nature genetics

Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants

Yingrui Li[1,19], Nicolas Vinckenbosch[2,19], Geng Tian[1,19], Emilia Huerta-Sanchez[2,3,19], Tao Jiang[1,19], Hui Jiang[1], Anders Albrechtsen[4], Gitte Andersen[5], Hongzhi Cao[1], Thorfinn Korneliussen[4], Niels Grarup[5], Yiran Guo[1], Ines Hellman[6], Xin Jin[1,7], Qibin Li[1], Jiangtao Liu[1], Xiao Liu[1], Thomas Sparsø[5], Meifang Tang[1], Honglong Wu[1], Renhua Wu[1], Chang Yu[1], Hancheng Zheng[1,7], Arne Astrup[8], Lars Bolund[1,9,10], Johan Holmkvist[5], Torben Jørgensen[11,12], Karsten Kristiansen[1,4], Ole Schmitz[13,14], Thue W Schwartz[15], Xiuqing Zhang[1], Ruiqiang Li[1,4], Huanming Yang[1], Jian Wang[1], Torben Hansen[5,16], Oluf Pedersen[5,17,18], Rasmus Nielsen[2–4] & Jun Wang[1,4]

Targeted capture combined with massively parallel exome sequencing is a promising approach to identify genetic variants implicated in human traits. We report exome sequencing of 200 individuals from Denmark with targeted capture of 18,654 coding genes and sequence coverage of each individual exome at an average depth of 12-fold. On average, about 95% of the target regions were covered by at least one read. We identified 121,870 SNPs in the sample population, including 53,081 coding SNPs (cSNPs). Using a statistical method for SNP calling and an estimation of allelic frequencies based on our population data, we derived the allele frequency spectrum of cSNPs with a minor allele frequency greater than 0.02. We identified a 1.8-fold excess of deleterious, non-syonomyous cSNPs over synonymous cSNPs in the low-frequency range (minor allele frequencies between 2% and 5%). This excess was more pronounced for X-linked SNPs, suggesting that deleterious substitutions are primarily recessive.

Next-generation technologies have reduced the costs of high-throughput sequencing by several orders of magnitude and have allowed for the whole-genome sequencing of several human genomes[1–3]. However, whole-genome sequencing of the large numbers of individuals needed for studies of population genetics or trait associations remains unaffordable. One alternate approach, taken by the the 1000 Genomes Project, is low-pass sequencing (with an average of fourfold genome sequencing depth per sample) of the whole genomes of a large number of individuals. This approach may be useful in identifying population genetics patterns by combining data from the whole sample population.

Exome sequencing through use of targeted sequencing based on array capture[4] is another alternative approach that allows researchers to concentrate their sequencing efforts on the complete set of coding exons of the human genome; using this approach, researchers are perhaps more likely to include functionally important regions. Recent studies have successfully applied targeted capture and exome sequencing to identify genetic changes involved in Mendelian diseases[5,6]. Such exome-capture sequencing data can now be generated from large population samples. This provides unprecedented opportunities to both characterize the impact of natural selection and to better understand the role of low-frequency variants in the pathogenesis of human diseases. Here we used modest-depth sequencing of the exomes of 200 individuals of European ancestry from Denmark with an average of 12-fold sequencing depth (defined as total number of uniquely mapped bases divided by the full length of target region, which was approximately 34 Mb) per sample to discover new, low-frequency variants by aggregating data from all 200 individuals. With this intermediate design between low-pass population sequencing and deep individual sequencing, we aimed to derive a high-resolution allele frequency spectrum of cSNPs with a minimum allele frequency of 0.02 to characterize the distribution of allele frequencies in a human population and to use this distribution to make inferences about the effect of natural selection in the human genome.

