Demographic history and rare allele sharing among human populations


©Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305-5120; dDepartment of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721; eDepartment of Biology, Boston College, Chestnut Hill, MA 02467; hDepartment of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853; and 1Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030

Edited by Michael Lynch, Indiana University, Bloomington, IN, and approved June 3, 2011 (received for review December 24, 2010)

High-throughput sequencing technology enables population-level surveys of human genomic variation. Here, we examine the joint allele frequency distributions across continental human populations and present an approach for combining complementary aspects of whole-genome, low-coverage data and targeted high-coverage data. We apply this approach to data generated by the pilot phase of the Thousand Genomes Project, including whole-genome 2–4x coverage data for 179 samples from HapMap European, Asian, and African panels as well as high-coverage target sequencing of the exons of 800 genes from 697 individuals in seven populations. We use the site frequency spectra obtained from these data to infer demographic parameters for an Out-of-Africa model for populations of African, European, and Asian descent and to predict, by a jackknife-based approach, the amount of genetic diversity that will be discovered as sample sizes are increased. We predict that the number of discovered nonsynonymous coding variants will reach 100,000 in each population after ~1,000 sequenced chromosomes per population, whereas ~2,500 chromosomes will be needed for the same number of synonymous variants. Beyond this point, the number of segregating sites in the European and Asian panel populations is expected to overcome that of the African panel because of faster recent population growth. Overall, we find that the majority of human genomic variable sites are rare and exhibit little sharing among diverged populations. Our results emphasize that replication of disease association for specific rare genetic variants across diverged populations must overcome both reduced statistical power because of rarity and higher population divergence.

demographic inference | genetic drift | population genetics | human evolution

The Thousand Genomes Project (1000G) is the most extensive study to date of human genomic diversity (1). The pilot phase of the project consisted of whole-genome, low-coverage sequencing of 179 samples from four HapMap populations at 2–4x coverage, an exon pilot experiment that targeted exons from over 800 genes in 697 samples across seven HapMap populations with ~50x coverage, and a trio pilot focusing on two mother–father–child trios (1). In this article, we present an approach for combining the low-coverage and the exon pilot data, and use it to estimate the joint allele frequency spectrum for individuals of European origin in Utah (CEU), Han Chinese individuals in Beijing (CHB), Japanese individuals in Tokyo (JPT), and Yoruba individuals in Ibadan, Nigeria (YRI). Our motivation for this analysis is that two pilot projects provide complementary information: the low-coverage pilot captures most of the common variation in the populations sequenced across the accessible human genome at the cost of missing some of the rarer variants, whereas the target capture data provide a more complete picture of rare variants on an interesting subset of the data. In this article, we are interested in leveraging the strengths of the exon and low-coverage pilots to obtain accurate estimates of population genetic parameters. We will focus in particular on the P-population site frequency spectrum Φ, a P-dimensional histogram that records the joint distribution of diallelic SNPs as displayed in Fig. 1.

More specifically, the value Φ(f1, f2, …, fP) of bin (f1, f2, …, fP) is the number of SNPs that occurs in f1 chromosomes from population 1, f2 chromosomes from population 2, etc. Because the allele frequency in diploid population i ranges from 0 to 2ni, where ni is the number of individuals sequenced in this population, Φ is a (2ni + 1) × (2ni + 1) × … × (2ni + 1) array. Because the number of individuals who are successfully sequenced at any given site may vary, ni is, in practice, chosen to be somewhat smaller than the total number of individual sequenced, and each site with n > ni sequenced individuals contributes to bin f in proportion to the probability that one finds derived alleles in a random selection of ni of the n samples.

The one-population site frequency spectrum (SFS) is a staple of population genetics and is commonly used to reveal broad patterns of selection (2, 3) and demography (3–5). The multiple-population SFS has received increased attention recently (6–10), because it provides additional information about between-population structure. Many standard population genetic statistics, such as FST and Tajima’s D, are summaries of the multiple-population SFS.

In this article, we study SFSs derived from the 1000G pilot project data. We develop methods to precisely estimate SFSs from high-throughput sequencing data and use this information to estimate demographic parameters for a detailed Out-of-Africa demographic model by using dadie (6), a software package that uses diffusion approximation to calculate expected SFSs across multiple populations (6, 10). We use these parameters to predict the number of variants to be discovered as the number of sequenced samples in the 1000G is increased. We also present a jackknife-based approach to the prediction of the number of undiscovered variants and compare the predictions of the two approaches.

Theory

Linear Error Model for SFSs from Low-Coverage Data. Data of the kind collected by the 1000G low-coverage pilot (1) (2–4x coverage across 179 individuals) provide a large volume of data from which precise demographic inference can be drawn. However, the low coverage leads to biases that must be addressed to ensure accuracy of the inference (11, 12). We use an empirical approach to tune an error model for low-coverage sequencing based on a direct comparison of the SFSs generated by whole genome and capture experiments on the part of the genome sequenced by both experiments.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence should be addressed. E-mail: cdbustam@stanford.edu.

This article contains supporting information online at www.pnas.orglookup/suppl/doi:10.1073/pnas.1019276108/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1019276108

PNAS | July 19, 2011 | vol. 108 | no. 29 | 11983–11988
The usefulness of this approach relies on two observations. First, demographic inference does not require the knowledge of the particular sites that are variable, but rather requires statistical averages over all sites. Although it is impossible to infer which variable sites were missed, the average number of missed sites can be estimated directly. Second, the most significant bias caused by low coverage in the 1000G data is an elevated false-negative rate for rare variant genotype calls (1). Because the majority of genetic variants are rare, it is possible to infer error rates for such variants based on high-quality sequence data from a relatively small subset of the genome.

