Estimation of demography and mutation rates from one million haploid genomes

Authors

Joshua G. Schraiber, Jeffrey P. Spence, Michael D. Edge

Correspondence

schraibe@usc.edu (J.G.S.), edgem@usc.edu (M.D.E.)

Samples of millions of genomes provide substantial information about recent demography and mutation, but standard population-genetic methods make assumptions not met in these data. We introduce DR EVIL to address this challenge, providing high-resolution estimates of mutation and demography and exploring consequences for selection.

Schraiber et al., 2025, The American Journal of Human Genetics *112*, 2152–2166
September 4, 2025
https://doi.org/10.1016/j.ajhg.2025.07.008

CellPress

Estimation of demography and mutation rates from one million haploid genomes

Joshua G. Schraiber,[1,*] Jeffrey P. Spence,[2] and Michael D. Edge[1,*]

Summary

As genetic sequencing costs have plummeted, datasets with sizes previously unthinkable have begun to appear. Such datasets present opportunities to learn about evolutionary history, particularly via rare alleles that record the very recent past. However, beyond the computational challenges inherent in the analysis of many large-scale datasets, large population-genetic datasets present theoretical problems. In particular, the majority of population-genetic tools require the assumption that each mutant allele in the sample is the result of a single mutation (the "infinite-sites" assumption), which is violated in large samples. Here, we present DR EVIL, a method for estimating mutation rates and recent demographic history from very large samples. DR EVIL avoids the infinite-sites assumption by using a diffusion approximation to a branching-process model with recurrent mutation. This approach results in tractable likelihoods that are accurate for rare alleles. We show that DR EVIL performs well in simulations and apply it to rare-variant data from one million haploid samples. We identify mutation-rate heterogeneity even after accounting for trinucleotide context and methylation status. We also predict that at modern sample sizes, the alleles at most polymorphic sites with high mutation rates represent the descendants of multiple mutation events.

Introduction

We now have datasets with genome or exome sequences from hundreds of thousands of individuals.[1–8] Large genomic datasets have the potential to shed light on the evolutionary forces that shape genetic variation by enabling a high-resolution view of recent demographic history,[9–12] mutation rates,[1,2,4,13–17] and natural selection.[4,18–20] Understanding these evolutionary forces is crucial for the interpretation of whole-genome sequencing data.

From a population-genetic perspective, the key advantage of ultra-large sequencing datasets is the information they provide about rare variants. Because the age of a variant is correlated with its frequency,[21–25] rare variants likely arose recently, and the frequencies of rare variants are informative about recent population history. This is especially relevant in humans, where recent explosive population growth has resulted in a large excess of rare variants.[9]

In addition, rare variants can provide substantial power to estimate mutation rates. Although average mutation rates can be estimated from *de novo* mutations in trios or large pedigrees,[26–28] there is substantial variation in mutation rate throughout the genome.[13,17,29–31] Many genomic features, such as flanking nucleotide sequence,[13,32] methylation level,[32] and replication timing,[31] are known to affect mutation rate. The many rare variants found in samples of hundreds of thousands of haplotypes provide a large pool of data from which to estimate mutation rates that depend on these features.[17,30] Moreover, because the full set of factors that affect mutation rates is unknown, there may be residual heterogeneity in mutation rates that can be identified in large datasets, potentially helping discover additional factors that influence mutation rates.

Natural selection prevents most deleterious variants from rising to high frequency. Thus, rare variants are enriched for deleterious mutations. This fact is commonly invoked in criteria to predict variant pathogenicity[33–37] or in efforts to enrich for variants that are more likely to play a causal role in complex traits and disease.[7,38–40] However, the interaction of mutation, selection, and demography means that simple rules for identifying putatively pathogenic variants, such as allele-frequency cutoffs, are not robust to variation in mutation rate.[31] Moreover, recent work showed that the distribution of allele frequencies, not merely the presence or absence of alleles, contributes substantially to improving estimates of selection.[19] It is therefore important to build models of rare variation that can account for natural selection and recurrent mutation to improve our understanding of selection in the human genome.

Despite the importance of rare variation in large datasets, most current population-genetic methods are unable to access the information contained in such variants. One of the key assumptions undergirding many population-genetic methods is that all copies of a given allele share a single mutational origin, known as the infinite-sites assumption.[41–43] In contrast, recurrent mutation, in which variants of a given type have multiple mutational origins, is detectable even in datasets with only tens of thousands of individuals.[1,28,44] In addition, although

[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA; [2]Department of Genetics, Stanford University, Stanford, CA, USA
*Correspondence: schraibe@usc.edu (J.G.S.), edgem@usc.edu (M.D.E.)
https://doi.org/10.1016/j.ajhg.2025.07.008.

there exist fast algorithms for computing the likelihood of genetic data under the infinite-sites model in the absence of natural selection,[45–48] computing likelihoods for variants subject to natural selection remains challenging.[47,49–51] Although simulation-based approaches, such as approximate Bayesian computation[52] or supervised machine learning using simulations,[53] can circumvent some of these limitations, the ways in which these approaches use the input data can be difficult to interpret, and the computational burden associated with simulation can make exploration of different models infeasible.

There has also been theoretical work about large samples, but again assuming the infinite-sites model. For example, for sample sizes that are bigger than the effective population size, early work showed that it may be possible to estimate mutation rates separately from the effective population size, which is not possible when the sample size is much smaller than the effective population size.[54] Moreover, the site frequency spectrum for large samples differs depending on whether one assumes a diffusion model (equivalently Kingman's coalescent) or the discrete-time Wright-Fisher model, even assuming the infinite-sites model.[55–57]

Nonetheless, recent approaches have made progress toward modeling rare variants subject to recurrent mutation[58] and selection[59,60] using coalescent and diffusion techniques, clarifying how recurrent mutation and rapid population growth shape the distribution of allele frequencies. Yet these approaches are not designed to estimate mutation rates and demography jointly, leaving a methodological gap.

To enable population-genetic inference from rare variation in ultra-large sequencing datasets, we present DR EVIL (Diffusion for Rare Elements in Variation Inventories that are Large). The core of our method is a rare-variant approximation of the usual diffusion approximation used in population genetics that incorporates both recurrent mutation and selection. We show that the resulting process falls into a model class for which a solution was recently found. We use this solution to arrive at approximate likelihoods for counts of rare alleles, which can then be optimized to estimate mutation and demographic parameters.

We compare the performance of our approach for estimation of mutation rates with existing methods and show that our method is more accurate and can correct for the presence of mutation-rate heterogeneity. Furthermore, we demonstrate the importance of correctly accounting for recurrent mutation when estimating recent demographic history.

We then apply our method to one million samples from gnomAD.[2] We detect mutation-rate heterogeneity that remains even after accounting for methylation status and trinucleotide context. Finally, we demonstrate the importance of accurate modeling of rare variation by exploring the impact of natural selection on the probability of observed allele counts as a function of mutation rate and the number of distinct mutations to which the copies of a rare allele in a very large sample trace their origin. ’

Methods

Approximate sampling formula for rare alleles

We use a standard Wright-Fisher model of allele-frequency dynamics in a population of time-varying size. Specifically, we assume mutations arise with rate μ per site per generation, the fitness of the heterozygote is $1 + hs$, and that the effective population size at time t is given by a function $N(t)$. Importantly, we do not make the infinite-sites assumption that each site is only mutated once. Instead, we allow for recurrent mutation.