We used the NimbleGen 2.1M Exon Capture Array to capture 34 Mb of the human genome, which included the coding sequences of 18,654

[1]BGI-Shenzhen, Shenzhen, China. [2]Department of Integrative Biology, University of California Berkeley, Berkeley, California, USA. [3]Department of Statistics, University of California Berkeley, Berkeley, California, USA. [4]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [5]Hagedorn Research Institute, Copenhagen, Denmark. [6]Department of Mathematics, University of Vienna, Vienna, Austria. [7]School of Bioscience and Biotechnology, South China University of Technology, Guangzhou, China. [8]Department of Human Nutrition, Faculty of Life Sciences, University of Copenhagen, Copenhagen, Denmark. [9]Institute of Human Genetics, University of Aarhus, Aarhus, Denmark. [10]Danish Center for Translational Breast Cancer Research, Copenhagen, Denmark. [11]Research Centre for Prevention and Health, Glostrup University Hospital, Glostrup, Denmark. [12]Faculty of Health Science, University of Copenhagen, Copenhagen, Denmark. [13]Department of Endocrinology and Diabetes, Aarhus University Hospital, Aarhus, Denmark. [14]Department of Clinical Pharmacology, University of Aarhus, Aarhus, Denmark. [15]Laboratory for Molecular Pharmacology, University of Copenhagen, Copenhagen, Denmark. [16]Faculty of Health Sciences, University of Southern Denmark, Odense, Denmark. [17]Faculty of Health Sciences, University of Aarhus, Aarhus, Denmark. [18]Institute of Biomedical Sciences, University of Copenhagen, Copenhagen, Denmark. [19]These authors contributed equally to this work. Correspondence should be addressed to Jun Wang (wangj@genomics.org.cn), R.N. (rasmus_nielsen@berkeley.edu) or O.P. (oluf@hagedorn.dk).

Received 8 February; accepted 8 September; published online 3 October 2010; doi:10.1038/ng.680

Table 1 Summary of data production of individual exome sequencing of 200 Danes

	Mean ± s.d.[a]	Total
Mapped data amount (Mb)	479.2 ± 40.3	95,818
Sequencing depth	14.1 ± 1.2	2,809
Coverage[b]	95.5 ± 2.1	–
Fraction of target covered ≥ ×5	80.5 ± 4.1	–
Fraction of target covered ≥ ×10	59.1 ± 3.9	–
Fraction of target covered ≥ ×15	40.0 ± 4.0	–
Fraction of target covered ≥ ×20	25.2 ± 4.0	–
Fraction of target covered ≥ ×25	15.3 ± 3.6	–
Fraction of target covered ≥ ×30	9.0 ± 2.9	–
Rate of nucleotide difference[b]	1.01 ± 0.11	–

The mapped data amount is the sum of read bases that were aligned to the 34-Mb target region; those bases aligned to the human genome, but those which were not in a targeted region were not considered in this study. Coverage is the proportion of the targeted region that was covered by at least one uniquely aligned read. The rate of nucleotide difference is the proportion of nucleotide mismatches in unique aligned read bases, which we then used for SNP calling. This rate of nucleotide difference was calculated by dividing the number of mismatched bases by the total number of bases in uniquely aligned reads.
[a]Of 200 total samples.[b]Values given in percentages.

(92.8%) well-annotated genes together with stretches of untranslated or intronic flanking sequences (Online Methods and **Supplementary Note**). Targeted regions primarily contained exonic sequences, which we collectively refer to here as an exome. We sequenced DNA from 200 individuals from Denmark with an average coverage of ×12 to ×18 for each individual exome (**Supplementary Table 1**). A total of 95.8 Gb of high-quality data aligned to the targeted regions with a per-base mismatch rate of 1% (**Table 1**) and covered the human exome with an average depth of approximately 2,800-fold. For each sample individual, on average, about 95% of the target regions were covered by at least one read and more than 60% of the target regions were covered by at least ten reads, which is in agreement with previous studies[5,6]. Only 8.8% of the total 34-Mb target regions were covered by less than 600 reads due to systematic bias in the hybrid capturing experiment. Going forward, we only used the regions passing this 600-fold read-depth criterion for estimation of allele frequencies. In addition, we only used SNPs with estimated minor allele frequency (MAF) > 0.02.