Because the SFS does not keep track of linkage information, we use an error model that acts independently on each genomic site. We suppose that the underlying true SFS $S_{ij}^p$ and the observed SFS $S_i^p$ for population $p$ are related by a linear error model:

$$S_i^p = \sum_j \sum_k A_{ijk}^p S_{jk}^p, \quad \sum_j \sum_k A_{ijk}^p = 1.$$  \[1\]

If we have $N_c$ frequency bins per population, the number of parameters in this model is $N_c^3 - N_c^2$. We therefore need to introduce additional simplifying assumptions (justifications are provided below):

i) The errors occur independently in each population:

$$A_{ijk}^p = A_{ij}^p A_{jk}^p A_{ik}^p.$$

ii) The probability $\varepsilon_i^p$ of missing a site decays exponentially with the number of variants present in the population:

$$\varepsilon_i^p = \left\{ \begin{array}{ll} 0 & \text{if } i' = 0 \\ \alpha^p e^{-r^p i'} & \text{otherwise} \end{array} \right.$$

iii) If a site is found to be variable in one population, its frequency is estimated accurately:

$$A_{ik}^p = \left\{ \begin{array}{ll} 1 - \varepsilon_i^p & \text{if } i = i' \\ \varepsilon_i^p & \text{if } i = 0, i'\neq0 \\ 0 & \text{otherwise} \end{array} \right.$$

The resulting model has six parameters (amplitudes $\alpha^p$ and error decay rates $r^p$) and captures the bulk of the discrepancy between the high- and low-coverage data. It is motivated by the following observations:

i) The low-coverage SNP calls were made population by population. We assume, first, that the leading source of error is an insufficient number of variant reads to confidently call a variant and, second, that in a low-coverage experiment, the uncorrelated sampling fluctuations in read numbers play the largest role in the variation in read numbers.

ii) Variant calls in 1000G require multiple independent observations of a variant across a population to rule out read errors and call a variant genotype. This stringency strongly reduces the rate of false-positive calls, but it results in missing actual variants at a rate that depends on the expected number of nonreference reads observed at a given position across a population (1). The decay in the probability of detecting less than a fixed, small number $c$ of reads for a variant present in $i$ of $N$ chromosome sequenced at depth $d$ is dominated by $(1-i/N)^{2d/2}e^{-d/2}$. Therefore we fit a heuristic exponential error model $\alpha^p e^{-r^p}$, where the effective read depth $2\beta$ accounts for fluctuations in read depth and read quality across the genome (Table 1).

iii) After enough variant reads are found in a 1000G population to justify a variant call, a single variant read is sufficient to make a genotype variant call, hence a reduced false-negative rate and low systematic bias in estimated frequencies when a variant has been identified.

Estimating the Parameters of the Error Model. Because errors are assumed to occur independently in each population, the error rates can be inferred directly from error rates in single-population SFSs. We compute the single-population SFSs at sites that are found to be variable in the exon pilot data and with at least 80 genotype calls in all three populations. The exon pilot and low-coverage pilot are then compared, and the optimal parameters $\alpha^p$ and $r^p$ are obtained through a linear fit using the first three frequency bins of the compared spectra.

Note that this error model can be inverted to give a correction model for the SFS, which does not require the knowledge of the number of fixed variants (SI Appendix). However, the correction model may involve the subtraction of large numbers and has non-Poisson uncertainties. When inferring demographic parameters by maximum likelihood of a Poisson Random Field (6), we therefore incorporate the error model in our demographic model rather than attempt to correct the SFS.

Prediction of the Rate of Variant Discovery. One practical use of inferring a demographic model is the ability to predict the number of variants that will be discovered in subsequent experiments. To study the impact of model choice on such predictions, we propose an alternate predictor of discovery rate based on sampling theory and inspired by an analogy with capture–recapture approaches to estimating animal population sizes (13–15) (let us consider rabbits for definiteness). In this analogy, a rabbit is akin to a SNP, a field trip is akin to an individual sequenced, and a rabbit capture is akin to the identification of a variant in a sequenced individual. In the absence of measurement errors, the probability of identifying a variant in a randomly chosen sequenced individual is proportional to the frequency of the variant in the population. This distribution of probabilities is akin to the variability in rabbit capture probability; a common SNP is akin to a trap-happy rabbit, and a rare SNP is akin to a trap-shy rabbit.

We propose a population genetics analog of the Burnham–Overton jackknife (16, 17) to estimate the total number $V(N)$ of segregating sites in a sample of $N$ chromosomes based on a subsample of $n$-sequenced chromosomes. This jackknife estimator uses the assumption that

$$V(N) = V(n) + \sum_{i=1}^n a_i^p \Delta(N,n),$$  \[2\]

where $\Delta(N,n) = \sum_{j=1}^n 1/j$ for a fixed jackknife order $p$. Explicit expressions for the $a_i^p$ as well as performance benchmarking.
and additional discussion of this estimator are provided in SI Appendix.

