Neutral genomic regions refine models of recent rapid human population growth

Elodie Gazave[a,1], Li Ma[a,1,2], Diana Chang[a,1], Alex Coventry[b], Feng Gao[a], Donna Muzny[c], Eric Boerwinkle[c], Richard A. Gibbs[c], Charles F. Sing[d], Andrew G. Clark[b], and Alon Keinan[a,3]

Departments of [a]Biological Statistics and Computational Biology and [b]Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853; [c]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030; and [d]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48105

Human populations have experienced dramatic growth since the Neolithic revolution. Recent studies that sequenced a very large number of individuals observed an extreme excess of rare variants and provided clear evidence of recent rapid growth in effective population size, although estimates have varied greatly among studies. All these studies were based on protein-coding genes, in which variants are also impacted by natural selection. In this study, we introduce targeted sequencing data for studying recent human history with minimal confounding by natural selection. We sequenced loci far from genes that meet a wide array of additional criteria such that mutations in these loci are putatively neutral. As population structure also skews allele frequencies, we sequenced 500 individuals of relatively homogeneous ancestry by first analyzing the population structure of 9,716 European Americans. We used very high coverage sequencing to reliably call rare variants and fit an extensive array of models of recent European demographic history to the site frequency spectrum. The best-fit model estimates ~3.4% growth per generation during the last ~140 generations, resulting in a population size increase of two orders of magnitude. This model fits the data very well, largely due to our observation that assumptions of more ancient demography can impact estimates of recent growth. This observation and results also shed light on the discrepancy in demographic estimates among recent studies.

Archeological and historical records reveal that modern human populations have experienced dramatic growth, likely driven by the Neolithic revolution about 10,000 y ago (1, 2). Since then, the worldwide human population size has increased at a fast pace, and faster yet in the last ~2,000 y, giving rise to today's population in excess of 7 billion people (3, 4). A central question in population genetics is how such demographic events affect the effective size (N_e) of populations over time, and as a consequence, how they have shaped extant patterns of genetic variation. [Effective population size, which is typically smaller than the census size, determines the genetic properties of a population (5).] Focusing often on human populations of European descent, estimates of N_e from genetic variation have been traditionally on the order of 10,000 individuals (6–11), although higher and lower estimates have also been obtained (12–16). More recent studies based on sequencing data from a relatively small number of individuals have considered recent population growth in fitting models to the observed site frequency spectrum (SFS) and reported as much as a 0.5% increase in N_e per generation, culminating in a N_e of a few tens of thousands today (13, 14). It has been recently hypothesized that these studies could not capture the full scope of population growth because a larger sample size of individuals is needed to observe single nucleotide variants (SNVs) that arose during the recent epoch of growth (4).

With extreme recent population growth as experienced by human populations, the vast majority of SNVs are expected to be very young and rare, i.e., of very low allele frequency (4). Indeed, several recent sequencing studies with very large numbers of individuals have observed an unprecedented excess in the proportion of rare SNVs (17–19). Fitting models to the SFS, these studies have captured a clearer, more rapid recent population

growth than earlier studies (17–19). At the same time, demographic estimates varied by as much as an order of magnitude between these studies (*SI Appendix*, Table S1).

Not all rare SNVs are as recent as others, and a variant's selective effect plays an important part in its frequency. Purifying (negative) selection acting on deleterious alleles is expected to give rise to an excess of rare variants, which has been demonstrated for human populations (17, 19, 20). Thus, the genetic signature left by purifying selection on the SFS confounds the signature left by recent growth (21). To minimize this confounding effect, recent studies based on protein-coding genes considered for modeling solely synonymous SNVs, which do not modify the amino acid sequence (17–19). However, synonymous mutations have been shown to be targeted by natural selection, e.g., due to their impact on translation efficiency and accuracy, splicing, and folding energy (17, 19, 22–26). Hence, to study recent human genetic history with minimal effects of selection, it is not only desirable to consider accurate sequencing data from a large number of individuals, but also to focus on genomic regions in which mutations are putatively neutral, i.e., not affected by natural selection. Another potential confounder of demographic inference is population structure. Because both

<div style="border:1px solid #000; padding:8px;">

Significance

Recent rapid growth of human populations predicts that a large number of genetic variants in populations today are very rare, i.e., appear in a small number of individuals. This effect is similar to that of purifying selection, which drives deleterious alleles to become rarer. Recent studies of the genetic signature left by rapid growth were confounded by purifying selection since they focused on genes. Here, to study recent human history with minimal confounding by selection, we sequenced and examined genetic variants far from genes. These data point to the human population size growing by about 3.4% per generation over the last 3,000–4,000 y, resulting in a greater than 100-fold increase in population size over that epoch.

