Supplementary Information

Contents

A. Sequencing target and acquisition 2

i. Choice of Neutral Regions to target 2

ii. Illumina Library Construction 4

iii. Sequence Capture Methods 4

iv. Illumina Sequencing 5

v. Illumina Primary Analysis 6

B. Variant and genotype calling 7

i. Variant Analysis 7

ii. Variant filters 7

C. Calling Validation 11

i. Calling validation by comparison to the site frequency spectrum (SFS) of the NHLBI Exome data 11

ii. Calling validation by comparison with data from the CHARGE-S project 12

D. Data homogeneity and coverage 20

i. Coverage effect on variant calling 20

ii. Assessing population homogeneity with POPRES sample 21

iii. The effect of homogeneity of the sampled population on the SFS 21

E. Demographic modeling and the SFS 23

i. Downsampling data for figures of the SFS 23

ii. SFS of simulated populations according to previously published demographic models 23

iii. Demographic inference 24

iv. Evaluation of the power of the maximum likelihood approach 26

v. Validation of improved accuracy of estimated SFS using smoothing spline 27

F. Difference between census population size and effective population size 29

G. *ms* command lines 32

H. Conservation analysis of the NR data 33

I. Figures 38

J. Tables 48

A. Sequencing target and acquisition

i. Choice of Neutral Regions to target

Regions for sequencing were targeted to be as neutral as possible with two constraints: 1) the target regions should consist of 5 to 20 kilobases (kb) of consecutive bases, and 2) the target regions should be arranged by groups of two or three (duplets or triplets) in partial linkage disequilibrium (LD) which is motivated by the design in (1). To achieve this goal, we designed the procedure outlined in the following.

To minimize the confounding by natural selection, we considered only genomic regions located at least 100,000 bp and 0.1 cM away from coding or potentially coding sequences (defined as the union of Known Genes and Genes Bounds UCSC tracks), and free from segmental duplications (http://humanparalogy.gs.washington.edu/build36/build36.htm) and known copy number variants (http://projects.tcag.ca/variation/downloads/variation.hg18.v9.mar.2010.txt). Genomic sequences that may be under recent positive selection were also excluded, based on the consensus set assembled by Akey (2). From the contiguous stretches that fulfilled the above criteria, we only kept those larger than 100kb since these can satisfy the duplet or triplet design criteria. To further minimize the effects of selection, we considered regions with the lowest content of conserved elements (Mammal El, phastConsElement44wayPlacental UCSC track) and repetitive elements (rmskRM327 UCSC track), and with no CpG islands. All these features are now implemented as part of our Neutral Regions Explorer (NRE) (3).

Despite the best of our efforts, some of these loci may have unannotated function. Though our filtering procedures aims to minimize the confounder of natural selection on loci in these regions based upon several criteria, loci in these regions may still be under functional constraint. Hence loci in our regions are putatively neutral based only on the above criteria.

A total of 15 target regions, between 5,340 bp to 20,000 bp long, spanning a total of 216,240 bp were used for sequencing (Table S2). They are arranged in 3 duplets and 3 triplets. Measured from their most distant 3' and 5' points, the duplets or triplets cover a physical distance of 195 kb on average (range: 93 kb and 430 kb) and are separated by an average genetic distance of 0.275 cM (range: 0.20cM-0.36 cM).

Target regions were also chosen to cover a range of diverse LD structures within regions and between the 2 regions in duplets or 3 regions in triplets. Within each duplet or triplet, regions are on average 0.154cM apart. Recombination varies across the regions from very low (0.0005 cM/Mb) to moderately high (5.07 cM/Mb), with an average of 1.05cM/Mb, close to the genome-wide average.

Mean physical and genetic distance between the neutral regions and the closest gene were 211 kb (range: 100 kb-577 kb) and 0.37 cM (range: 0.12 cM-1.37 cM), respectively. They contain on average 0.81% (0.27%-1.2%) of conserved elements (less than 1/5 of the genome average). Their GC content (39.34%) is close to the overall genomic level.

ii. Illumina Library Construction

Genomic DNA samples were constructed into Illumina paired-end pre-capture libraries according to the manufacturer's protocol (Illumina Multiplexing_SamplePrep_Guide_1005361_D) with modifications as described in the *BCM-HGSC Illumina Barcoded Paired-End Capture Library Preparation* protocol. Libraries were prepared using Beckman robotic workstations (Biomek NXp and FXp models). The complete protocol and oligonucleotide sequences are accessible from the HGSC website (https://hgsc.bcm.edu/sites/default/files/documents/Illumina_Barcoded_Paired-End_Capture_Library_Preparation.pdf).

Briefly, 1 µg of genomic DNA in 100µl volume was sheared into fragments of approximately 300-400 base pairs in a Covaris plate with E210 system (Covaris, Inc. Woburn, MA) followed by end-repair, A-tailing and ligation of the Illumina multiplexing PE adaptors. Pre-capture Ligation Mediated-PCR (LM-PCR) was performed for 7 cycles of amplification using the 2X SOLiD Library High Fidelity Amplification Mix (a custom product manufactured by Invitrogen). Universal primer IMUX-P1.0 and a pre-capture barcoded primer IBC were used in the PCR amplification. In total a set of 12 such barcoded primers were used on these samples. Purification was performed with Agencourt AMPure XP beads after enzymatic reactions. Following the final XP beads purification, quantification and size distribution of the pre-capture LM-PCR product was determined using the LabChip GX electrophoresis system (PerkinElmer).

iii. Sequence Capture Methods

Twelve pre-capture libraries were pooled together (approximately 84 ng/sample, 1µg total library DNA per pool) and hybridized in solution to the custom probe design (1.1Mb capture

region, manufactured by NimbleGen) according to the *NimbleGen SeqCap EZ Exome Library SR User's Guide (Version 2.2)* with minor revisions. Human COT1 DNA (100 µg) and 3'-dideoxycytosin (ddC)-modified blocking oligonucleotides (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT/3ddC,
5'-CAAGCAGAAGACGGCATACGAGAT/3ddC and 5'-ACTGGAGTTCAGACGTGTGCTCTTCCGATCT/3ddC; 650 pmoles each) were added into the hybridization to block repetitive genomic sequences and the Illumina adaptor sequences. Post-capture LM-PCR amplification was performed using the 2X SOLiD Library High Fidelity Amplification Mix with 14 cycles of amplification. After the final AMPure XP bead purification, quantity and size of the capture library was analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500. The efficiency of the capture was evaluated by performing a qPCR-based quality check on the four standard NimbleGen internal controls. Successful enrichment of the capture libraries was estimated to range from a 6 to 9 of ΔCt value over the non-enriched samples. Aliquots of 10 nM concentration of enriched libraries were submitted for sequencing.

iv. Illumina Sequencing

Library templates were prepared for sequencing using Illumina's cBot cluster generation system with TruSeq PE Cluster Generation Kits (Part no. PE-401-3001). Briefly, these libraries were denatured with sodium hydroxide and diluted to 3-6 pM in hybridization buffer in order to achieve a load density of ~800K clusters/mm^2. Each library pool was loaded in a single lane of a HiSeq flow cell, and each lane was spiked with 2% phiX control library for run quality control. The sample libraries then underwent bridge amplification to form clonal clusters, followed by

hybridization with the sequencing primer. Sequencing runs were performed in paired-end mode using the Illumina HiSeq 2000 platform. Using the TruSeq SBS Kits (Part no. FC-401-3001), sequencing-by-synthesis reactions were extended for 101 cycles from each end, with an additional 7 cycles for the index read. Real Time Analysis (RTA) software was used to process the image analysis and base calling. Sequencing runs generated approximately 250-400 million successful reads on each lane of a flow cell, averaging yield of 2.2 Gb per sample. With these sequencing yields, samples achieved an average of 95% of the targeted bases covered to a depth of 20X or greater.

v. Illumina Primary Analysis

Illumina sequence analysis was done using the HGSC Illumina analysis pipeline, moving data step by step through various analysis tools from the initial sequence generation on the instrument to finished duplicate-marked BAMs. Firstly, the primary analysis software on the instrument produces .bcl files that are transferred off-instrument into the HGSC analysis infrastructure by the HiSeq Real-time Analysis module. Once the run is complete and all .bcl files are transferred, the pipeline runs the vendor's primary analysis software (CASAVA v1.7), which demultiplexes pooled samples and generates sequence reads and base-call confidence values (qualities). The next step is mapping of reads to the hg18 Human reference genome using the Burrows-Wheeler aligner (BWA (4), http://bio-bwa.sourceforge.net/) to produce a BAM (binary alignment/map) file (5). We then mark duplicates (using Picard and SAMtools), and where necessary merge separate sequence-event BAMs into a single sample-level BAM. BAM sorting, duplicate read marking, and BAM format validation all occur at this step.

B. Variant and genotype calling

i. Variant Analysis

In each individual, aligned reads were subjected to "duplicate removal" using PICARD-1.66 (http://picard.sourceforge.net). Downstream analyses were performed with the Genome Analysis ToolKit, GATK-1.5-31 (6, 7). Reads were locally realigned (GATK IndelRealigner). Variant detection and genotyping were performed using the UnifiedGenotyper (UG) tool from the GATK exclusively on the targeted regions. The UG tool generated an initial variant dataset for each sample (in variant call format, VCF) used as "raw" calls. The extreme depth of coverage of our regions allowed us to repeat this procedure three times, generating 3 VCF files based on a randomly chosen subset of reads for each sample, with a maximum of 250x depth per sample. An identical SNV filtering procedure was applied to the three replicates. Only the SNVs that passed the filters in all three replicates were kept for the genotype calling. These replicates can be seen as a partial validation.

ii. Variant filters

A threshold on the minimum base quality and the minimum mapping quality of the reads (40) was set to generate the raw calls. The maximum number of alternate alleles to genotype was set to 2. The threshold for 6 quality control variables (HS, QD, FS, MQRS, RPRS, Hrun) was empirically determined by plotting the distribution of the aforementioned variables. Raw calls were filtered out if they showed a high Haplotype Score (HS > 50), low average quality (QD < 2), if they presented a strong strand bias (FS > 900), a strong correlation between mapping

quality and alternate allele (|MQRS| > 40), and a strong bias position along the read (|RPRS| > 40) or fell in a homopolymer run (HRun > 5). In addition, all variants falling within 10 bp of an indel were filtered out. Triallelic positions were also removed from the dataset.

In parallel, we ran the UG with the -A RMSMappingQuality and --output_mode EMIT_ALL_SITES options. This outputs the distribution of QUAL (Phred scaled probability that REF/ALT polymorphism exists at this site) and mrsMQ (mean root square of Mapping Quality of the reads across all samples) of the entire targeted region (independently on whether the sites are polymorphic or not). Threshold values for the variants were determined by plotting the empirical distributions of mrsMQ and QUAL values (Fig. B1). QUAL provides an overall quality score for the existence of an alternate allele in each genomic position. Variants with QUAL < 250 and mrsMQ < 30 were filtered out.