To model the site frequency spectrum of a large sample, we require the probabilities $p_{n,k}$ that an allele in a sample of n haploid genomes will be observed k times. To obtain these probabilities for rare alleles (i.e., those for which $k \ll n$), we make two approximations in addition to the standard Wright-Fisher diffusion approximation. First, we approximate the typical Wright-Fisher diffusion using Feller's diffusion approximation of the branching process, sometimes called Feller's continuous-state branching process.[61] Branching-process approximations to the dynamics of rare alleles have a long history in population genetics.[62,63] Intuitively, the branching process assumes that the allele is sufficiently rare that it is only found in heterozygous individuals, and additionally that heterozygous individuals are rare enough that they are unlikely to mate with other heterozygotes. Thus, they can be modeled as expanding in an unconstrained population (Figure 1, top left). Next, we approximate the binomial sampling of alleles in a finite sample with Poisson sampling.[64] Here, we rely on the fact that a binomial sample with a small probability of success in a large number of trials is approximated well by a Poisson distribution. Defining the relative population size $\rho(t) = N(t)/N_0$ for some arbitrary N_0, the population-scaled mutation rate $\theta = 4N_0\mu$, and the population-scaled heterozygous selection coefficient $\gamma = 4N_0hs$, we show in the supplemental information that the sampling probabilities satisfy a system of differential equations, with

$$\frac{d}{dt}p_k(t) = (f + (k-1)b)p_{k-1}(t) - (f + k(b+d))p_k(t) + (k+1)dp_{k+1}(t), \quad \text{(Equation 1)}$$

where, in the language of birth-death processes, $f = \frac{n\theta}{2}$ is the immigration rate, $b = \frac{n}{2\rho(t)}$ is the birth rate, and $d = \frac{n}{2\rho(t)} - \frac{\gamma}{2}$ is the death rate. Equation 1 can then be recognized as an inhomogeneous linear birth-death process with immigration. Although birth-death processes with immigration have been studied for many years,[65] to our knowledge the first analytic solution with inhomogeneous coefficients was presented recently.[66] In the supplemental information, we show that this solution can be written as

$$p_k(t) = e^{-\xi_0}\frac{B_k(\xi_1, \xi_2, \ldots, \xi_k)}{k!}, \quad \text{(Equation 2)}$$

where $B_k(x_1, \ldots, x_k)$ is a complete Bell polynomial,[67,68] and

$$\xi_i = \begin{cases} \int_0^t f(1 - \alpha(t;s))ds, & \text{if } i = 0 \\ i!\int_0^t f q_i(t;s)ds, & \text{if } i > 0 \end{cases},$$

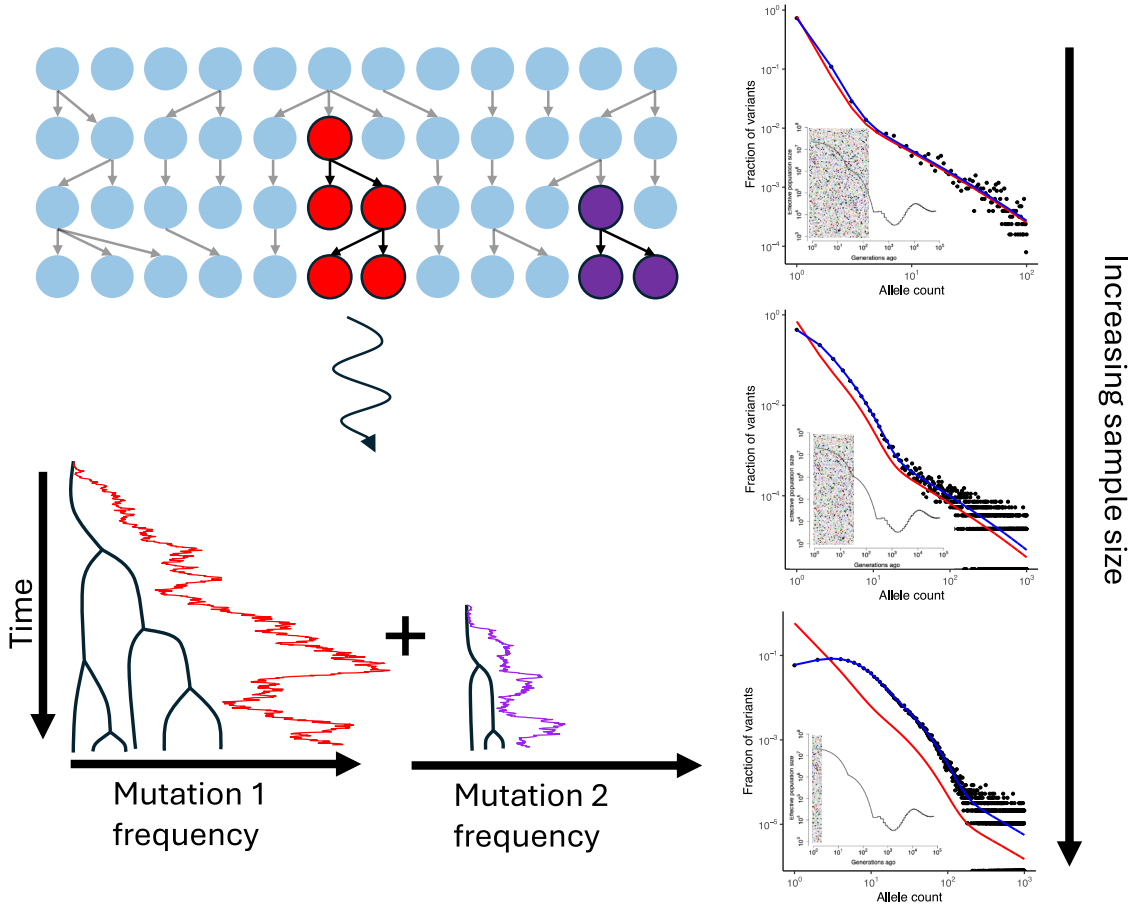

Figure 1. Modeling rare variation and recurrent mutation
On the left side, we consider a discrete population model in which a rare variant expands into a population that is dominated by another allele (indicated in blue). Two lineages of the rare variant arise, one shown in red and one shown in purple. We make a diffusion approximation to this process, as indicated by the curved arrow. In the diffusion approximation, we find that the genealogy of a sample of rare alleles is modeled by a birth-death process, and that the sampling distribution can be obtained by adding the sampling distribution of the two different origins. On the right side, we demonstrate how simulated data (shown as filled circles) departs from an infinite-sites model (shown in red) as sample sizes increase. Our method (shown in blue) accurately matches the site frequency spectrum by modeling recurrent mutation. In each inset, we qualitatively indicate that as sample sizes become increasingly large, we reveal more recent demographic history.

where $\alpha(t;s)$ is the probability that a birth-death process (without immigration) started with one copy at time s is not extinct at time t, and $q_i(t;s)$ is the probability that a birth-death process (without immigration) started with one copy at time s is at i copies at time t. Explicit formulae for $\alpha(t;s)$ and $q_i(t;s)$ are given in the supplemental information. This formula can be interpreted as the sum of birth-death genealogies arising from each mutational origin (Figure 1, bottom left) and is explored further in the supplemental information. We also describe a reparameterization used to improve computational efficiency and numerical stability in the supplemental information. In simulations, we find that our analytic model matches simulated site frequency spectra, whereas classic approaches using the infinite-sites model fail as sample sizes become larger (Figure 1, right side).