</div>

GENETICS

large-scale and fine-scale population structures exist in European populations (27–30), pooling individuals of different European ancestries can lead—when not accounted for in modeling—to biases in the observed SFS and, consequently, in estimates of recent history (*SI Appendix*).

In this study, we aim to capture recent demographic history and estimate the magnitude of the recent growth experienced by humans while limiting the confounding by natural selection and population structure. For this purpose, we selected a small set of genomic regions that are putatively neutral based on a wide array of criteria; the SFS of mutations in these regions is likely to reflect historical changes in N_e rather than selection. We sequenced these regions in a large sample of individuals that share a relatively similar genetic ancestry within Europe. Very deep sequencing coverage allowed us to reliably observe even singletons (SNVs with an allele appearing in a single copy in the sample). Based on this dataset, we explored different models of recent European demographic history. Our best-fit model estimates a growth of 3.4% per generation over the last 141 generations, which is more rapid than estimated in recent large-scale studies (17, 19). We show that differences among previous models (17, 19)—and between these models and ours—can be partially explained by a priori assumptions about more ancient demographic events.

Results

To estimate recent European demographic history with minimal confounding of natural selection, we sequenced genomic regions that we selected such that mutations therein would be as neutral as possible. Thus, the neutral regions dataset (NR) consists of loci that are at least 100,000 bp and at least 0.1 cM away from coding or potentially coding sequences. The loci are also free of known copy number variants, segmental duplications, and loci shown to have been under recent positive selection, and have minimal amounts of conserved and repetitive elements (*Materials and Methods* and *SI Appendix*). We sequenced a final set of 15 loci that met all these criteria, spanning a total of 216,240 bp (*SI Appendix*, Table S2).

In modeling a population's recent demographic history, if the sample of individuals trace their recent ancestry to different populations, even different European populations, the number of singletons and other rare variants can be biased (*SI Appendix*, Fig. S1). To sequence the neutral regions in a sample that minimizes such population structure, we applied principal component analysis (PCA) to whole-genome genotyping data of 9,716 individuals of European ancestry from the Atherosclerosis Risk in Communities (ARIC) cohort (31). This analysis revealed extensive population structure, and we consequently sequenced 500 individuals that form a relatively homogenous cluster (Fig. 1A). Validating the choice of samples by comparing to a diverse panel of European populations from POPRES (32), these 500 individuals indeed show much less heterogeneity than the broader ARIC sample (Fig. 1B) and share a northwestern European ancestry that is similar to the POPRES UK sample (Fig. 1C).