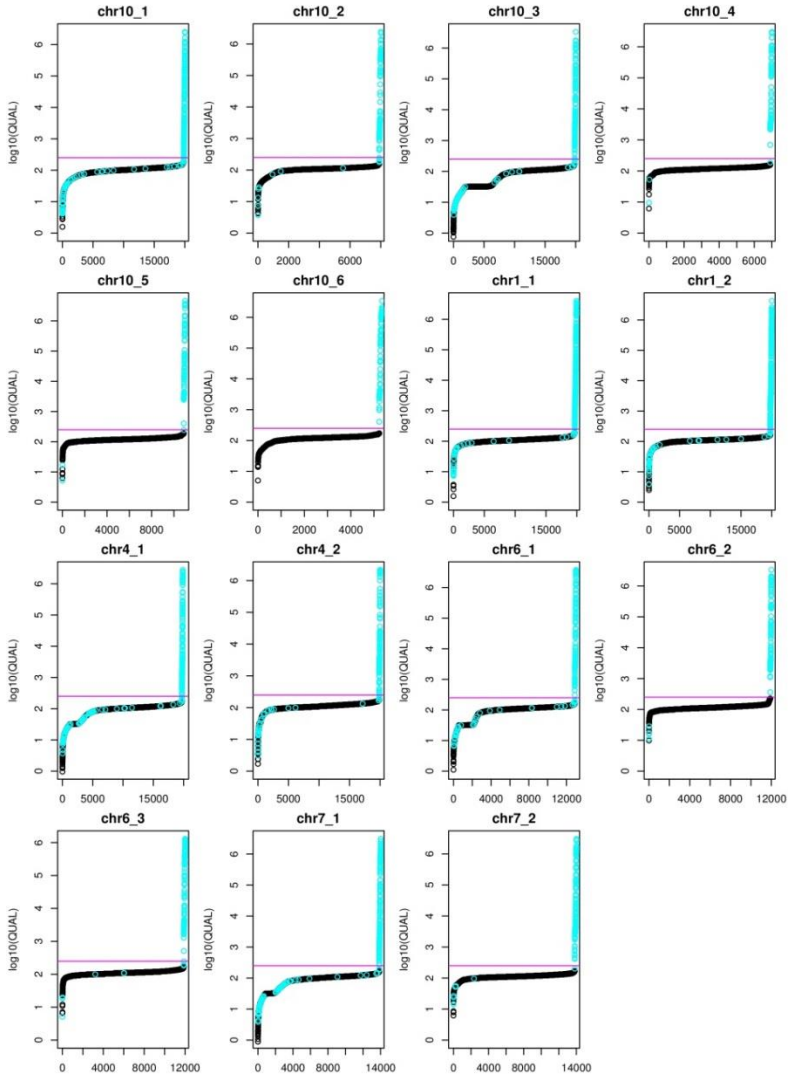

Figure B1. Distribution of the QUAL scores (on a $\log_{10}$ scale) for each variant before filtering. Black points represent sites where no call was made by the UnifiedGenotyper tool, while cyan points are genomic positions marked as polymorphic by the software. Polymorphic positions are expected to have higher QUAL values, while low QUAL values indicate a low probability of the existence of at least one copy of an alternate allele, as expected in non-polymorphic positions. The red line shows the cut-off value we used to filter out calls (cyan points) with low confidence.

Individual genotype filters

Individual genotypes were marked as missing if they presented strong allelic imbalance (>20%), a depth of coverage < 20x or a low individual genotype quality (QC < 30), resulting in only a small amount of missing data. We established that 95% of SNVs have successful calls for at least 900 individual chromosomes out of the possible 986.

Variant calling quality control

Several quality control steps were applied to the final set of variant calls. When compared to dbSNP135, 62.5% of the variants called in our regions were new. Ti/Tv ratio showed no indication of bias calling or sequencing error, neither for all SNVs (Ti/Tv=2.22) nor for novel SNVs alone (Ti/Tv = 2.29). No SNV presented significant departure from Hardy-Weinberg equilibrium.

C. Calling Validation

i. Calling validation by comparison to the site frequency spectrum (SFS) of the NHLBI Exome data

The NHLBI exome sequencing project (ESP) data was obtained from the Exome Variant Server of the NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS/). Data was downloaded from the server in December, 2012. We first filtered out indels, variants with an average read depth of less than 20 and variants that were monomorphic in Europeans. Due to increased noise at lower number of variants, functional categories with less than 1000 SNVs were not analyzed. In order to compare the ESP data to the neutral regions (NR) data, we probabilistically downsampled both datasets to a sample size of 900 chromosomes.

NR data plotted against the ESP data from different functional categories reveals that the NR SFS is more similar to the SFS of less conserved functional categories (near genes, intergenic and intronic regions, Fig. S4a). In addition, the NR data has the lowest fraction of singletons when compared to categories that are more conserved (missense, UTR, splice sites, synonymous, synonymous near splice sites, stop-gain, Fig. S4b), reflecting that fewer variants in the NR dataset are under selection.

To test whether the SFS of the NR data and the SFS of a particular functional category of the ESP data were significantly different, we additionally calculated the chi-square test statistic as $\sum_i \frac{(e_i - o_i)^2}{e_i}$. Let the expected SFS be the SFS obtained from combining the ESP data and the

NR data under the null hypothesis that the two datasets have the same SFS. Then e_i is a vector of size 2, equal to the product of the number of variants in either the NR data or the ESP data and the proportion of variants in minor allele count bin i of the estimated SFS. o_i is also a vector of size 2, equal to the observed number of variants in each minor allele count category in either the NR data or the ESP data. To increase the consistency of the result, we binned categories with more than 10 minor alleles into one bin such that the chi-square test statistic has 10 degrees of freedom. We performed a chi-square test and calculated a p-value for each of the pairwise comparisons. Similar to what we showed on Figure S4, we find that categories with greater functional constraint deviate further from the NR data (Table S3).

ii. Calling validation by comparison with data from the CHARGE-S project

In order to test for overall calling quality, false positive and false negative rates in the NR data, we compared our data to other sequencing data. An appropriate dataset for this comparison is the 962 individuals with whole-genome sequencing data in the CHARGE-S project (8).

In order to compare the SFS of both the CHARGE-S and the NR data, we subset the CHARGE-S data to an equal sample size as sampled in the NR data (n = 493 individuals). These individuals were chosen either to be samples that were also in the NR data, or individuals from a homogeneous cluster on a principal component analysis (PCA) of the CHARGE-S data. Out of the 493 individuals chosen, 395 are part of the ARIC cohort, of which 20 were also sequenced in the NR analysis. Thus, we had sequences that were obtained both by the CHARGE-S project and by the present study, for the same genomic regions and individuals. Using the list of genotype calls for the 20 overlapping individuals in each study, we could use the genotype consistency as

a measure of calling quality. In calculating the proportion of genotypes ascertained in the other dataset, we considered only positions where the derived allele is present in at least one of the 20 individuals and only heterozygous genotypes.

One important property of this validation procedure is the asymmetry of the two datasets. The NR dataset has an average median coverage of 295X, while the average median coverage in CHARGE-S for the corresponding regions in the subset of 493 individuals is 4.6X. This fundamental difference in coverage represents a limitation for the estimation of false positives in the NR dataset, as a call missed in CHARGE-S does not necessarily indicate a calling error in the high coverage NR data. It also represents to a lesser extent a limitation in estimating the false negative rate since the low-coverage data are more prone to sequencing errors.

An estimate of the proportion of calls missed due to low coverage was computed as part of the CHARGE-S project. The adjustment is based on 886 individuals that have both high coverage exome-sequence data and low coverage whole genome sequence data. An exponential decay function was used to fit the discovery rate of the first 20 bins, similar the one used in Gravel *et al.* (9). This adjustment represents a deviation, for each minor allele count (MAC), from a perfect calling with 100% power. We apply this adjustment to the CHARGE-S data to estimate false positive rate of the NR data.

Overall calling quality

First, for each dataset of 493 individuals, we counted the number of copies of the minor allele at all the positions that were polymorphic. Eliminating 8 cases where our data have more than 80%

missing genotypes, the consistency in minor allele frequency (MAF) between both datasets is very high (Spearman rho=98.56) for the 1115 polymorphic positions called both by CHARGE-S and the NR (Fig. C1). This shows that the overall NR calling does not present a strong bias related to variant frequency.

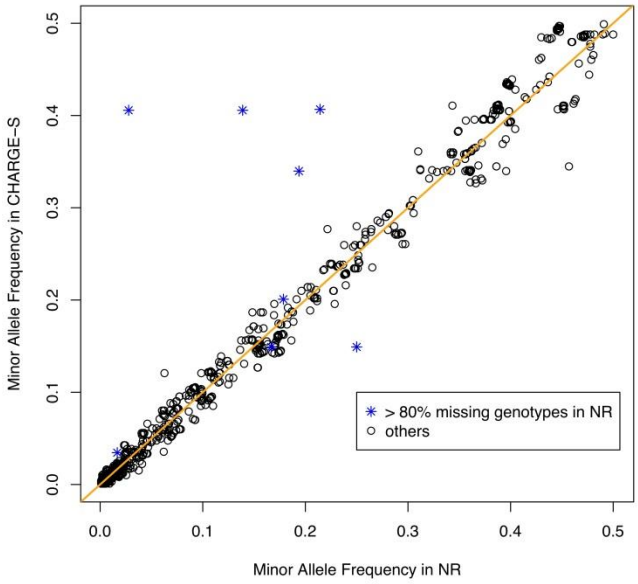

Figure C1. Concordance of the minor allele frequency (MAF) of the NR and the CHARGE-S subset. Outliers, marked with a star, are SNVs for which more than 80% of the individuals have missing data in the NR data. Individual genotypes of the CHARGE-S were imputed and have therefore no missing data.

Second, the average median coverage of the positions that show perfect consistency between CHARGE-S and NR is 4.96X and 336.4X, for CHARGE-S and NR respectively, which is above each dataset's average. Similarly, the average median coverage of positions with one or more genotype found in NR but not called in CHARGE-S is 4.76X and 317.5X, for CHARGE-S and NR respectively. In contrast, the average median coverage of positions with one or more genotype found in CHARGE-S but not called in NR is below average: 4.27X and 223X for CHARGE-S and NR respectively. This suggests that these positions are more likely to be

enriched for calling errors. However, the increase in calling error is probably very low for data of 223X coverage, suggesting that the genotypes with discrepancy are more likely to be CHARGE-S calling errors.

False positive rate in NR

As a first step, we considered the proportion of concordant genotypes in the subset of 20 individuals sequenced independently by the CHARGE-S project and the present study (black in Fig. C2). We observe that a low proportion of genotypes in the NR validation subset are consistent with the CHARGE-S genotypes. For example, only 37% of singletons in the NR validation subset are also called by CHARGE-S. The genotype consistency between both datasets increases for increasing MAC. For MAC=8, this proportion reaches 85% and for MAC=41 the proportion is above 95%. Given the high coverage of the NR dataset and the low coverage in CHARGE-S, we expect a difference in the power to call variants, especially for low MAC categories. In particular, we expect a fraction of the NR variants to be absent in CHARGE-S without necessarily being NR false positives. To explore this idea, we estimated for each MAC the proportion of SNVs that the CHARGE-S data had the power (yellow in Fig. C2) to detect by fitting an exponential decay function to the first 20 MAC categories, similar to Gravel *et al*. (9). The discovery power can be interpreted as an expected false negative rate in CHARGE-S. We observed that the proportion of NR genotypes observed in CHARGE-S follows very closely the theoretical maximum power line in CHARGE-S. In other words, the proportion of NR calls not observed by CHARGE-S follows closely the expected proportion of false negatives in CHARGE-S. As a consequence, this suggests that the false positive rate in NR is not very high,

although the difference in coverage prevents us from providing a reasonable quantitative estimate.