Results

After filtering away the exon pilot calls based on less than 15x coverage and individuals with substantial discrepancy with HapMap (SI Appendix), we compared the joint SFSs with the expected spectra obtained if each individual had been assigned to a population randomly in the independent sites model. That is, given an $N \times N$ spectrum $\phi(i,j)$, we have an expected spectrum of

$$\phi^*(i,j) = \frac{N}{\binom{N}{i+j}} \sum_{i'+j' \geq 0} \phi(i', j').$$

[Figs. 2 and 3 indicate that, even for pairs of closely related populations, we find a substantial reduction in allele sharing for rare variants compared with a single randomly mixing population. In particular, Fig. 2, Right shows the Anscombe residuals between expectation and data. Blue strips along the axis correspond to a significant excess of variants private to one panel in the data, and they are accompanied by a reduction in shared variants (red). The residuals are larger (darker colors) for rare variants not only because of a larger number of sites but also because of reduced sharing. Indeed, we see in Fig. 3 that the amount of sharing, expressed as a proportion of the expectation in a panmictic population, is only a few percent between continental populations for variants present at 2% minor allele frequency (MAF) and about 60% for variants at 20% MAF. More closely related populations, such as CHB and JPT, still exhibit a 50% reduction in sharing at 2% MAF but barely any reduction for variants at 20% MAF. Interestingly, even closely related populations, such as CHB and CHD, exhibit a 20% reduction in sharing for 2% MAF. This finding is consistent with recent population structure, but although this analysis used only genotype calls with high-coverage data, such a reduction in sharing could also be partly explained by differences in the sequencing platform between the two populations.

To increase the number of sites available for estimating joint SFS, we turned to the low-coverage pilot data. Direct comparison of low-coverage and exon capture genotype calls at sites called in the exon capture pilot shows a significant discrepancy for rare variants (SI Appendix, Figs. S1–S3) because of elevated rates of false-negative variant calls in low-coverage data. This finding results in biased estimates of the distribution of allele frequencies.

The bulk of the systematic discrepancy between high- and low-coverage SNP calls could be described using the simple false-negative model described above (Table 1 and SI Appendix, Figs. S1–S4). The most substantial discrepancy between this model and the data is in the CHB + JPT, possibly because this group was the metapopulation with the lowest coverage. In this case, the high-coverage singleton counts are 634% higher than the uncorrected low-coverage counts. After error correction, a discrepancy of 19% remains, with the corrected low-coverage site predicting more counts (SI Appendix, Fig. S3). Despite the high false-negative rate for singleton calls in the low-coverage data, its sheer volume provides an advantage in estimation precision over the much smaller exon pilot dataset. Similarly, the false-negative model for multiple-population SFSs (Eq. 1) was found to account for the bulk of the discrepancy between the multiple-population SFS derived from low- and high-coverage SNP calls (SI Appendix, Fig. S4).

We modeled the joint SFS for synonymous sites in African (YRI), Asian (CHB and JPT), and European (CEU) data sequenced in the low-coverage pilot using the 13-parameter demographic model used in ref. 6 (Fig. 4 and Table 2), taking into account the expected error model. The SFS was calculated using $n = 40$ samples per panel (80 chromosomes). We obtained maximum composite likelihood estimates for the 13 parameters using $\text{ddal},$ a diffusion-approximation-based package for estimating expected SFSs resulting from various demographic models (6) (Table 2). Our maximum likelihood parameters are broadly consistent with previously reported values using National Institute on Environmental Health Sciences (NIEHS) data (6). However, the resulting confidence intervals, determined by conventional bootstrap (likelihood profiles are provided in SI Appendix), are substantially narrower than those intervals resulting from NIEHS or high-coverage data alone. As an example, using a 25 y generation time, we find a time of split between African and Eurasian populations of $T_h = 51$ thousand years ago (kya; 95% confidence

![Fig. 2. Joint allele SFSs (all sites) for selected pairs of populations from the exome sequencing panel (Left) compared with expected spectra under site by site population label permutation (Center). Shown are sites occurring in, at most, 15 of 100 chromosomes. All population pairs, including two different panels of Chinese individuals sampled in Beijing, China and Denver, Colorado (CHB and CHD), as well as two groups of European origin (CEU and Tuscans from Italy, or TSI), show substantial residuals for rare variants (Right), consistent with reduced sharing. White bins contain less than one count.](Image)

### Table 1. Parameter values in the error model (Eq. 1)

<table>
<thead>
<tr>
<th>Population</th>
<th>$\alpha^\phi$</th>
<th>$\beta^\phi$</th>
<th>Mean mapped depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEU</td>
<td>1.466</td>
<td>0.737</td>
<td>4.62x</td>
</tr>
<tr>
<td>CHB + JPT</td>
<td>1.855</td>
<td>0.754</td>
<td>2.65x</td>
</tr>
<tr>
<td>YRI</td>
<td>1.220</td>
<td>0.551</td>
<td>3.42x</td>
</tr>
</tbody>
</table>

Note that the parameter values differ substantially between populations and that error rates decrease with increased coverage.

Gravel et al.

PNAS | July 19, 2011 | vol. 108 | no. 29 | 11985
We also show how sequencing a large number of individuals at low coverage is an efficient strategy not only for discovering the maximum number of variable sites but also for estimating demographic parameters, at least when error rates can be estimated. Different statistical methods have been proposed that include read depth information and models of sequencing errors to reduce biases in allele frequency estimation (11, 12). Because of the availability of high-coverage data for a subset of the genome in the populations studied here, we used direct comparison with high-coverage data to estimate and correct biases caused by low coverage. A significant advantage of the direct comparison approach is simplicity and computational efficiency; it can use existing curated genotype calls rather than require a full analysis of an error model at the individual read level. This advantage is particularly useful for data generated by 1000G, because multiple sequencing platforms and calling pipelines with different error modes have been used jointly. In general, the two approaches are not mutually exclusive, and when practical, a statistically corrected low-coverage SFS could be further corrected by comparison with targeted high-coverage data. Here, we used, as a reference, an exon capture dataset with >50x coverage and validation rates of 96.8% overall and 93.8% for singletons. We also restricted our analysis to a high-quality subset of the data (by selecting individuals with good coverage and HapMap concordance and selecting sites with sufficient coverage). The false-negative rate in the exon capture data was estimated to be below 5% for variants of at least 1% in frequency and 26% for variants below 1% in frequency. To avoid resulting biases in the frequency-dependent false-negative estimates for the low-coverage data, we restricted the comparison with sites where a high-coverage variant call had been made.

