Supplementary Tables

Table S1. Detailed data production information for each sample.

See supplementary data set: TableS1_DataProductionSummary.xls

Table S2. Genotyping validation of novel SNPs identified in this study.

Minor allele count	# of SNPs tested	# validated	# not validated	Validation rate (%)
1	74	67	7	91
2	19	19	0	100
3	9	9	0	100
4	4	4	0	100
5	5	5	0	100
6	7	7	0	100
7	1	1	0	100
>=8	21	21	0	100
Total	**140**	**133**	**7**	95

Table S3. Sequenom iPex genotyping results and sequencing results of each sample individual at genotyped sites. Only Q20 bases were counted (format: "genotyping allele|sequencing allele"; '-' will be given if failure genotyping, missing data or genotype called with quality less than 20).

See supplementary data set: Table S3_ArrayGenotypingResult.xls

Table S4. Summary of annotated SNPs that are discovered by aggregating data from 200 individuals. Note that the exome in this study refers to the full target region, which also contained a part of intronic and intergenic regions. CDS stands for coding sequences, UTR for untranslated regions.

Exon				
CDS				
synonymous	non-synonymous	UTR	intron/intergenic	Total
25,275	27,806	5,967	62,822	121,870

Table S5. Putative extrapolation estimation of SNP counts in each individual.

See supplementary data set: Table S5_IndividualSNPsummary.xls

Table S6. 20 genes with highest HK score for positive selection. P and F indicate the number of fixed and polymorphic substitutions observed. We excluded *SPTAN1*,

Nature Genetics: doi: 10.1038/ng.680

BPTF, and *TEX15*, as their substitution patterns are suggestive of biased gene conversion.

Gene Symbol	Description	F	P	F/P	Score
KIR3DP1	killer-cell Ig-like receptor	82	10	8.20	>7
LILRA1	leukocyte immunoglobulin-like receptor,	60	7	8.57	7
TPTE	transmembrane phosphatase with tensin homology	86	16	5.38	7
KIR2DL1	killer cell immunoglobulin-like receptor, two	40	3	13.33	6.05
VPS13D	vacuolar protein sorting 13D isoform 1	39	4	9.75	5.19
FLG	filaggrin	99	28	3.54	5.03
CES2	carboxylesterase 2 isoform 1	22	0	∞	4.95
TPRX1	tetra-peptide repeat homeobox	22	0	∞	4.95
HMCN1	hemicentin 1	62	15	4.13	4.12
TRPM2	transient receptor potential cation channel,	32	4	8.00	3.92
KIR2DL3	killer cell immunoglobulin-like receptor, two	34	5	6.80	3.76
KIAA1199	KIAA1199	21	1	21.00	3.75
SORBS2	sorbin and SH3 domain containing 2 isoform 2	24	2	12.00	3.62
TTC26	tetratricopeptide repeat domain 26 isoform 1	16	0	∞	3.60
SULT1C3	sulfotransferase family, cytosolic, 1C, member	33	5	6.60	3.59
HERC2	hect domain and RLD 2	43	9	4.78	3.50
SGTA	small glutamine-rich tetratricopeptide	15	0	∞	3.37
DYNC1H1	cytoplasmic dynein 1 heavy chain 1	47	11	4.27	3.37
CBWD2	COBW domain-containing protein 2	19	1	19.00	3.33
CSHL1	chorionic somatomammotropin hormone-like 1	22	2	11.00	3.24

Table S7. GO categories showing an excess of fixed (F) compared to polymorphic (P) substitutions. The 10 terms with F+P>600 and the highest scores are shown. The score is the –log10(p-value), calculated using the FUNC package with a correction for multiple testing (Supplementary Note).

GO terms	F	P	F/P	Score
muscle contraction	449	239	1.88	2.20
positive regulation of macromolecule metabolic process	970	589	1.65	1.86
regulation of transcription, DNA-dependent	4836	3130	1.55	1.85
sodium ion transport	588	337	1.74	1.81
regulation of localization	879	539	1.63	1.66
cell migration	762	465	1.64	1.63
regulation of locomotion	384	222	1.73	1.57
defense response	1648	966	1.71	1.56
positive regulation of cellular metabolic process	1026	607	1.69	1.39
immune system process	2087	1318	1.58	1.31