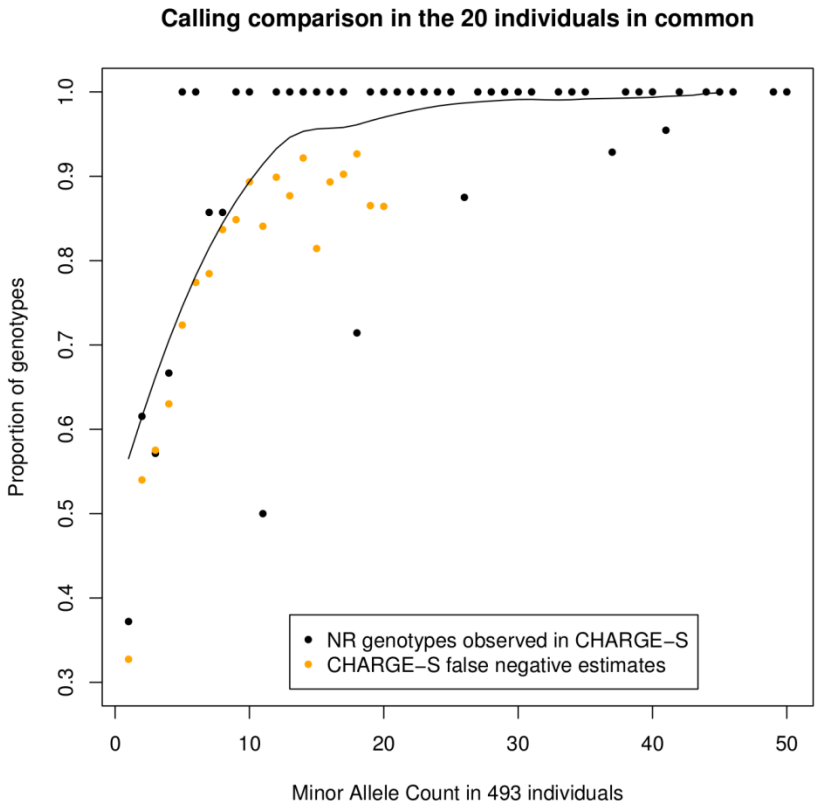

Figure C2. Proportion of concordant genotypes in the subset of 20 individuals sequenced independently by the CHARGE-S project and the present study. The y-axis presents the proportion of NR genotypes that are concordant with CHARGE-S genotypes (black points). The x-axis shows the minor allele count (MAC) of the CHARGE-S SNVs. Data are shown for the first 50 categories of MAC. Note that MAC values go beyond 20 because although only 20 individuals were involved in the genotype comparison, we considered the minor allele count of the SNVs in the whole sample of 493 individuals. We removed part of the noise due to the low sample size involved in the comparison (20 individuals) by fitting a loess function to the data (line). Yellow points show the estimated discovery power in the low coverage CHARGE-S data. The discovery power provides an approximation to the false negative rate in CHARGE-S. The NR genotypes discordant with the CHARGE-S genotypes are more often found in low MAC categories of the CHARGE-S data, for which the discovery power is also lower. The overall agreement between the black and yellow points shows that the SNVs in NR not found in CHARGE-S can be almost entirely explained by the limited discovery power in the lower coverage CHARGE-S data.

False negative rate in NR

In another approach to identify potential false negatives in the NR dataset, we counted the genotypes called by CHARGE-S and not discovered by the NR in the 20 individuals sequenced by both studies. The proportion obtained varied between 0% and 50% depending on the MAC. The principal difference with the previous comparison is that the proportion does not increase with the MAF, but is equally distributed between rare and common variants of the NR data, ruling out a systematic bias related to the allele frequency (Fig. C3a). In contrast, the NR genotypes discordant with CHARGE-S are mainly rare variants in CHARGE-S (Fig. C3b).

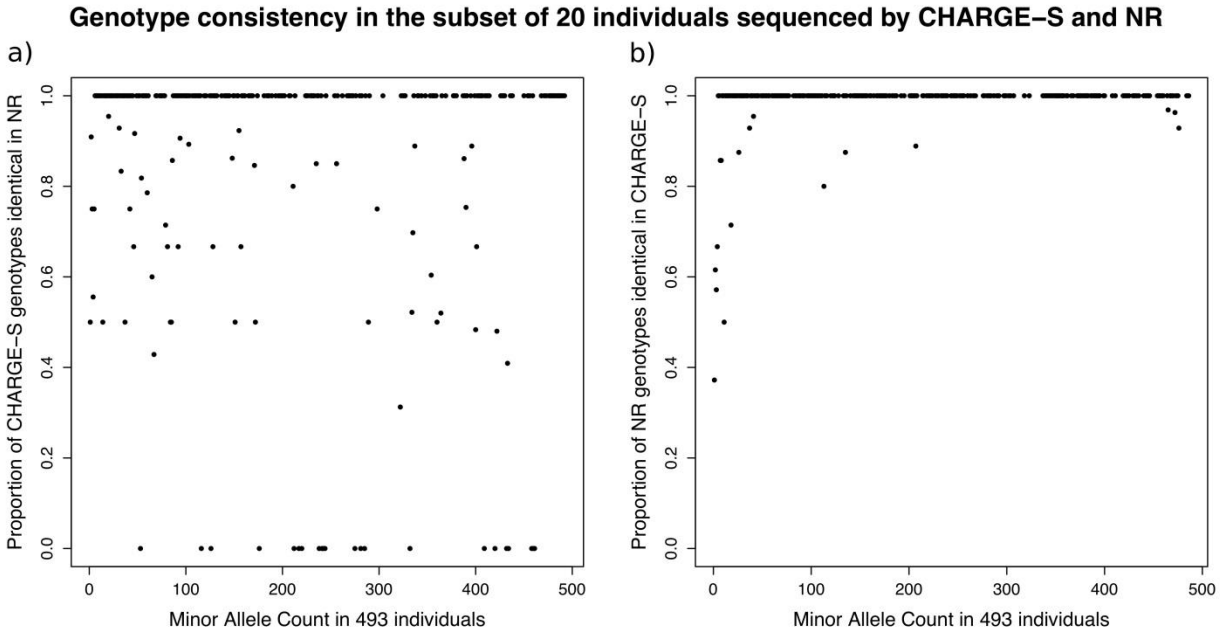

Figure C3. Genotype consistency among the 20 individuals sequenced both by CHARGE-S and the present study. a) The y-axis presents the proportion of CHARGE-S genotypes that are concordant with NR genotypes. b) The y-axis presents the proportion of NR genotypes that are concordant with CHARGE-S genotypes. On both panels a) and b), the x-axis shows the minor allele count of the CHARGE-S SNVs in the whole sample. Panel b) data are the same as in Figure C2 and is shown for comparison purposes. We observe that the proportion of CHARGE-S genotypes discordant with NR genotypes are equally distributed among rare and common variants of the NR data (panel a), indicating no systematic bias in the NR genotypes, whereas the NR genotypes discordant with CHARGE-S are mainly rare variants in CHARGE-S (panel b).

In order to understand the source of the genotype discrepancies, we checked the NR calling before quality control and filtering. We observe that 77.3% (92/119) of CHARGE-S genotypes not observed in NR are either positions not called due to low base quality, low mapping quality or extremely low coverage (12 cases), or positions called in the raw NR that were subsequently filtered in the QC step (19 cases). This includes triallelic positions (9 cases), proximity to indels (14 cases), coverage < 20 X (38 cases), and positions with low quality control variables (Supplementary Note B). The remaining 22.7% (27/119) of inconsistencies are CHARGE-S genotypes absent in NR, though located in regions with good coverage and QC values. The BAM files for these 27 individuals were visually examined at the ambiguous positions with the Integrative Genome Viewer application (IGV version 2.0.23), enabling the filter to remove duplicated reads. The NR genotypes in these 27 cases do not present any sign of a polymorphism.

From this analysis we conclude that our strict filtering may have generated some false negatives, notably due to strict filtering of QC variables and filtering around indels. Filtering around indels is often justified and has become a common procedure (10-12). In addition, it only affects 1.9% of all the raw calls of polymorphic positions in the NR data and does not systematically affect variants of a particular frequency (the average MAC of SNVs around indels is 138, varying from MAC 1 to 350).

Altogether, comparing the genotype calls in the 20 individuals present in both datasets, only 119 heterozygous genotypes in CHARGE-S were not found in the 3929 heterozygous genotypes in the NR data, representing 3.03% of the false negatives. However, this value is an upper bound,

with the assumption that all CHARGE-S calls not found in NR were true calls. Due to the difference in coverage, and some visual validation, we can safely assume that the real false negative rate is probably much lower. Importantly, we showed that the CHARGE-S calls not found in NR are spread over all frequency categories, showing that our filtering process does not generate a strong bias in the SFS and therefore in the demographic estimates.

D. Data homogeneity and coverage

i. Coverage effect on variant calling

We empirically investigated the effect of coverage by running the main GATK pipeline 3 times, modifying only the maximum number of reads considered per position. This was achieved by changing the option -dcov to 111, 27 and 15. The two first values (111X and 27X) were chosen from the average median depth reported in Tennessen *et al.*(13) and Nelson *et al*. (12), respectively. The 15X maximum depth was chosen arbitrarily to represent lower coverage. The rest of the pipeline is identical to that applied to the NR dataset, with the exception of filtering of individuals with less than 20X, which was only applied to the 111X set.

The SFS derived from the pipeline run on a maximum of 111X does not show any perceptible difference from the NR SFS. In fact they only differ by one SNV (Fig. S3). The SFS from the 27X coverage subset presents a lower proportion of singletons (Fig. S3), though we expect that a relaxed set of filters would allow partial recovery of true singletons. For this reason, we interpret that the comparison between the NR dataset and the 27X subset does not show that 27X coverage represents insufficient coverage, but rather shows that high coverage allows application of very strict quality filters (leading to a reduced false positive rate), which have to be relaxed to call SNVs on data with lower coverage. Finally, we observe that the 15X coverage presents a substantial deviation from the high coverage dataset (Fig. S3). Interestingly, the change in the shape of the SFS from 27X to 15X is more accentuated than the change from 111X to 27X, suggesting that coverage depths below 20X are insufficient for high quality calling.

ii. Assessing population homogeneity with POPRES sample

We additionally compared the sequenced individuals to individuals from POPRES (14). POPRES genotyping data was obtained from dbGaP. To ensure similar sample sizes across European populations, 10 individuals were chosen from each POPRES population (with the exception of Croatia and Greece, where only 7 and 8 individuals could be chosen). Following Novembre *et al.* (15), ancestry for each individual was assigned as the reported ancestry of the maternal and paternal grandparents (after removing individuals with mixed grandparental ancestry). Data for POPRES and ARIC individuals were merged, with 369,145 overlapping sites after removing sites with a joint missingness rate above 10%. Linkage disequilibrium based pruning was applied with a maximum r^2 threshold of 0.5 in PLINK (16) in addition to removing sites with potential allele flips and regions of extended LD (15). We computed principal components on the remaining 150,497 autosomal variants using EIGENSOFT (17). The PCA of POPRES individuals are qualitatively similar to those previously published (15). We next computed PCA using samples from the POPRES cohort and ten randomly chosen samples from the NR project. We found that individuals from the NR had North-Western European ancestry, similar to that of individuals from the POPRES UK sample. Similar results are observed other random NR samples.

iii. The effect of homogeneity of the sampled population on the SFS

In order to test for the importance of using a single homogeneous population to model recent demography, we simulated two populations and compared their SFS (Fig. S1). The first population followed the history of Model II described in Table 1. The other population consisted of 8 subpopulations that mimic European populations and split 400 generations ago: all 8

subpopulations share a common history and have the same initial population size (5633) at the time of split. After the split, all 8 subpopulations follow the same demographic history of Model II in Table 1. We sampled 900 individuals from both populations. For the population with 8 subpopulations, we performed two sampling strategies. In the first case, we sampled a balanced number of individuals in each subpopulation. In the second case, the number of samples taken from each subpopulation was 308, 124, 32, 104, 116, 32, 36 and 48, in proportion to the sample sizes from different European populations in Table S10 of (12). When sampling a balanced number of individuals in each subpopulation, we used 112 samples from 6 populations and 114 samples from 2 populations because we could not equally divide 900 samples into 8. All simulations were done using the software ms (18). The results show that although each subpopulation had the same demographic history as the homogeneous population, the SFS derived from a non-homogeneous sample shows a higher proportion of rare variants than the SFS from the homogeneous sample. This is true both in the case of balanced or unbalanced sampling of 900 individuals from 8 subpopulations. These results demonstrate that using non-homogeneous samples can bias demographic estimates.