We found that the bulk of the discrepancy between high- and low-coverage data could be described by a simple model that uses only two parameters per population, and in which error rates decay exponentially with MAF, frequencies of detected variant sites are accurately determined, and errors occur independently in each population. The latter assumption is perhaps the most debatable: we expect at least some correlations in the coverage at a given site for different populations. An error model taking into account such correlations would, therefore, be desirable. However, the NIEHS data (6) resulted in a maximum likelihood estimate of \( T_B = 140 \) kya (95% confidence interval = 40–270 kya). The inference based on the exon pilot alone yields \( T_B = 98 \) kya (95% confidence interval = 43–210 kya).

In general, the gain in precision was strongest for the parameters involved in more ancient events. Inference based on uncorrected low-coverage data yielded an unrealistic \( T_B = 14 \) kya split.

Beyond their fundamental interest as descriptors of human history, these parameters allow for a number of experimental predictions; given a demographic model, we can predict, for example, the number of synonymous variants to be discovered in samples of larger size that are currently in the process of being sequenced. We predicted the number of variants to be discovered in each of the three population considered (CEU, CHB + JPT, and YRI) as the sample size is increased using both the inferred demographic model and the jackknife estimator of the number of undiscovered variants presented in Methods (Fig. 5). Because the jackknife does not rely on assumptions about demography and selection, we also used it to predict the number of non-synonymous sites to be discovered. The jackknife approach predicts that, as sample size is increased, the total number of segregating sites in CEU and CHB + JPT panels should overtake the number of segregating sites in the YRI population.

**Discussion**

Our results illustrate that the vast majority of human variable sites are rare and that the majority of rare variants exhibit, at most, very little sharing among continental populations. We also find reduced sharing for rare variants compared with common variants among more closely related populations, such as CHB-JPT, CEU-TSI, and CHB-CHD. This lack of sharing can be explained by population divergence, and we expect that the fraction of newly discovered variable sites that are population-specific will keep increasing with sample size. This finding poses a formidable challenge for the reproduction of genome-wide association studies for rare functional variants across diverse populations, because the statistical difficulties caused by variant rarity within a population combine with increased between-population divergence.

interval = 45–69 kya). By contrast, the NIEHS data (6) resulted in a maximum likelihood estimate of \( T_B = 140 \) kya (95% confidence interval = 40–270 kya). The inference based on the exon pilot alone yields \( T_B = 98 \) kya (95% confidence interval = 43–210 kya).

In general, the gain in precision was strongest for the parameters involved in more ancient events. Inference based on uncorrected low-coverage data yielded an unrealistic \( T_B = 14 \) kya split.

Beyond their fundamental interest as descriptors of human history, these parameters allow for a number of experimental predictions; given a demographic model, we can predict, for example, the number of synonymous variants to be discovered in samples of larger size that are currently in the process of being sequenced. We predicted the number of variants to be discovered in each of the three population considered (CEU, CHB + JPT, and YRI) as the sample size is increased using both the inferred demographic model and the jackknife estimator of the number of undiscovered variants presented in Methods (Fig. 5). Because the jackknife does not rely on assumptions about demography and selection, we also used it to predict the number of non-synonymous sites to be discovered. The jackknife approach predicts that, as sample size is increased, the total number of segregating sites in CEU and CHB + JPT panels should overtake the number of segregating sites in the YRI population.

**Discussion**

Our results illustrate that the vast majority of human variable sites are rare and that the majority of rare variants exhibit, at most, very little sharing among continental populations. We also find reduced sharing for rare variants compared with common variants among more closely related populations, such as CHB-JPT, CEU-TSI, and CHB-CHD. This lack of sharing can be explained by population divergence, and we expect that the fraction of newly discovered variable sites that are population-specific will keep increasing with sample size. This finding poses a formidable challenge for the reproduction of genome-wide association studies for rare functional variants across diverse populations, because the statistical difficulties caused by variant rarity within a population combine with increased between-population divergence.

interval = 45–69 kya). By contrast, the NIEHS data (6) resulted in a maximum likelihood estimate of \( T_B = 140 \) kya (95% confidence interval = 40–270 kya). The inference based on the exon pilot alone yields \( T_B = 98 \) kya (95% confidence interval = 43–210 kya).

In general, the gain in precision was strongest for the parameters involved in more ancient events. Inference based on uncorrected low-coverage data yielded an unrealistic \( T_B = 14 \) kya split.

Beyond their fundamental interest as descriptors of human history, these parameters allow for a number of experimental predictions; given a demographic model, we can predict, for example, the number of synonymous variants to be discovered in samples of larger size that are currently in the process of being sequenced. We predicted the number of variants to be discovered in each of the three population considered (CEU, CHB + JPT, and YRI) as the sample size is increased using both the inferred demographic model and the jackknife estimator of the number of undiscovered variants presented in Methods (Fig. 5). Because the jackknife does not rely on assumptions about demography and selection, we also used it to predict the number of non-synonymous sites to be discovered. The jackknife approach predicts that, as sample size is increased, the total number of segregating sites in CEU and CHB + JPT panels should overtake the number of segregating sites in the YRI population.