Sequencing was carried out with Illumina HiSeq: 100-bp paired-end for a median average coverage depth of 295X per individual
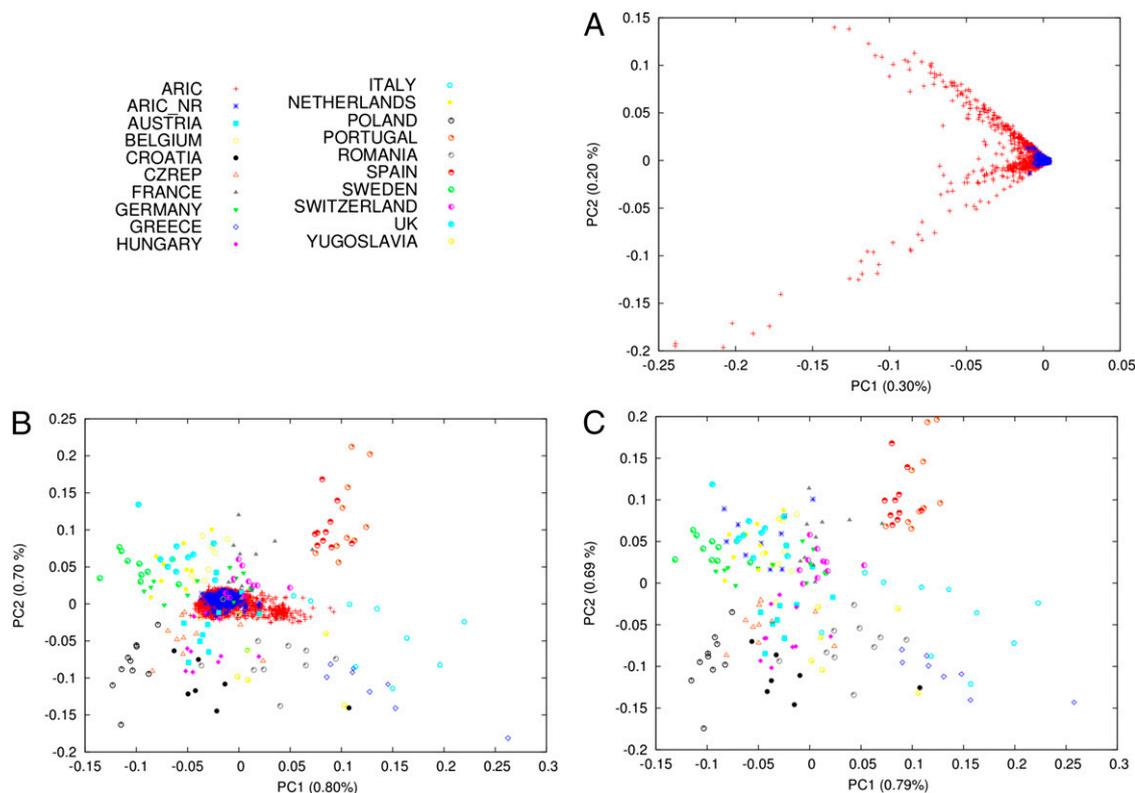
Fig. 1. Genetic ancestry of the NR sequenced individuals. (*A*) PCA of all individuals of European ancestry from the ARIC cohort, with the exception of outlier individuals (*SI Appendix*). Principal components 1 and 2 are plotted and reveal extensive population structure. Individuals chosen for sequencing in the present study are denoted in blue. (*B*) PCA of individuals from the POPRES cohort, with the individuals from *A* subsequently projected onto the resulting principal components (*SI Appendix*). Principal component 1 (*x* axis) appears to capture southern vs. northern European ancestry, whereas principal component 2 (*y* axis) captures western vs. eastern European ancestry, in line with the results of Novembre et al. (27). The full set of individuals from ARIC (*ARIC*) show greater variability, mostly across the first principal component, than the set of individuals sequenced in this study (*ARIC_NR*). (*C*) Same as *B*, except that the PCA includes 10 randomly chosen individuals from the present study rather than projecting all of them as in *B*. These 10 individuals cluster with individuals from the POPRES UK and related populations, which is also the case for other random sets of 10 individuals.

(after filtering of duplicated reads; *SI Appendix*, Fig. S2). We used UnifiedGenotyper (GATK) for variant detection and genotype calling (33, 34), and after applying strict filters to its calls, we obtained 1,834 high quality SNVs (*SI Appendix*). Of these variants, 62.5% have not been reported in dbSNP 135 (novel SNVs). The Ti/Tv (transition/transversion) ratio showed no indication of biases, for both all SNVs (Ti/Tv = 2.22) and for novel SNVs alone (Ti/Tv = 2.29). None of the called SNVs presented significant departure from Hardy–Weinberg equilibrium. We further validated the quality of variant and genotype calling by comparing to those from whole-genome sequencing of the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE-S) project, which overlaps with a few individuals from our NR dataset (35). This validation supports the high quality of the NR dataset due to the very high sequencing coverage (*SI Appendix*). High-quality genotypes for at least 450 individuals were obtained for 95% of SNVs, which are used for presentation of the SFS throughout while probabilistically subsampling 900 random chromosomes for each SNV (*SI Appendix*). We used the full set of SNVs for demographic modeling, where our approach accounts for missing data (*Materials and Methods*).

We compared the SFS from the NR dataset to that from the Exome Sequencing Project (ESP), which is of equivalent deep sequencing (*SI Appendix*, Fig. S3). We randomly subsampled the latter to 900 chromosomes and stratified by functional annotation. The more likely to be functional an annotation is, the more the ESP SFS deviates from that of the neutral regions: the highest agreement is observed for intronic and intergenic annotations from the ESP data, although these are still significantly different from the NR ($P < 10^{-3}$ for a goodness of fit test; *SI Appendix*, Table S3 and Fig. S4). Agreement is much worse between synonymous SNVs and the NR data ($P = 5.4 \times 10^{-19}$) and worst for missense, splice, and nonsense SNVs ($P < 10^{-30}$ for each; *SI Appendix*, Table S3 and Fig. S4). This lack of agreement is due to a higher proportion of very rare variants in the ESP data (*SI Appendix*, Fig. S4), which is consistent with purifying selection playing a larger role in maintaining alleles at lower frequencies in and around genes. We further examined the NR SNVs in comparison with synonymous SNVs via genomic evolutionary rate profiling (GERP) scores (36), showing our data to be much more narrowly distributed around a score of 0, which corresponds to the absence of functional constraint (*SI Appendix*, Fig. S5). These low GERP scores might simply reflect the NR regions being chosen to minimize conserved elements, but applying the same conservation criteria to synonymous SNVs leads to similar

results (*SI Appendix*, Fig. S5). This set of results combined support our unique choice of regions and the relatively neutral nature of SNVs in these regions, as well as further validate our SNV and genotype calling pipeline.