E. Demographic modeling and the SFS

i. Downsampling data for figures of the SFS

In order to compare SNVs with differing numbers of successful genotype calls, it was necessary to estimate the allele counts given a lower sample size. Employing the strategy implemented in (19) for a folded site frequency spectrum (SFS), we estimated allele counts for a sample size of 900 chromosomes (as this constituted 95% of our variants). All variants with less than 900 successful genotype counts were removed. Thus, for each SNV, the probability it is of a certain allele count is given by the following formula.

$$P[\min(i-d,\ n-i-(n-m-d))] = \frac{\binom{i}{d}\binom{n-i}{n-m-d}}{\binom{n}{n-m}}$$

where i is the original allele count, m is the number of chromosomes to downsample to, n is the number of successful genotype calls for this SNV and $d = 0,\ldots,\min(i, n\text{-}m)$. When presenting the SFS, variants are binned into the following minor allele count categories: (1, 2, 3, 4, 5, 6-10, 11-20, 21-50, 51-100, 101-200, 201-450). The proportion of variants is then weighted by the number of minor allele count categories in each bin.

ii. SFS of simulated populations according to previously published demographic models

We compared the SFS of the NR data to other SFS of previously published models (11, 12). For each of these published scenarios, we simulated 10,000 independent regions for 900

chromosomes (to match the sample size of the NR data) with the coalescent simulator ms (18). We observe that the fraction of singletons in the NR data is below the estimates given by the two models (Fig. S6). The models over-predict the number of singletons (55.65% and 54.23%, respectively, compared to 38.4% in the SFS of the NR data) and extremely rare variants, a pattern that could be expected if less purifying selection acted on the loci sequenced in the NR data. Though these previous studies modeled recent demography based on the SFS derived solely from synonymous variants in order to minimize the confounding effect of selection, the data of these are nonetheless derived from sequencing of genic regions (11, 12), which may partially explain the differences in estimates with the NR data.

iii. Demographic inference

In all demographic inference models, we fixed ancient demographic history following previously described scenarios (9, 20, 21), and focused on the parameters associated with recent human population growth. When using the ancient demography model of (9), we assumed an instantaneous recovery from the second bottleneck instead of a gradual recovery. We estimated parameters of interest using the maximum likelihood approach of Keinan *et al.* (20). For each of the models considered, we profiled the likelihood surface as a function of the multiple parameters using predefined grid points, which corresponds to a particular combination of multiple parameters. Grid points were carefully selected to cover the range of most likely parameter values with better resolution through an initial set of broader grid points and later zooming in (Table S6). For each grid point, we calculated the expected SFS of the model using simulated genealogies from ms (18). For each grid point, 200,000 ms simulations were conducted and the length of branches that lead to a different allele count was calculated for each.

Branch lengths were then summed up across the 200,000 genealogies and standardized to obtain the expected SFS.

Given the expected SFS, we calculated the likelihood of the observed SFS as follows. Let *s* denote the number of SNVs and *n* denote the number of chromosome samples. Also let x_i and m_i denote the number of minor alleles and the total number of successful genotypes at SNV *i* ($x_i \leq m_i/2 \leq n/2$). Then, a composite likelihood for the data is given by $L = \prod_{i=1}^{s} [P(x_i \text{ of } m_i) + P(m_i - x_i \text{ of } m_i)]$, where *P(a of b)* is the probability of observing *a* alleles out of *b*, conditioned on demography. In order to account for the sampling of the m_i successful genotypes out of the possible *n*, we derived *L* by the law of total probability to be

$$L = \prod_{i=1}^{s} \left(\sum_{j=x_i}^{n-m_i+x_i} \frac{\binom{j}{x_i}\binom{n-j}{m_i-x_i}}{\binom{n}{m_i}} P(j \text{ of } n) + \sum_{j=m_i-x_i}^{n-x_i} \frac{\binom{n-j}{x_i}\binom{j}{m_i-x_i}}{\binom{n}{m_i}} P(j \text{ of } n) \right)$$

To explore models of recent history, we started with a simple recent growth model with two parameters capturing the time the growth started and final N_e ('Model I'). Since we assume an exponential growth with constant rate, the growth rate is no longer a free parameter and can be derived for any given time of growth and final N_e. We also studied two three-parameter models: similar to Model I, with the ancestral N_e before growth as an additional free parameter ('Model II'), and a model with two epochs of growth with the start time of the first growth fixed at 400 generations ago ('Model III'). We additionally explored a four-parameter model by allowing the time of the first growth in Model III to be another free parameter ('Model IV').

To improve resolution, we applied a smooth spline approach to estimate the frequency of SNVs belonging to a MAC category as a function of each parameter (see below). In all models where the smooth spline approach was used, the 95% CI for a given parameter is defined as all values of the parameter for which the log-likelihood is comprised between the maximum value minus 1.92 (half of the 95% percentile of the $\chi^2_{(1)}$ distribution) and the maximum, while fixing all other parameters at the maximum likelihood estimates. CI for the 4-parameter models, where the smooth spline approach was not applied for computational efficiency, were obtained from the original grid points.

iv. Evaluation of the power of the maximum likelihood approach

One epoch of growth

To evaluate the power of our maximum-likelihood approach, we used the coalescent simulator ms (18) to simulate a population with an ancient history as described in Keinan *et al.* (20), and with a single epoch of growth starting 400 generations ago. During this time, the population grows from N_e = 10,000 to N_e = 1,100,000, representing a growth rate of 1.18% per generation. We chose a value of 1,100,000 for the final N_e as described in (22) because it is between the other two recent estimates of European N_e: 512,000 7(11) and 4,000,000 (12). For this scenario, we simulated 986 chromosomes with 500,000 SNVs. We randomly subsampled 1800 SNVs (comparable to the number of SNVs in the NR dataset), and applied the likelihood method to estimate three parameters: initial N_e before the growth, time of the onset of growth, and growth rate. We repeat the above procedure and define power as the proportion of 10,000 replicates of resampling for which the three parameter estimates are a minimum of one grid point away from

the true value. The result shows a reasonable power of ~80% to capture the three parameters for the scenario we tested.

Two epochs of growth

As above, we simulated a population with ancient demography as described Keinan *et al.* (20) with a 2-epoch growth model. We varied the rate of the first epoch of growth while fixing all other parameters, including the length of the two epochs of growth (280 and 120 generations respectively), the rate of the second growth (3%), N_e before growth (5,633), and ancient demographic parameters (20). For each parameter combination, we simulated a sample of 986 chromosomes with 500,000 SNVs, from which we randomly subsampled 1800 SNVs and conducted a likelihood ratio test between a reduced Model II (2 parameters with fixed ancestral population size of 5633) and Model III (three parameters). We repeated this procedure and defined power as the proportion of 1000 replicates for which the likelihood ratio test p-value is less than 0.05. Power was estimated for each parameter combination.

As shown in Fig. S9, we have reasonably good power (over 60%) for detecting the first epoch of growth when the rate of growth falls between 0.6~1.8%.

v. Validation of improved accuracy of estimated SFS using smoothing spline

A smoothing spline (smooth.spline in R (23)) was used to estimate the frequency of SNVs belonging to a specific MAC category as a function of one of the parameters each time. To validate the performance of the smoothing spline, we conducted a simulation study using one

specific demographic history with one epoch of growth at 100 generations ago and 24 grid parameter values for final N_e varying from 2 to 200 million. For each of the two-parameter combinations, we performed 10,000 ms simulations to get the estimated SFS and an additional 1.2 million ms simulations to get another estimate of the SFS as the "true underlying" SFS. Although the estimate from 1.2 million simulations may not be equal to the true underlying SFS, this was only used as a standard to validate the performance of the smoothing spline by comparing the estimate from 10,000 simulations, the estimate after applying the smoothing spline, and the estimate from 1.2 million simulations.

The squared distance from the 1.2 million simulations SFS was consistently smaller for smoothing spline estimates than the original ones (Fig. S10), indicating the superior performance of the smoothing spline approach. Moreover, by using smoothing spline, we can obtain the estimated SFS for any grid point from a fitted line, which will increase the accuracy of our approach, in addition to saving computing time.

In order to further reduce noise and increase accuracy, we binned SNVs with MAC greater than 50 into one category. Smoothing spline and likelihood calculations were applied to the data after binning (22).

F. Difference between census population size and effective population size

We show that a model with two epochs of growth (Model III) does not provide a better fit to our data than a model with a single epoch of growth (Model II). Besides power considerations, this result suggests that although the census population size (N_c) started to increase with the invention of agriculture 400 generations ago (24, 25), the effective population size (N_e) only began to grow recently (140 generations ago, according to Model II). A high variance in reproductive success is one known factor that explains $N_c > N_e$ (26). Here, we explore how large the variance in reproductive success should be in order to explain alone a constant N_e during the Neolithic Europe.

According to Crow and Kimura (27), for diploid species with two sexes under the assumption of random mating, the effective population size of the t^{th} generation is a function of the census population size of t^{th} generation $N_c(t)$, the average number of offspring per individual of $(t - 1)^{th}$ generation $\bar{k}(t - 1)$ and the variance of the number of offspring per individual $V_k(t - 1)$. $\bar{k}$=2 corresponds to the population maintaining a constant census population size and $V_k = 2$ for a population of constant population size with a Poisson distribution of offspring number. Thus, we have:

$$N_e(t) \approx \frac{2N_c(t)}{\bar{k}(t - 1) - 1 + \frac{V_k(t - 1)}{\bar{k}(t - 1)}}$$

which further leads to the following equation,

$$\frac{N_e(t+1)}{N_e(t)} = \frac{2N_c(t+1)}{2N_c(t)} \frac{\bar{k}(t - 1) - 1 + \frac{V_k(t - 1)}{\bar{k}(t - 1)}}{\bar{k}(t) - 1 + \frac{V_k(t)}{\bar{k}(t)}}$$

If the growth rate of the N_c is (α - 1), i.e., $N_c(t + 1) = \alpha N_c(t)$, then the average number of offspring per individual $\bar{k} = 2\alpha$ is a constant, and the growth rate of the effective population size is (β - 1), i.e., $N_e(t + 1) = \beta N_e(t)$. Solving for the variance of the number of offspring per individual of t^{th} generation at time growth begins, $V_k(t)$ gives:

$$V_k(t) = \left(\frac{\alpha}{\beta}\right)^t V_k(0) + 2\alpha(2\alpha - 1)\left[\left(\frac{\alpha}{\beta}\right)^t - 1\right] \text{ (Eq.1)}$$

Archeological data can provide estimates for α during the time frame we attempted to detect the first growth epoch (260 generations, from 400 to 140 generations ago). Around 1000 BC, Northern Europe population size estimates vary between 100,000 for the British Islands to roughly 200,000 for a territory overlapping part of the current Poland and Germany (28). Growth in census population size may have been as high as ~1% per generation at that time according to archeological data (24, 29, 30), although other evidence suggests it could be around 0.4% (9). This leads to estimates of α ranging from 1.004 to 1.01 and $N_{c, 140}$ (the census population size 140 generations ago) ranging from 2.82x to 13.29x of $N_{c, 400}$ (the census population size 400 generations ago). As an illustration, if $N_{c, 400}$ is 5633 (as in Model II), population growth results in an $N_{c, 140}$ between 15,904 and 74,868 for an α of 1.004 and 1.01, respectively. If we assume that the reproductive success 400 generations ago was Poisson distributed, i.e., $V_{k, 400} = 2$ (the variance in reproductive success 400 generations ago), and there was actually no effective population growth ($\beta = 1$), the above scenarios result in a $V_{k, 140}$ (the variance in reproductive success 140 generations ago) between 9.34 to 51.91 for α of 1.004 and 1.01, respectively. These values are in the reasonable range for historical civilizations (31). For example, historical data on intensive agriculturalist civilizations report as many as hundreds of children for a very few elite males. Estimates based on contemporary pre-industrialized societies show that the variance in male reproductive success in pastoralists and horticulturalist cultures varies between 8 and 86

(31). Thus, the increase in V_k may explain why the increase in N_c was not accompanied with an increase in N_e.