**Discussion**

Our results illustrate that the vast majority of human variable sites are rare and that the majority of rare variants exhibit, at most, very little sharing among continental populations. We also find reduced sharing for rare variants compared with common variants among more closely related populations, such as CHB-JPT, CEU-TSI, and CHB-CHD. This lack of sharing can be explained by population divergence, and we expect that the fraction of newly discovered variable sites that are population-specific will keep increasing with sample size. This finding poses a formidable challenge for the reproduction of genome-wide association studies for rare functional variants across diverse populations, because the statistical difficulties caused by variant rarity within a population combine with increased between-population divergence.
ever, given the limited data available to infer the parameters of the error model, the independence assumption is a reasonable tradeoff that allows for the capture of the bulk of the error patterns. Finally, the error rates likely differ between different genomic regions (such as coding vs. noncoding DNA), motivating our focus on exonic regions where high-coverage data were available. This finding emphasizes the importance of obtaining high-quality genotype data through sequencing or chip genotyping for representative noncoding regions.

The demographic model discussed in this paper was introduced in Gutenkunst et al. (6), where it was used to analyze the NIEHS intergenic data. Despite differences in putatively neutral sites (selected intergenics vs. synonymous), sequencing technology (Sanger vs. high throughput), and panel choice (CHB only vs. CHB + JPT), the inferred parameters are in broad agreement (Table 2). Inference based only on capture data provides overlapping 95% confidence intervals, with the single exception of Europe–Asia migration rate (1.8 $\times$ 3.9 $\times$ 10$^{-5}$ vs. 4.1 $\times$ 8.2 $\times$ 10$^{-5}$).

The main difference between these three sets of parameter estimates is the width of the confidence intervals. The inference based on exon capture data provides reduced uncertainty compared with the NIEHS data, despite a comparable number of variable sites in the SFS; the additional number of samples per site results in more accurate frequency estimates that further constrain the demographic model. A much greater reduction in the confidence intervals is obtained by considering the low-coverage and exon capture data jointly (a 90% reduction of the confidence interval for the Out-of-Africa split time compared with a 27% reduction with the exon data only). Our estimate of the Out-of-Africa split time using the low-coverage data, 51 kya, is also in better agreement with both prior genetic and archaeological estimates of the modern human expansion out of Africa (18). It should be emphasized that, because we use a single Western African population as our African panel, the divergence described by our model might have occurred earlier than the actual Out-of-Africa event.

The narrow confidence intervals on some of the parameters should not obscure the fact that the parameter estimates are model-dependent. As a simple example, a model that does not allow for migration would require more recent split times to produce similar levels of population divergence. The demographic history of the four populations considered is much more eventful than what is accounted for by our model. Additional geographically intermediate populations from the Near East and Central Asia that were not included in our analysis might contribute significantly to the allele frequency distribution as ghost populations (19). Incorporating an appropriate number of source populations for estimates of migration has been a general limitation of two- and three-population models under isolation migration coalescent, approximate Bayesian computation, and diffusion-based approaches. This limitation might explain why our estimate of the divergence between East Asians and Europeans is more recent than estimates based on archaeological evidence (18), but is comparable with estimates of 23 kya (20) under an approximate Bayesian computation approach and 25 kya under an isolation migration approach with mtDNA X and Y sequence data (21).

Similarly, the current population sizes inferred from our model (15,500, 35,900, and 49,000 for YRI, CEU, and CHB,
respectively) are still significantly lower than census sizes. Because our model accounts for some population size changes, these are expected to be in closer relationship to census sizes compared with the classical effective population size, but additional model refinement [such as structure within populations, generation overlap, and a recent increase in growth rate, which was observed in the work by Coventry et al. (22), in a sample of 10,422 European-Americans] will be needed to close the gap.

Predictions based on the demographic model and the jackknife approach differ as to the number of new variants to be discovered, particularly for CHB + JPT (Fig. 5). This difference is easily understood by considering the differences in the two approaches. The demographic model attempts to fit the complete SFS at the cost of model assumptions that might bias the results. By contrast, the jackknife approach focuses on the rare variants, and the model assumptions are weaker. The difference can be traced to the fact that the maximum likelihood demographic model predicts a number of singletons somewhat lower than the observed number (SI Appendix, Fig. S5). If this discrepancy is due to limitations in the model that fail to account for an excess of rare variants, we expect the jackknife estimator to be more accurate. By contrast, if the difference is because of inaccurate singleton frequency estimation (from sequencing errors leading to 6.2% of false-negative variants in the high-coverage data) or limitations of our correction model (SI Appendix, Figs. S1–S3), the demographic model is expected to provide more robust estimates.

Nonetheless, both methods predict at least 50,000 synonymous variants in the human genome when sequencing 1,000 individuals for the CEU and CHB populations, substantially more than would be predicted from population genetic models of constant size. The jackknife approach applied directly to the seven target capture populations shows similar patterns, with some variation within continents, in JPT samples showing less rare variants than the Chinese populations, and in TSI samples showing more rare variants than CEU (SI Appendix, Fig. S7). These results highlight the importance, for the planning of medical sequencing experiments, of accurate demographic models of human populations and the dramatic impact that recent human population growth has had on the structure of genetic variation. Specifically, our prediction that most genetic variants are rare and highly diverged suggests that genome-wide association studies aiming to correlate common disease susceptibility with rare variants may need extraordinarily large sample sizes and precise definitions of population samples to accurately compare frequencies in cases and controls. Eventually, a clear tradeoff will ensue between cataloging variants and genotyping vs. completely sequencing human genomes and comparing them among populations of cases and controls.