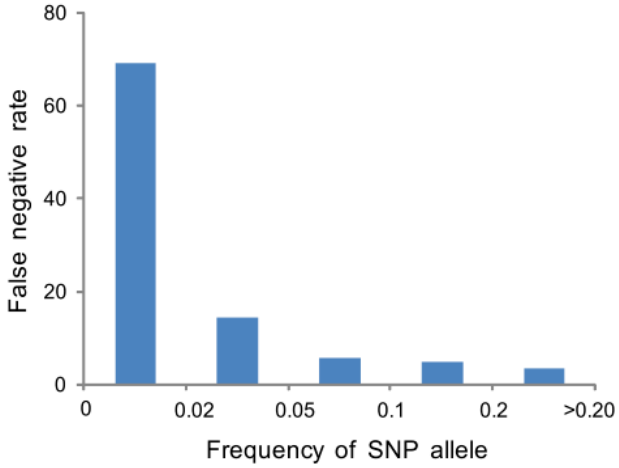
Supplementary Figures

Figure S1. False negative rate varies with SNP frequencies in (a) target region and (b) target region where we calculate allele frequency (depth >600). (c) Number of newly discovered SNPs when sequentially adding sampled individuals. Blue for SNPs with MAF >0.02; red for all. A diminishing increment is observed for identification

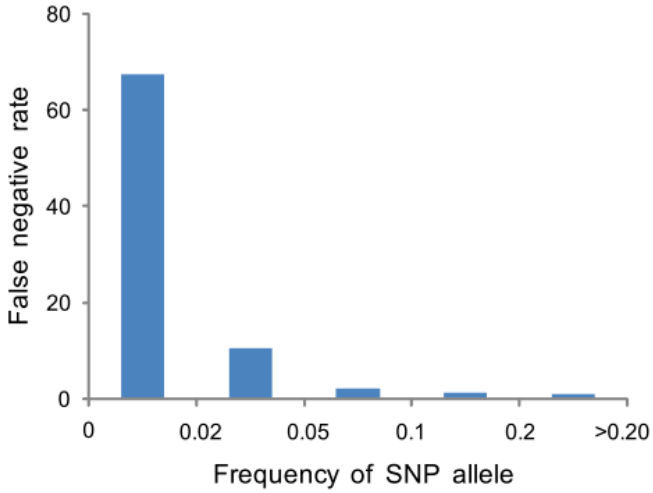
Nature Genetics: doi: 10.1038/ng.680

of SNPs with MAF>0.02, whereas discovery of all SNPs does not show a saturation as a function of the number of sampled individuals.

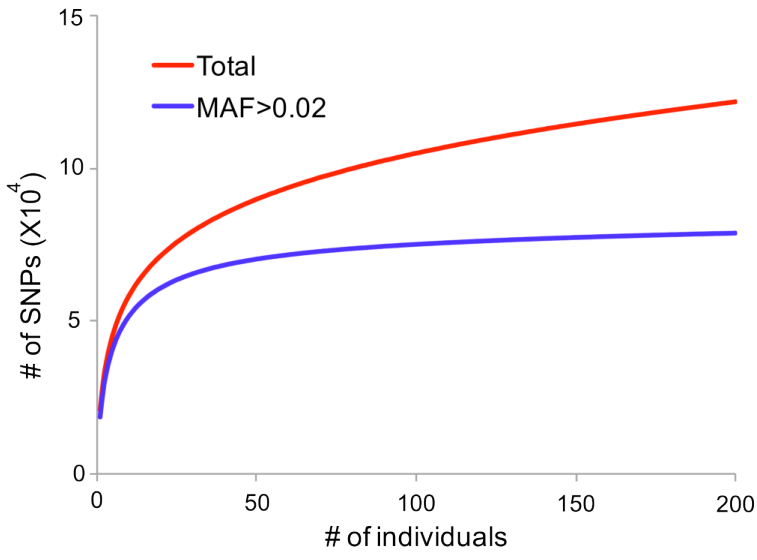
a



b



c



Nature Genetics: doi: 10.1038/ng.680

Figure S2. Site frequency spectra (SFS) for non-synonymous (Nsyn.) and synonymous (Syn.) SNPs were compared to the neutral expectation in A. The frequencies plotted in this panel are from reads with quality value >30. Panel B displays the SFS for 2 categories of non-synonymous SNPs with physicochemical distances between the two amino acid variants >95 or ≤95, respectively.

