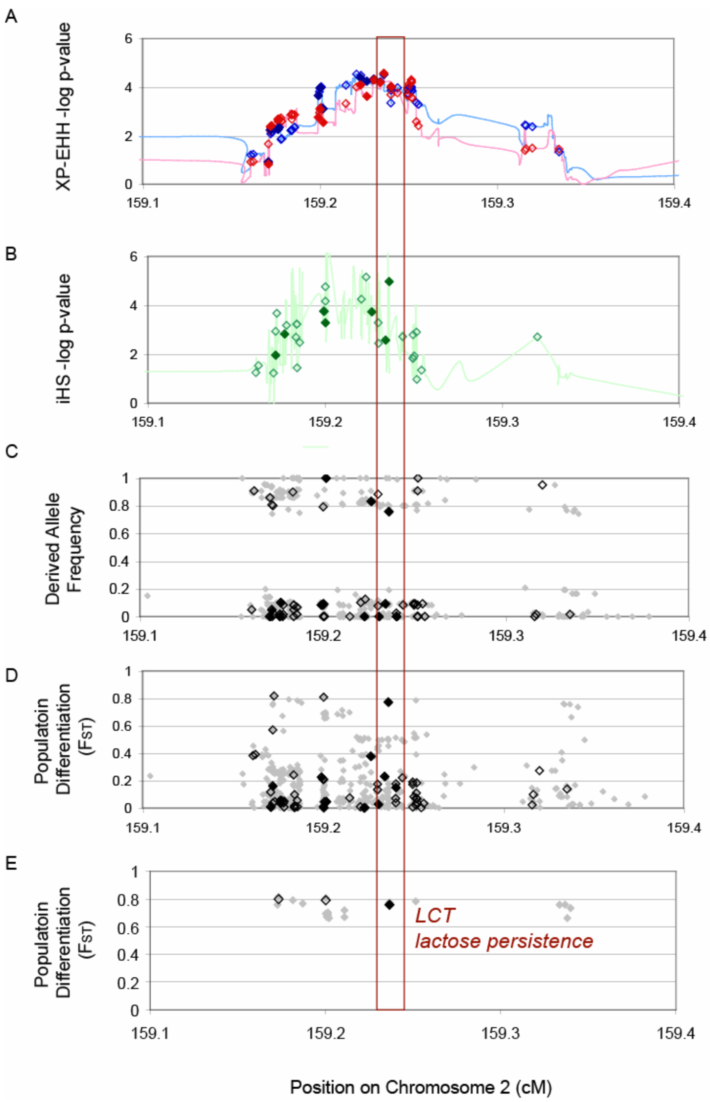

Figure S4 *LCT* region of positive selection. We found strong evidence for selection based on XP-EHH, LRH, and iHS tests at the locus near *LCT* (A) We examined XP-EHH between CEU and JPT+CHB (blue), CEU and YRI (red), and YRI and JPT+CHB (gray), and found strong evidence of recent selection in CEU. (B) We also identify strong evidence based on the iHS tests. We classified potential functional SNPs into lower probability (bordered diamonds) and high probability (filled diamonds). We examined SNPs for our 3 criteria for a target of selection based on (B) the frequency of derived alleles, (C) differences between poplations and (D) differences between populations for high frequency derived alleles less than 20% in non-selected populations. The lactose persistence allele at *LCT* is one of 24 polymorphisms that are high frequency derived and only common in CEU.

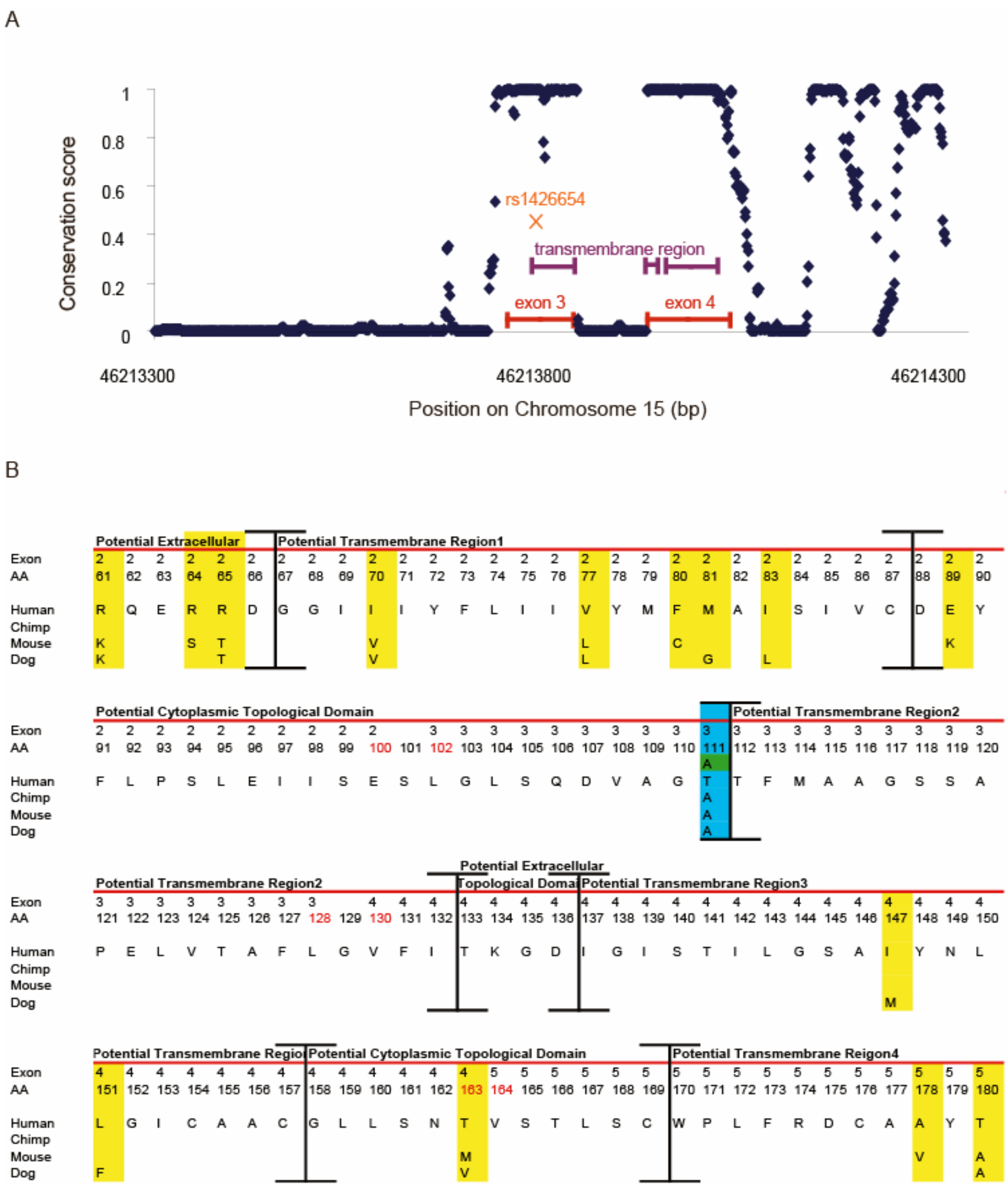

Figure S5 *SLC24A5* Ala111Thr in highly conserved transmembrane region. A) Conservation score (blue diamonds) around exon 3 of *SLC24A5* on Chromosome 3. The Ala111Thr polymorphism (rs1426654) lies within a highly conserved potential trasmembrane region in exon 3. B) A closer view of the amino acid sequence in *SLC24A5*. The exon and amino acid number is shown at the top. Red lines indicate high conservation based on PhastCon (Methods). Amino acids with substitutions between the 4 species are highlighted in yellow. Ala111Thr is indicated in blue.