As the sequencing depth was not adequate to call genotypes of each individual accurately, we developed a SNP calling and frequency estimation method based on population data that simultaneously used genotype likelihood information from all individuals. This approach significantly increased the statistical power to detect SNPs and frequency estimations (**Supplementary Note**). Applying this method, we detected 121,870 high-quality SNPs with a false positive rate of 2% per site (Online Methods and **Supplementary Note**) from the sample population in the 34-Mb target region. Among these SNPs, 53,707 (44.1%) had not been previously reported (dbSNP, version 129), and 80% of these newly identified SNPs had MAF < 5%. We assessed the accuracy of our genotype calls on a randomly chosen subsample of the new SNPs and found inconsistent genotypes for 9% of the very rare SNPs (defined as MAF < 0.02) but no inconsistent genotypes for the more common SNPs, indicating a low overall false positive rate (Online Methods, **Supplementary Note** and **Supplementary Tables 2** and **3**). The false negative rate of SNPs with MAF > 0.02 was estimated to be 5.1% using SNPs that had been genotyped in the HapMap European CEU population (**Supplementary Note** and **Supplementary Fig. 1**). In all, 25,275 synonymous and 27,806 non-synonymous coding SNPs were identified (**Supplementary Tables 4** and **5**), of which 22,216 (42.6%) were new.

Our large population sample provided the opportunity to characterize the impact of natural selection on protein coding genes, (**Supplementary Note**, **Supplementary Tables 6** and **7**) something that has been difficult in previous studies of Europeans due to limitations based on either the number of genes analyzed or the sample size[7–10]. The distribution of allele frequencies can reveal signatures of natural selection, but an accurate estimation of allele frequencies using next generation sequencing data is challenging due to high error rates and varying coverage. Relying on such genotypes to compute allele frequencies thus leads to biased estimates, as even the best method for calling genotypes results in a biased measure of population genetic variability[11–14].

We circumvented this pitfall by developing an unbiased (minimum variance) estimator of allele frequencies, which directly relies on the nucleotide reads observed in each individual exome and accounts for error rates and coverage variation. Simulations and stringent quality thresholds indicated that our estimator was reliable for derived alleles with frequencies of over 2% in our dataset (Online Methods and **Supplementary Fig. 2**). According to above genotyping validations results, we observed the false positive results only in SNPs with MAF < 0.02 (though the false positive rate in this frequency range was still <10%). Hence, we only included SNPs with MAF > 0.02 in our subsequent analyses to ensure the validity of our conclusions.

To quantify the effect of selection on deleterious mutations, we compared the distribution of allele frequencies among non-synonymous (that is, amino acid changing) and synonymous cSNPs (**Fig. 1a**). Synonymous cSNPs closely followed the distribution expected in the absence of natural selection. In contrast, non-synonymous cSNPs showed a much larger proportion of low frequency alleles, indicative of a strong purifying selection[10,15–17], showing that a large proportion of these mutations were likely to be deleterious. We observed a notable 1.8-fold excess in the proportion of non-syonomyous to synonymous cSNPs for alleles with low (2–4%) frequencies. Although this excess is not incompatible with findings in previous studies[10], our study included a larger sample size and suggests that the excess of low frequency non-synonymous mutations predominantly comes from very rare mutations (that is, mutations with frequencies <4%) and not from higher frequency mutations. To further demonstrate that the pattern we observed was due to the presence of deleterious mutations and not to sequencing artifacts, we categorized sites with non-synonymous cSNPs according to their conservation across species. As expected, substitutions on very conserved sites tended to segregate at much lower frequencies than mutations on less conserved sites (**Fig. 1b**). We obtained similar results when classifying non-synonymous cSNPs on the basis of the amino acid change induced (**Supplementary Fig. 2b**).

We assessed the amount of selection required to explain the observed differences between the distributions of synonymous and non-synonymous cSNPs. Our estimator, γ, was an effective population-scaled selection coefficient that may differ from actual selection coefficients, as we ignored demographic effects and linkage among sites. However, as the synonymous cSNPs closely followed the neutral expectation in our data (**Fig. 1a**), we expect our method to provide estimates that are in agreement with previous work, which also ignored linkage but relied on demographic models[10,17]. We assumed that γ f ollowed a mixture distribution, with a proportion (k) of non-synonymous cSNPs being neutral ($\gamma = 0$) and other non-synonymous cSNPs having γ values that follow a gamma distribution with parameters α and β (Online Methods). We estimated $\alpha = 4$, $\beta = 0.09$ and $k = 0.2$, which indicated a larger proportion of weakly deleterious and mildly deleterious (~80%) mutations than did previous estimates (54%)[10] (**Supplementary Fig. 3**). The difference in the estimates was likely due to our discovery of a larger proportion of low frequency non-synonymous mutations than was previously identified in smaller