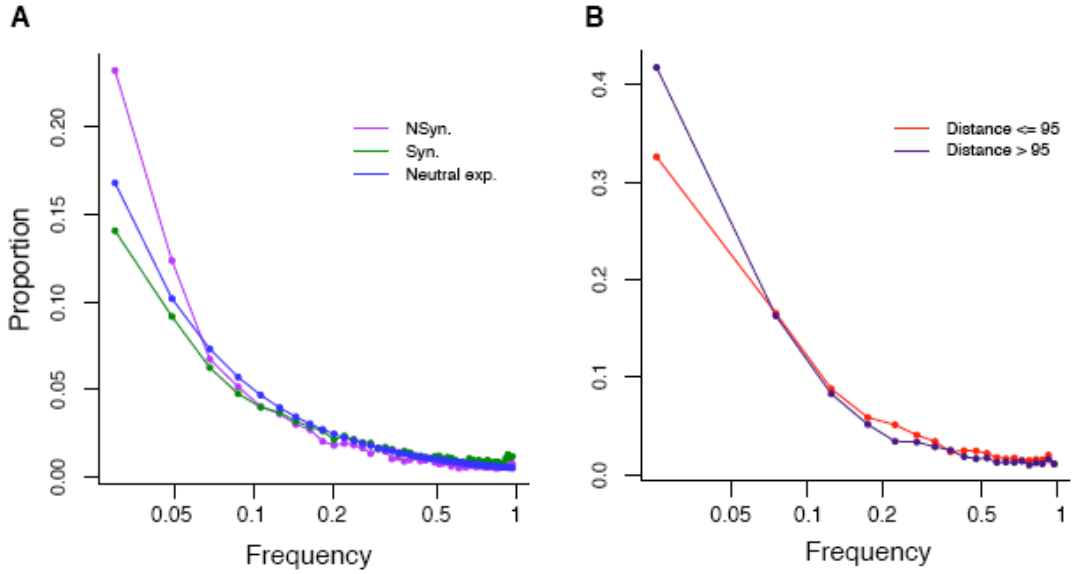
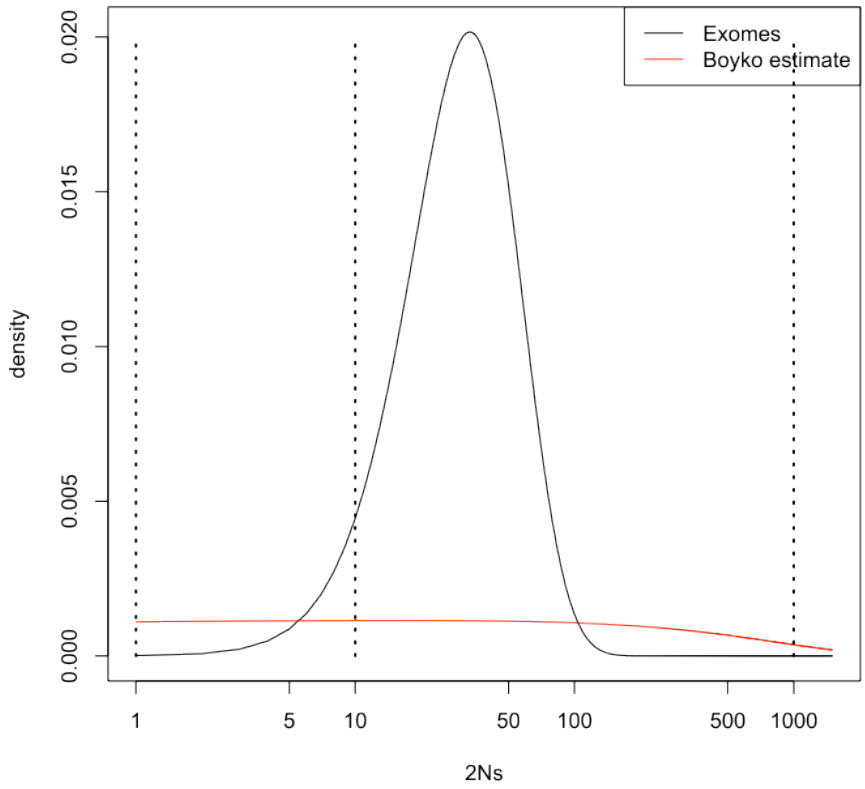


Figure S3. Comparison of the gamma distributions of selective effects obtained from our data (black) and from previous estimates (red). Note the log scale on the x-axis. The black dotted lines represent the break points s=0.00001 (2Ns ~ 1), s=0.0001 (2Ns ~ 10) and s=0.01 (2Ns ~1000) where N = 52907 and is the estimate of the effective population size in Europeans (ref. 10).