Although the SFS of the NR dataset has a lower proportion of singletons compared with sites under purifying selection, it still exhibits a marked enrichment in this proportion (Fig. 2) compared with the expectation for a population that has remained at constant size throughout history (38.4% vs. 13.6%). This enrichment is consistent with the impact of recent population growth on the distribution of allele frequencies (4, 17–19). However, the SFS predicted by recently published demographic models of European populations that include recent exponential population growth (17, 19) do not closely match the SFS of the NR data (*SI Appendix*, Fig. S6). These models predict a higher proportion of rare variants and a smaller proportion of common variants than we observed, which is consistent with these models being based on synonymous SNVs with increased effect of purifying selection (17, 19). Hence, we next estimated the magnitude and duration of population growth by fitting the SFS of the NR to several different models of recent history.

The first demographic model that we fit to the SFS consists of a recent exponential population growth with two free parameters: the time growth started and the extant N_e, with the growth assumed to continue into the present. More ancient demographic events, before the epoch of growth, including two population bottlenecks, were assumed to follow the model and estimates of ref. 9 (*SI Appendix*, Table S1). The resulting model (model I) estimates the extant population size and, as a consequence, the growth rate, with very large uncertainty, as evident from the 95% CIs (Table 1), and the model does not fit the data very accurately (Fig. 2).

Model I, similar to other recent models of population growth (17–19), assumed more ancient demography as fixed, with the intention of obtaining better resolution for estimating the two parameters of recent history. However, different recent models of population growth have assumed different models of ancient demography (*SI Appendix*, Table S1): Tennessen et al. (19) assumed what had been estimated previously by Gravel et al. (14), Nelson et al. (17), and Coventry et al. (18), what has been estimated by Schaffner et al. (15), and here what has been estimated by Keinan et al. (9). To test how sensitive model I is to the details of assumed ancient history, we repeated fitting model I while assuming each of the above ancient demographic models. Our results point to large differences among the three resulting models, with some parameters for recent demography being as
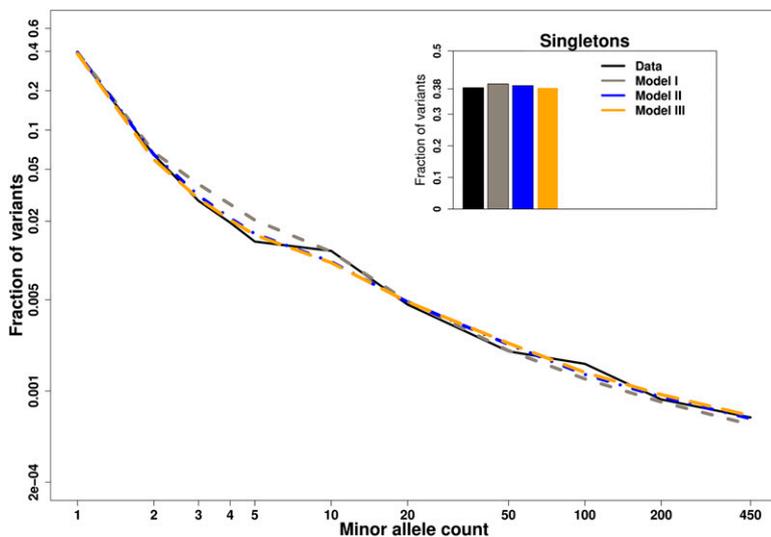
Fig. 2. SFS of the NR dataset and demographic models. *x* axis represents, in log scale, a partition of the number of copies of the minor allele with each number indicating the upper bound of a bin. (Minor allele counts of 1–5 are not binned.) *y* axis presents, in log scale, the proportion of SNVs that fall into each bin. (*Inset*) Fraction of variants that are singletons (*y* axis is presented in linear scale). Data, empirical SFS of NR; model I, a two-parameter model with one epoch of growth, where the duration of growth and final N_e were estimated; model II, an extension of Model I with a third parameter corresponding to N_e before the growth epoch; model III, a model with two separate epochs of growth (Table 1). For clarity, model IV is not presented because its SFS is very similar to that of model III.