In addition to the case that the increase in V_k leads to a constant N_e, it is also possible that the increase in V_k causes the growth rate of N_c to be accompanied with a milder growth rate of effective population size, which might be in the range for which we have limited statistical power. The following table shows V_k under different combinations of growth rates of N_c (in the range of 0.6% to 1%) and growth rates of N_e (in the range of 0.3% to 0.6%, corresponding to statistical power of 25% and 60% respectively).

β \ α	1.006	1.007	1.008	1.009	1.01
1.003	6.74	9.34	12.70	17.06	22.71
1.004	4.74	6.74	9.33	12.70	17.05
1.005	3.19	4.74	6.74	9.33	12.69
1.006	2.00	3.19	4.74	6.74	9.33

Most of the variances above are within reasonable ranges of estimates for human populations (31). Even if the census population size has grown by 1% per generation over this period of 260 generations, V_k of 9.3 would lead N_e to only grow by 0.6% per generation, which we only have a statistical power of 60% to detect.

G. *ms* command lines

Commands for running *ms* for the four models (I through IV, respectively) from Table 1 in the main text, using best-fit parameters, are reported below:

Model I: ./ms 986 25000 -T -G 0.11531883E+07 -eG 0.54230771E-05 0 -eN 0.2980769E-04 0.1056635E-03 -eN 0.3461538E-04 0.1923077E-02 -eN 0.2221154E-03 0.3642163E-04 -eN 0.2269231E-03 0.1923077E-02

Model II: ./ms 986 25000 -T -G 0.88335800E+05 -eG 0.53822631E-04 0 -eN 0.2370031E-03 0.8401376E-03 -eN 0.2752294E-03 0.1529052E-01 -eN 0.1766055E-02 0.2895910E-03 -eN 0.1804281E-02 0.1529052E-01

Model III: ./ms 986 25000 -T -G 0.10703884E+06 -eG 0.45445210E-04 -0.53861189E+03 -eG 0.13698630E-03 0 -eN 0.2123288E-03 0.7526712E-03 -eN 0.2465753E-03 0.1369863E-01 -eN 0.1582192E-02 0.2594418E-03 -eN 0.1616438E-02 0.1369863E-01

Model IV: ./ms 986 25000 -T -G 0.65788146E+05 -eG 0.70000000E-04 -0.11920425E+04 -eG 0.17000000E-03 0 -eN 0.3100000E-03 0.1098900E-02 -eN 0.3600000E-03 0.2000000E-01 -eN 0.2310000E-02 0.3787850E-03 -eN 0.2360000E-02 0.2000000E-01

H. Conservation analysis of the NR data

We chose the loci with the aim of minimizing the confounding of selection in our data. To investigate the level of selection, beyond comparisons to other types of loci (Figure S4; Table S3), we utilized GERP (32) to assess the amount of selection or constraint loci in our regions were under. GERP measures the difference between the expected substitution rate under neutrality and the observed rate, where positive scores are given to constrained elements or variants suspected of having undergone purifying selection (32). We lifted-over our data from hg18 to hg19 (33) and used the GERP conservation track on the UCSC genome browser (33) to obtain the score for each of our 1834 SNVs that passed quality control filtering. For comparison, we obtained GERP estimates for synonymous SNVs from the Exome Sequencing Project (ESP) data (34). Comparing the distribution of variants from the NR data to those from the synonymous category within the ESP data, we found that variants from the NR data are more closely distributed around zero (Fig. S5). We additionally filtered synonymous SNVs in the ESP data that lay in conserved elements (Mammal El, phastConsElement44wayPlacental UCSC track) and still observed a wider distribution as compared to variants from the NR data. These results support that the putatively neutral loci in our regions are less confounded by selection than the synonymous SNPs used in some of the previous studies that modeled recent growth in population size. This is reasonable given the criteria we filtered upon when choosing regions to sequence (Supplementary Note A).

We next partitioned SNVs to those with a GERP score between -4 and 0, and those with a positive score. The SFS of both sets are very similar to the SFS of the original data, though the

set with positive GERP scores deviates slightly in the direction that the ESP synonymous data deviate from our data (Fig. H1).This is in line with this set being more affected by selection, but still to a much smaller extent than synonymous SNPs are. We additionally re-estimated demographic parameters from the best-fit model II (Table 1 in main text) using each of the two subsets of the NR data. We found that the 95% confidence interval for both subsets include the original estimates on the full data, though estimates from the GERP score > 0 subset were somewhat closer to demographic estimates from previous studies utilizing synonymous variants (13) (Table H1).

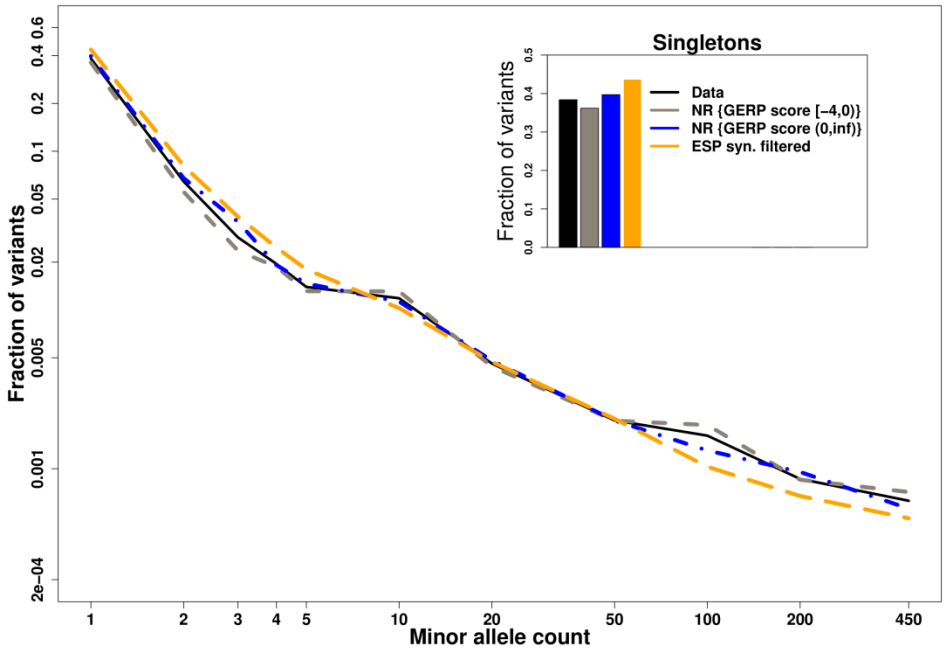

Figure H1. SFS of the variants from the NR data separated into variants with a GERP score between -4 and 0 (not including 0), and those with a positive GERP scores. For comparison, the SFS of synonymous variants from the Exome Sequencing Project after filtering SNVs in conserved elements is also presented. The SFS of variants with a positive GERP score deviates slightly from the SFS of the full NR data in the direction of the synonymous variants, consistent with more purifying selection on these sites, but not as extensive as purifying selection on synonymous sites.

	GERP_Score>0	GERP_Score<0
Ne after earlier growth	6519.0 (3949.4, 9000.0)	4924.0 (3240.5, 6962.0)
Duration of growth (generations)	156.7 (108.8, 216.6)	108.8 (76.9, 148.7)
Ne after growth (millions)	0.65 (0.2, 112.0)	1.33 (0.22,284.0)

Table H1. Demographic inference of Model II on subsamples of data with different GERP scores (Supplementary Note H). 95% confidence intervals are shown in parenthesis. In all cases, 95% confidence intervals include the estimates derived from the full NR dataset.

REFERENCES

1. Wall JD, *et al.* (2008) A novel DNA sequence database for analyzing human demographic history. *Genome Res* 18(8):1354-1361.
2. Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res* 19(5):711-722.
3. Arbiza L, Zhong E, & Keinan A (2012) NRE: a tool for exploring neutral loci in the human genome. *Bmc Bioinformatics* 13.
4. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
5. Li H, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
6. McKenna A, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297-1303.
7. DePristo MA, *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491-498.
8. Psaty BM, *et al.* (2009) Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet* 2(1):73-80.
9. Gravel S, *et al.* (2011) Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A* 108(29):11983-11988.
10. Genomes Project C, *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56-65.
11. Tennessen JA, *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y* 337(6090):64-69.
12. Nelson MR, *et al.* (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337(6090):100-104.
13. Tennessen JA, *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64-69.
14. Nelson MR, *et al.* (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* 83(3):347-358.
15. Novembre J, *et al.* (2008) Genes mirror geography within Europe. *Nature* 456(7218):98-101.
16. Purcell S, *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559-575.
17. Price AL, *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904-909.
18. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337-338.
19. Marth GT, Czabarka E, Murvai J, & Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166(1):351-372.

20. Keinan A, Mullikin JC, Patterson N, & Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* 39(10):1251-1255.
21. Schaffner SF, *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15(11):1576-1583.
22. Coventry A, *et al.* (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* 1:131.
23. R Core Team (2013) R: A Language and Environment for Statistical Computing.
24. Biraben J-N (2003) L'évolution du nombre des hommes. *Population et Sociétés, bulletin mensuel d'information de l'Institut national d'études démographiques* 394:1-4.
25. Keinan A & Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336(6082):740-743.
26. Felsenstein J (1971) Inbreeding and variance effective numbers in populations with overlapping generations. *Genetics* 68(4):581-597.
27. Crow JF & Kimura M (1970) *An introduction to population genetics theory* (Harper & Row, New York,) pp xiv, 591 p.
28. Harding AF (2000) *European Societies in the Bronze Age* (Cambridge University Press) p 572.
29. Kremer M (1993) Population-Growth and Technological-Change - One Million BC to 1990. *Q J Econ* 108(3):681-716.
30. Haub C (1995) How many people have ever lived on earth? *Popul Today* 23(2):4-5.
31. Betzig L (2012) Means, variances, and ranges in reproductive success: comparative evidence. *Evolution and Human Behavior* 33(4):309-317.
32. Cooper GM, *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15(7):901-913.
33. Kent WJ, *et al.* (2002) The human genome browser at UCSC. *Genome Res* 12(6):996-1006.
34. NHLBI GO Exome Sequencing Project (ESP) Exome Variant Server. 2012(December).