Nature Genetics: doi: 10.1038/ng.680



Supplementary Note

Sample acquisition

Genomic DNA was purified from blood leukocytes from 200 individuals of Danish nationality: 108 men (50.8 +/- 5.9 years (mean+/- SD) and Body Mass Index (BMI) 24.5 +/- 1.9 kg/m2; and 92 women (53.7 +/- 4.3 years and BMI: 23.6 +/- 2.2 kg/m2) were recruited at random through the central person register in the northern part of the Copenhagen region. The study subjects were examined at the Research Centre for Prevention and Health, Glostrup Hospital, and Steno Diabetes Center, Gentofte, Denmark. All study participants provided written informed consent and the study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Copenhagen County, Denmark.

Exon-capturing and sequencing

Nature Genetics: doi: 10.1038/ng.680

Following the manufacturer protocol, genomic DNA of each individual was hybridized with NimbleGen 2.1M-probe sequence capture array[1] to enrich the exonic DNA in each library. The array was able to capture 18,654 (92%) of the 20,091 genes that have been deposited in Consensus Coding Sequence Region (CCDS) database. We constructed a secondary library from the primary captured DNA, which enabled the Illumina Genome Analzyer (GA) II platform, as described[2] with adaptations. In brief, we broke down the DNA sample input into fragments with several hundred base-pairs in length and ligated each fragment with PCR linkers. Single-stranded fragments were hybridized to NimbleGen 2.1M HD sequence capture array. Fragments that did not hybridize with array probes were washed out. The probe-hybridized fragments were then eluted and subjected to ligation-mediated PCR (LM-PCR). We defined the current library the "primary library". DNA fragments in the primary library were concatenated by DNA ligase to 2k-3k fragments. We then sheared them to ~200 bp small molecules and ligated them with Illumina sequencing adaptors. The secondary library was subjected to Illumina Genome Analyzer (GA) sequencing. We performed GA sequencing for each sample independently to ensure each sample had at least 12-fold coverage. Raw image files were processed by Illumina Pipeline (version 1.3.4) for base-calling and to generate the reads set. SOAPaligner[3,4] (v2.01) was used to align the sequencing reads to the NCBI human genome reference assembly (build 36.3) with parameters set to "-a -D -o -r 1 -t -c -f 4". Reads that aligned to the designed target region (TR) were collected for SNP identification and subsequent analysis.

Genotype calling and SNP calling

To take advantage of aggregating sequencing data from the whole population, we simultaneously check the genotype likelihoods at a site of all individuals. The

Nature Genetics: doi: 10.1038/ng.680

population-based integrative SNP ascertainment has significantly larger statistical power to detect SNP site than individual-based methods. For each genomic site, a heuristic value S_k was calculated for each alternative (different from reference) allele count k over all sample individuals as following:

$$S_k = \prod_{i=1}^{N}\left(\sum_{j=0}^{2}\left(P(o_i|j)\times\frac{(2N)!}{k!(2N-k)!}\times\left(\frac{k}{2N}\right)^{j}\times\left(\frac{2N-k}{2N}\right)^{2-j}\right)\right)$$

where N is the total number of individuals (200 in our study), o_i is the observed read bases in the i-th individual, and j is the number of alternative alleles in a diploid genome, i.e. j=0 denotes homozygote reference alleles, j=1 denotes heterozygotes and j=2 for alternative homozygotes. $P(o_i|j)$ is the likelihood of genotype j on each individual, which was calculated by SOAPsnp[5] with parameter "-F 1" set. Note that the heuristic formula was not mathematically justified as typical maximum likelihoods estimation, which examines all cases. It assumed the 400 haploids from all individuals were independent and used $k/2N$ to approximate the population minor allele frequency.