Likelihood

A key application of sampling formulae in population genetics is to construct a likelihood of the observed data for estimating evolutionary parameters. Here, we are focused on estimating demographic parameters and mutation rates, leaving inference of natural selection to future work.

The data are the observed allele counts from a sample of haploid size n. Because mutation rates vary by genomic context,[13,30] we assume the data are stratified into J mutational contexts, where $c_{j,k}$ is the number of sites of context j observed k times. We then compute a sampling probability for each context, $p_{j,k}$, by modifying Equation 2 to have a different θ_j for each context, although demographic history is shared across contexts. Finally, because our approximate sampling formula is only valid for rare alleles, we set a maximum allele count $K \leq n$ and renormalize the sampling probability, $\tilde{p}_{j,k} = \frac{p_{j,k}}{\sum_{k=0}^{K} p_{j,k}}$. Then, the log likelihood is

$$\log(L) = \sum_{j=1}^{J} \sum_{k=0}^{K} c_{j,k} \log\left(\tilde{p}_{j,k}\right). \qquad \text{(Equation 3)}$$

In practice, we choose $K = \min(0.01n, 1{,}000)$, which we find to be a good combination of accuracy and computational speed for larger samples. Because we ignore correlated allele frequencies

among sites, this is a composite likelihood. Nonetheless, for rare variants, linkage disequilibrium can be low.[69] In the supplemental information, we prove that this likelihood is valid even when there are more than two alleles segregating at a single locus, as long as all alleles but one are rare. Loosely, this is because in the branching-process approximation the rare alleles do not interfere with each other, just as the different mutational origins of a single rare allele do not interfere with each other

Demographic history estimation

To obtain maximum-likelihood estimates of mutation rates, we use the fact that Equation 3 factorizes by context when the demographic history is fixed. Thus, we can estimate the θ_c for each context independently, which is accomplished by using the one-dimensional Brent optimization algorithm[70] for each context.

To estimate demographic parameters, we leveraged the fact that we only model very rare alleles and thus only have the ability to make precise population-size estimates at very recent times. We estimate demographic parameters on the diffusion timescale (i.e., time measured in units of $2N_0$ generations and population sizes measured relative to N_0) by first setting a cutoff time in the past, more anciently than which we assume the population size was constant. We then specify a number of pieces more recent than that cutoff in which to estimate the effective population size and space them evenly on a log scale.

To jointly estimate demographic history and mutation rates, we take a multi-step approach. In the first step, we optimize demographic parameters using the L-BFGS-B algorithm[71] with numerical gradients. During each evaluation of the likelihood, we set θ_c for each context to be equal to the method-of-moments estimator, described in detail below. This ensures that, given the updated demographic history, the mutation-rate parameters are near their maximum-likelihood estimates. Because similar likelihoods are known to have complicated geometries,[72] we initialize optimization from multiple random locations in demographic parameter space.

Following this initial likelihood optimization, the parameters are near the maximum-likelihood estimates, but not at them, because the method-of-moments estimate of the mutation rate generally differs from the maximum-likelihood estimate. Hence, we perform two additional rounds of coordinate ascent, in which we alternately fix demographic parameters and find the maximum-likelihood estimates of the mutation rate given the demography, and subsequently fix mutation rates and estimate demographic parameters given the mutation rates.

Estimating a distribution of mutation rates for each context

We fix the demographic parameters at those estimated assuming a single mutation rate per context and fit a gamma distribution of mutation rates to each context in turn. For each context, we compute the likelihood-ratio test statistic $\Lambda = 2\log(L_{\text{heterogeneity}}/L_{\text{no heterogeneity}})$. Standard theory shows that under the null hypothesis of no heterogeneity, the likelihood-ratio statistic is distributed as $\chi^2(1)$.[73]

Estimating sequencing error rates

We fix the demographic parameters at those estimated assuming a single mutation rate per context and fit a sequencing error rate to each context in turn. For each context, we compute the likelihood-ratio-test statistic $\Lambda = 2\log(L_{\text{error}}/L_{\text{no error}})$. Standard theory shows that under the null hypothesis of no heterogeneity, the likelihood-ratio statistic is distributed as $\chi^2(1)$.[73]

Rescaling parameters

To scale the parameters, which are inferred on the diffusion timescale (i.e., time measured in units of $2N_0$ generations and population sizes measured relative to N_0), we match the average mutation rate in our dataset to the average mutation rate for the same sites in the gnomAD dataset. That is, we compute $\overline{\theta} = \frac{1}{n}\sum_c n_c\theta_c$, where n_c is the number of sites in context c, θ_c is the estimated $4N_0\mu_c$ for context c, and n is the total number of sites, as well as $\overline{\mu} = \frac{1}{n}\sum_c n_c\mu_c$, where μ_c is the gnomAD mutation rate for context c; we then compute $N_0 = \frac{\overline{\theta}}{4\overline{\mu}}$. This value of N_0 can then be used to rescale from the diffusion scale to an interpretable scale.

Simulations

We developed a custom Wright-Fisher model simulator to simulate unlinked variants evolving subject to genetic drift, natural selection, and recurrent mutation.

To simulate data for testing mutation-rate estimation, we fixed the demographic history at that of the Schiffels-Durbin-Agarwal model.[18] For each mutation rate and sample size, we assumed a target size of 100,000 and imposed a cutoff of $K = \min(0.005 \times n, 1,000)$ for each sample size n for maximum-likelihood analyses, although the method-of-moments and first-order estimators used the entire dataset. We simulated each combination of mutation rate and sample size 45 times and averaged simulations to obtain root-mean-squared error (RMSE). When simulating data to test our power to infer mutation-rate heterogeneity, we used the same general scheme but with the haploid sample size fixed at 1,000,000 and simulated mutation rates for each locus from a gamma distribution with the desired mean and coefficient of variation. We simulated 40 replicates per combination of mean and coefficient of variation and computed the likelihood-ratio test statistic Λ for each. The power estimate is the fraction of tests for which a χ^2 test with 1° of freedom would reject the null hypothesis of no heterogeneity at the 5% level.

To simulate data for estimation of demographic parameters, we began with the Schiffels-Durbin-Agarwal model and modified it to reflect different demographic histories. To simulate two phases of exponential growth, we simulated 227 generations of growth at a rate of 2% starting 252 generations ago, followed by an additional 25 generations of growth at a rate of 12%. To simulate a population that expands and contracts, we simulated 227 generations of growth at a rate of 3% starting 252 generations ago followed by 12 generations of decline at a rate of $-$ 10%. For each demographic history, we simulated sample sizes of 1,000, 100,000, and 1,000,000. For each demography and sample-size pair, we simulated 50 replicates. Each replicate consisted of a simulation of 1,500,000 independently evolving positions, with 500,000 having a mutation rate of 10^{-9} per site per generation, 500,000 having a mutation rate of 10^{-8} per site per generation, and 500,000 having a mutation rate of 10^{-7} per site per generation.