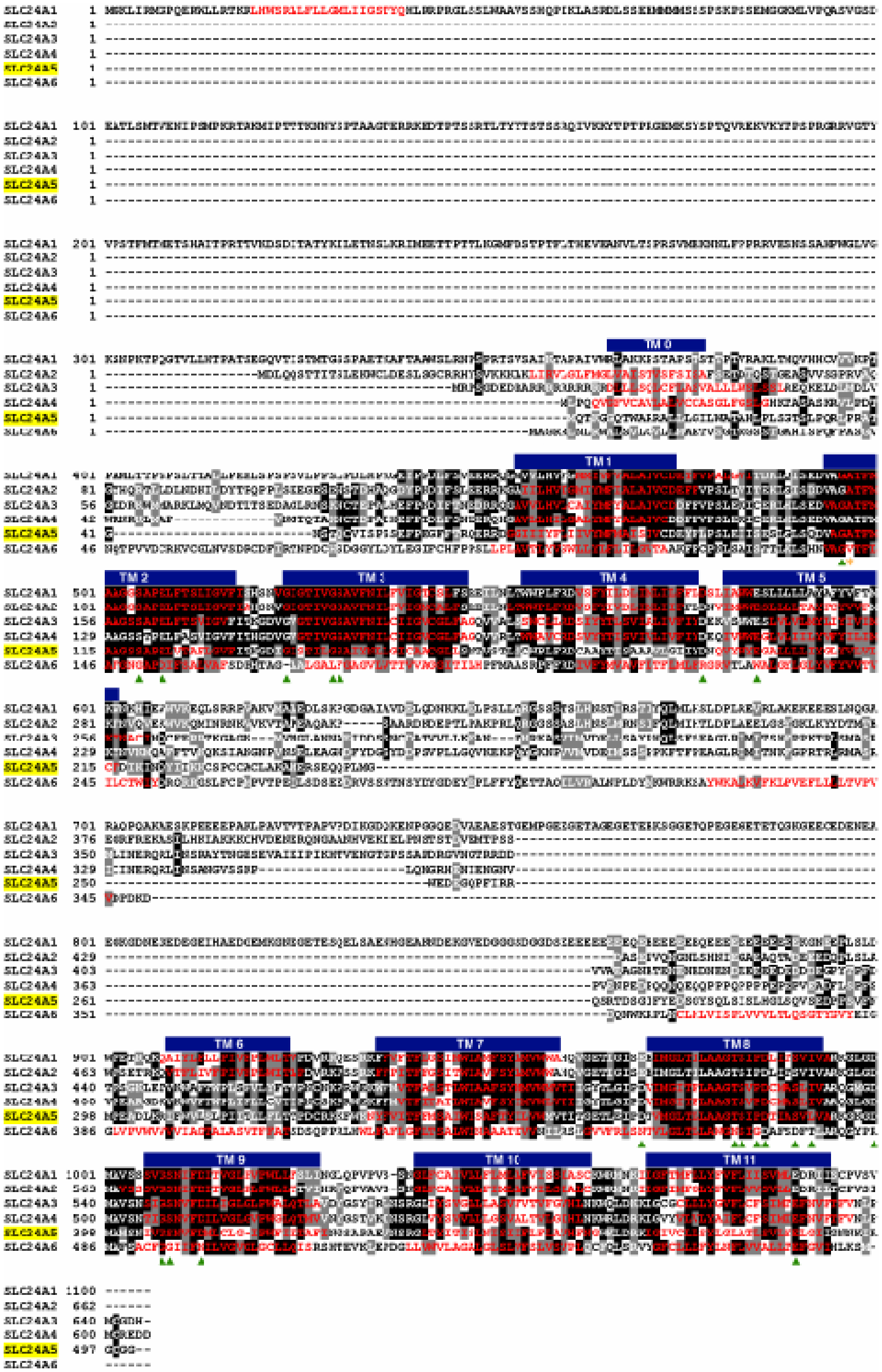

Figure S6 Annotated Alignment of human SLC24 amino acid sequences. Legend on next page.

Figure S6 Annotated Alignment of human SLC24 amino acid sequences. Alignment of the amino acid sequences of the six human SLC24, K^+-dependent Na^+/Ca^{2+} exchanger family members. Aligned residues identical or similar in greater than three sequences are shadowed in black and grey, respectively. Residues predicted to be in transmembrane regions (TMs) are red. Blue boxes above the alignment represent consensus TM regions in which three or more residues are predicted to be in a TM region. The polymorphic SLC24A5 residue, A111, encoded by candidate SNP rs1426654 is marked by an orange asterisk. Green triangles indicate mutations that lead to a >70% decrease in transporter activity, as part of a scanning mutagenesis study of residues 172-212 and 536-575 in SLC24A2 [1]. Notably the G176A mutation of SLC24A2, corresponding to G110 in SLC24A5, leads to one of the most severe reductions in SLC24A2 activity, >85%, and the A177S mutation, corresponding to A111 in SLC24A5, leads to a ~40% reduction in SLC24A2 transporter activity [1].

1 R. J. Winkfein, R. T. Szerencsei, T. G. Kinjo et al., Biochemistry 42 (2), 543 (2003).

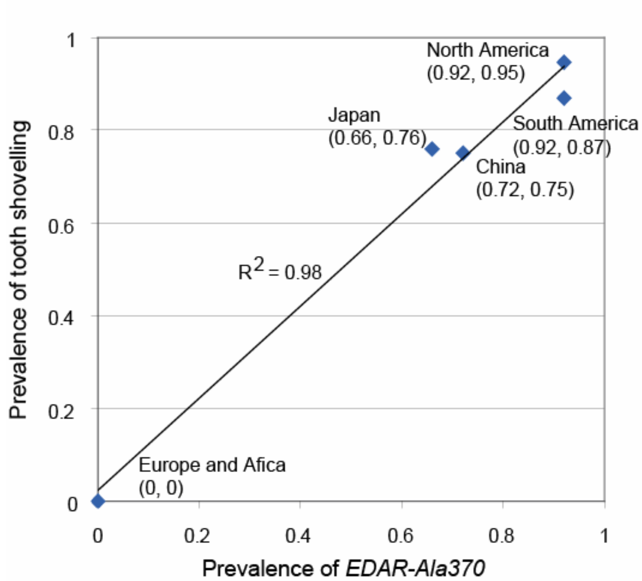

Figure S7 Prevalence of tooth shovelling and *EDAR-Ala370* allele in 4 Sinodont populations. A great deal is known from the anthropological record about the physical traits regulated by the EDA pathway, particular teeth and to less extent hair, in human populations. There are two distinct tooth patterns common to Asia [1], defined by a phenomenon called "tooth shoveling," in which the back surface of the upper incisors has a "shovel" appearance.[1] Shoveling consists of a "combination of a concave lingual surface and elevated marginal ridges enclosing a central fossa in the upper central incisor teeth." [2]. The pattern is particular among the Sinodonts, a population that evolved from the Sundadonts (the original inhabitants of Asia) as they moved north and inland into Asia. Sinodonts evolved in present-day China, and they also migrated from the Asian mainland into Japan around 2,000 years ago. Native American populations came from Asia in at least two waves of migration,[3] and may be in part populated by Sinodonts. High tooth shoveling frequencies have accordingly been reported in Sinodont populations in China-Mongolia, Japan, NE Siberia-Amur, Aleut-Eskimo, Greater NW Coast, North America, and South America. We had *EDAR-Ala370* allele frequency data for four Sinodont populations, where tooth shovelling frequencies have been determined and examined the correlation. There are many limitations to this analysis. Only 4 populations (as well as Europe and Africa) frequencies are known. Moreover the samples are not the same and may reflect different subpopulations.

1. Turner, C. G., 2nd. Teeth and prehistory in Asia. Sci Am 260, 88-91, 94-6 (1989).
2. Hsu, J. W. et al. Ethnic dental analysis of shovel and Carabelli's traits in a Chinese population. Aust Dent J 44, 40-5 (1999).
3. Karafet, T. M. et al. Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. Am J Hum Genet 64, 817-31 (1999).