I. Figures

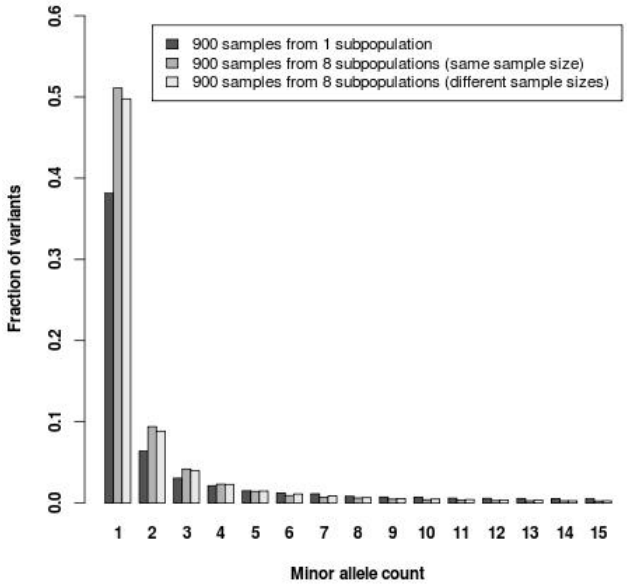

Figure S1. Effect of population homogeneity on the SFS. Samples of 900 chromosomes were taken from either one single population, or from 8 subpopulations. The subpopulations are derived from a population that split 400 generations ago into 8 subpopulations. Each of the 8 subpopulations had the same initial population size (5,633) and the same demographic history after the split as Model II in Table 1. When using unbalanced sample size, we sampled 308, 124, 32, 204, 116, 32, 36 and 48 chromosomes from each subpopulation (Supplementary Note D). When sampling a balanced number of individuals in each subpopulation, we used 112 samples from 6 populations and 114 samples from 2 populations.

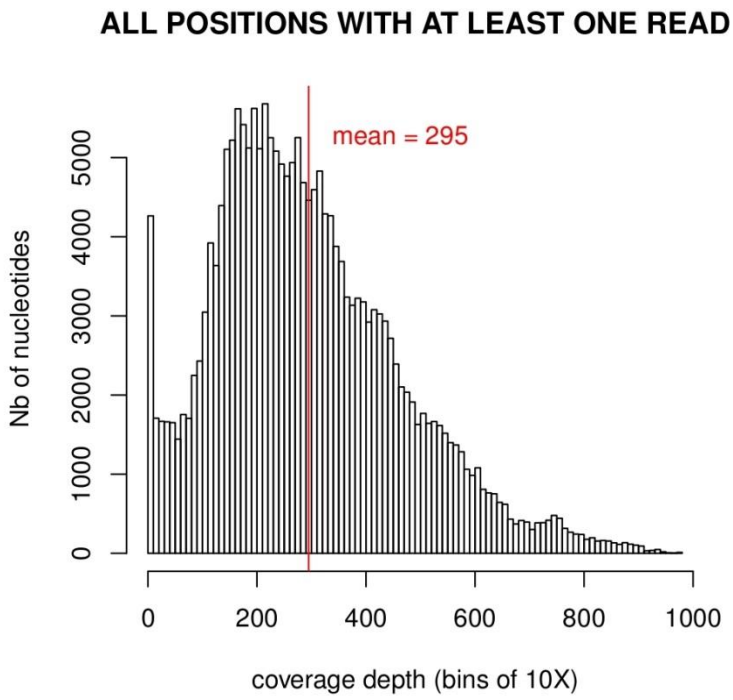

Figure S2: Distribution of the median read coverage calculated for each position of the 216,240 bp sequenced across all 493 individuals. Positions with no reads were not plotted.

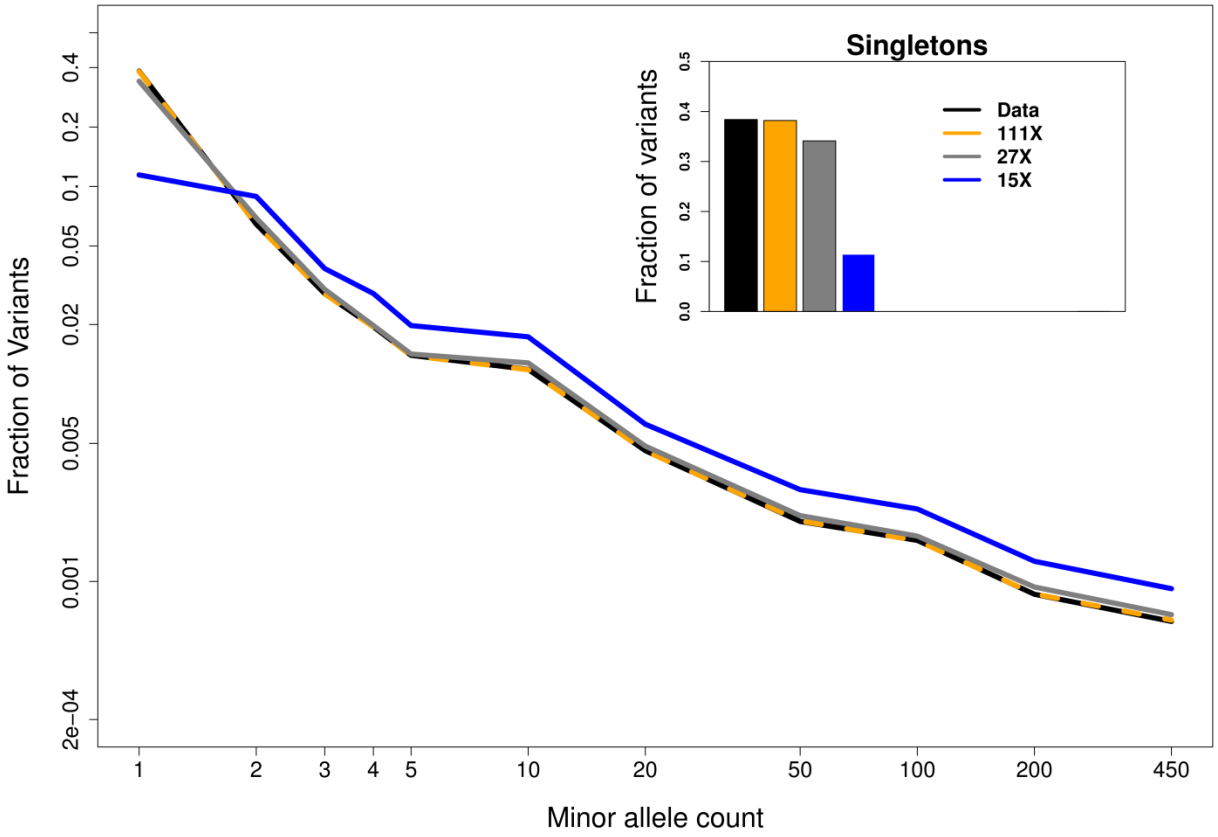

Figure S3. SFS based on different maximum depth of coverage. SFS for 111X and 27X coverage are close to the NR data, while SFS for the 15X coverage is especially different in its much lower proportion of singletons. See Supplementary Note D for a further description of these results.

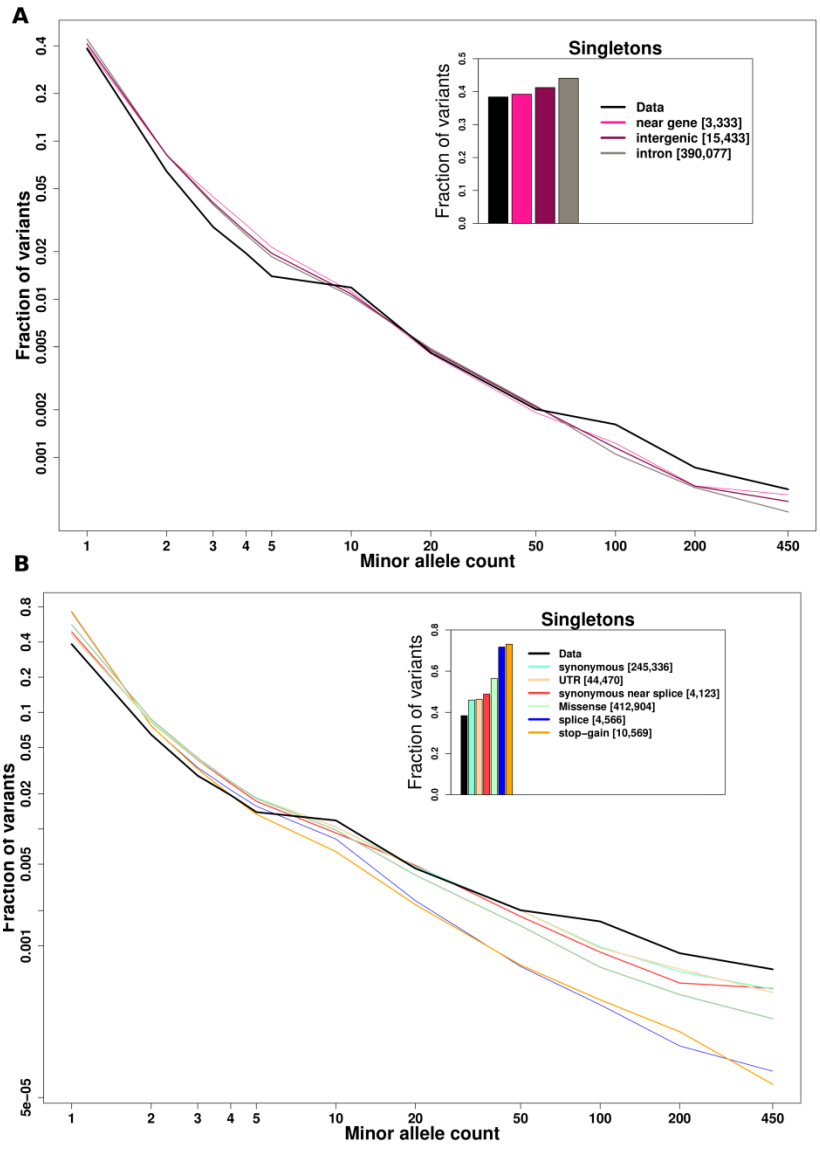

Figure S4. SFS of the ESP data compared to SFS of NR data. Minor allele count of the ARIC neutral region data plotted against data from the NHLBI exome data of various functional categories that are (A) less conserved and (B) more conserved. 'Near gene 3' and 'near gene 5' were combined into the single category of near gene, UTR 3', UTR 5' were combined into the category UTR and splice 3' and 5' were combined into functional category splice. Minor allele counts were binned for counts greater than 5. 900 chromosomes were probabilistically subsampled for each variant (Supplementary Note E). The y-axis is presented on log scale. Inset zooms in on the fraction of variants that are singletons for each category (the y-axis here is in linear scale).