Data

gnomAD data processing

To generate gene annotations, we used the hg38 knownCanonical table from the UCSC genome browser.[74] We obtained gene

symbols using the kgXref table and the exon and coding sequence positions for the canonical transcripts from the knownGene table. We joined tables in the UCSC genome browser by selecting fields from primary and related tables prior to downloading. To ensure that we used only coding exons and not untranslated exons, we restricted only to exons that overlapped with the coding sequence. This resulted in a list of putative genes. Next, we downloaded the 30-way phastcons track bigWig file from the UCSC genome browser[75] and used bigWigAverage OverBed to calculate the average phastcons score per exon. We then averaged exons together and only retained genes with an average phastcons score greater than 0.3. This resulted in a dataset of 16,207 genes. We then computed variant effect predictions (synonymous, non-synonymous, and stop gain) and trinucleotide contexts for every mutation using a custom script. As a final level of quality control, we joined with gnomAD v4.1 allele number counts and restricted to sites with a total allele number greater than 1,100,000 that were in the UK Biobank capture region. We further restricted our analyses to only autosomal loci. Any sites that did not pass the quality filters described here were excluded from all subsequent analyses.

We then downloaded variant information from the gnomAD v4.1 dataset from the gnomAD website. For positions without a variant in the gnomAD variant calls, we counted them as having a non-reference frequency of 0. Next, we used a hypergeometric distribution to subsample all sites observed in the non-Finnish European ("nfe") subset of gnomAD down to a haploid sample size of 1,000,000.

B-score analyses

We downloaded B scores from https://github.com/sellalab/HumanLinkedSelectionMaps/tree/master/Bmaps.[76] We lifted over coordinates from hg19 to hg38 and subsequently intersected B scores with our exonic dataset. We then stratified sites by B-score decile and ran our entire inference pipeline, including mutation-rate rescaling, on each decile separately.

Results

Maximum-likelihood estimation of demography and mutation rates

In the methods section, we describe how to use the sampling formula (Equation 2) as part of a maximum-likelihood estimation procedure for mutation rates and demographic history. Here, we assess the performance of our method and compare it with other approaches.

Estimation of mutation rates

To examine the performance of our estimation procedure, we first explore the estimation of mutation rates when the demographic model is fixed. Although our method allows maximum-likelihood estimation of mutation rates, an alternative method-of-moments procedure is suggested by the solution in Equation 2. In empirical data, the probability that a site is not variable, p_0, can be estimated as one minus the fraction of sites that are segregating in the sample,

$$\widehat{p}_0 = 1 - \frac{\text{Number of observed mutations}}{\text{Number of potential mutations}}, \equiv 1 - \widehat{p}_s$$

where we use $\widehat{p}_s$ to indicate proportion of segregating sites in the sample. Then, noting that the immigration rate $f = \frac{\theta}{2}n = 2N_0\mu n$, a method-of-moments estimator of the mutation rate can be computed,

$$\widehat{\mu} \propto -\log(1-\widehat{p}_s) \quad \text{(Equation 4)}$$

$$= \widehat{p}_s + \frac{1}{2}\widehat{p}_s^2 + \frac{1}{3}\widehat{p}_s^3 + \ldots. \quad \text{(Equation 5)}$$

This method-of-moments estimator has been previously used, although it was derived via a different method.[17,58] The second line follows from a Taylor expansion of the logarithm, and the constant of proportionality can be determined by constraining the average mutation rate in the genome to match an independent observation, as described in methods.[2,4,16] The first-order term, $\mu \propto \widehat{p}_s$, corresponds to an infinite-sites estimator of the mutation rate, and is commonly used to estimate mutation rates from population-genetic data.[2,4,37] However, it is clear from Equation 5 that the first-order approximation is poor when the fraction of segregating sites is appreciable, underestimating the mutation rate. The full method-of-moments estimator also fails when a mutational context is fully saturated: in that case, $\widehat{p}_s = 1$, and the mutation-rate estimate is infinite. For this reason, mutation-rate estimates in high-mutation-rate contexts are typically made by downsampling data,[2,4] discarding information.

With DR EVIL, we use maximum likelihood to estimate the mutation rate. Maximum likelihood has the advantage of using allele-frequency information in addition to presence/absence information. When mutation rates are low, the shape of the site frequency spectrum is not affected by the mutation rate; instead, the mutation rate only scales the total number of segregating variants. Thus, allele frequencies provide little information above presence/absence in the low-mutation-rate regime. However, when there are recurrent mutations, the shape of the site frequency spectrum is influenced by the mutation rate; for example, two independent singletons of the same allele will look like a doubleton (Figure 2A). Particularly notable is that as the mutation rate increases, the site frequency spectrum becomes non-monotonic and non-convex; under the infinite-sites model, the frequency spectrum must be decreasing and convex in the absence of population structure.[77]

One downside of common approaches and the DR EVIL maximum-likelihood estimation is that they require an accurate estimate of the mutational target size. However, due to the subtleties of read mapping and genotype calling, assessing which sites could have potentially been called as variable is surprisingly difficult.[78–80] Because the shape of the site frequency spectrum is influenced by the mutation rate, it provides information to estimate mutation rates without an estimate of the mutational target size, providing a method for mutation-rate estimation that avoids some technical challenges. Thus, we