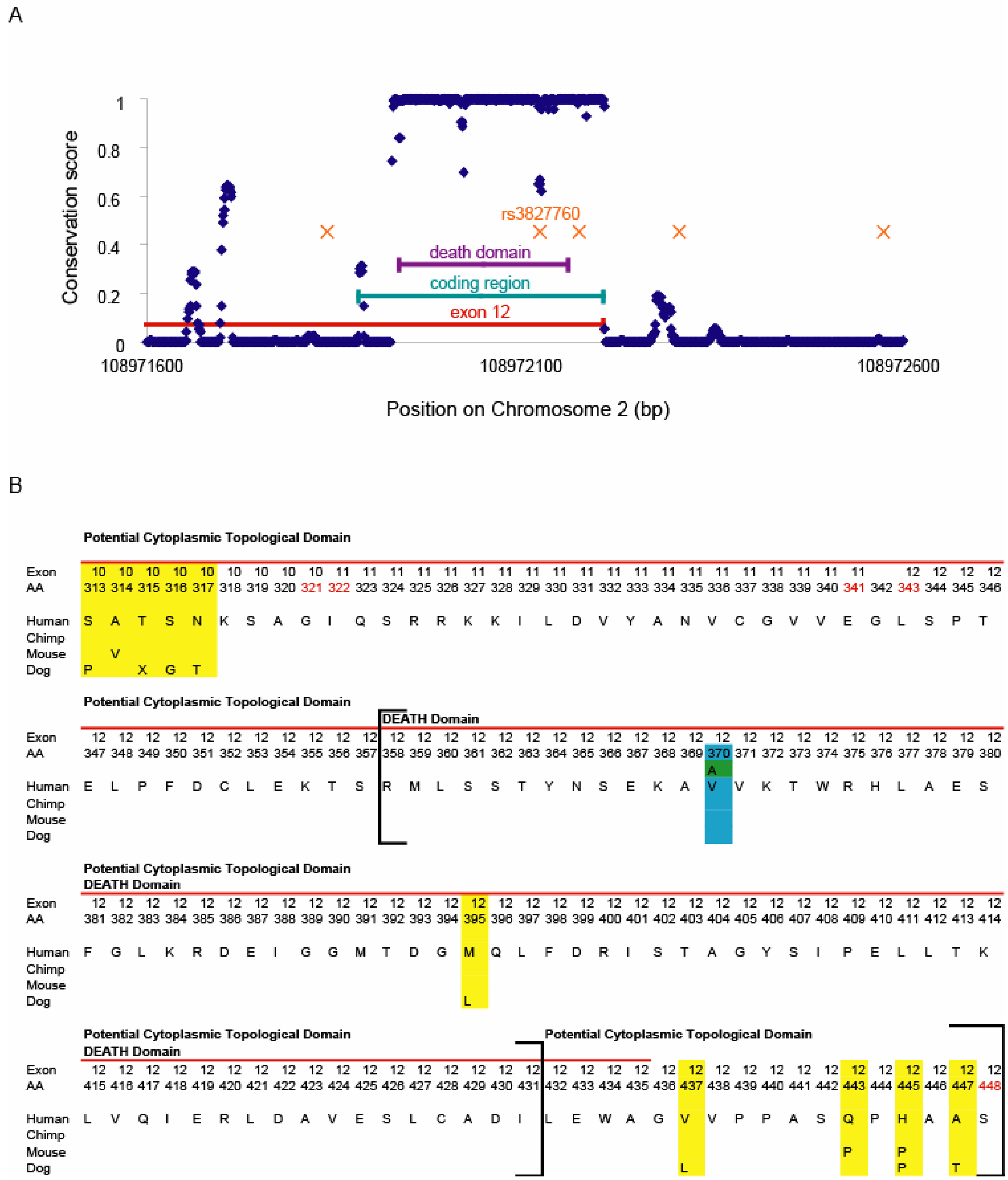

Figure S8 *EDAR* Val370Ala in highly conserved death domain. A) Conservation score (blue diamonds) around Exon 12 of *EDAR* on Chromosome 2. The Val370Ala polymorphism (rs3827760) lies within the highly conserved death domain in the coding porrtion of exon 12. B) A closer view of the amino acid sequence in *EDAR*. The exon and amino acid number is shown at the top. Red lines indicate high conservation based on PhastCon (Methods). Amino acids with substitutions between the 4 species are highlighted in yellow. Val370Ala is indicated in blue.

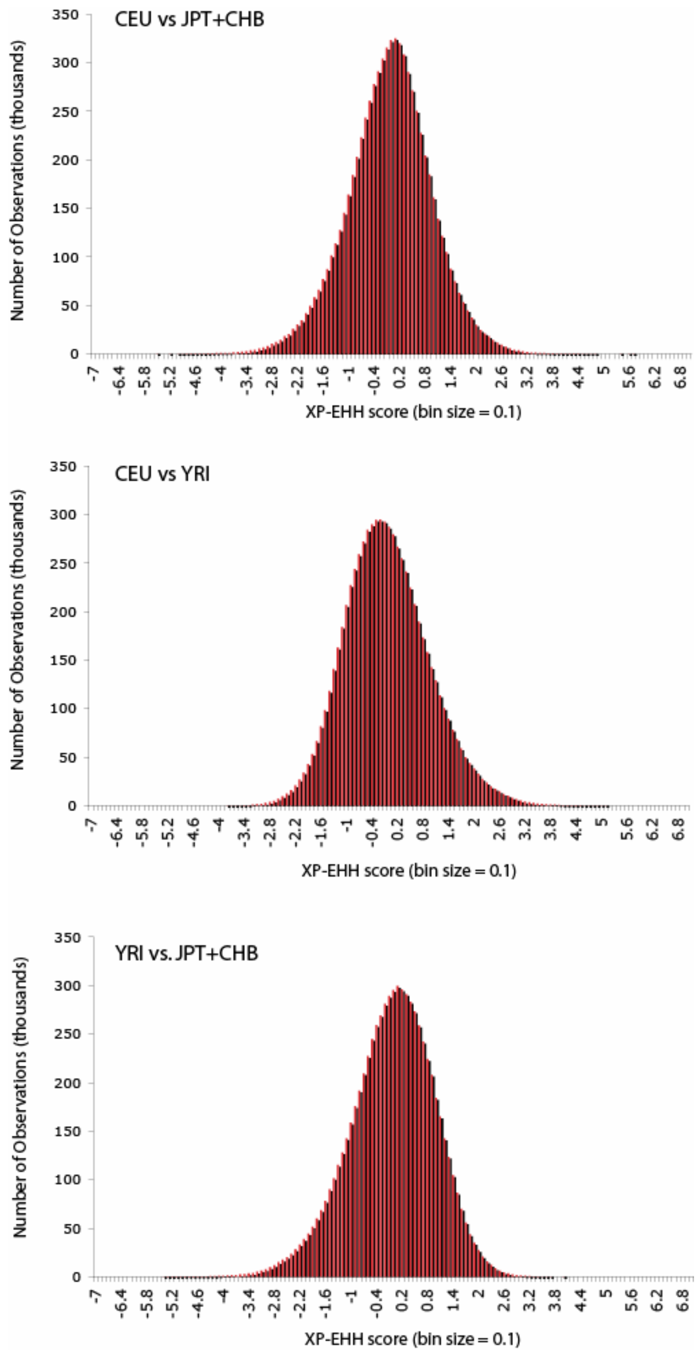

Figure S9 Distribution of XP-EHH scores for each population comparison in the from the HapMap Phase 2 dataset. For CEU vs. CHB+JPT, a score of 4.34 is in the 99.943 percentile, and 5.1 is in the 99.988 precentile. For CEU vs YRI, a score of 4.34 is in the 99.970 percentile, and 5.1 is in the 99.998 precentile. For YRI vs JPT+CHB, a score of 4.34 is in the 99.942 percentile, and 5.1 is in the 99.987 precentile.

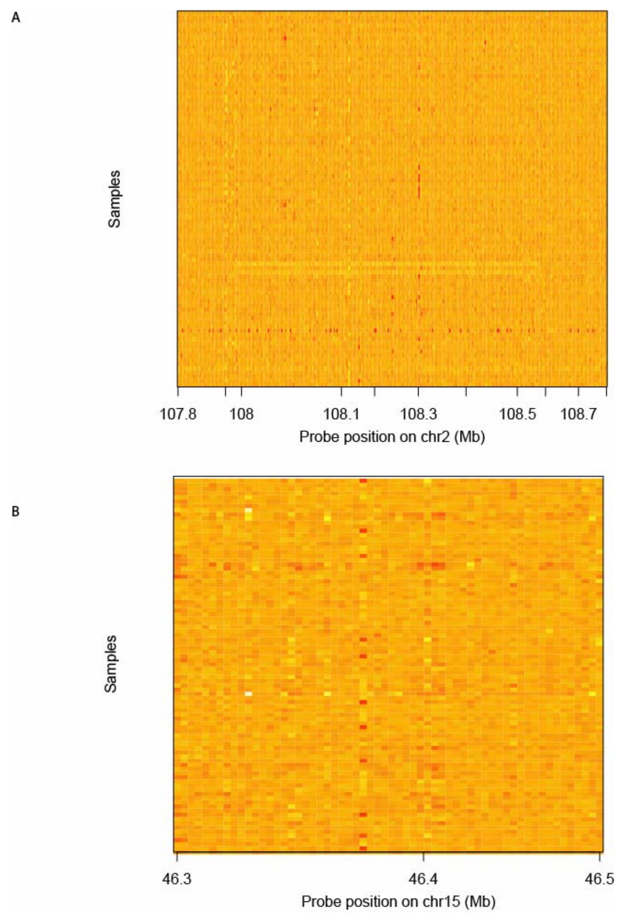

Figure S10 Analysis of oligonucleotide array data to assess CNVs in the candidate regions containing *EDAR* and *SLC24A5*. A) CNV was previously reported overlapping the *EDAR* region on BAC probes spanning 108.392-108.536 Mb[1] and 107.908-108.682 Mb [37]. Analysis of oligonucleotide array data showed that these observations were due to a 600-kb duplication variant spanning the 600kb region between segmentally duplicated sequence at 107.951-107.977 and 108.568-108.594 Mb (and therefore likely to have resulted from non-allelic homologous recombination between those sequences). The duplication allele was observed in two related YRI individuals (NA18870 and NA18872) but in no other HapMap samples, and is therefore unlikely to explain the signature of selection in this region. B) CNV was previously reported overlapping the *SLC24A5* region, by a single BAC probe spanning 46.296-46.451 Mb[2]; however, despite the fact that this region contained 60 probes on the oligonucleotide array, we observed no evidence for a CNV in any of the HapMap samples in this region, and suggest that the earlier report is a false discovery.