Then, for all k=0,1,2,...,400, we picked the k with the largest S_k value as an approximate estimate of minor allele count. The true minor allele count k was defined as k_{max}. If k_{max} =0, this site was not considered a SNP. Otherwise (k_{max} >0), the site was considered a potential SNP. In total, approximately 530,000 SNPs were potential polymorphic. To evaluate the confidence of a SNP call, we used a PHRED scaled value ratio:

Nature Genetics: doi: 10.1038/ng.680

$$Q = 10\log_{10}\left(\frac{S_{k_{\max}}}{S_0}\right)$$

We used a Q20 threshold to filter unreliable SNP calls and the SNPs passing this cutoff were ascertained. Overall, by the heuristic formula, higher true alternative allele frequency and/or more sequencing reads resulted in larger value ratio favoring a SNP call, which is pertinent for polymorphism identification. In testing the method, we found that with such scale of sample individuals, it was more conservative to choose the highest-scored minor allele count than the true maximum likelihood estimate (biased to underestimation). Thus it was only used for confident SNP calls, but not for subsequent analysis.

The Q20 threshold filtered the potential SNP set down to approximately 170k. We then filtered the SNPs based on the following criteria: 1) The SNP should be observed in at least one individual in a way that the number of reads containing mutant alleles was larger than the reads containing reference alleles; 2) The SNPs should not be significantly enriched in heterozygous state. Use of both criteria were to avoid possible reproducible errors in the exome capturing process. We checked whether the SNPs followed Hardy-Weinberg law and found only 1.9% SNPs had H-W statistics <0.01, which indicated the SNP calling was accurate after using these filtering thresholds, and, thus, provided a solid basis for subsequent population analysis.

We then assigned the SNP alleles back to each individual by independent SNP calling using SOAPsnp, with the prior probability set to fit the population allele frequency. The parameter was set to "-i -d -o -r 0.00005 -e 0.0001 -M -t -u -L -s -2 -T", where option "-s" input the SNPs ascertained from population SNP calling. Thus, we also obtained the genotype calls for each individual.

Nature Genetics: doi: 10.1038/ng.680

Evaluation of SNP detection power

To assess the accuracy of SNP calls on the sample population, we randomly selected 140 novel SNPs for genotyping validation by Sequenom iPlex platform. Per each genotyping test, the signal was analyzed by the manufacturer's software to obtain an evaluation of genotyping quality. Per each site, if no fewer than 90% of the individuals had their genotyping quality at "A" (most confident) level, we took the site as having an overall high quality and used it in comparison.

Of all tested sites, 133 were true positives. As novel SNPs composed 44% of all identified SNPs, we estimated that the overall false positive rate was about (140-133)/140*44%=2%. The SNPs projected for subsequent analysis had MAF>0.02, and thus were of high quality.

We also used SNPs that had been genotyped in the HapMap CEU population to evaluate the false negative rate at each minor allele frequency (Fig. S1a), and we estimated that 5.1% of SNPs with a minor allele frequency (in HapMap CEU population) >0.02 were not observed in our study. In regions with a minimum total depth (summed from 200 individuals) of over 600-fold, the rate was estimated to 2.1% (Fig. S1b). This finding indicates that the estimated false negative rate could be attributed to regions that tend to have a lower sequencing depth. A diminishing increment for discovering SNPs with minor allele frequency >0.02 was observed, which suggests that we have successfully identified most of the SNPs with minor allele frequency >0.02 (Fig. S1c). Then we estimated exonic SNP counts for each individual samples was about 13,210 ~ 31,025 (Table S5).

Evidence that our estimator is unbiased:

It is enough to show that $E(p_i) = 0$ (if the individual is homozygote for major allele), $E(p_i) = 1$ (if the individual is homozygote for the minor allele) or $E(p_i) = .5$ (if the individual is heterozygote). Without loss of generality, assume that we have two alleles A (minor allele) and T (major allele). Assuming n_{iT} is fixed, we have to show

Nature Genetics: doi: 10.1038/ng.680

that the expectation of p_i from (1A) reduces to finding the expectation of the number of reads showing an A, E(n_i) . Consider the following cases:

Case 1: *Individual is homozygous for major allele T*

Let n_i be the number of reads that are A's . Notice that E(p_i) = 0 if and only if E(n_i) = en_{iT} . Therefore we need to show that E(n_i) = en_{iT} .

Pf: Let the indicator variable I_i = 1 if read i is an A, and 0 otherwise. Therefore, we can write $n_i = I_1 + ... + I_{n_{iT}}$. Taking expectations, $E(n_i) = E(I_1) + ... + E(I_{n_{iT}})$, where $E(I_i) = P(I_i = 1)$ = Probability that read i is an A. Since the individual is homozygous for allele T, the only way to get a read that is an A is if we have an error. The probability that we have an error, $P(T \rightarrow A) = e$.