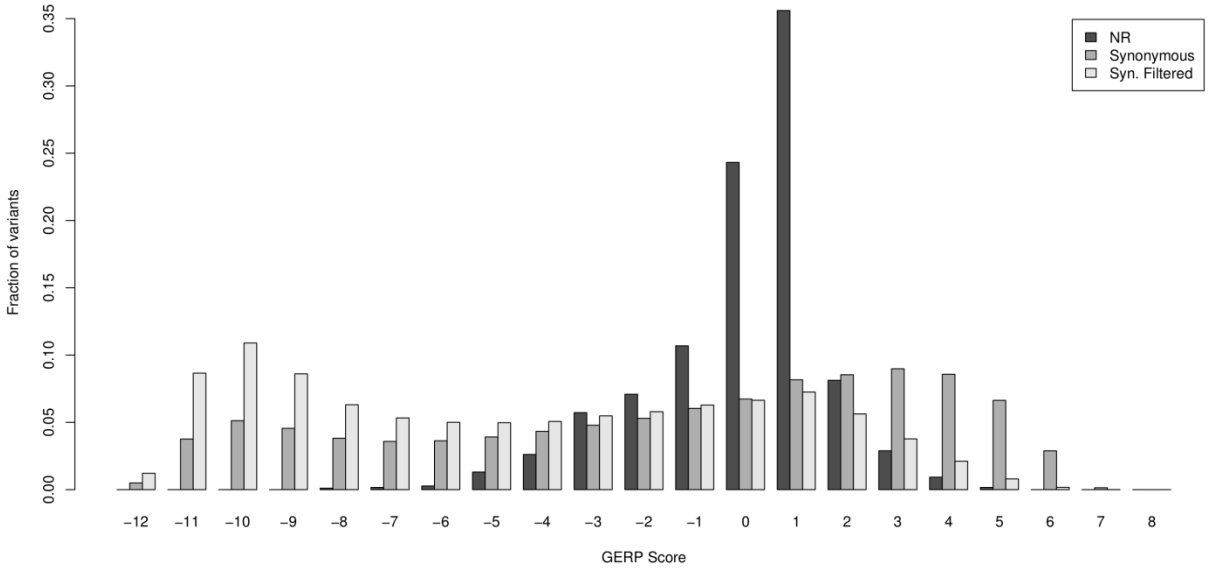

Figure S5. The distribution of GERP scores for variants in the NR data, and synonymous variants in the ESP data. Variants from the NR project ('NR') are more tightly distributed around zero and have a smaller portion of variants with positive GERP > 0, which points to evolutionary conservation. Synonymous variants from the ESP data ('Synonymous') and synonymous variants after filtering out variants in conserved elements ('Syn. Filtered') are more widely distributed around zero. Details regarding GERP scores are provided in Supplementary Note H.

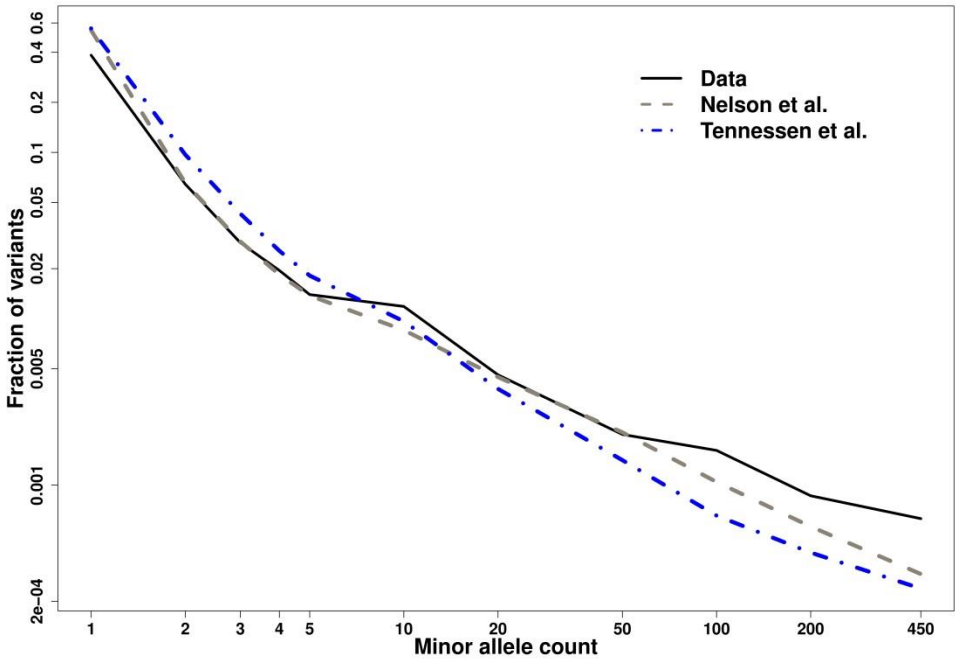

Figure S6. Comparison between the SFS of the NR data to simulated SFS for a matching sample size (900) of various published models. The y-axis is presented in log scale. Simulated data follow two previously published models of recent growth: Nelson *et al.* (2012), and Tennessen *et al.* (2012). The fraction of singletons in the NR data is below the estimates given by the other models. For further details see Supplementary Note E.

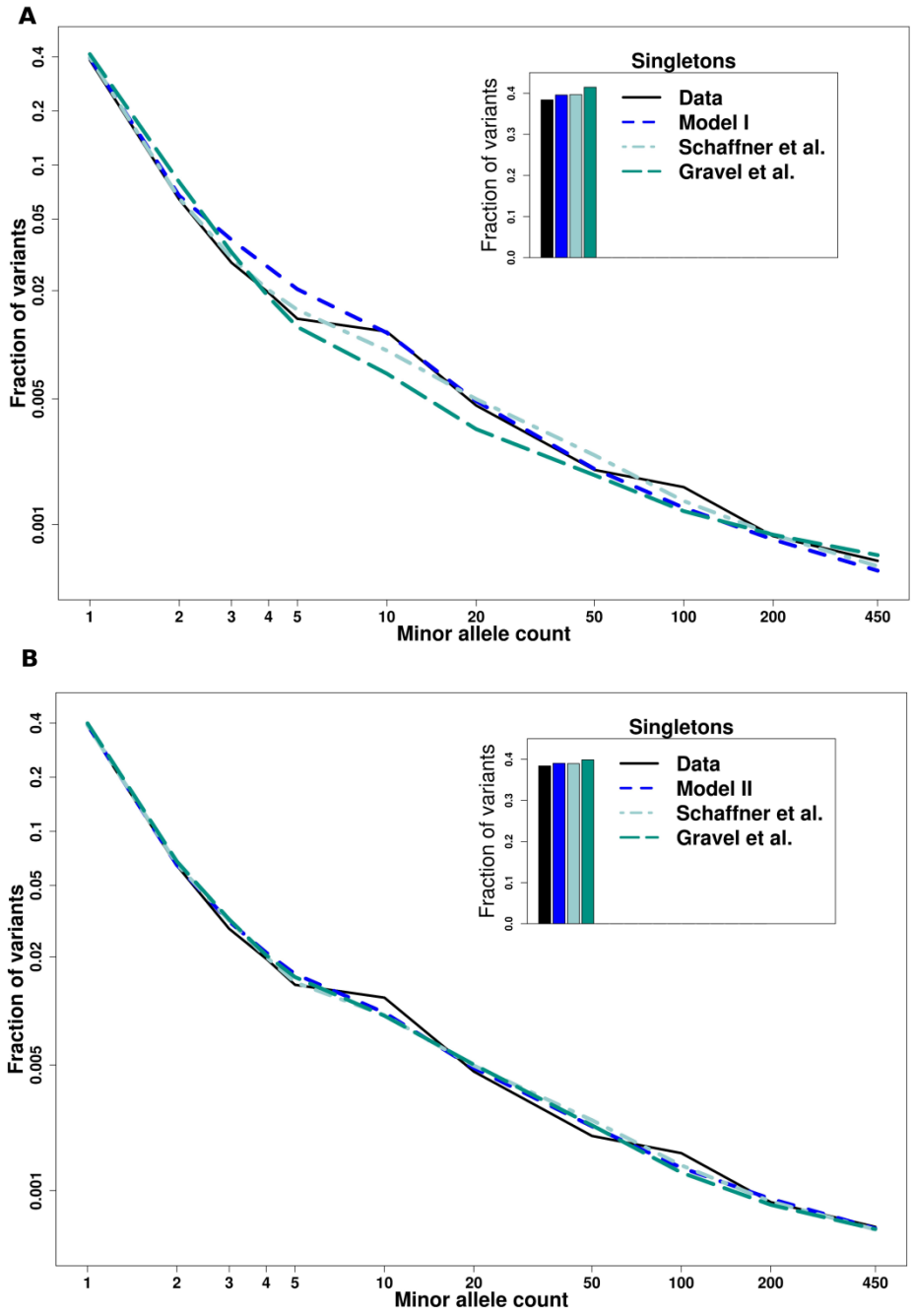

Figure S7: Impact of the ancestral demography on the modeling of recent demography. (A) Model I fit with different scenarios of ancient history. (B) Model II, where population size before growth is added as a free parameter. The y-axis is presented in log scale. All three models are fitted to the NR data and estimate parameter values for a single epoch of growth, with the only difference between the three models being the ancient history scenario assumed before the growth. The sample size and number of SNVs is consistent across all SFS.

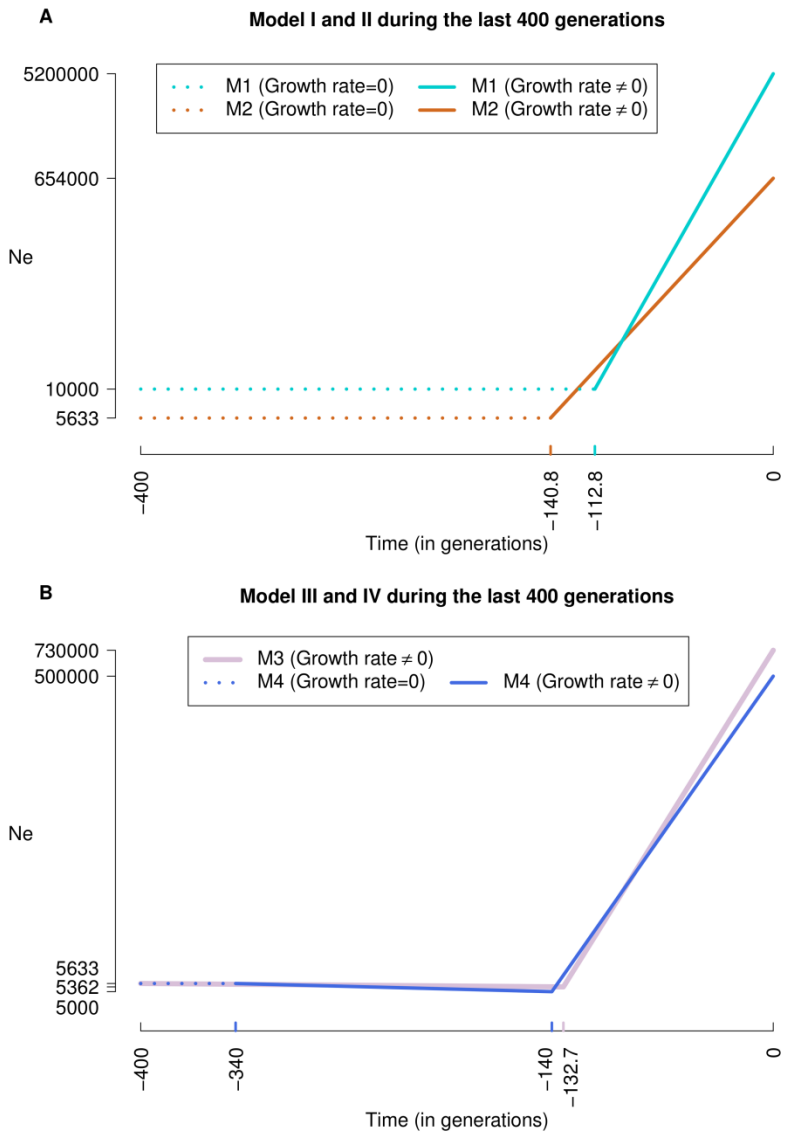

Figure S8. Comparison of the 4 models of recent human history. On both panels, Ne is represented on a log scale. (A) Models with 1 epoch of growth (Model I and Model II; Table 1) represented for the last 400 generations. Dotted lines denote epochs where a model assumes a constant size, and solid lines denote an estimated change in population size. (B) Models with 2 epochs of growth (Model III and Model IV; Table 1) during the last 400 generations.