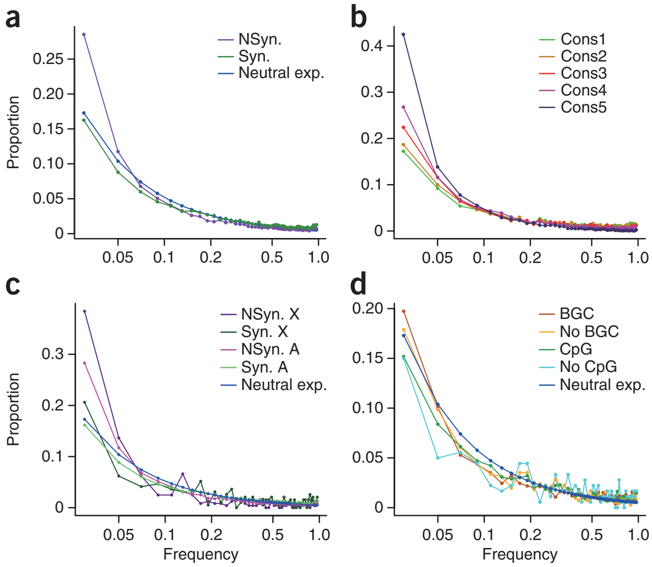

Figure 1 Comparision of site frequency spectra for SNPs in different annotation categories with functional consideration. (**a**) Unfolded site frequency spectra (SFS) for non-synonymous (Nsyn) and synonymous (Syn) SNPs compared to the neutral expectation. (**b**) Non-synonymous SFS for sites having increasing conservation scores, with Cons5 being the most conserved category. (**c**) SFS for X-linked (X) and autosomal (A) SNPs. (**d**) SFS for substitutions potentially influenced by biased gene conversion (BGC) or cytosine deamination at CpG sites (CpG) compared to unaffected substitutions. Dots in all panels indicate the proportion of SNPs within a given frequency bin. All bins have the same width, and the first bin contains frequencies from 2–4%.

sample sizes. Previous studies based on smaller sample sizes did not have the statistical power to detect these low-frequency mutations.

Notably, the number of rare non-synonymous SNPs differed between the X chromosome and autosomes, with a higher excess of these SNPs present in the X chromosome (**Fig. 1c**). The X chromosome is well known to have less variation than autosomes[18]. Many factors may contribute to this effect, including mutation rate variation, but the lower effective population size on the X chromosome may be the main underlying cause. Another contributing factor may be the presence of recessive deleterious mutations exposed to selection on the X chromosome in males[19]. If selected mutations are recessive, selection will have a stronger effect on the X chromosome than on autosomes. Depending on the true dominance relationships and distribution of selection coefficients, this effect is also a possible explanation for the excess of rare non-synonymous mutations relative to synonymous mutations on the X chromosome observed in this study. Negative selection acting on recessive mutations may prevent more of these mutations from reaching intermediate frequencies on the X chromosome than on the autosomes, leading to a relatively higher excess of rare non-synonymous mutations on the X chromosome. However, other selective factors may affect the frequency spectrum in non-synonymous sites, including positive selection and Hill-Robertson effects[20]. Explanations based solely on mutation rate variation, such as a higher or lower mutation rate in males, cannot alone explain the pattern observed, as changes in the mutation rate alone have no effect on the frequency spectrum.

Processes unrelated to selection may also affect the observed frequencies of alleles. Methylation leads to hypermutability in CpG sites[21], which would promote recurrent mutations at the same sites. Likewise, biased gene conversion from A and T to G and C has been proposed to be common in mammals, especially in regions with high recombination rates[22]. This type of gene conversion would result in a segregation distortion that favors the transmission of C or G alleles. If this effect is strong and pervasive, the site frequency spectrum of mutations from A and T to G and C may differ from that of other mutations. Using fourfold degenerate sites, we compared the frequency distributions for sites potentially affected by CpG methylation or gene conversion versus unaffected sites (**Fig. 1d** and Online Methods). The distributions we obtained were not significantly different from each other ($P > 0.05$; Mann-Whitney U test), indicating that, when considered at a genome-wide scale, neither CpG mutation nor biased gene conversion had a strong effect on the frequency spectrum.