1 D. P. Locke, A. J. Sharp, S. A. McCarroll et al., Am J Hum Genet 79 (2), 275 (2006).
2 R. Redon, S. Ishikawa, K. R. Fitch et al., Nature 444 (7118), 444 (2006).

Supplemental Tables

Table S1 Power of the LRH test to detect a selected allele that emerged at 5 different points in time, and rose to 5 different frequencies in 3 different populations, given a 1% false positive rate (FPR).

LRH (1% FPR)

Allele Freq	5kya	10kya	15kya	20kya	30kya
Europe					
20	0.16	0.15	0.14	0.17	0.15
40	0.69	0.53	0.48	0.43	0.33
60	0.79	0.66	0.54	0.64	0.54
80	0.78	0.59	0.52	0.46	0.43
100	0.53	0.32	0.30	0.28	0.20
Asia					
20	0.16	0.22	0.09	0.14	0.10
40	0.58	0.51	0.35	0.48	0.50
60	0.76	0.62	0.53	0.51	0.46
80	0.80	0.52	0.48	0.43	0.40
100	0.53	0.40	0.27	0.21	0.24
Africa					
20	0.44	0.16	0.15	0.15	0.15
40	0.94	0.71	0.66	0.49	0.50
60	0.98	0.87	0.73	0.75	0.73
80	1.00	0.98	0.76	0.71	0.72
100	0.65	0.86	0.79	0.75	0.69

Table S2 Power of the iHS test to detect a selected allele that emerged at 5 different points in time, and rose to 5 different frequencies in 3 different populations, given a 1% FPR.

iHS (1% FPR)

Allele Freq	5kya	10kya	15kya	20kya	30kya
Europe					
20	0.10	0.12	0.05	0.07	0.12
40	0.55	0.38	0.29	0.28	0.38
60	0.87	0.7	0.44	0.51	0.46
80	1.00	0.87	0.57	0.39	0.54
100	0.59	0.65	0.48	0.43	0.38
Asia					
20	0.04	0.16	0.04	0.02	0.11
40	0.48	0.5	0.28	0.32	0.29
60	0.84	0.72	0.48	0.43	0.35
80	0.97	0.79	0.54	0.40	0.49
100	0.52	0.66	0.50	0.41	0.23
Africa					
20	0.33	0.12	0.06	0.12	0.14
40	0.88	0.7	0.52	0.40	0.53
60	0.94	0.88	0.60	0.67	0.69
80	1.00	0.95	0.78	0.68	0.74
100	0.78	0.92	0.89	0.89	0.81

Table S3 Power of the XP-EHH test to detect a selected allele that emerged at 5 different points in time, and rose to 5 different frequencies in 3 different populations, given a 1% FPR.

XP-EHH (1% FPR)

Allele Freq	5kya	10kya	15kya	20kya	30kya
Europe					
20	0.02	0.00	0.01	0.01	0
40	0.21	0.11	0.03	0.03	0.01
60	0.80	0.35	0.20	0.21	0.09
80	0.96	0.68	0.48	0.30	0.27
100	1.00	1.00	0.95	0.81	0.53
Asia					
20	0.00	0.00	0.01	0.00	0
40	0.08	0.04	0.04	0.01	0.01
60	0.67	0.29	0.23	0.12	0.05
80	0.96	0.58	0.38	0.19	0.2
100	0.99	1.00	0.89	0.76	0.4
Africa					
20	0.00	0.00	0.03	0.00	0.02
40	0.08	0.02	0.07	0.02	0.03
60	0.68	0.31	0.17	0.14	0.16
80	0.96	0.74	0.39	0.27	0.35
100	1.00	1.00	0.97	0.89	0.62

Table S4 Power of the LRH, iHS, and XP-EHH tests to detect a selected allele that emerged at 15 thousand years ago (15kya) and rose to 100% frequency in 3 different populations for 7 different FPR.

			15 kya	
FPR	**Parameters**	**Europe**	**Asia**	**Africa**
LRH				
2	10% LRH in 100kb >3.5	0.44	0.42	0.84
1	10% LRH in 100kb >3.85	0.30	0.27	0.79
0.8	10% LRH in 100kb > 4	0.26	0.23	0.75
0.6	10% LRH in 100kb > 4.2	0.24	0.16	0.70
0.4	10% LRH in 100kb > 4.4	0.20	0.15	0.66
0.2	10% LRH in 100kb > 4.65	0.11	0.13	0.56
0	10% LRH in 100kb > 5.4	0.07	0.02	0.29
iHS				
2	30% iHS in 100kb >2.8	0.54	0.56	0.89
1	30% iHS in 100kb >3.0	0.48	0.50	0.89
0.8	30% iHS in 100kb >3.1	0.47	0.47	0.88
0.6	30% iHS in 100kb >3.15	0.43	0.43	0.86
0.4	30% iHS in 100kb >3.4	0.32	0.36	0.83
0.2	30% iHS in 100kb >3.75	0.24	0.25	0.71
0	30% iHS in 100kb >5.9	0.01	0.00	0.16
XP-EHH				
2	1 SNP Xpop > 4	0.98	0.93	0.98
1	1 SNP Xpop >4.4	0.97	0.93	0.99
0.8	1 SNP Xpop>4.5	0.96	0.89	0.98
0.6	1 SNP Xpop>4.6	0.95	0.89	0.96
0.4	1 SNP Xpop>4.65	0.95	0.87	0.95
0.2	1 SNP Xpop>4.8	0.94	0.80	0.94
0	1 SNP Xpop>5.3	0.81	0.61	0.86

Table S5 The FPR of XP-EHH tests under several population demographic scenarios. We first used a previously validated demographic model for the 3 HapMap populations, CEU, YRI, and CHB+JPT [24] and obtained an 'Overall FPR' for these. We then compared to 4 demographic models: Bottleneck 0.0, 0.1, 0.2 and 0.3, refer to increasing intensities of bottleneck, as measure by the inbreeding coefficient, 0.0, 0.1, 0.2, and0.3, respectively (Methods).

Overall FPR	**1**	**0.4**	**0**
Asia	0.9	0.3	0
Europe	1.4	0.6	0
Africa	0.8	0.3	0
Bottleneck 0	0.7	0.5	0
Bottleneck 0.1	0.5	0.3	0
Bottleneck 0.2	0.3	0.3	0
Bottleneck 0.3	0	0	0

Table S6 The mean and standard deviation for the AllEHH logratio with increasing strength of bottleneck.

Bottleneck vs. YRI		
Inbreeding Coefficient	**Mean AllEHH logratio**	**Std Dev All EHH logratio**
0	-0.032389697	0.38512096
0.1	0.28246412	0.42372218
0.2	0.60321206	0.50540924
0.3	0.8909626	0.5617148
Bottleneck vs. JPT+CHB		
Inbreeding Coefficient	**Mean AllEHH logratio**	**Std Dev All EHH logratio**
0	-0.002054234	0.3377865
0.1	-0.3157318	0.39682296
0.2	-0.6341556	0.4830781
0.3	-0.9136418	0.54949707

Table S7 Candidate Regions for recent selective sweeps using XP-EHH test. The top regions for the LRH and iHS tests are given in the HapMap Phase 2 paper, a companion paper in this issue.