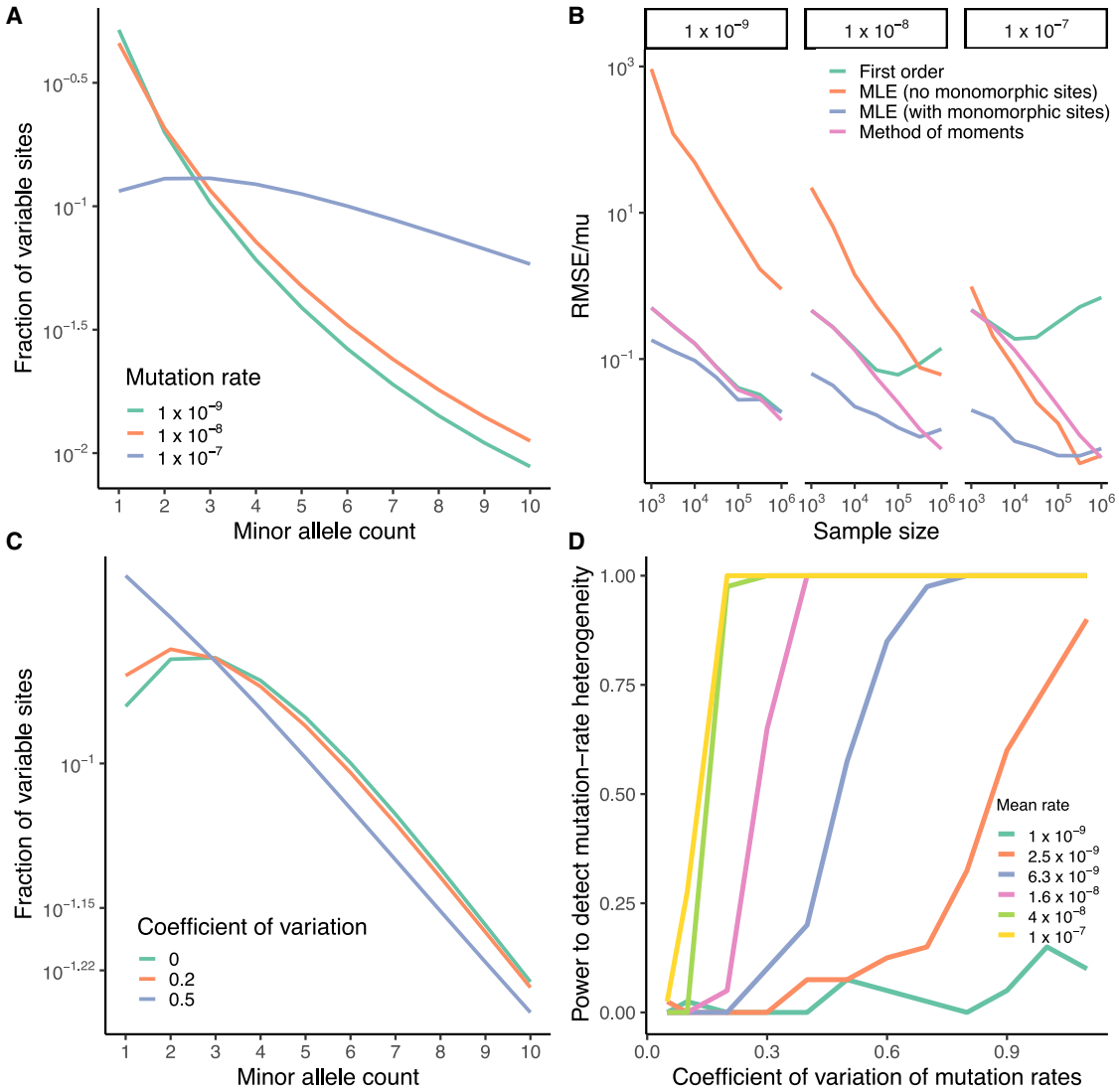

Figure 2. Mutation-rate estimation from site frequency spectra with recurrent mutation
(A) Three example site frequency spectra, simulated with a published demography[18] and three mutation rates, indicated by color. All spectra are normalized to sum to 1.
(B) Performance of different methods of mutation-rate estimation as sample size changes. The horizontal axis shows the sample size and the vertical axis the relative root-mean-squared error (RMSE). Different methods are indicated with different colors. Each panel shows the results of a simulation, with the mutation rate indicated at the top of the panel.
(C) The impact of mutation-rate heterogeneity on the site frequency spectrum. Each site frequency spectrum has an average mutation rate of 10^{-7}, but mutation rates are distributed according to a gamma distribution with the indicated coefficient of variation.
(D) Power to detect mutation-rate heterogeneity. The horizontal axis shows the coefficient of variation of the distribution of mutation rates, and the vertical axis shows the power to infer mutation-rate heterogeneity. Each line represents a different mutation rate.

also explored the possibility of estimating mutation rates without target sizes.

To test these different methods for estimating mutation rates, we simulated data under a recently proposed demographic model[18] while varying sample size and mutation rate. Figure 2B shows that the maximum-likelihood estimator of the mutation rate is consistently more accurate than any alternative; Figure S2 shows that the other approaches are biased downward. As expected, for small mutation rates and sample sizes, the method-of-moments and first-order estimator perform equivalently. However, as the sample size increases, the method-of-moments and maximum-likelihood estimators become increasingly accurate, but the first-order estimator begins to increase in relative RMSE. This is because it does not properly account for recurrent mutation. Moreover, although the maximum-likelihood estimator without a target size performs poorly for low mutation rates and small sample sizes, it becomes comparable to the full maximum-likelihood estimator for high mutation rates at large sample sizes.

Finally, although it is common practice in human genetics to stratify variants by factors that influence their mutation rate (for example, trinucleotide context and methylation level), there is substantial evidence for residual

mutation-rate variation.[13,17,30] To model mutation-rate variation that is not captured by the chosen stratification strategy, we incorporated a gamma distribution of mutation rates (supplemental information). We emphasize that we chose the gamma distribution for analytical convenience and flexibility and do not claim a biological motivation. Figure 2C shows that the coefficient of variation, which measures the amount mutation-rate heterogeneity, has a substantial impact on the shape of the frequency spectrum, resulting in an excess of low-frequency variants. To test the ability to detect residual mutation-rate variation, we simulated data under a gamma distribution of mutation rates. Figure 2D shows statistical power as a function of the coefficient of variation and the mutation rate. Unsurprisingly, there is more power to detect mutation-rate heterogeneity for higher average mutation rates and larger coefficients of variation. Figure S3 shows that not accounting for mutation-rate heterogeneity results in a biased estimate of the average mutation rate. However, when we reject the null hypothesis of a single mutation rate, we accurately estimate the mean mutation rate and coefficient of variation.

Joint estimation of demography and mutation

In most applications, the exact demography is unknown and must also be estimated. This is particularly true when working with very large datasets: a key advantage of using extremely large sample sizes is that they reveal more recent demographic history because they contain rare variants that arose recently. However, in large samples, recurrent mutation distorts the site frequency spectrum compared with what would be expected under an infinite-sites model, as seen in Figure 1. Therefore, inferences about demographic history made using tools that rely on the infinite-sites assumption may be biased. On the other hand, because the likelihood approach developed here only applies to rare variants, which arose recently, it is not suited to estimate very ancient demographic history.

Thus, because DR EVIL models rare variants, we only infer very recent demography and fix more ancient demography. The ancient demographic history can be fixed by, for example, inferring demographic parameters using infinite-sites site frequency spectrum methods on a smaller subsample, or using complementary methods, such as coalescent hidden Markov models.[81–83] We then infer a piecewise-constant demography for recent periods. In the methods section, we describe a multi-step approach to infer demographic history and mutation rates jointly.

To test DR EVIL, we simulated data under two different demographic models across sample sizes and compared our estimates with those made using an infinite-sites model. First, we explored a model with two phases of exponential growth (Figure 3). Increasing sample sizes allows for inference of increasingly recent demographic history. With 1,000,000 samples, we see accurate estimation of the effective population size in the last ten generations. We also see the critical importance of modeling recurrent mutation: comparing the estimates of effective population size when modeling rare variation (Figure 3A) with estimates that use the infinite-sites model (Figure 3B) shows that infinite-sites estimates display increased variance in the more distant past and substantial bias in the recent past. Figure S4 shows a similar pattern under a model of population expansion followed by decrease, in which estimates made using the infinite-sites model are both higher variance and more biased across sample sizes.