We report, to our knowledge, the largest dataset of directly sequenced human exomes published to date. Although we only sequenced the exome regions and the sequence coverage per individual was not adequate to ascertain genotypes for each individual with high confidence, our data can be used for accurate estimation of the allele frequencies of cSNPs with MAF > 0.02. These data are therefore useful for studying the allele frequency spectrum, as well as other population genetics patterns in this sample cohort. We find that coding regions harbor a larger proportion of low-frequency deleterious mutations than previously anticipated. Our analyses indicate that most of these mutations are recessive, which partly explains their segregation in humans despite strong purifying selection. Our observations are also consistent with recent claims that only a very small proportion of the heritable variation associated with common polygenic traits has been identified in association studies[23]. Based on our findings, we support the idea that much of the heritable variation affecting fitness is caused by low-frequency mutations, which are often overlooked in studies based on genotyping rather than resequencing. Further, we demonstrate that the use of next-generation sequencing is an important tool in population genetics studies. Future analyses of non-coding regions and ethnically diverse samples will help build a complete picture of human genomic variation and an understanding of the interaction between genetic drift, mutation, recombination and selection in the human genome.

URLs. NimbleGen 2.1M array, http://www.nimblegen.com/products/seqcap/; CCDS database, http://www.ncbi.nlm.nih.gov/projects/CCDS/; Sequenom iPlex platform, http://www.sequenom.com/iplex/; UCSC Genome Browser, http://genome.ucsc.edu/; SOAP, http://soap.genomics.org.cn/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. The sequence data has been deposited in the NCBI Short Read Archive with accession number SRA009884.

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

ACKNOWLEDGMENTS
This project was funded by the Lundbeck Foundation and produced by The Lundbeck Foundation Centre for Applied Medical Genomics in Personalised Disease Prediction, Prevention and Care (LuCAMP). The project was also supported by a National Basic Research Program of China (973 program no. 2011CB809200; 2007CB815703; 2007CB815705; and 863 program no. 2006AA02Z177; 2006AA02Z334; 2006AA02A302; 2009AA022707), the Chinese Academy of Science (GJHZ0701-6), the National Natural Science Foundation of China (30725008; 30890032; 30811130531; 30221004), the Chinese Academy of Science (GJHZ0701-6), the Chinese 973 program (2007CB815703; 2007CB815705), the Chinese 863

program (2006AA02Z177), the Danish Platform for Integrative Biology and the Ole Rømer grant from the Danish Natural Science Research Council. The Shenzhen Municipal Government and the Yantian District Local Government of Shenzhen additionally funded the project (grants JC200903190767A; JC200903190772A; ZYC200903240076A; CXB200903110066A; ZYC200903240077A; ZYC200903240076A and ZYC200903240080A). N.V. and E.H.-S. were supported with fellowships from the Swiss and American National Science Foundations. We are indebted to T. Lauritzen and K. Borch-Johnsen for their contribution to LuCAMP.