Region	Chromosome	Start	Stop	Maximum XP-EHH score	Peak SNP ID	Selected Population	Compare Population	Identified in 2 comparisons	Genes in region	(x) identified at 0.4% FPR threshold in simulations, all others identified at 1% FPR
1	1	30359744	30366699	4.854	rs4949250	CEU	JPT+CHB			
2	1	35109198	35164815	5.316	rs11804392	CEU	YRI		ZMYM6	x
3	2	9315721	9315721	4.79	rs875053	JPT+CHB	YRI		DDEF2	
4	2	72305454	72927242	5.53	rs6717899	JPT+CHB	CEU, YRI	x		x
5	2	108408653	108971124	5.684	rs1105109	JPT+CHB	CEU, YRI	x	SULT1C2, GCC2, FLJ38668, LIMS1, RANBP2, FLJ32745, EDAR	x
6	2	135663041	136424290	5.513	rs3795901	CEU	YRI, JPT+CHB	x	RAB3GAP1, ZRANB3, R3HDM1, UBXD2, LCT	x
7	2	177317730	178285258	5.412	rs1534679	CEU, JPT+CHB	CEU, YRI	x	HNRPA3, NFE2L2, AGPS, FLJ30990	x
8	2	206028349	206043667	5.029	rs1511873	CEU	YRI		ALS2CR19	
9	2	238144810	238161079	4.893	rs9287620	CEU	YRI			
10	3	26230802	26239053	4.966	rs11918137	JPT+CHB	YRI			
11	3	108754249	108994687	5.534	rs9883282	JPT+CHB	CEU		BBX	x
12	4	41984060	41989630	4.811	rs6826469	JPT+CHB	YRI		CCDC4	
13	5	11886256	11893734	4.858	rs12521011	CEU	YRI		CTNND2	
14	5	117381470	117679927	5.876	rs11241446	JPT+CHB	CEU			x
15	5	142119542	142125869	4.866	rs764387	CEU	YRI			
16	10	2986576	2988247	4.812	rs2454822	CEU	YRI			
17	10	22642019	22798204	5.978	rs12241555	CEU, JPT+CHB	YRI	x	COMMD3, BMI1, SPAG6	x
18	10	55541277	55543799	4.926	rs7074276	JPT+CHB	CEU	x	PCDH15	
19	10	118258077	118276595	5.009	rs10885979	CEU	JPT+CHB			
20	10	127865903	127865903	4.933	rs2927508	CEU	JPT+CHB		ADAM12	
21	11	131440546	131443589	4.958	rs11828462	JPT+CHB	CEU		HNT	
22	12	64360488	64364566	4.972	rs10878314	CEU	JPT+CHB			
23	12	78757457	78827321	5.1	rs7305173	CEU	JPT+CHB		PPP1R12A	x
24	13	73770157	73770157	4.858	rs17062507	CEU	YRI			
25	15	26064184	26088260	4.862	rs10438451	CEU	YRI		HERC2	
26	15	29003953	29073042	4.943	rs7170710	JPT+CHB	YRI		KIAA1018, MTMR10	
27	15	46155214	46657748	6.413	rs1559857	CEU	YRI, JPT+CHB	x	SLC24A5, MYEF2, SLC12A1, DUT, FBN1	x
28	15	61748992	61848071	5.251	rs16947373	JPT+CHB	YRI		HERC1	x
29	16	64165845	64452865	5.287	rs410941	JPT+CHB	CEU, YRI	x		x
30	16	77061737	77089133	5.178	rs16947649	CEU	YRI		WWOX	x
31	17	53305194	53357191	5.984	rs9898004	JPT+CHB	CEU, YRI	x	CUEDC1	x
32	17	56419222	56515445	5.385	rs8073202	CEU	YRI		BCAS3	x
33	22	45109651	45133715	4.843	rs16995204	JPT+CHB	YRI		CELSR1	
34	23	18881880	19138487	5.637	rs7341964	CEU	YRI		GPR64, PDHA1, MAP3K15	x
35	23	35759035	35939638	6.065	rs5973574	CEU	YRI		CXorf22, RP13-11B7.1	x

Table S7 continued.

Region	Chromosome	Start	Stop	Maximum XP-EHH score	Peak SNP ID	Selected Population	Compare Population	Identified in 2 comparisons	Genes in region	(x) identified at 0.4% FPR threshold in simulations, all others identified at 1% FPR
36	23	36476826	36521901	5.248	rs5973753	JPT+CHB	YRI			x
37	23	37069665	37555024	5.844	rs17144310	CEU, JPT+CHB	YRI	x	PRRG1, LANCL3, LOC644106, XK, CYBB, DYNLT3	x
38	23	109767056	111117626	6.392	rs10521530	CEU, JPT+CHB	YRI	x	CHRDL1, PAK3, CAPN6, DCX, GLT28D1, CXorf45, TRPC5	x
39	23	113291719	113296616	4.938	rs12389690	JPT+CHB	YRI			
40	23	141796760	141804088	4.868	rs5953797	CEU	YRI			
41	23	147341578	147421230	5.107	rs956659	JPT+CHB	CEU		AFF2	x
42	23	150287808	150488109	5.082	rs12860832	JPT+CHB	YRI		PASD1	

Table S8. Fraction of SNPs estimated to be genotyped in the HapMap and to be identified in dbSNP. We estimated these numbers using full sequence data from the ENCODE project, assuming it is representative of the true genome, and applied a correction for those SNPs likely missed by ENCODE (only important for very low frequency SNPs, < 5%).

% of SNPs in HapMap	YRI	CEU	JBT+CHB
MAF > 5%	43	46	49
MAF > 20%	56	50	51
% of SNPs in dbSNP			
MAF > 5%	66	81	79
MAF > 20%	86	90	88

Table S9. Forty-one polymorphisms with multiple lines of evidence for selection.

Region	SNP ID	Gene	SNP Class
1	rs1028180	*BLZF1*	amino acid: Q > R
1	rs3862937	*SLC19A2*	conserved intron
3	rs3827760	*EDAR*	amino acid: V > A
3	rs17261772	*RAB3GAP1*	conserved 3' UTR
4	rs1446585	*R3HDM1*	conserved intron
4	rs4988235	*LCT*	promoter
5	rs1513875		conserved intron
5	rs6706063		conserved intron
5	rs6706426		conserved intron
5	rs6758766		conserved intron
5	rs2037044		conserved noncoding
5	rs13005005		conserved noncoding
5	rs17626597		conserved noncoding
5	rs17627058		conserved noncoding
5	rs3770005	*PDE11A*	conserved noncoding
7	rs1047626	*SLC30A9*	amino acid: M > V
7	rs2660326	*SLC30A9*	conserved intron
7	rs3827590	*SLC30A9*	conserved intron
7	rs3827591	*SLC30A9*	conserved intron
7	rs4861155	*SLC30A9*	conserved intron
7	rs13756		conserved noncoding
8	rs11100128		conserved noncoding
11	rs10903929		conserved noncoding
13	rs16905686	*PCDH15*	transcription factor
13	rs4935502	*PCDH15*	amino acid: D > A
16	rs1426654	*SLC24A5*	amino acid: T > A
17	rs10851731	*HERC1*	conserved coding
17	rs2229749	*HERC1*	amino acid: E > D
17	rs2272209	*HERC1*	conserved intron
17	rs2228511	*HERC1*	conserved coding
17	rs6494428	*HERC1*	conserved intron
17	rs16947373	*HERC1*	conserved intron
19	rs2242406	*CHST5*	conserved 5' UTR
19	rs3743599	*ADAT1*	amino acid: T > N
19	rs6834	*KARS*	amino acid: T > S
21	rs9303429	*BCAS3*	conserved intron
21	rs6504005	*BCAS3*	conserved intron
21	rs6504010	*BCAS3*	conserved intron
24	rs1573662	*LARGE*	conserved 5' UTR
24	rs5999077	*LARGE*	conserved intron
24	rs1013337	*LARGE*	conserved 5' UTR

Table S10: Nonsynonymous, derived, differentiated alleles in HapMap2

SNP ID	Chromosome	HG17 position	Gene	CEU Derived Allele Freq	YRI Derived Allele Freq	JPT+CHB Derived Allele Freq	Fst percentile	Tested population	Long Haplotype Signal
rs12142199	1	1289110	CPSF3L	0.775	0.0083	0.0278	1	C	
rs2072994	1	11513736		0.6667	0.0167	0.2667	1	C	
rs2296224	1	20756858	KIF17	0.9917	0.0083	0.0444	1	C	
rs7537203	1	35895041	CLSPN	0.125	0.7083	0.7944	1	YJ	
rs2056899	1	47319871	CYP4A22	0.6167	0.0167	0	1	C	
rs1288389	1	53256618	PODN	0.1	0.6417	0.6722	1	YJ	
rs4915691	1	65579540	DNAJC6	0.975	0.325	0.3833	1	C	
rs1137100	1	65748462	LEPR	0.3417	0.1167	0.8278	1	J	
rs3819946	1	74887907	CRYZ	0.8833	0.3917	0.2222	1	C	
rs12041465	1	75321070	LHX8	0.1833	0.1167	0.8333	1	J	
rs2815413	1	93384744	CCDC18	0.85	0.2583	0.9611	0.99	CJ	
rs2229496	1	149695683	IVL	0.9167	0.3	0.3722	1	C	
rs2061690	1	151732153	PBXIP1	0.4333	0.0833	0.8944	1	J	
rs6682716	1	153364921		0.7333	0.0333	0.3056	1	C	
rs926103	1	153598055	SH2D2A	0.2583	0.2333	0.8722	1	J	**
rs12075	1	155988427	DARC	0.4833	0	0.9056	1	J	
rs1028180	1	166077526	BLZF1	0.0083	0.05	0.7	1	J	**
rs6020	1	166250770	F5	0	0.3167	0.7	1	J	
rs6696455	1	171819386	TNN	0.5333	0.0833	0.9222	1	CJ	
rs155443	1	186295442		0	0.65	0	1	Y	
rs6003	1	193762678	F13B	0.925	0.275	0.9611	1	CJ	
rs1361754	1	202533529	FLJ32569	0.6	0.3333	0	0.99	C	
rs291102	1	203494873	PIGR	0.025	0.85	0.1222	1	Y	
rs2070065	1	211199639	CENPF	0.9417	0.2083	0.8333	1	CJ	
rs2666839	1	211204619	CENPF	0.9417	0.2083	0.8333	1	CJ	
rs335524	1	211214591	CENPF	0.375	0.25	0.9056	1	J	
rs2275303	1	228845954	SIPA1L2	0	0	0.6	1	J	
rs2642992	1	243477598	ZNF695	0.3917	0.775	0.0056	1	Y	
rs7555046	1	244264946		0.8917	0.1083	0.9778	1	CJ	
rs7567833	2	3184917	COLEC11	0.9583	0.1333	0.9889	1	CJ	
rs2715860	2	9479134	DDEF2	0.6333	0.0667	0.8444	1	CJ	**
rs2288709	2	43915661	DYNC2LI1	0.6417	0.0333	0.8667	1	CJ	**
rs3813227	2	73563622	ALMS1	0.8	0.075	0.9889	1	CJ	
rs6546837	2	73589553	ALMS1	0.8	0.075	0.9889	1	CJ	
rs6546838	2	73590935	ALMS1	0.8	0.0917	0.9889	1	CJ	
rs6724782	2	73591645	ALMS1	0.8	0.075	0.9889	1	CJ	
rs6546839	2	73592163	ALMS1	0.8	0.075	0.9889	1	CJ	
rs2056486	2	73629222	ALMS1	0.8	0.0917	0.9889	1	CJ	
rs10193972	2	73629311	ALMS1	0.8	0.0917	0.9889	1	CJ	
rs1063588	2	74602033	GCS1	0.0917	0.8667	0.8333	1	YJ	
rs1047911	2	74611433	MRPL53	0.0917	0.85	0.8333	1	YJ	
rs6707475	2	74622146	FLJ12788	0.9	0.0083	0.1667	1	C	**
rs17009998	2	74636833	LBX2	0.0917	0.15	0.8333	1	J	
rs2231250	2	74667831	AUP1	0.1167	0.725	0.8333	1	YJ	