Application to gnomAD

We applied our approach to estimate mutation rates and recent demographic history from the gnomAD v4.1 dataset. We took advantage of the massive sample size afforded by the inclusion of the UK Biobank exome data and computed the synonymous site frequency spectrum of non-Finnish European samples, downsampled to 1,000,000 haploids. We imposed an allele-frequency cutoff, only analyzing alleles appearing in the sample fewer than $K = 1,000$ times. To account for the context dependence of mutation rates, we stratified the site frequency spectrum by trinucleotide context and methylation level. Because some trinucleotide contexts do not result in synonymous mutations, we observed 92 out of the 96 possible trinucleotide contexts. Combined with methylation status at CpG transitions,[2] we analyzed 144 total contexts.

To estimate demographic history, we divided the recent past into bins that were evenly spaced on the log scale and estimated the effective population size in each bin (methods and Table S1). We also fit both a single mutation rate and a gamma distribution of mutation rates to each context (Table S2). Figures 4A and 4B show the fit of the model to the observed data for two trinucleotide contexts, the low-mutation-rate CAA-to-CTA context and the high-mutation-rate CGT-to-CAT context at methylation level 6, respectively. The CAA-to-CTA context appears consistent with an infinite-sites model, but the CGT-to-CAT context can only be fit well by allowing for recurrent mutation. Moreover, we show in Figure S5 that the fit is significantly improved by allowing a gamma distribution of mutation rates. Thus, we see that our model can fit the data well and that it is important to model recurrent mutation and residual mutation-rate variation to do so.

Figure 4C shows our estimated demographic history. Consistent with expectations, we see explosive population growth and infer an effective population size of approximately 31 million over approximately the last seven generations. To ensure that our results were robust to the hyperparameter choices we made (such as the number of bins in which to infer effective population size and the cutoff in the past up to which we inferred population size), we explored a number of different hyperparameter choices, finding that inference of recent effective population sizes were similar across different choices, despite some noise in the ancient past (Figure S6).

Background selection—the impact of the selective removal of deleterious variants on genetic diversity at linked neutral loci—has been shown to affect demographic inference.[84–87] To determine the sensitivity of our results to

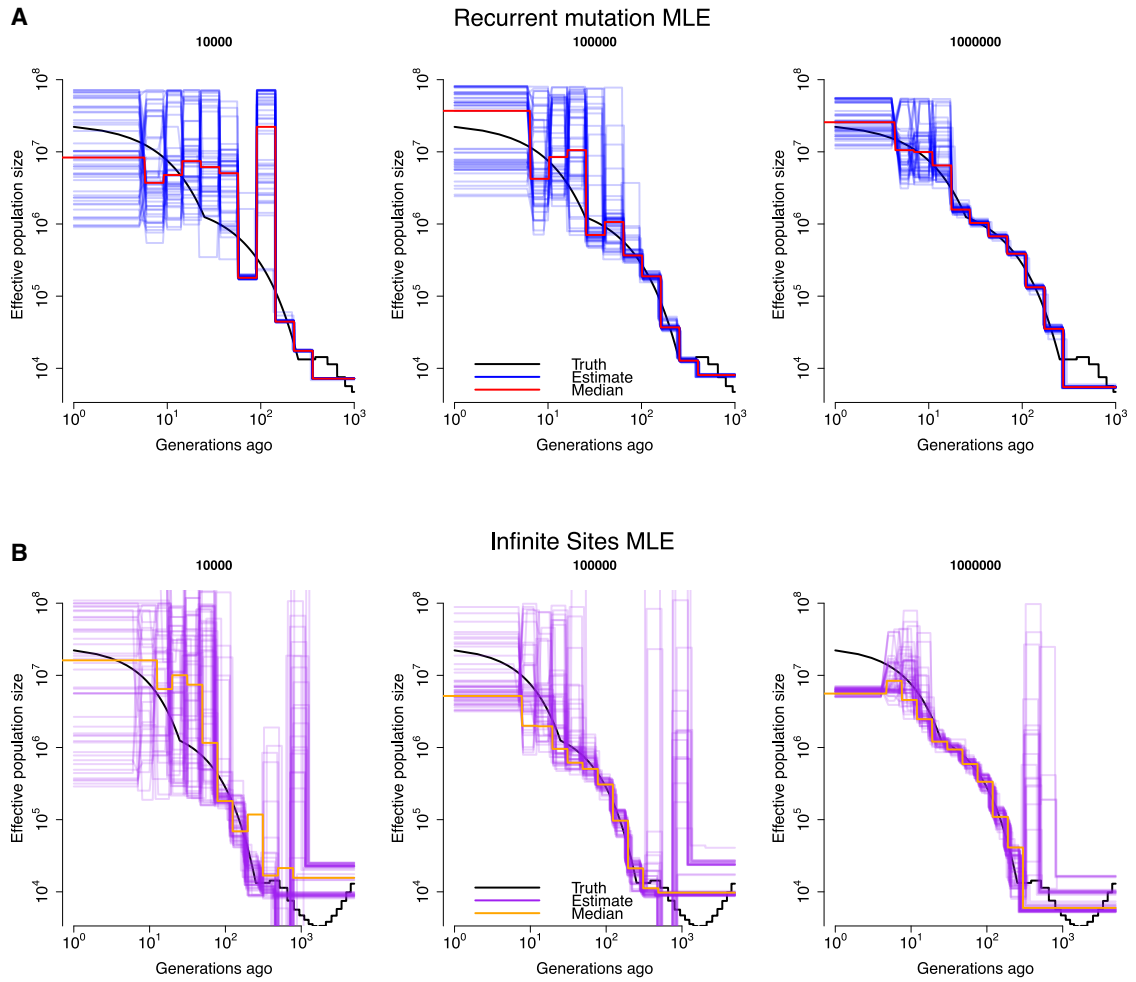

Figure 3. The importance of sample size and recurrent mutation for demographic inference
In each panel, the horizontal axis shows the time in generations and the vertical axis the effective population size. The solid black lines show the simulated demography, and each blue line shows the estimate from a single simulation replicate. The red line shows the median across simulations. Population sizes and generation times are scaled according to the median rescaling across simulations. The number on top of each subpanel represents the sample size.
(A) Estimation accounting for recurrent mutation.
(B) Estimation under the infinite-sites model.

the effects of background selection, we stratified the exome by decile of B score,[76] which quantifies the expected effect of background selection. We inferred demographic history from each decile independently. Despite the additional noise due to each bin having a tenth as many variants as were included in the full analysis, the demographic histories estimated in different B-score deciles were broadly similar to each other and to the estimates from the full dataset, suggesting that our overall conclusions are robust (Figure S7A). Nonetheless, we do see that higher B scores (i.e., lower levels of background selection) are associated with larger estimates of population size, consistent with expectations (Figure S7B).