Table 1. Four models of recent demographic history and population growth

Model	Number of free parameters	Ne before growth	Duration of earlier growth (generations)	Ne after earlier growth	Duration of recent growth (generations)	Ne after recent growth (millions)	Growth rate during recent growth	Log likelihood
Model I	2	*10,000*	NA	NA	**112.8 (92.9, 136.8)**	5.2 (0.8, 300)	5.54% (3.2, 11.1)	−3,595.141
Model II*	3	**5,633 (4,400, 7,100)**	NA	NA	**140.8 (116.8, 164.7)**	0.654 (0.3, 2.87)	3.38% (2.4, 5.1)	−3,583.975
Model III	3	*5,633*	267.3	**5,362 (3,614, 7,955)**	**132.7 (101.8, 165.5)**	0.73 (0.3, 5.7)	3.70% (2.6, 6.4)	−3,584.178
Model IV	4	*5,633*	**200 (200, 600)**	**5,000 (3,000, 15,000)**	**140 (80, 160)**	0.5 (0.3, 50)	3.84% (0.36, 4.44)	−3,583.501

The table describes the recent history estimated by four models, as well as the goodness of fit of each (log likelihood). Bold values denote the maximum likelihood estimates (and 95% CIs) of the free parameters estimated by each model. Italicized values are assumed as fixed in the model, and regular font denotes parameters that are a direct function of estimated parameters. All four models assume the model of ancient demographic history as estimated in ref. 9, which includes two population bottlenecks (Fig. 3). Results based on other models of ancient demographic history are provided in *SI Appendix*.
*This model is considered as the best-fit model because none of the models with additional parameters provide a significantly improved fit.

much as an order of magnitude different (*SI Appendix*, Table S4). We stress that both data and methodology underlying these inferences are identical and hence conclude that the assumption of ancient demography has a major effect on estimating the timing and magnitude of recent population growth, which explains some of the differences among the recently published models (17–19).

To alleviate some of the sensitivity to ancient demography assumptions, we fit model II that extends model I by adding an additional parameter for the effective population size just before exponential growth. This three-parameter model fits the NR data significantly better than model I ($P = 2.3 \times 10^{-6}$; Fig. 2). It estimates the ancestral N_e before the growth to be 5,633 (CI of 4,400–7,100), markedly lower than the fixed value of 10,000 in model I (Table 1). It estimates growth starting 141 (117–165) generations ago, which is a little earlier than in model I, with a less rapid growth rate of 3.4% (2.4–5.1%) per generation, which culminates in an extant N_e of 0.65 (0.3–2.87) million individuals (Table 1 and Figs. 3 and 4).

To test whether this improved model II can explain the differences between different assumed ancient demographic histories, we repeated fitting its three parameters to the NR similarly to above with model I. When ancient demography is assumed from Gravel et al. (14), this model fits the data significantly better ($P = 2.9 \times 10^{-7}$) than the equivalent of model I with the same ancient demography. Parameter estimates become practically identical to those of the above model II based on Keinan et al. (9) (Fig. 4 and *SI Appendix*, Tables S4 and S5 and Fig. S7). Model II based on Schaffner et al. (15) does not fit the data better than the respective model I ($P = 0.52$) and provides the poorest fit of all three models. One notable difference in the model of Schaffner et al. (15) is that the timing of the second European population bottleneck is assumed as fixed at over twofold that estimated in the other two models (9, 14). The extant N_e is almost identical for all these three models based on

model II, with best-fit estimates varying between 0.47 and 0.77 million individuals (Fig. 4 and *SI Appendix*, Table S5). We conclude that by modeling the epoch before growth, the improved model II is much less sensitive to assumptions about more ancient demography, and it goes a long way in closing the gap between different published models of recent growth (17, 19) and between these and model I.

Archeological and historical records suggest that the growth of the human census population size has accelerated over time (3, 4). Our models thus far considered a single epoch of exponential growth, estimated to have started ~3,500 y ago (assuming 25 y per generation). Hence, we considered several additional models in which growth can span two separate epochs with a different growth rate in each: model III with three parameters, model IV with four parameters, and all other possible four-parameter models (Table 1). None of these more detailed models fit the data better than model II ($P > 0.86$; Table 1 and Fig. 2). They all estimate the second, more recent epoch of growth to be practically identical to the one estimated in model II and the earlier epoch of growth to be equivalent to an epoch of constant population size (Table 1 and *SI Appendix*, Fig. S8).