SNP ID	Chromosome	HG17 position	Gene	CEU Derived Allele Freq	YRI Derived Allele Freq	JPT+CHB Derived Allele Freq	Fst percentile	Tested population	Long Haplotype Signal
rs2305160	2	101049822	NPAS2	0.3167	0.0333	1	1	J	
rs1402467	2	108453326	SULT1C2	0.8083	0.0583	0.8833	1	CJ	**
rs3827760	2	108972119	EDAR	0	0	0.8667	1	J	**
rs9287519	2	132111197		0.1667	0	0.6778	1	J	
rs1438307	2	136332898		0.8167	0.05	0.4389	1	C	**
rs10186922	2	159556973		0.125	0.8167	0.8722	1	YJ	
rs6738031	2	167105429	SCN7A	0.3167	0	0.75	0.99	J	
rs10497520	2	179470361	TTN	0.8917	0.3417	0.2111	1	C	
rs4667001	2	185627253	C2orf10	0.6167	0.025	0.8667	1	CJ	
rs1366842	2	185627749	C2orf10	0.6167	0.025	0.8667	1	CJ	
rs13396213	2	201593744	NIF3L1	0.7667	0.2333	0.9944	1	CJ	
rs11890512	2	215736230	ABCA12	0.025	0.65	0	1	Y	
rs586194	2	219435938	TTLL4	0.6083	0.025	0.8556	1	CJ	
rs3731892	2	219961856		0.9	0.2333	0.3056	1	C	
rs394558	3	10277172	TATDN2	0.5833	0.6333	0.0389	1	CY	
rs1839022	3	27022506		0.3	1	1	1	YJ	
rs1126478	3	46476217	LTF	0.7333	0.0167	0.35	1	C	
rs887515	3	52498445	NISCH	0.8333	0.275	0.9944	0.99	CJ	
rs1131356	3	58084202	FLNB	0.2167	0.4667	0.9222	1	J	
rs12632456	3	58093595	FLNB	0.2083	0.575	0.9722	1	YJ	
rs9868484	3	109671683		0.7583	0.675	0	1	CY	
rs9288952	3	113667715	BTLA	0.9583	0.125	0.7278	1	CJ	
rs2306857	3	114209874	C3orf17	0.8417	0.1333	0.3833	1	C	
rs11539377	3	120702263	C3orf1	0.9917	0.375	0.9889	1	CJ	
rs17310144	3	125148592	CCDC14	0.675	0.075	0.0944	1	C	**
rs641320	3	139830655	FAIM	0.9583	0.1833	0.9889	1	CJ	
rs13043	3	139830686	FAIM	0.0083	0.6667	0	1	Y	
rs11499	3	182176757	FXR1	0.9917	0	0	1	C	
rs734312	4	6421426	WFS1	0.7167	0	0.8556	1	CJ	
rs2227852	4	9460634	DRD5	0.9917	0	0	1	C	**
rs3733591	4	9598399	SLC2A9	0.1917	0.0333	0.7056	0.99	J	
rs4590080	4	41475705	KFZP686A012	0.9583	0.225	0.9833	1	CJ	
rs1047626	4	41844599	SLC30A9	0.7333	0.0583	0.9667	1	CJ	**
rs5825	4	46567077		0.9917	0	0	1	C	
rs2289443	4	75388744	MTHFD2L	0.9583	0.1583	0.85	1	CJ	
rs17014118	4	89676474	HERC6	0.7917	0.1417	0.3222	1	C	
rs1229984	4	100596497	ADH1B	0	0	0.7556	1	J	
rs10009368	4	135479206		0.7167	0.2583	0.0056	1	C	
rs11559290	4	159959281	ETFDH	0.8167	0.125	0.9833	1	CJ	
rs2438652	5	10292261	LOC134145	0.1333	0.0833	0.8667	1	J	
rs16891982	5	33987450	SLC45A2	1	0	0	1	C	**
rs37369	5	35072872	AGXT2	0.1083	0.7	0.5944	1	YJ	
rs1864183	5	81584972	ATG10	0.4583	0.175	0.9333	1	J	
rs1864182	5	81584996	ATG10	0.4167	0.825	0.05	1	Y	

SNP ID	Chromosome	HG17 position	Gene	CEU Derived Allele Freq	YRI Derived Allele Freq	JPT+CHB Derived Allele Freq	Fst percentile	Tested population	Long Haplotype Signal
rs12515587	5	141229147	PCDH1	0.8583	0.1917	0.9778	1	CJ	
rs7709485	5	145875089	GPR151	0.1583	0.2	0.7778	1	J	
rs2256966	5	178346422	GRM6	0.2333	1	0.9722	1	YJ	
rs10060182	5	179218358	LOC51149	0.175	0.15	0.7944	1	J	
rs11738161	5	180052616		0.75	0.15	0.2444	1	C	
rs1042391	6	16398740	GMPR	0.65	0.0917	0.0944	1	C	
rs2274305	6	24399182	DCDC2	0.675	0.0083	0.7333	1	CJ	
rs2229642	6	33767450	ITPR3	0.575	0.85	0.1056	1	CY	
rs4713668	6	33798774	IHPK3	0.3833	0.0833	0.8222	1	J	
rs9349180	6	41293452		0.775	0.1583	0.8944	1	CJ	
rs239798	6	54913647	FAM83B	0.825	0.225	0.9889	1	CJ	
rs7383447	6	80077256		0.6083	0.0167	0.1778	1	C	
rs7745023	6	121619069	C6orf170	0.6167	0.0583	0.0833	1	C	
rs675531	6	128082532	C6orf190	0.2167	0.275	0.8611	1	J	
rs1044498	6	132214061	ENPP1	0.8667	0.075	0.9389	1	CJ	
rs6926101	6	133146813	C6orf192	0.95	0.3167	0.9889	1	CJ	
rs4236176	6	169888355	WDR27	0.3	0.4083	0.9222	1	J	
rs1078211	6	170018932	C6orf208	0.55	0.0667	0.0167	1	C	
rs2301721	7	26969353	HOXA7	0.85	0.0667	0.8444	1	CJ	
rs11765552	7	97466766	LMTK2	0.525	0	0.0722	1	C	
rs542137	7	100017728	ZAN	0.3667	0.1583	0.85	1	J	
rs539445	7	100018018	ZAN	0.6333	0.8417	0.15	1	CY	
rs1627354	7	107271935	LAMB4	0.0417	0.7167	0	1	Y	
rs10260756	7	107283795	LAMB4	0	0.7	0.0056	1	Y	
rs2908004	7	120563720	WNT16	0.6333	0.0667	0.8444	1	CJ	
rs10265	7	138483407	HSPC268	0.8083	0.0917	0.7833	1	CJ	
rs7781826	7	143569849		0.25	0.35	0.9056	1	J	
rs2948305	8	8135987		0.4	0.0083	0.8222	1	J	
rs6601495	8	10517787	RP1L1	0	0.8583	0	1	Y	
rs7461273	8	11815386		0.575	0.1417	0.9611	1	CJ	
rs4871857	8	23115269	TNFRSF10A	0.5917	0.7583	0.0222	1	CY	
rs6557634	8	23116201	TNFRSF10A	0.4083	0.25	0.9778	1	J	
rs323344	8	30822067	TEX15	0.1167	0.9	0.05	1	Y	
rs323345	8	30822144	TEX15	0.8833	0.1	0.95	1	CJ	
rs323346	8	30822973	TEX15	0.1833	0.8417	0.0889	1	Y	
rs323347	8	30825766	TEX15	0.8167	0.1	0.9111	1	CJ	
rs3924999	8	32572900	NRG1	0.3583	0.0167	0.7944	1	J	
rs7818806	8	50816259		0.8167	0.1083	0.8333	1	CJ	
rs6987308	8	144847859	ZNF707	0.0917	0.5	0.75	1	J	
rs1871534	8	145610489	SLC39A4	0	0.9833	0	1	Y	
rs3747532	9	14712477	CER1	0.6417	0.125	0.9556	1	CJ	
rs10972048	9	34300927		0.7083	0.0333	0.3111	1	C	
rs2282192	9	97751893	C9orf156	0.275	0.85	0.8	1	YJ	**
rs1265891	9	114189666	AKNA	1	0.275	0.9556	1	CJ	