We then examined the distribution of estimated mutation rates. Figure 4D shows that, consistent with previous analyses, CpG transitions are estimated to have mutation rates 10–100 times larger than in other contexts. Figure S8A shows that that our estimates are largely concordant with those of gnomAD, although for some contexts we find substantial disagreement. To ensure our estimation of mutation rates was robust to our estimation of the target size (which relies on several assumptions about where mutations can be reliably called), we verified that our estimated mutation rates were concordant with and without information about the number of monomorphic sites (Figure S8B). We have power to estimate mutation rates without including target size because, due to explosive population growth, all but the lowest mutation-rate categories show substantial deviations from the infinite-sites model, and fits are improved by incorporating recurrent mutation (Figure S9). Within CpG contexts, we find that both context and methylation level are associated with mutation-rate differences (Figure 4E). Further, although we fail to reject the null hypothesis of no mutation-rate heterogeneity for most contexts (49/144 contexts have support for heterogeneity at a 10% false discovery rate), Figure 4F

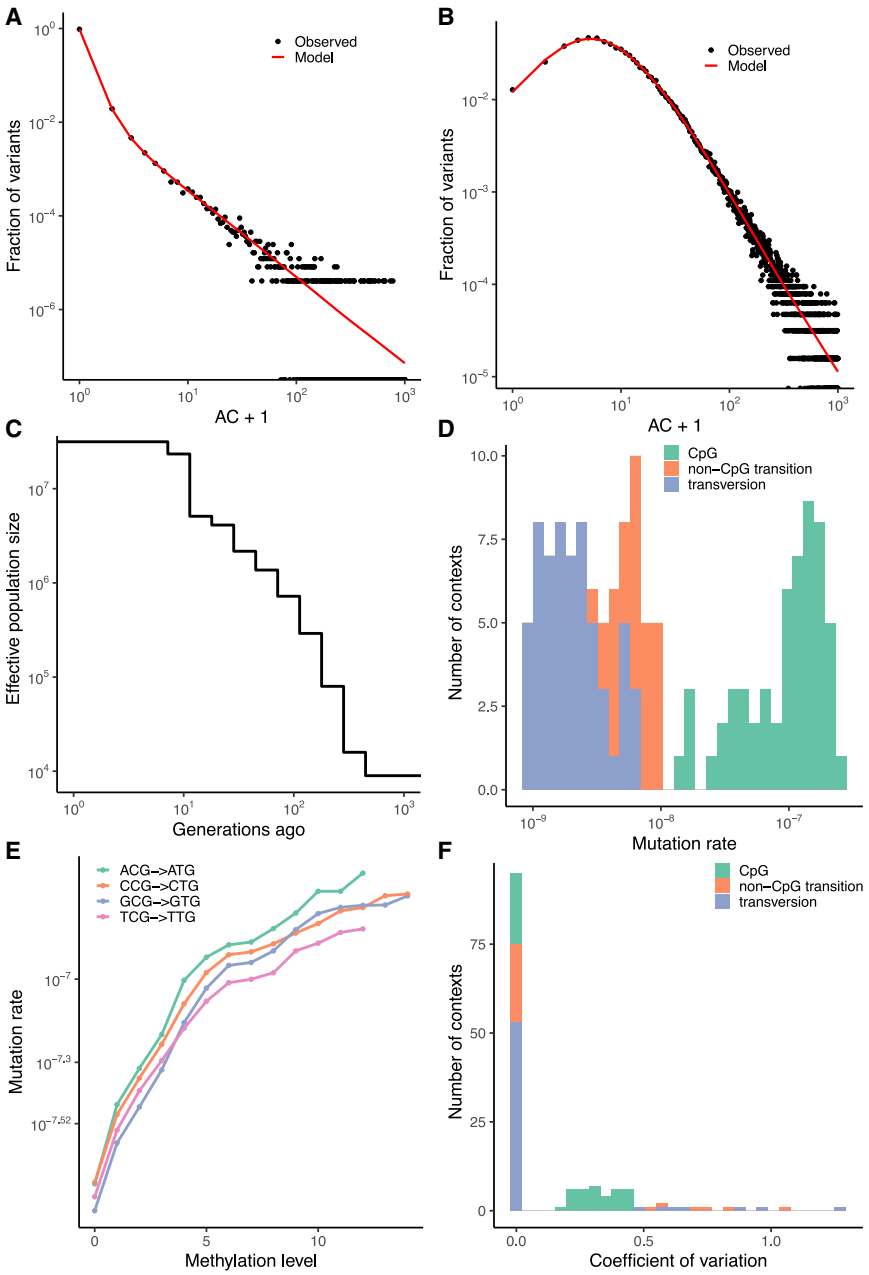

Figure 4. Estimation using 1,000,000 samples from gnomAD v4.1

(A) The fit of the model for the low-mutation-rate CAA-to-CTA context.

(B) The fit of the model for the high-mutation-rate CGT-to-CAT context at methylation level 6. In (A) and (B), the horizontal axis shows the number of minor alleles plus 1 to account for the log scale.

(C) Estimated demographic history.

(D) Estimated distribution of mutation rates across contexts and methylation status.

(E) Mutation rate as a function of methylation level across CpG contexts.

(F) Estimated coefficients of variation across contexts and methylation status. Contexts in which the null hypothesis of no heterogeneity is not rejected at a 10% false discovery rate are assigned a coefficient of variation of 0.

shows that we find significant residual mutation-rate variation in many categories, even after accounting for methylation level. Interestingly, while CpG contexts have an appreciable coefficient of variation, our largest estimates of residual variation occur in non-CpG contexts, in line with the observation that there are features beyond methylation and trinucleotide context that influence mutation rates.[17,30]

A possible concern in large sample sizes is that sequencing error could create a substantial number of false-positive variants as well as possibly confounding our estimates of mutation-rate heterogeneity. We derived a model of sequencing error for rare alleles in large datasets and found a closed-form solution that we implemented in a maximum-likelihood approach to estimate sequencing error rates (supplemental information and methods). We found substantial evidence for an effect of sequencing error in low-mutation-rate contexts (Figure S10 and Table S3); however, we found that contexts with sequencing error effects were not the same as the contexts with mutation-rate heterogeneity (Figure S10B), suggesting that these two processes do not confound each other. Interestingly, our estimates of sequencing error rate are strongly correlated with mutation rate (Figure S10C), which may be consistent with deviations from the breeding structure assumed by the diffusion model[54,60] or may be due to ascertainment, because sequencing errors must be larger to have a noticeable impact on contexts with higher mutation rates. Nonetheless, we see that, as predicted by analyses shown in the supplemental information, sequencing error has a relatively small effect on the predicted frequency spectrum, only resulting in a small increase in singletons and a small decrease in monomorphic sites for the lowest-mutation-rate contexts (Figure S10D).