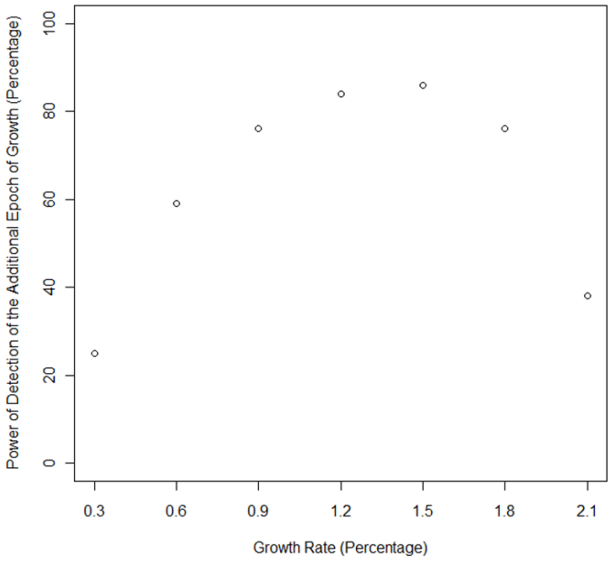

Figure S9: Power to detect an additional epoch of growth for different growth rates. We vary the growth rate of the first epoch of growth while fixing all other parameters. Power is estimated as the proportion of replicates for which a likelihood ratio test p-value is less than 0.05. We performed 1,000 replicates to estimate the power under each parameter combination. (Supplementary Note E).

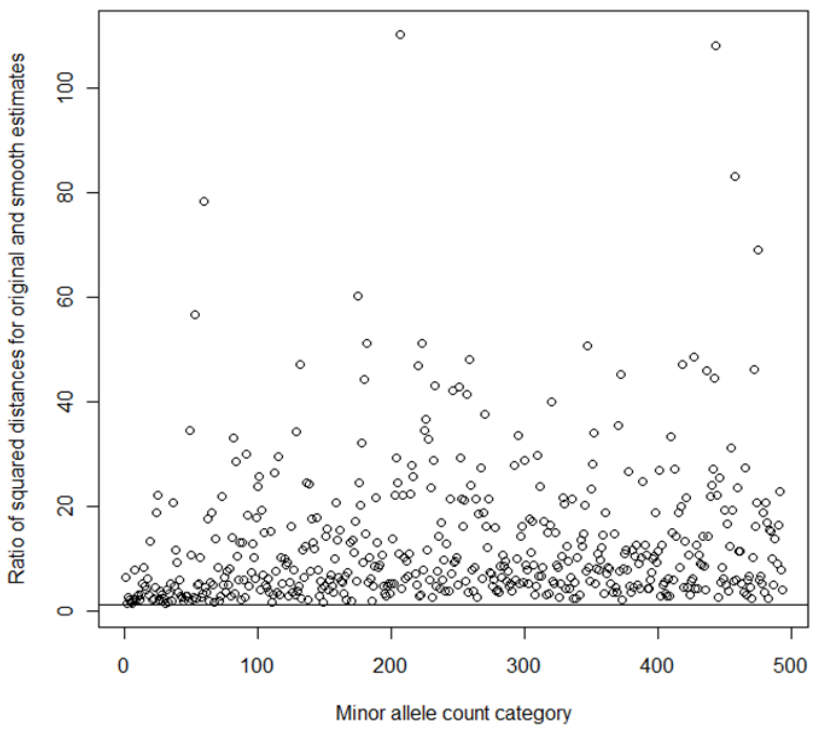

Figure S10. Ratio between distances from "true" SFS for the original and smoothing spline estimated SFS across minor allele count categories using simulations. The solid line denotes the ratio of 1. All points are above the solid line, which means the smoothing spline estimates are closer to the "true" SFS than the original SFS, thus indicating the superior performance of the smoothing spline approach.

J. Tables

Model of recent history	Tennessen *et al.* (2012)	Nelson *et al.* (2012)	Coventry *et al.* (2010)	This paper Model I	This paper Model II	This paper Additional Models
Ancient demography	Gravel *et al.*	Schaffner *et al.*	Schaffner *et al.*	Keinan *et al.*	Keinan *et al.*	Keinan *et al.*
Ancestral Pop Size	14,470	14,000	14,000	10,000	10,000	10,000
Time Out of Africa population bottleneck (kya)	51	87.5	87.5	118	118	118
Pop size after the first bottleneck	1,861	7,700	7,700	10,000	10,000	10,000
Time second bottleneck	23	47.5	47.5	18	18	18
Pop size after the second bottleneck	1,032	7,700	7,700	10,000	5,633	5,633
Growth Duration (generations)	715	NA	NA	NA	NA	see Table 1
Size	9,208*	NA	NA	NA	NA	see Table 1
Growth rate (per generation)	**0.307% (0.301,0.313)**	NA	NA	NA	NA	see Table 1
Growth Duration (generations)	**205**	370	**56 (36,112)**	**112.8 (92.9,136.8)**	**140.8 (116.8,164.7)**	see Table 1
Final size	512,000	**4,000,000 (2.5M,5M)**	1,100,000	**5,200,000 (0.8M,300M)**	**654,000 (0.3M, 2.87M)**	see Table 1
Growth rate (per generation)	**1.95% (1.89,2.01)**	**1.70% (1.2,2.3)**	**9.40% (4.5,14.5)**	**5.54% (3.2,11.1)**	**3.38% (2.4,5.1)**	see Table 1

Table S1. Estimates of recent demography by previously published models compared to the models estimated in the present study. The estimates based on the two-parameter model (Model I in Table 1) are given. Bold values indicated estimated parameters. The other scenarios that have been explored are detailed in Table 1. The asterisk (*) indicates that in the model of Tennessen *et al.* (11), the population does not have an instantaneous recovery of its original size after the second bottleneck: it starts growing at a slow exponential rate of 0.307% per generation after the bottleneck to reach an initial size (before the recent fast growth) of 9028 individuals. For other models, bottleneck intensities are given in Schaffner *et al.* (21) for Nelson *et al.* (12) and Coventry *et al.* (22) models, and in Keinan *et al.* (20) for the current study.

Chr	Target start	Target end	Target length	Base-pairs with coverage > 20X in at least 1 individual	Base-pairs with coverage > 20X in at least 450 individuals	# of SNVs after QC
1	237,146,097	237,165,997	19,900	19,902	19,764	180
1	237,360,801	237,380,801	20,000	20,004	19,539	197
4	164,000,000	164,020,000	20,000	17,962	16,117	96
4	164,115,673	164,135,673	20,000	19,768	19,443	119
6	165,342,001	165,355,001	13,000	11,098	10,659	86
6	165,413,304	165,425,304	12,000	12,002	12,002	103
6	165,470,500	165,482,500	12,000	12,000	12,000	122
7	49,565,734	49,579,734	14,000	12,577	11,183	115
7	49,645,200	49,659,200	14,000	14,000	13,779	172
10	82,981,023	83,001,023	20,000	19,872	18,278	126
10	83,207,715	83,215,715	8,000	8,004	7,378	79
10	83,388,407	83,408,407	20,000	14,534	12,773	131
10	133,143,109	133,150,109	7,000	7,003	7,003	81
10	133,231,083	133,242,083	11,000	11,004	11,004	117
10	133,266,053	133,271,393	5,340	5,285	5,066	110
Total			*216,240*	*205,015*	*195,988*	*1834*

Table S2. Genomic locations of the neutral regions (NR) sequenced in this analysis. All positions are given for the human genome assembly hg18.

Functional_Category	# SNVs	$\chi 2$	P value
Near gene	3,353	31.6	4.60E-04
Intergenic	15,433	43.9	3.46E-06
Intron	390,077	77.4	1.64E-12
Synonymous	245,336	110	5.44E-19
UTR	44,470	119	8.07E-21
Synonymous near splice	4,123	109.6	6.31E-19
Missense	412,904	521.1	1.33E-105
Splice	4,566	1018.2	2.27E-212
Stop gained	10,569	1531.4	4.45E-323

Table S3. Chi-square statistic and test for goodness of fit of the ESP SFS to the SFS of the NR data. Results are presented for pairwise comparisons between each functional category of the ESP data and the NR data. No functional category shows evidence of fitting the NR data. Details regarding the goodness of fit test are presented in Supplementary Note C.

Ancient Demography	Number of free Parameters	Initial Pop Size	Growth Duration (generations)	Final Pop Size (millions)	Growth Rate	log Likelihood
Schaffner *et al.* (2005)	2	*7700*	**276.5 (236.6,320.5)**	**0.81 (0.46,1.8)**	1.68% (1.3,2.2)	-3586.254
Gravel *et al.* (2011)	2	*1032*	**168.7 (144.8,196.7)**	**0.15 (0.1,0.24)**	2.60% (2.1,3.3)	-3597.474
Keinan *et al.* (2007)	2	*10,000*	**112.8 (92.9,136.8)**	**5.2 (0.8,300)**	5.54% (3.2,11.1)	-3595.141

Table S4. Additional two-parameter models based on the NR SFS, varying ancient history scenario. Estimates of the free parameters are shown in bold, and fixed parameters are shown in italicized character and regular font denotes parameters that were derived from other estimated parameters. 95% confidence intervals are also shown in parenthesis. Model I (Table 1) is also given and highlighted in grey for easy comparison.

Ancient demography	Number of free parameters	Initial Pop Size t2	Growth2 Duration (generations)	Final Pop Size (millions)	Growth Rate	log Likelihood
Schaffner *et al.*	3	**7494 (6000,9000)**	**268.6 (228.6,312.5)**	**0.771 (0.4,1.95)**	1.73% (1.3,2.4)	-3586.051
Gravel *et al.*	3	**4126 (3300,5000)**	**128.8 (112.8,148.8)**	**0.471 (0.23,1.26)**	3.68% (2.7,5.1)	-3584.336
Keinan *et al.*	3	**5633 (4400,7100)**	**140.8 (116.8,164.7)**	**0.654 (0.3,2.87)**	3.38% (2.4,5.1)	-3583.975

Table S5. Additional three-parameter models based on the NR SFS, varying ancient history scenario. The fixed and free parameters for the modeling of one recent epoch of growth with three parameters. 95% confidence intervals are also shown in parenthesis. Bold characters are estimated parameters and regular font denotes parameters that were derived from other estimated parameters. For comparison, Model II with three parameters (Table 1) is also given in grey.

Model	Total # grid points	Pop Size t_1	t_1	Pop Size t_2	t_2	Pop Size t_3
Model I	240	NA	NA	NA	5-600 (15)	0.05-30 (16)
Model II	1920	2k-9k (8)	NA	NA	5-600 (15)	0.05-30 (16)
Model III	1920	NA	NA	0.002-0.1 (12)	200-380 (10)	0.005-100 (16)
Model IV	7840	NA	200-600 (7)	0.003-0.1 (10)	50-200 (8)	0.01-100 (14)

Table S6. Grid points used in parameter estimation. The ranges of values used for the four models are listed below. Values in parenthesis indicate the number of points. For each parameter, we used 7-16 grid points with slightly more points for parameters associated with the second growth and fewer points for the first one. For all models, we ran ~200K ms simulations for each of the grid points. In order to validate that 200K ms simulations are sufficient for our estimates, we rerun 10 million ms simulations for a portion of the grid and find similar results.