SNP ID	Chromosome	HG17 position	Gene	CEU Derived Allele Freq	YRI Derived Allele Freq	JPT+CHB Derived Allele Freq	Fst percentile	Tested population	Long Haplotype Signal
rs10985704	9	122410232	OR1L8	0.5083	0.4333	0.0056	0.99	C	
rs1476859	9	122470681	OR1B1	0.6667	0.975	0.2056	1	CY	
rs1572912	9	128645108	TBC1D13	0.7417	0.0417	0.6833	1	CJ	
rs2966332	9	131212933	PPAPDC3	0	0.4083	0.7278	1	J	
rs543573	9	132232383	SETX	0.9	0.3083	0.3222	1	C	**
rs1183768	9	132232785	SETX	0.9	0.3083	0.3222	1	C	**
rs602990	9	133673548	VAV2	0.475	0.0833	0.9889	1	J	
rs15772	10	15185861	RPP38	0.75	0.05	0.9167	1	CJ	
rs7074847	10	22715863	SPAG6	0.0083	0.675	0	1	Y	**
rs4935502	10	55625450	PCDH15	0.1583	0.1667	0.8944	1	J	**
rs4536103	10	71002210	NEUROG3	0	0.9833	0.9889	1	YJ	
rs10785923	10	91728536		0.525	0.3833	0.9667	1	CJ	
rs2862954	10	101902054	SPFH1	0.5333	0.0083	0.0667	1	C	
rs7099565	10	116709533	TRUB1	0.6083	0.2	0.9722	1	CJ	
rs10794208	10	126905364		0.6167	0.0917	0.9056	1	CJ	
rs331537	11	4427852	OR52K2	0.0333	0.775	0.0333	1	Y	
rs1462983	11	6086413	OR56B4	0.4833	0.05	0.8667	1	J	
rs7130656	11	45789085	SLC35C1	0.325	0.775	1	1	YJ	
rs3736508	11	45931706	PHF21A	0.0167	0	0.5889	1	J	
rs2260655	11	60865550	DAK	1	0.3417	0.9667	1	CJ	
rs7103126	11	68819969	MYEOV	0.8417	0.1167	0.5111	0.99	CJ	
rs557881	12	189386	SLC6A12	0.5333	0.325	1	1	CJ	
rs12319376	12	1424058	ERC1	0.9167	0.275	0.9889	1	CJ	
rs1984564	12	6960454	MBOAT5	0.95	0.3083	0.9722	1	CJ	
rs1124164	12	10640842	KLRA1	0	0.6167	0	1	Y	
rs708167	12	27126266		0.5083	0.0333	0.0278	1	C	
rs7133970	12	51412341		0.7667	0.2583	0.9944	0.99	CJ	
rs2171497	12	53630400		0.1083	0.0417	0.6778	1	J	
rs939875	12	63555314		0.9583	0.25	1	1	CJ	
rs7978197	12	67612814	CPM	0.9833	0.3917	1	0.99	CJ	
rs10777084	12	86882562	C12orf50	0.1083	0.8083	0.0389	1	Y	
rs4964460	12	105207441	TCP11L2	0.1417	0	0.7	1	J	
rs3742000	12	110801259		0.775	0.1833	0.1111	1	C	
rs12231744	12	110939775	C12orf30	0	0.0583	0.6	0.99	J	
rs7318174	13	19036252		0.7333	0.1333	0.9333	1	CJ	
rs7995033	13	24729888	MTMR6	0.8667	0.0583	0.5333	1	CJ	
rs1056820	13	40413286	ELF1	0.7167	0.0917	0.2222	1	C	
rs17099455	14	23492847	DHRS4	0.9417	0.2833	0.9833	1	CJ	
rs2229309	14	23908923	NFATC4	0.6333	0.0417	0.1722	1	C	
rs7149586	14	23915681	NFATC4	0.3	0.8417	0.8278	1	YJ	
rs2274068	14	35222928	GARNL1	0.8583	0.275	0.3111	1	C	
rs2274271	14	54725445	DLG7	0.9083	0.15	0.7444	0.99	CJ	
rs3742578	14	56742468	EXOC5	0.1333	0.8833	0.0667	1	Y	
rs11844594	14	76913567	C14orf174	0.4917	0.0917	0.9222	1	J	

SNP ID	Chromosome	HG17 position	Gene	CEU Derived Allele Freq	YRI Derived Allele Freq	JPT+CHB Derived Allele Freq	Fst percentile	Tested population	Long Haplotype Signal
rs2193595	14	76914874	C14orf174	0.4833	0.0917	0.9222	1	J	
rs3742728	14	77020877	THSD3	0.5833	0.1417	0.9444	1	CJ	
rs1800414	15	25870632	OCA2	0	0	0.5833	1	J	
rs8040932	15	27134311	APBA2	0.85	0.175	0.8667	1	CJ	
rs936212	15	38368835	PLCB2	0	0	0.5889	1	J	
rs12911738	15	38690976	CASC5	0.1083	0.075	0.6944	1	J	
rs8040502	15	38702482	CASC5	0.1083	0.075	0.6944	1	J	
rs3816533	15	39921389	PLA2G4B	0.1583	0	0.7056	1	J	
rs1456235	15	39936764	SPTBN5	0.6	0.0667	0.9667	1	CJ	
rs7181742	15	40430821	GANC, CAPN3	0.9	0.225	0.8778	0.99	CJ	
rs1801449	15	40468491	CAPN3	0.9417	0.2333	0.9111	1	CJ	
rs12917189	15	40810774	CDAN1	0.75	0.0083	0.7167	1	CJ	
rs689647	15	41549488	TP53BP1	0.0667	0.8	0.4778	1	Y	
rs2245715	15	41605344	MAP1A	0.075	0.7417	0.5167	1	YJ	
rs1704792	15	43008164		0.95	0.0417	0.55	1	CJ	
rs269868	15	43179367	DUOX2	0.9833	0.1667	0.9333	1	CJ	
rs11854484	15	43332770	SLC28A2	0.7	0.0917	0.0667	1	C	
rs1060896	15	43341559	SLC28A2	0.7333	0.0833	0.0667	1	C	
rs1288775	15	43448970	GATM	0.825	0.1167	0.1833	1	C	
rs1426654	15	46213776	SLC24A5	1	0.025	0.0111	1	C	**
rs2229749	15	61724262	HERC1	0.0417	0.1833	0.9056	1	J	**
rs2010875	15	62944535	PLEKHQ1	0.075	0.175	0.7944	1	J	**
rs5742915	15	72123686	PML	0.55	0.0167	0.0056	1	C	
rs1036938	15	77024302	CTSH	0.2917	0.8583	0.8722	1	YJ	
rs2242046	15	83279733	SLC28A1	0.5083	0	0.0778	0.99	C	
rs2106673	15	89253599	MAN2A2	0.775	0.7333	0.0111	1	CY	
rs11073964	15	89344765	VPS33B	0.7	0.0167	0	1	C	
rs3747579	16	4385328	CORO7	0.7333	0.05	0.8333	1	CJ	
rs749670	16	30996126	ZNF646	0.375	0	0.9167	1	J	
rs7193955	16	46680083	ABCC12	0.8167	0.0667	0.9222	1	CJ	
rs17822931	16	46815699	ABCC11	0.125	0	0.9333	1	J	
rs11860295	16	65873735	PLEKHG4	0.075	0.8	0.0056	1	Y	
rs3868142	16	65877724	PLEKHG4	0.075	0.7833	0.0056	1	Y	
rs8052655	16	65966681	LRRC36	0.0417	0.7083	0.0056	1	Y	
rs3743599	16	74204077	ADAT1	0.05	0	0.7611	1	J	**
rs3743598	16	74204186	ADAT1	0.8167	0.175	0.1556	1	C	**
rs11640912	16	75917420	ADAMTS18	0.6417	0.8667	0.1444	1	CY	
rs12918952	16	76978276	WWOX	0.6333	0.0833	0.0944	1	C	
rs16956174	16	80591113	HSPC105	1	0.3333	0.9889	1	CJ	
rs462769	16	88290764	C16orf76	0.3667	0.8083	0.9833	1	YJ	
rs7195066	16	88363824	FANCA	0.7083	0.0167	0.0056	1	C	
rs7190823	16	88393544	FANCA	0.5667	0.15	0.0111	1	C	
rs703903	17	3142253	OR3A1	0.5583	0.275	1	1	CJ	
rs224534	17	3433451	TRPV1	0.2667	0.0083	0.8222	1	J	