We investigated whether low statistical power due to a limited amount of data could explain the lack of an acceleration of growth. We simulated a scenario that is identical to model II, except for the addition of an earlier epoch of milder growth, and estimated how often model III provides a significantly better fit than model II (*SI Appendix*). Our modeling has a nonnegligible statistical power of capturing two separate epochs of growth (*SI Appendix*, Fig. S9). Power is 86% when the earlier growth rate is about half that of the more recent epoch. It decreases for milder growth during the first epoch—being as low as 25% for a growth rate of only 0.3% per generation—as well as when growth during the earlier epoch becomes similar to the recent rate (*SI Appendix*, Fig. S9). Overall, power is >60% for detecting two distinct epochs of growth as long as the growth rate during the earlier epoch is in the range of 0.6–1.8% (compared with 3.4% in the recent epoch), which is consistent with archaeological data (*SI Appendix*).

Discussion

Recent studies have provided clear evidence that human populations have experienced recent explosive population growth, although detailed estimates of growth varied greatly (17–19). These were in the context of medical genetic studies, hence based on the sequencing of protein-coding genes. Here we generated a dataset with the sole purpose of accurately capturing rare putatively neutral variants for studying recent human history and population growth. As such, our NR data consists of several characteristics. First, because both demography and natural selection shape the distribution of allele frequencies, loci for the NR data were carefully chosen to minimize the influence of natural selection. Second, because population structure also affects the distribution of allele frequencies, the data consist of

Fig. 3. Schematic representation of the best fit model. N_e is shown in log scale during the last 5,000 generations, with the last 620 generations as estimated by model II (Table 1), and the preceding period following ref. 9.

Gazave et al.

Fig. 4. Log-likelihood surface of model II using three different models of ancient history. (*A*–*C*) Log-likelihood as a function of two parameters—the time growth started and the final N_e—with the third parameter (N_e before the growth, in thousands) fixed on the maximum likelihood estimate. (*D*–*F*) Similarly, with N_e before the growth and final N_e presented and the time of growth fixed. The model was estimated by fitting all three parameters concurrently (*Materials and Methods*). *A* and *D* are for the model based on ancient demography from Keinan et al. (9) (Table 1), *B* and *E* are from Gravel et al. (14), and *C* and *F* are from Schaffner et al. (15) (*SI Appendix*, Table S5). Colored contours are at intervals of 2 log-likelihood. (Note the different scales between panels.) Red crosses denote the maximum-likelihood estimates.

individuals with a homogenous European ancestry, similar to that of individuals from the United Kingdom. Third, because the genetic signature of growth is in rare variants, the data consist of a relatively large sample of 500 individuals. Although some recent studies have considered a larger sample size, this was at the cost of studying a medical cohort with more heterogeneous ancestry. Fourth, to deal with the relatively high error rate of next-generation sequencing, we sequenced the neutral loci in all individuals to a very high coverage, which allows strict filtering and yields a set of very high quality SNVs. These characteristics combine to make the NR dataset ideally suited for population genetic studies of rare variants and recent history.

We used the NR data to consider an array of models of recent human demographic history while showing that our results are consistent with being less confounded by natural selection. The best-fit model points to Europeans having experienced recent growth from an effective population size of about 4,000–7,000 individuals as recently as 120–160 generations (3,000–4,000 y) ago. Growth over the last 3,000–4,000 y is estimated at an average rate of about 2–5% per generation, resulting in an overall increase in effective population size of two orders of magnitude. This model fits the data very well, but only after the realization that assumptions of ancient demography impact the estimate of recent population growth. We hypothesize that this is the case because previous models of ancient demographic history resulted in parameters that confound more recent and more ancient history (37), with the recent growth indirectly affecting them in a manner dependent on sample size. This realization leads to the model we report here fitting much better than previous models of recent growth, and it sheds light on the discrepancies among the latter.