Thus,

$$E(n_i) = E(I_i) + ... + E(I_{n_{iT}}) = P(I_1 = 1) + ... + P(I_{n_{iT}} = 1) = en_{iT}$$

Case 2: *Individual is homozygous for major allele A*

Let n_i be the number of reads that are A's . Notice that E(p_i) = 1 if and only if E(n_i) = n_{iT} - en_{iT} . Therefore we need to show that E(n_i) = n_{iT} - en_{iT} .

Pf: Let the indicator variable I_i = 1 if read i is an A, and 0 otherwise. Therefore, we can write $n_i = I_1 + ... + I_{n_{iT}}$. Taking expectations, and $E(n_i) = E(I_1) + ... + E(I_{n_{iT}})$, where $E(I_i) = P(I_i = 1)$ = Probability that read i is an A. Since the individual is homozygous for allele A, the only way **not** to get a read that is an A is if there was an error. The probability that there was an error, $P(A \rightarrow T) = e$, so $P(I_i = 1) = 1 - e$. Thus

Nature Genetics: doi: 10.1038/ng.680

$$E(n_i) = E(I_1) + ...+ E(I_{n_{iT}}) = P(I_1 = 1) + ...+ P(I_{n_{iT}} = 1) = n_{iT}(1-e) = n_{iT} - en_{iT}$$

Case 3: *Individual is heterozygous A/T*

Let n_i be the number of reads that are *A*'s . Notice that E(p_i) = .5 if and only if E(n_i) = (.5)n_{iT} . Therefore we need to show that E(n_i) = (.5)n_{iT} .

Pf: Let the indicator variable I_i = 1 if read *i* is an *A*, and 0 otherwise. Therefore we can write $n_i = I_1 + ...+ I_{n_{iT}}$. Taking expectations, $E(n_i) = E(I_1) + ...+ E(I_{n_{iT}})$, and $E(I_i) = P(I_i = 1)$ = Probability that read *i* is an *A*. Since the individual is heterozygous for allele *A*, then

$$P(I_i = 1) = .5 + .5\ P(T \rightarrow A) - .5\ P(A \rightarrow T) = .5 + .5\ e - .5\ e = .5,$$

Therefore,

$$E(n_i) = E(I_1) + ...+ E(I_{n_{iT}}) = P(I_1 = 1) + ...+ P(I_{n_{iT}} = 1) = \frac{1}{2} n_{iT}$$

Selection tests

Multiple studies have focused on detecting genes affected by positive selection[6-10]. We scanned our large dataset for candidate genes by examining whether the proportion of fixed substitutions observed in a given gene significantly deviated from the genome-wide expectation. This test has recently been shown to have a high statistical power to detect positive selection, but it has not yet been applied genome-wide in the human genome[11]. For each gene, we tested the longest coding mRNA variant for signature of selection using a Hudson-Kreitman-Aguadé test (HKA). We counted the number of human specific fixed substitutions in gene *i*, F_i, as all positions where the macaque and chimpanzee had a common variant not shared with human

Nature Genetics: doi: 10.1038/ng.680

and the major human allele had a frequency >99%. We counted the number of polymorphic positions in gene *i*, P_i, where the derived allele had frequency >= 2%. For both the population genetic and the comparative data, we only included data in regions where there were at least 600 reads and comparative data available from both the chimpanzee and the macaque. Our implementation of the HKA test evaluated whether the number of fixed differences observed per gene, given the total number of fixed and polymorphic sites observed, was higher or lower than expected given the genome wide proportion of fixed differences, $p = \frac{\sum_i F_i}{\sum_i (F_i + P_i)}$ using a binomial null model. For gene *i* we report a score measuring the relative evidence for an excess of fixed differences in gene *i* as -log(Prob($X > F_i$)) where X is distributed Binomial[P_i+F_i, p]. A high value of this score was indicative of an excess of fixed differences. The motivation for reporting a log scaled score rather than a p-value was that the p-value could not be interpreted literally, as the true conditional distribution of F_i depends on local recombination rate and the population demography. Our approach was, therefore, an outlier approach that cannot be used to determine the total amount of positive selection in the human genome, but will be useful in ranking genes according to evidence for an excess of substitutions compared to polymorphisms. Our major objective was not to determine the amount that positive selection has affected the human genome, but rather to identify the best candidates for positive selection and compare these to results obtained using previous methods.