AUTHOR CONTRIBUTIONS

LuCamp was founded and is managed by O.P., Jun Wang, R.N., T.H., G.A., L.B., O.S., T. Lauritzen, K.K., T. Jørgensen, A. Astrup, T.W.S. and A. Albrechtsen. Y.L., N.V., G.T., E.H.-S. and T. Jiang contributed equally to this work. H.Y., Jian Wang, O.P. and Jun Wang managed the present project. Jun Wang, R.N., O.P. and Y.L. designed the analyses. O.P., T.H. and T. Jørgensen recruited the volunteers and prepared the DNA samples. Jun Wang, R.N., Y.L., N.V., E.H.-S., T. Jiang, A. Albrechtsen, H.C., T.K., Y.G., X.J., Q.L., H.W., C.Y., H.Z. and O.P. performed the data analyses. G.T., H.J., J.L., X.L., M.T., R.W. and X.Z. performed sequencing and Sequenom genotyping. Jun Wang, R.N., O.P., N.V., E.H.-S. and Y.L. wrote the first manuscript. All authors contributed to the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
2. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
3. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
4. Albert, T.J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).
5. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
6. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 19096–19101 (2009).
7. Leabman, M.K. *et al.* Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. *Proc. Natl. Acad. Sci. USA* **100**, 5896–5901 (2003).
8. Bustamante, C.D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
9. Nielsen, R. *et al.* Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* **19**, 838–849 (2009).
10. Boyko, A.R. *et al.* Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4**, e1000083 (2008).
11. Johnson, P.L. & Slatkin, M. Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res.* **16**, 1320–1327 (2006).
12. Johnson, P.L. & Slatkin, M. Accounting for bias from sequencing error in population genetic estimates. *Mol. Biol. Evol.* **25**, 199–206 (2008).
13. Lynch, M. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol. Biol. Evol.* **25**, 2409–2419 (2008).
14. Lynch, M. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* **182**, 295–301 (2009).
15. Sunyaev, S.R., Lathe, W.C. III, Ramensky, V.E. & Bork, P. SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet.* **16**, 335–337 (2000).
16. 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).
17. Keightley, P.D. & Eyre-Walker, A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**, 2251–2261 (2007).
18. Hammer, M.F. *et al.* Heterogeneous patterns of variation among multiple human X–linked loci: the possible role of diversity-reducing selection in non-Africans. *Genetics* **167**, 1841–1853 (2004).
19. Vicoso, B. & Charlesworth, B. Evolution on the X chromosome: unusual patterns and processes. *Nat. Rev. Genet.* **7**, 645–653 (2006).
20. Hill, W.G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
21. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
22. Meunier, J. & Duret, L. Recombination drives the evolution of GC content in the human genome. *Mol. Biol. Evol.* **21**, 984–990 (2004).
23. McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).

ONLINE METHODS

Sample acquisition and exon sequencing. Genomic DNA was purified from blood leukocytes from 200 individuals of Danish nationality. Exome capture was performed on a NimbleGen 2.1M HD array (Roche). Exon-enriched DNA libraries were then subjected to a secondary library construction for Illumina GA sequencing and were sequenced by the Illumina Genome Analyzer II platform following the manufacturer's instructions (**Supplementary Note**).

Genotype calling and SNP calling. SOAPaligner[24,25] was used to align the sequencing reads to the NCBI human genome reference assembly (build 36.3). The likelihood of possible genotypes at each site in each individual was calculated using SOAPsnp[26]. For each site, the likelihoods from all sample individuals were then integrated using a heuristic formula to give a score that measured the confidence to ascertain SNPs. Highly confident SNPs were then extracted for subsequent analysis. A randomly selected subset of SNPs that were not deposited in the dbSNP database (v129) was genotyped using a Sequenom iPlex array for validation (**Supplementary Note**).

Estimation of allele frequencies. Our aim was to obtain unbiased estimates of allele frequencies. This is difficult to do using SNP calling, as most procedures for SNP calling tend to lead to either a deficiency or an excess of rare variants depending on how conservative the applied method is[12]. Although it is possible to correct the sample frequency spectrum obtained on the basis of SNP calling statistically[12], we instead chose to pursue an arguably more direct approach for estimating allele frequencies for each SNP. To do this, we first eliminated all reads with Q score < 20. The mismatch rate in these reads was 0.41%. We then determined which two nucleotides were most common among A, C, T and G. We let the set of these nucleotides be S, meaning if there were 400 As, 42, Cs, 13 Ts and 9 Gs, then $S = \{A, C\}$. We then eliminated all reads that were not elements of S. This was done under the assumption that all SNPs are truly diallelic. Because we eventually used a frequency cutoff of 2%, ignoring the third and fourth most common bases had no effect on our inferences.

We let n_i be the number of reads of the minor allele in S in individual i. We let the total number of reads in S in individual i be n_{iT}. Then we calculated p_i, an estimator for the probability of the minor allele in individual i. If $S = \{A, C\}$, where C is the minor allele, we get

$$n_i = n_C^{\text{true}} - en_C^{\text{true}} + en_A^{\text{true}}$$

where n_C^{true} is the true number of reads that should be a C, but due to sequencing errors, we observed n_i instead. Because we assume sites are biallelic, $n_A^{\text{true}} = n_{iT} - n_C^{\text{true}}$. If we divide by n_{iT}, we get