Methods

Numerics. The unprecedented size of the 1,000 genomes data created challenges for the numerical solution of the diffusion equation. Namely, the number of grid points required to accurately estimate the three population SFS grows rapidly with the number of samples in each population. We optimized $\lambda_{\text{salt}}$ and released version 1.5.0, in which the number of grid points necessary to achieve a given accuracy is reduced. As in ref. 6, we obtained SFSs with three different grid sizes (60, 70, and 80) and extrapolated to infinite grid size. Each likelihood evaluation took between 1 and 2 min on a 2.26-GHz processor. Optimization required hundreds to thousands of likelihood evaluations. Likelihoods were computed using the folded SFS to avoid biases caused by ancestral misidentification. Convergence of the maximum likelihood optimization process was ensured by restarting the search with modified initial conditions. The maximal likelihood parameters were chosen, but differences in parameter estimates from the different restarts were, on average, much smaller than the reported confidence intervals.

Conversion from Genetic to Physical Units. The different parameters involved in the diffusion equation solved by $\lambda_{\text{salt}}$ are normalized by the ancestral population size $N_a$ during the likelihood maximization. The optimal value of $N_a$ is calculated using the fact that the total number of segregating sites in a sample of $n$ individuals is proportional to $N_a \mu$, where $\mu$ is the mutation rate and $L$ is the effective length sequenced. For this analysis, we used $L = 5,007,837$, the number of autosomal fourfold degenerate sites that passed quality control in all three populations, and $\mu = 2.36 \times 10^{-9}$. For estimates based on exon data alone, we fixed the effective length sequencing to 68% of the target length by requesting equal values for $N_a$ in the corrected low-coverage and exon pilot estimates. The remaining 32% is composed of called sites that failed quality controls and sites for which no genotype call has been made. When performing bootstrap analysis, the total number of fourfold degenerate sites varied from bootstrap sample to bootstrap sample and was adjusted accordingly. Finally, to convert generation time to years, we used a generation time of 25 y. Estimated parameters are shown in Table 2.

Supporting information for *Demographic history and rare allele sharing among human populations.*

Simon Gravel, Brenna M. Henn, Ryan N. Gutenkunst, Amit R. Indap, Gabor T. Marth, Andrew G. Clark, The 1000 Genomes consortium, Carlos D. Bustamante

I. ESTIMATING PARAMETER CONFIDENCE INTERVALS VIA BOOTSTRAP

In order to obtain confidence intervals on the inferred demographic parameters, we used a bootstrap approach. Since linkage effectively reduces the number of independent measurements, we re-sampled genomic regions rather than individual SNPs. More specifically, we considered regions with a coding sequence (CDS) annotation in Gencode [1] version 3b, and merged regions with 50kbp of each other. We were then left with 9328 distinct and well-separated regions. We resampled these with replacement, forming a set of overlapping genomic regions. For each of these bootstrapped sets, we extracted SNPs generated from both the high- and low-coverage data. We obtained inferred values for both the error rates due to low coverage and the resulting inferred demographic parameters.

II. SAMPLE SELECTION

Short read sequencing results in high site-to-site and sample-to-sample variability in coverage beyond the independent draw. Poisson sampling [Marth et al., in submission]. Since we are interested in obtaining high quality reads from the capture data, we discarded individuals for which the median coverage per exon was below 15. We also discarded individuals with unusually high discrepancy with validation or HapMap data. As a result, we removed samples [NA18524, NA18529, NA18530, NA18531, NA18543, NA18557, NA18599, NA18615, NA18620, NA18627, NA18628, NA18635, NA18642, NA18749, NA18773] (15 samples) from CHB, [NA06985, NA06994, NA11840, NA11995, NA12004, NA12006, NA12156, NA12414, NA12763, NA12815, NA12829, NA12842, NA12872, NA12873, NA12874, NA12889] (16 samples) from CHD, and [NA19046, NA19307, NA19310, NA19312, NA19317, NA19321, NA19441, NA19453, NA19456] (9 samples) from LWK.

III. ERROR CORRECTION

The matrix $A$ describing the error model can be inverted to give a correction model

$$S^0_{ijk} = \sum_{i',j',k'} (A^{-1})_{ijk;i'j'k'} S_{i'j'k'},$$

with

$$(A^{-1})_{ijk;i'j'k'} = \begin{cases} \frac{1}{1-p} & \text{if } i = i' \\ \frac{-e_i}{1-p} & \text{if } i = 0, i' \neq 0 \\ 0 & \text{otherwise.} \end{cases} \quad (S1)$$

Note that $A_{ij'k'}^{-1} = 0$ for $i' = 0, i \neq 0$, implying that it is not necessary to know the number of observed invariant sites to estimate the corrected SFS for variable sites. The same holds true for $A$, obtaining $S^0_{ijk}$ for $(i, j, k) \neq (0, 0, 0)$ does not require the knowledge of $S_{0,0,0}$.

In practice, this correction model should be used with caution; in some cases, estimated entries in the corrected SFS result from a difference of two large numbers, and can even be negative. The Poisson random field approximation, used in the maximum likelihood calculation, would not hold for such cases. However, the corrected model is a sensible approach for calculating basic population demographic estimates (such as $F_{ST}$) that are not based on likelihood estimates.

Comparison of exon data, low-coverage data, and this error-correction model are shown on Figures S1, S2, S3.