Motivated by archeological evidence of growth starting with the Neolithic revolution ~10,000 y ago and accelerating in the Common Era, we considered models that allow for acceleration of the rate of growth, but none supported such acceleration. One recent model considered two separate epochs of exponential growth (19). However, the first captures a slow recovery from the Eurasian population bottleneck ~23,000 y ago, with a weak growth rate of 0.3% that leads to an N_e of only 9,208, which is similar to the instantaneous recovery from the population bottleneck in other models (14). Thus, to date no recent acceleration in the rate of growth that is along the lines suggested by archeological evidence has been observed in genetic data. Power calculations showed that with our data size and modeling framework, the results do not support a milder growth before it accelerated with >60% certainty. Not capturing two separate

epochs of growth can be due to limited statistical power or overly simplified models. However, another potential explanation is that effective population size increases extremely slowly with the census population size, at least initially. Although several factors contribute to these phenomenon, the particular increase in census population size with the Neolithic revolution has been accompanied by changing social structure that has led to increased variability in reproductive success; the advent of agriculture led to differential accumulation of richness, more notably in males, resulting in differential access to females compared with a hunter-gatherer lifestyle (38). Increased variance in reproductive success results in relatively decreased effective population size. Perhaps jointly with other population processes associated with this social shift, e.g., changing generation time, this can explain either a lack of growth in effective population size initially or a much milder one than in census size. This increased variance, in turn, can lead to our models only capturing the more recent and more rapid growth.

In conclusion, we presented refined models of the recent explosive growth of a European population. These models can inform studies of natural selection (21, 39–41), the architecture of complex diseases, and the methods that should best be used for genotype-phenotype mapping. We hope that these models and the public availability of the NR dataset will facilitate additional such studies. (Data available in dbSNP, with more detailed data at http://keinanlab.cb.bscb.cornell.edu.) However, models of recent demographic history are still limited to Europeans (17–19) and African Americans (19), and there is a need to extend them to additional populations. As the vast majority of rare variants are population specific (24, 42, 43), such studies of additional populations will also facilitate better consideration of the replicability of genome-wide association studies results across populations.

Materials and Methods

Selection of Individuals with Shared European Genetic Ancestry. PCA was run on 9,716 European Americans from the ARIC cohort (31), using EIGENSOFT (44) on whole-genome genotyping data from the Affymetrix 6.0 genotyping array (dbGaP accession phs000090.v1.p1). Outliers of inferred non-European ancestry were removed, in addition to regions of extended linkage disequilibrium such as inversions (*SI Appendix*). A total of 500 individuals that were densely clustered together based on the first four principal components (PCs) were then chosen for sequencing. We tested for plate effect, which showed no correlation with the localization of the individuals on these PCs. We validated the ancestry of the 500 individuals by merging with data from the Affymetrix 500k genotyping array of 10 individuals from each POPRES population (32) and repeating PCA (*SI Appendix*).

GENETICS

Choice of Target Putatively Neutral Regions. To minimize the effect of selection, we considered genomic regions located at least 100,000 bp and 0.1 cM from any coding or potentially coding loci. We excluded genomic sequences containing segmental duplications and known copy number variants (*SI Appendix*), as well as regions under recent positive selection (45). Among contiguous genomic regions of at least 100 kb that satisfy these criteria, we then ranked targets for sequencing by their content of conserved and repetitive elements and removed CpG islands. The resulting NR dataset spans a total of 216,240 bp across 15 regions, each between 5,340 to 20,000 bp long (*SI Appendix*, Table 2). These and additional criteria for selecting regions are implemented in the Neutral Regions Explorer web-server at http://nre.cb.bscb.cornell.edu (46).

Sequencing, Mapping, and Variant Calling. Illumina HiSEq. 2000 with 100-bp paired-end reads was used for sequencing. Reads were mapped to hg18 human reference genome using Burrows-Wheeler Aligner (BWA) (47) (*SI Appendix*). For each individual, aligned reads were subjected to duplicate removal using Picard v.1.66 (http://picard.sourceforge.net). Subsequent SNV calling, quality control filtering, and genotype calling were performed with the Genome Analysis ToolKit, GATK-1.5–31 (33, 34), as detailed in *SI Appendix*. Analyses are based on 493 individuals that were successfully sequenced.

Demographic Inference. To obtain estimates of recent demographic history, we calculated the likelihood of the observed SFS conditioned on several demographic models. To reduce parameter space, we fixed ancient history as estimated by previous studies (9, 14, 15) and only estimated parameters of more recent history. These models have either a single epoch of recent growth or an additional, earlier epoch of growth. Models include different combinations of the following parameters: the time recent growth started, final N_e after recent growth, N_e before growth, start time of earlier growth, and N_e after earlier growth (Table 1). We tested whether a model provides a better fit than another using Vuong's test (48). For each model, we estimated the SFS at different grid points using ms (49), with each grid point being a particular combination of parameter values. We then calculated the composite likelihood of the model following the approach of ref. 9, as the probability of the observed minor allele (the less common of the two alleles) counts of all SNVs while accounting for missing data (*SI Appendix*). We profiled the likelihood surface using for each parameter 7–16 predefined grid points that span a range of plausible values. We increased the number of grid points for each parameter 10-fold by fitting a smooth spline function for the proportion of SNVs of each allele count as a function of all parameters, which improved accuracy (*SI Appendix*, Fig. S10), with only a minor increase in computational burden. Two-sided 95% CIs for each parameter was estimated following a $\chi^2(1)$ distribution that accounts for variation across SNVs (*SI Appendix*).