SNP ID	Chromosome	HG17 position	Gene	CEU Derived Allele Freq	YRI Derived Allele Freq	JPT+CHB Derived Allele Freq	Fst percentile	Tested population	Long Haplotype Signal
rs9899177	17	4999532	USP6	0.6667	0.1667	0.0944	1	C	
rs2189335	17	5266869	RPAIN	0.1583	0.1667	0.7722	1	J	
rs2287499	17	7532893	WDR79	0.8417	0.0833	0.7222	1	CJ	
rs11649804	17	17637480	RAI1	0.2583	0.3083	0.8556	0.99	J	
rs3818717	17	17647830	RAI1	0.7083	0.1333	0.0556	1	C	
rs3183702	17	17688014		0.2667	0.375	0.9389	1	J	
rs7225888	17	26322430	RNF135	0.9917	0.175	0.9944	1	CJ	
rs6505228	17	26399303		0.9917	0.1833	1	1	CJ	
rs1003645	17	31364397	CCL23	0.1917	0.9667	0.4167	0.99	Y	
rs1058808	17	35137563	ERBB2	0.7083	0	0.1667	1	C	
rs9891361	17	36913439	KRT13	0.925	0.2	0.9222	1	CJ	
rs2074158	17	37510689	LGP2	0.1917	0.85	0.1389	1	Y	
rs9909488	17	40695542		0.975	0.2583	0.9944	1	CJ	
rs550510	17	44281614	CALCOCO2	0.15	0	0.7056	1	J	
rs3760413	17	45807775	EME1	0.1	0	0.7444	1	J	
rs2643103	17	56141407	BCAS3	1	0.2083	0.8722	1	CJ	**
rs6504233	17	59918244	POLG2	0.0167	0.6667	0	1	Y	
rs6504234	17	59918318	POLG2	0.075	0.8083	0.0333	1	Y	
rs1427463	17	59923044	POLG2	0.075	0.8083	0.0333	1	Y	
rs4581	17	61641219	APOH	0.15	0.6083	0.7722	1	YJ	
rs2056439	17	76893612	C17orf55	0.7833	0.075	0.7222	1	CJ	
rs4891392	18	65869668	RTTN	0.05	0.725	0	1	Y	
rs3911730	18	66022323	RTTN	0.8917	0.0667	1	1	CJ	
rs687320	19	6024382	RFX2	0.9917	0.35	1	1	CJ	
rs2240227	19	15713242	OR10H3	0.0583	0	0.6	0.99	J	
rs2608738	19	16760865	LOC284434	0.9667	0.375	1	0.99	CJ	
rs2302970	19	37790472	ANKRD27	0.6417	0.0167	0.0778	1	C	
rs6510426	19	39727713		0.8583	0.25	0.9889	1	CJ	
rs30461	19	44480955	IL29	0.8833	0.25	0.9611	1	CJ	
rs8110904	19	47723209	CEACAM1	0.0083	0.5667	1	1	YJ	
rs7260180	19	49720009	CEACAM20	0.6583	0.0833	0.1722	1	C	
rs447802	19	53808171	FAM83E	0.7667	0.2	0.2111	1	C	
rs601338	19	53898486	FUT2	0.5417	0.5417	0.0111	1	CY	
rs602662	19	53898797	FUT2	0.6083	0.5417	0.0111	1	CY	
rs1559155	19	54324584	PPFIA3	0.2917	0.0083	0.8	1	J	
rs7246479	19	60516144		0.5	0	0.8222	1	J	
rs2076015	20	7911041	TXNDC13	0.075	0.6917	0.55	1	YJ	
rs947310	20	29912986	DUSP15	0.6	0.1833	1	1	CJ	
rs4911287	20	31090952	BPIL3	0.6917	0.0083	0.2833	1	C	
rs2274934	20	60330882	LAMA5	0.65	0.05	0.7889	1	CJ	
rs3810548	20	60339273	LAMA5	0.025	0.6333	0	1	Y	
rs2071152	21	44327549	TMEM1	0	0.7833	0.0056	1	Y	
rs8131523	21	45666871	.18A1, C21ori	0.1083	1	0.15	1	Y	
rs2073748	22	18343525	ARVCF	1	0.375	0.2222	1	C	

SNP ID	Chromosome	HG17 position	Gene	CEU Derived Allele Freq	YRI Derived Allele Freq	JPT+CHB Derived Allele Freq	Fst percentile	Tested population	Long Haplotype Signal
rs2236005	22	24747534	MYO18B	0.8417	0.0667	0.7389	1	CJ	
rs743920	22	27945682	EMID1	0.925	0.1833	0.4611	1	C	
rs1812240	22	41236604	CGI-96	0.0333	0.15	0.7	1	J	
rs137055	22	41294530	SERHL2	0.0583	0.5667	0.6611	1	YJ	
rs138993	22	41934705	SCUBE1	0.2417	0.6417	0.9278	1	YJ	
rs7410764	22	44794719		0.6667	0.0833	0.8389	0.99	CJ	
rs4044210	22	45106834	CELSR1	0.225	0.775	0.0222	0.99	Y	**
rs6008794	22	45108213	CELSR1	0.7833	0.225	0.9944	1	CJ	**
rs910799	22	48599429	ZBED4	0.1917	0.875	0.1722	1	Y	
rs1321	22	48618296	ALG12	0.8	0.125	0.8278	1	CJ	
rs8139422	22	48636224	CRELD2	0.0167	0.6917	0.0556	1	Y	
rs8142477	22	49301520	CPT1B, CHKB	0.9333	0.2583	0.5222	1	CJ	
rs3747295	23	17505901	NHS	1	0.0556	0.5259	1	CJ	
rs1385699	23	65608007	EDA2R	0.7889	0	1	0.99	CJ	**
rs1343879	23	74787550	MAGEE2	0.0222	0.0111	0.9111	1	J	
rs3115758	23	128507399	APLN	0.8889	0.0778	0.2593	1	C	
rs1059702	23	152805039	IRAK1	0	0.0111	0.8074	1	J	

Table S11 Reported copy-number- variant regions (CNVs), in our top 22 candidates for selection (Methods).

Candidate Region Chr:Position (MB,HG17)	CNV Start	CNV End	CNV Length (kb)	Reference
chr2:108.7	108391709	108536212	145	Locke
chr2:108.7	107908395	108681855	773	Redon
chr2:136.1	137031522	137074036	43	Redon
chr4:33.9	34631676	34916911	285	Sebat
chr4:33.9	34599887	34749291	149	Iafrate
chr4:33.9	34599808	34652577	47	McCarroll
chr4:33.9	33971807	34297097	325	Redon
chr4:33.9	34459337	34779109	320	Redon
chr4:33.9	34597970	34664260	66	Conrad
chr4:33.9	33183778	33220671	37	Conrad
chr4:159	159085344	159091948	7	Conrad
chr4:159	158907924	158955096	47	Conrad
chr10:22.7	2650802	3188773	538	Redon
chr12:78.3	78657407	78662732	5	Tuzun
chr15:46.4	46296330	46451070	155	Redon
chr16:74.3	73977115	74134472	157	Redon
chr16:74.3	74567282	74579010	12	Redon
chr19:43.5	44919744	45212970	293	Redon
chr22:32.5	32494349	32528033	34	Redon
chr23:63.5	63787842	63841340	47	McCarroll
chr23:63.5	62130137	62294591	164	Locke
chr23:63.5	61984759	62329082	344	Redon
chr23:63.5	63292574	63919710	627	Redon

Conrad, D. F. et al. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38, 75-81 (2006)
Iafrate, A. J. et al. Detection of large-scale variation in the human genome. Nat Genet 36, 949-51 (2004).
Locke, D. P. et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79, 275-90 (2006).
McCarroll, S. A. et al. Common deletion polymorphisms in the human genome. Nat Genet 38, 86-92 (2006).
Redon, R. et al. Global variation in copy number in the human genome. Nature 444, 444-54 (2006).
Sebat, J. et al. Large-scale copy number polymorphism in the human genome. Science 305, 525-8 (2004).
Tuzun, E. et al. Fine-scale structural variation of the human genome. Nat Genet 37, 727-32 (2005).