IV. COMPARISON BETWEEN MODEL AND DATA FOR MARGINAL FREQUENCY SPECTRA.

We compare in Figure S5 the model predicted and marginal site frequency spectra for the CEU, CHB+JPT, and YRI panels.
counts in target region

counts in target region

counts in target region

400
400
400
800
800
800
600
600
600
200
200
200

observed after

quencies.

between the exon and low coverage data across the population frequencies.

are polarized using the March 2006 assembly of the chimp genome (PanTro2). The corrected SFS captures the bulk of the differences between the exon and low coverage data across the population frequencies.

FIG. S1. Site frequency spectra for the YRI panel for sites with variant calls in the exon pilot. Shown are the exon pilot data, the low-coverage data, and a corrected SFS using the exponentially decaying error model, for derived allele frequency below 25%. All spectra are polarized using the March 2006 assembly of the chimp genome (PanTro2). The corrected SFS captures the bulk of the differences between the exon and low coverage data across the population frequencies.

FIG. S2. Site frequency spectra for the CEU panel for sites with variant calls in the exon pilot. Shown are the exon pilot data, the low-coverage data, and a corrected SFS using the exponentially decaying error model, for derived allele frequency below 25%. All spectra are polarized using the March 2006 assembly of the chimp genome (PanTro2). The corrected SFS captures the bulk of the differences between the exon and low coverage data across the population frequencies.

FIG. S3. Site frequency spectra for the CHB and JPT panels for sites with variant calls in the exon pilot. Shown are the exon pilot data, the low-coverage data, and a corrected SFS using the exponentially decaying error model, for derived allele frequency below 25%. All spectra are polarized using the March 2006 assembly of the chimp genome (PanTro2). The corrected SFS captures the bulk of the differences between the exon and low coverage data across the population frequencies.

\[ \hat{V}(\infty) = V(n) + \sum_{i=1}^{p} \frac{a_{i}^{p} \Delta^{i}(N, n)}{n^{i}} \]

The Burnham-Overton jackknife estimator has been shown to perform well in a variety of situations [3, 4], even though it has been shown that, since an arbitrary number of very rarely observed objects might eventually be discovered, no point estimator can be generally unbiased in such infinite extrapolation [5, 6].

The genetics context requires a more modest extrapolation to a finite number of observations. In this case, the potential problem of an infinite number of rarely observed objects is eliminated: based on the random sampling assumption, we know that the rate of new discoveries per sample cannot increase with the number of samples. The number of discoveries can therefore be bounded with good confidence above and below. This does not guarantee the existence of an unbiased point estimator, but supports the idea that a jackknife estimator for finite extrapolation has the potential to be more reliable than one for infinite extrapolation.

With this in mind, we write the expansion

\[ \hat{V}_{n}(N) = V(n) + \sum_{i=1}^{p} a_{i}^{p} \Delta^{i}(N, n) \]  

where \( \Delta(N, n) = \sum_{j=m}^{N-1} 1/j \) for a fixed jackknife order \( p \). This assumption provides the expected behavior as \( n \to N \), and is exact for \( p \geq 1 \) in the case of a constant-size, neutrally evolving population. The parameters \( a_{i}^{p} \) are calculated by equating estimators \( \hat{V}_{n}(N) = \hat{V}_{n-1}(N) = \cdots = \hat{V}_{n-p}(N) \),

V. JACKKNIFE EXPANSION

The Burnham-Overton jackknife was introduced to predict the number of unobserved objects (typically, animals) based on a finite number of mark-and-recapture field trips. The basic assumption of this model is a postulated relation between the total number of objects \( V(\infty) \), the number \( V(n) \) of objects observed after \( n \) field trips, and \( n \) itself:
and solving the resulting $p$ linear equations. Explicit expressions are given below.

Figure S6 shows that, based on subsampled simulated data from our demographic model, our jackknife estimator could predict with good accuracy the number of segregating sites in the full data.

A. Parameters of the jackknife expansion

The jackknife expansion parameters for the number of undiscovered variants in a sample of size $N$, based on the site frequency spectrum $\phi(i)$ from a sample of size $n$, are easily derived using symbolic software using $V_n(N) = V(n) + \sum_{j=1}^{p} a_i^p \Delta^i(N,n)$, together with the equality $\hat{V}_n(N) = \hat{V}_{n-1}(N) = \cdots = \hat{V}_{n-p}(N)$, and

$$\sum_{j=1}^{k} \binom{k}{j} \Phi(j) = V(n) - V(n-k)$$

(S4)

$$= \sum_{j=1}^{p} a_i^p \left( \Delta^j(N,n) - \Delta^j(N,n) \right).$$

The resulting equations for the number of missed variants $m_p$ are bulky, but are simple rational expressions of $n$, $N$, and the $\phi(i)$. The predicted number $m_p$ of missing variants at second and third order are

$$m_2 = \frac{(-1 + 5n - 6n^2 + 2n^3)\Delta(N,n) + (-1 + n)(2 - 3n + n^2)\Delta(N,n)^2}{n(3 - 5n + 2n^2)} \phi(1) + \frac{(-2 - 2 + n)^2\Delta(N,n) - 2(-2 + n)(2 - 3n + n^2)\Delta(N,n)^2}{n(3 - 5n + 2n^2)} \phi(2)$$