1. Cohen JE (1996) *How Many People Can the Earth Support?* (W. W. Norton & Company, New York), 1st Ed.
2. Roberts L (2011) 9 billion? *Science* 333(6042):540–543.
3. Haub C (1995) How many people have ever lived on earth? *Popul Today* 23(2):4–5.
4. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336(6082):740–743.
5. Hartl D, Clark A (2007) *Principles of Population Genetics* (Sinauer, Sunderland, MA).
6. Erhlich HA, Bergström TF, Stoneking M, Gyllensten U (1996) HLA sequence polymorphism and the origin of humans. *Science* 274(5292):1552b–1554b.
7. Garrigan D, Hammer MF (2006) Reconstructing human origins in the genomic era. *Nat Rev Genet* 7(9):669–680.
8. Harding RM, et al. (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet* 60(4):772–789.
9. 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.
10. Takahata N (1993) Allelic genealogy and human evolution. *Mol Biol Evol* 10(1):2–22.
11. Yu N, et al. (2001) Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Mol Biol Evol* 18(2):214–222.
12. Tenesa A, et al. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Res* 17(4):520–526.
13. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5(10):e1000695.
14. Gravel S, et al.; 1000 Genomes Project (2011) Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA* 108(29):11983–11988.
15. Schaffner SF, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15(11):1576–1583.
16. Melé M, et al.; Genographic Consortium (2012) Recombination gives a new insight in the effective population size and the history of the old world human populations. *Mol Biol Evol* 29(1):25–30.
17. 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.
18. Coventry A, et al. (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* 1:131.
19. Tennessen JA, et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64–69.
20. Kiezun A, et al. (2012) Exome sequencing and the genetic basis of complex traits. *Nat Genet* 44(6):623–630.
21. Gazave E, Chang D, Clark AG, Keinan A (2013) Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. *Genetics* 195(3):969–978.
22. Chamary JV, Hurst LD (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* 6(9):R75.
23. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* 7(2):98–108.
24. Fu W, et al.; NHLBI Exome Sequencing Project (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493(7431):216–220.
25. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. *Nat Genet* 40(3):340–345.
26. Waldman YY, Tuller T, Keinan A, Ruppin E (2011) Selection for translation efficiency on synonymous polymorphisms in recent human evolution. *Genome Biol Evol* 3:749–761.
27. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456(7218):98–101.
28. Humphreys K, et al. (2011) The genetic structure of the Swedish population. *PLoS ONE* 6(8):e22547.
29. Price AL, et al. (2009) The impact of divergence time on the nature of population structure: An example from Iceland. *PLoS Genet* 5(6):e1000505.
30. Di Gaetano C, et al. (2012) An overview of the genetic structure within the Italian population from genome-wide data. *PLoS ONE* 7(9):e43759.
31. The ARIC Investigators (1989) The Atherosclerosis Risk in Communities (ARIC) Study: Design and objectives. The ARIC investigators. *Am J Epidemiol* 129(4):687–702.
32. 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.
33. 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.
34. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–498.
35. Morrison AC, et al.; Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) Consortium (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet* 45(8):899–901.
36. Cooper GM, et al.; NISC Comparative Sequencing Program (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15(7):901–913.
37. Myers S, Fefferman C, Patterson N (2008) Can one learn history from the allelic spectrum? *Theor Popul Biol* 73(3):342–348.
38. Betzig L (2012) Means, variances, and ranges in reproductive success: Comparative evidence. *Evol Hum Behav* 33(4):309–317.
39. Boyko AR, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 4(5):e1000083.
40. Yu F, et al. (2009) Detecting natural selection by empirical comparison to random regions of the genome. *Hum Mol Genet* 18(24):4853–4867.
41. Ayodo G, et al. (2007) Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am J Hum Genet* 81(2):234–242.
42. Altshuler DM, et al.; International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311):52–58.
43. Abecasis GR, et al.; 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.
44. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909.
45. Akey JM (2009) Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res* 19(5):711–722.
46. Arbiza L, Zhong E, Keinan A (2012) NRE: A tool for exploring neutral loci in the human genome. *BMC Bioinformatics* 13:301.
47. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
48. Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57(2):307–333.
49. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.