$$p_C^{\text{obs}} = (1-e)p_i + e(1-p_i)$$

If we multiply by n_{iT},

$$n_i = p_i n_{iT}\left(1-e\right) + (n_{iT} - p_i n_{iT})e$$

Now we can solve for p_i

$$p_i = \frac{n_i - en_{iT}}{n_{iT}\left(1-2e\right)} \quad \text{(1A)}$$

for $i = 1, 2,\ldots, k$. This is an error-corrected estimate of the allele frequency in individual I_i, obtained as the solution for p_i to the equation. Note that it is possible for p_i to be negative (for example, if $n_i = 0$). The parameter e is the error rate and is considered a fixed parameter. On the basis of previous analyses, we assume $e = 0.004$. This estimate is, if anything, an overestimate because it represents the error rate between the major and minor allele only. For example, suppose the nucleotide read at a site for an individual is A, then due to a sequencing error, an A can be changed to a T, C or G. We assume sites are biallelic, therefore we only consider one change, which implies that 0.4% is an overestimate of the error rate (as a 1% per-read-base mismatch rate divided by 3 is less than 0.4%). We then calculate $w_i = \dfrac{2n_{iT}}{n_{iT}+1}$ (see derivation below), the inverse of the variance of p_i (up to a scalar). Note that w_i lies between (0,2), takes the value 0 when there are no reads, 1 when there is 1 read (when there was only one chromosome sampled) and 2 as the number of reads tends to infinity (corresponding to certainty that the 2 chromosomes have been sampled).

The estimate of the minor allele frequency is then calculated as

$$\hat{p} = \max\left\{0, \sum_{i=1}^{k} p_i w_i \middle/ \sum_{i=1}^{k} w_i\right\}$$

This estimate of allele frequencies was combined among autosomal sites with population coverage of at least 600-fold (the coverage thresholds for X and Y chromosome sites were 438(X) and 162(Y)) in accordance with the number of X chromosomes and closely approximated the true distribution of parameter values; we conducted several simulation studies. The method clearly estimated the distribution accurately. The performance of the method was aided by the use of a large sample size. For smaller sample sizes, the use of a distribution of estimates to approximate the estimate of a distribution would lead to a larger discrepancy between the estimated and true distribution.

Derivation of weights, w_i. Without loss of generality, we assumed that we have two alleles A (minor allele) and T (major allele). Let

$$Yi = \frac{I_1 + \ldots.. + I_{n_{iT}}}{n_{iT}}$$

which is the proportion of reads with allele A, where I_i takes the value 1 when the read is an A and 0 when the read is not an A.
We now need to find the variance of Y_i.

$$\begin{aligned}\text{Var}(Y_i) &= \frac{\text{Var}(I_1 + \ldots. + I_{n_{iT}})}{n^2_{iT}} \\ &= \frac{1}{n^2_{iT}}\text{Var}(I_1 + \ldots. + I_{n_{iT}}) \\ &= \frac{1}{n^2_{iT}}\left(\text{Var}(I_1) + \ldots. + \text{Var}(I_{n_{iT}}) + 2\sum_{i<j}\text{Cov}(I_i, I_j)\right)\end{aligned}$$

The variance terms, Var(I_i), are easily computed and are equal to

$$\text{Var}(I_i) = p_i(1-p_i)$$

where p_i is the population probability of observing a read with an A. The covariance term,

$$\text{Cov}(I_i, I_j) = \text{E}(I_i I_j) - \text{E}(I_i)\text{E}(I_j)$$

can also be computed assuming Hardy-Weinberg equilibrium.

The second term is $\text{E}(I_i)\text{E}(I_j) = p_i^2$. To compute the first term, let $M = I_i I_j$. Then by the law of total probability,

$$\begin{aligned}\text{P}(M=1) &= \text{P}(M=1\,|\,\text{AA})\text{P}(\text{AA}) \\ &+ \text{P}(M=1\,|\,\text{AT})\text{P}(\text{AT}) + \text{P}(M=1\,|\,\text{TT})\text{P}(\text{TT})\end{aligned}$$

So the probability P(M = 1) depends on the genotype of the individual. Assuming H-W equilibrium,

$$\begin{aligned}\text{P}(M=1) &= p_i^2 + \frac{1}{4}\left(2p_i(1-p_i)\right) + 0 \\ &= p_i^2 + \frac{1}{2}p_i(1-p_i)\end{aligned}$$