Among four candidates that had the strongest excess of fixed versus polymorphic substitutions, we found three genes involved in immune modulation (*KIR3DP1*, *LILRA1* and *KIR2DL*) (Table S6), which is consistent with previous studies that showed that some immunity genes evolve extremely rapidly. Other

Nature Genetics: doi: 10.1038/ng.680

interesting genes identified included *CES2*, which is involved in cocaine and heroin metabolism, and *FLG,* which harbors variants associated with Ichthyosis Vulgaris (Table S6).

We used a similar approach based on tail probabilities of the binomial distribution to assess whether the substitutions observed in a given gene had a pattern likely driven by gene conversion and/or deamination of methylated cytosine at CpG sites. In this application of the binomial test, the p-values could be interpreted literally as they were based on comparing different classes of mutations rather than polymorphsims versus fixed mutations. Although gene conversion appeared to have a very limited effect on allele frequencies genome-wide, we did note that it could produce strong local signals that mimicked positive selection[12]. Among the top 20 genes with an HK signal for positive selection (excess of fixed differences), we discarded 3 genes (SPTAN1, BPTF and TEX15) with substitutions that were likely due to gene conversion.

GO ontology analysis

We studied GO category enrichment using the binomial test available in the FUNC package, which includes a correction for multiple tests[13]. We used the refinement algorithm from the same package to limit redundancy among enriched terms. To avoid categories with very few sites (and/or genes) contributing to a signal, we only presented the results obtained for GO categories with at least 600 polymorphic and fixed differences. Among the GO categories presented, we found no evidence for biases relating to gene conversion or cytosine substitutions at CpG sites. Most of the tests we performed compared a gene (or a GO category) with a genome-wide genic pattern and did not provide an absolute measure of selective constraints. Therefore,

Nature Genetics: doi: 10.1038/ng.680

we presented the data using an outlier approach where genes or GO categories were sorted according to their respective score (score = -log10(test Pvalue)).

GO term analyses detected positive selection for genes related to defense and immune responses (Table S7), supporting the idea that environmental changes can cause rapid evolution in the genes that interact with the environment (e.g. Ref. 6). Interestingly, we also found positive selection on genes involved in muscle contraction and in the regulation of metabolic processes, which we theorize might be related to evolutionary changes in the human diet.

References

1 Albert, T. J. et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 4, 903-905 (2007).

2 Bentley, D. R. Whole-genome re-sequencing. Curr Opin Genet Dev 16, 545-552 (2006).

3 Wang, J. et al. The diploid genome sequence of an Asian individual. Nature 456, 60-65, doi:nature07484 [pii] 10.1038/nature07484 (2008).

4 Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics (2009).

5 Li, R. et al. SNP detection for massively parallel whole-genome resequencing. Genome Res 19, 1124-1132 (2009).

6 Nielsen, R. et al. Darwinian and demographic forces affecting human protein coding genes. Genome Res 19, 838-849, doi:gr.088336.108 [pii] 10.1101/gr.088336.108 (2009).

7 Sabeti, P. C. et al. Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832-837, doi:10.1038/nature01140 nature01140 [pii] (2002).

8 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. PLoS Biol 4, e72, doi:05-PLBI-RA-1239R2 [pii] 10.1371/journal.pbio.0040072 (2006).

9 Carlson, C. S. et al. Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15, 1553-1565, doi:15/11/1553 [pii] 10.1101/gr.4326505 (2005).

Nature Genetics: doi: 10.1038/ng.680

10 Williamson, S. H. et al. Localizing recent adaptive evolution in the human genome. PLoS Genet 3, e90, doi:06-PLGE-RA-0365R2 [pii]
10.1371/journal.pgen.0030090 (2007).
11 Zhai, W., Nielsen, R. & Slatkin, M. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. Mol Biol Evol 26, 273-283 (2009).
12 Galtier, N. & Duret, L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet 23, 273-277, doi:S0168-9525(07)00113-8 [pii]
10.1016/j.tig.2007.03.011 (2007).
13 Prufer, K. et al. FUNC: a package for detecting significant associations between gene sets and ontological annotations. BMC Bioinformatics 8, 41, doi:1471-2105-8-41 [pii]
10.1186/1471-2105-8-41 (2007).

Nature Genetics: doi: 10.1038/ng.680