(S5)

at second order, and

$$m_3 = \frac{127 - 501n + 885n^2 - 837n^3 + 432n^4 - 114n^5 + 12n^6}{n(-165 + 521n - 637n^2 + 377n^3 - 108n^4 + 12n^5)} \Delta(N,n) \phi(1) + \frac{2(2 - 2 + n)\Delta(N,n)^2 (-5 - 51n + 81n^2 - 39n^3 + 6n^4 + 2(-2 + n)^2 (3 - 4n + n^2)\Delta(N,n))}{n(165 - 356n + 281n^2 - 96n^3 + 12n^4)} \phi(1) + \frac{2(-1 + n)^2n(165 - 356n + 281n^2 - 96n^3 + 12n^4)}{(-1 + n)^2n(165 - 356n + 281n^2 - 96n^3 + 12n^4)} \Delta(N,n) \phi(2) + \frac{2(-560 + 1130n - 731n^2 + 55n^3 + 127n^4 - 51n^5 + 6n^6)\Delta(N,n)^2}{n(-165 + 521n - 637n^2 + 377n^3 - 108n^4 + 12n^5)} \phi(2) - \frac{4(-2 + n)^2 (-39 + 52n - 22n^2 + 3n^3)\Delta(N,n)^3}{n(165 - 356n + 281n^2 - 96n^3 + 12n^4)} \phi(2) + \frac{6(-3 + n)^3\Delta(N,n) (-3 + 2n + (5 - 8n + 3n^2)\Delta(N,n))}{(-1 + n)^2n(-55 + 82n - 39n^2 + 6n^3)} \phi(3) + \frac{6(-3 + n)^3(-2 + n)\Delta(N,n)^3}{n(-55 + 82n - 39n^2 + 6n^3)} \phi(3) \quad \text{(S6)}$$

at third order.

A limit of interest occurs when $n \gg 1$, when the estimates simplify to

$$m_1 = \Delta(N,n) \phi(1)$$

$$m_2 = \left( \Delta(N,n) + \frac{\Delta^2(N,n)}{2} \right) \phi(1) - \frac{\Delta^2(N,n)}{2} \phi(2)$$

$$m_3 = \left( \Delta(N,n) + \frac{\Delta^2(N,n)}{2} + \frac{\Delta^3(N,n)}{6} \right) \phi(1) - \left( \Delta^3(N,n) + \Delta^2(N,n) \right) \phi(2) + \Delta^3(N,n) \phi(3).$$

Note that the number of terms of the site frequency spectrum that are used in the estimate is equal to the order of the jackknife expansion, and the magnitude of the coefficient of each SFS term increases with the perturbation order, in a way that may diverge for sufficiently large $\Delta$. As a result, for large values of $\Delta$, the jackknife expansion becomes unstable and sensitive to noise: the choice of the jackknife expansion order is therefore always a compromise.
B. Jackknife estimates for all exon pilot populations

In figure S7, we show the jackknife projected number of sites to be discovered in all 7 populations from the exon pilot project. In Figure S6, we compare the results of the jackknife estimator applied to the expected SFS resulting from sampling 80 individuals from each of the three panels in our analysis.

VI. LIKELIHOOD PROFILES AND BOOTSTRAP DISTRIBUTION

Figure S8 provides a comparison of bootstrap estimates to likelihood profiling. We provide the results in genetic units, as likelihood profiling requires holding parameters constant while optimizing other parameters, a task much simplified if carried in the units used in the simulation. Given $N_a$ and generation time $g$ (see Methods), the transformation from genetic to physical units is

$$T_{Eu,As} = 2N_ag\tau_{Eu,As}$$
$$T_B = 2N_ag(\tau_{Eu,As} + \tau_B)$$
$$T_{Af} = 2N_ag(\tau_{Eu,As} + \tau_B + \tau_{Af})$$
$$N_{Af} = N_a\nu_{Af}$$
$$N_B = N_a\nu_B$$
$$N_{EU0} = N_a\nu_{EU0}$$
$$N_{AS0} = N_a\nu_{AS0}$$
$$r_{AS} = (\nu_{AS}/\nu_{AS0})^{g/T_{Eu,As}} - 1$$
$$r_{EU} = (\nu_{EU}/\nu_{EU0})^{g/T_{Eu,As}} - 1$$
$$m_i = M_i/2N_a.$$  \hspace{1cm} (S8)

As can be seen in Figure S8, likelihood profiles are in qualitative agreement with the bootstrap estimates. Bootstrap 95% confidence interval correspond roughly to the range 10-20 log-likelihood units (in base $e$). Note that the profiles display composite likelihoods since linkage between neighboring sites in the Poisson Random Field approach creates correlations between neighboring sites. As a result, we expect that likelihood-based confidence intervals provide an underestimate of the true variance, better captured by the bootstrap analysis.

FIG. S4. Two-dimensional marginal frequency spectra for (a) the low-coverage pilot, (b) the exon pilot, and (d) the exon pilot once the error model has been applied. (c) shows the Anscombe residuals between the two pilots, whereas (e) shows the residuals after the error model has been applied to the exon pilot. All spectra are polarized using the March 2006 assembly of the chimp genome (PanTro2).
FIG. S5. Comparison of model predictions, based on our maximum-likelihood parameters, and data from 4-fold synonymous sites polarized using chimp as an outgroup. The discrepancy for very common variants is due to cases where the chimp allele is not the ancestral allele [2].

FIG. S6. To control for bias of the jackknife estimator, we used our demographic model to generate "True" SFS for 1000 chromosomes in our 3 population panels. We then calculated average jackknife projections based on "observed" subsamples of 80 chromosomes for each population. In the absence of sampling, we find little bias for the first decade of extrapolation.

FIG. S7. Observations and third-order jackknife predictions of the number of variants to be discovered in the exon capture panels as sample size is increased, based on 80 chromosomes per population.
FIG. S8. Bootstrap and likelihood profiles for the Out-Of-Africa model for parameters expressed in genetic units. Likelihood profiles are obtained in the usual way by fixing one parameter and optimizing the likelihoods over the other parameters.