Therefore,

$$\begin{aligned}\text{Var}(I_1 + \ldots. + I_{n_{iT}}) &= n_{iT}\,p_i(1-p_i) \\ &+ 2\left(\frac{p_i(1-p_i)}{2}\right)\left(p_i^2 + \frac{1}{2}p_i(1-p_i) - p_i^2\right)\end{aligned}$$

$$\text{so, } \mathrm{Var}(Y_i) = p_i(1-p_i)\left(\frac{n_{iT}+1}{2n_{iT}}\right)$$

$$\text{If we let } w_i = \frac{1}{\mathrm{Var}(Y_i)} \Bigg/ \sum_k \frac{1}{\mathrm{Var}(Y_k)}$$

then we get an estimate with minimum variance (this can be verified using Lagrange-constrained minimization of the weights).

Estimation of selection coefficients. We fit a model for the density of non-synonymous mutations, f(x), observed across the whole genome as:

$$\mathrm{f}(x) = k(1/x) + (1-k)\int \mathrm{fsel}(x,-\gamma)\,\mathrm{Gamma}(\gamma;\alpha,\beta)\mathrm{d}\gamma$$

where the first term is a contribution from a fraction, k, of the sites having a neutral distribution, and the second term describes negative selection acting on the remaining sites. The first term in the integral, fsel, is the theoretical limiting density of frequencies of mutations in a large population (the mean density of the PRF). The negative selection parameter γ follows a gamma distribution with parameters α and β. f(x) is evaluated for 50 points evenly spaced between frequencies 0.02 and 0.98, and the counts of non-synonymous mutations were calculated in the same range in 50 equally spaced bins. The model was then fit using least squares. We defined a grid of points from $k = 0,\ldots,1$ in steps of 0.02, $\alpha = 2,\ldots,6$ in steps of 0.2 and $\beta = 0.05,\ldots,0.77$ in steps of 0.04. We then evaluated the least-squares function over the grid. We found the best fitting parameter values to be $k = 0.2$, $\alpha = 4$ and $\beta = 0.09$, with a least-squares cost of 0.28. This indicated that most of the sites were under negative selection, with a mean selection parameter of $(\alpha/\beta) = 45$. Previous estimates based on the same model[9] found $k = 0.23$, $(\alpha/\beta) = 816$ (with $\alpha = 1.02$) and predicted that 54% of all mutations were either weakly or moderately deleterious. Using the same assumptions regarding effective population size and definitions of weakly deleterious and mildly deleterious mutations, we found (by integrating our estimated distribution over values corresponding to 2Ns = 1 to 2Ns = 1,000) that the proportion of such mutations was 80%.

Supplementary Figure 3 compares the gamma distributions of selective effects obtained here and previously, indicating that we detected proportionally more moderate and mildly deleterious mutations. This is likely due to the fact that we had a larger sample size, which facilitated the detection of very rare mildly deleterious variants.

SNP annotation. We polarized SNPs into ancestral and derived variants using the multiz28way alignments available from UCSC (see URLs). For all detected SNPs covered with a level 1 chimpanzee net alignment, we extracted the corresponding chimpanzee and macaque nucleotide from these alignments. When chimpanzee and macaque shared the same nucleotide at a given SNP position, the human variant shared with these two other species was called ancestral. SNPs with no inferred ancestral state were discarded from all downstream analyses. Using RefSeq annotation, we classified SNPs into five functional categories (non-synonymous, synonymous, substitutions at fourfold degenerate sites, 5′ untranslated region and 3′ untranslated region) according to their genic location and their expected effect on the encoded gene products. We also classified SNPs into substitutions that are potentially affected by gene conversion (AT→GC) or not (A↔T and G↔C). Similarly, we distinguished between SNPs at CpG sites that were potentially due to deamination of methylated cytosines (CpG→TpG, CpG→CpA) or not (CpG→ApG, CpG→GpG, CpG→CpC and CpG→CpA).

We assessed the potential deleterious effect of non-synonymous SNPs using placental mammal conservation scores computed by Phylop (available at UCSC, see URLs). As an alternative approach, we used the physicochemical distances between pairs of amino acids as previously described[27] (**Supplementary Fig. 2b**).

24. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).

25. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).

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

27. Sawyer, S.A. & Hartl, D.L. Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176 (1992).