Finally, we sought to understand the effect of strong selection on variants in large-population-variation datasets. Using the demographic model we estimated from gnomAD, we explored the effect of selection on mutation saturation, allele frequencies, and recurrent mutation. First, we examined the information about selection conveyed by presence or absence of a variant (Figure 5A). Selection affects the ability to observe variants across a range of sample sizes. Unsurprisingly, increasing the sample size results in more observed variants, and increasing

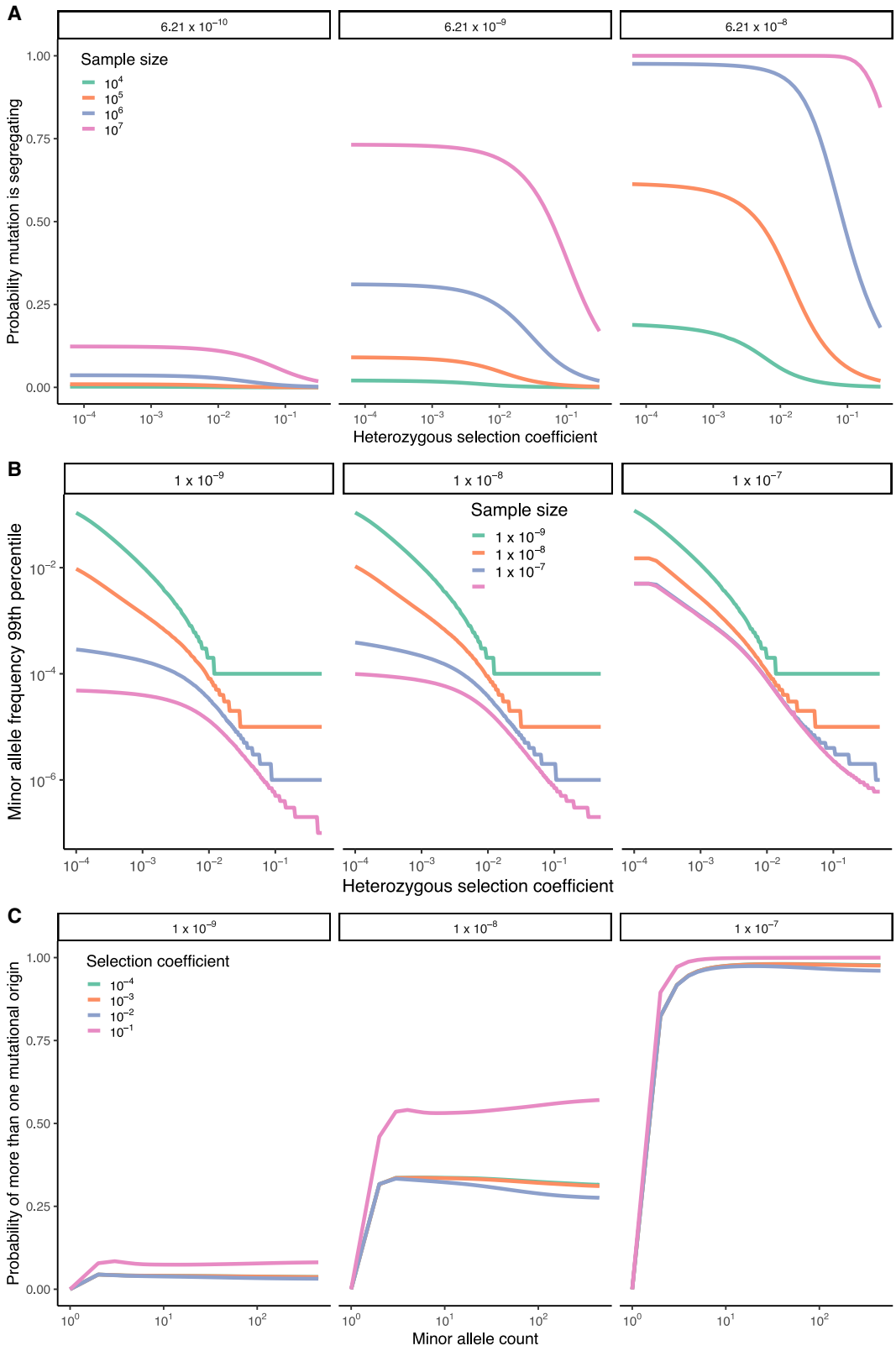

Figure 5. The effect of selection in large samples

All selection coefficients shown indicate the strength of negative selection.

(A) Strong selection reduces mutational saturation. The horizontal axis shows the heterozygous selection coefficient, while the vertical axis shows the proportion of all possible variants that are observed. Each subpanel corresponds to a different mutation rate, indicated at the top of the panel, and each line corresponds to a sample size, given by the color.

(legend continued on next page)

the heterozygous selection coefficient results in fewer observed variants. Consistent with previous work,[28,44] positions with high mutation rates are nearly saturated at current sample sizes and essentially completely saturated at samples around 10 million. In contrast, the vast majority of sites with mutation rates between 10^{-9} and 10^{-8} will be non-segregating in samples smaller than 10 million, meaning that presence or absence in a population variation database is only weakly informative about selection on those mutations.

Next, we examined the information available in frequencies of segregating alleles (Figure 5B). At each selection coefficient, we examined the 99th percentile of the allele frequency distribution conditioned on the allele segregating; thus, this is the largest selection coefficient that could be rejected at the 99% level. High mutation rates can result in mutations being an order of magnitude more common at a given selection coefficient compared with lower mutation rates. In samples of size 10 million, CpGs with selection coefficients between 10^{-3} and 10^{-2} will be an order of magnitude more common than slower-mutating alleles with the same heterozygous selection coefficient.

We also examined the effect of recurrent mutation (Figure 5C). The prevalence of recurrent mutation has implications for phasing and imputation, which typically assume an infinite-sites model. While alleles with mutation rates on the order of 10^{-9} are unlikely to have multiple origins, strongly selected mutations with mutation rates on the order of 10^{-8} or larger have a substantial probability of having arisen multiple times, suggesting that deleterious CpGs may be difficult to phase or impute. As an alternative view of this question, Figure S11 shows the expected number of mutational origins, which is close to one for low mutation rates but can be greater than four for high mutation rates and strong selection. Perhaps surprisingly, we find a non-monotonic relationship between the probability of multiple origins and the selection coefficient, demonstrating the importance of explicit population-genetic modeling in understanding how evolutionary forces affect genetic diversity (Figure S12 and supplemental information).

Discussion

Here, we presented DR EVIL, a method for population-genetic estimation of demographic parameters and mutation rates applicable to samples of millions of genomes. Because most of the information contained in extremely large sequencing studies that is unavailable in smaller datasets is contained in the rare variants they uncover, our approach focuses on modeling rare variation. A focus on rare variation allows computationally efficient handling of recurrent mutation and selection, enabling analyses of large datasets in which the standard assumptions of neutrality and infinite sites are violated.

Our approach allows for estimation of mutation rates with up to an order of magnitude lower RMSE than standard population-genetic approaches.[2,17] This is because in large sequencing studies, mutation affects not just the proportion of segregating sites but also the shape of the allele frequency distribution. Leveraging this fact, we found that in sufficiently large sequencing studies, mutation rates can be estimated without an estimate of the target size; that is, it is unnecessary to account for invariant sites. Computing the number of invariant sites can be difficult: it is important to know whether a site is invariant because it is truly invariant or because it was too difficult to genotype. We also characterized a method-of-moments estimator of the mutation rate that is substantially more robust to recurrent mutation and achieves impressive efficiency at large sample sizes. Further, our method provides a likelihood-ratio test for heterogeneity of mutation rates, and we find evidence of such heterogeneity in many sequence contexts, motivating future work uncovering the sources of heterogeneity.

Sequencing errors are a major concern when examining rare variation.[88,89] We built a theoretical model of sequencing error and found that errors create artificial singletons in large datasets, with a more pronounced effect in larger datasets. Strikingly, when we estimate sequencing error rates, we find that they are strongly correlated with mutation rates. Although this may reflect some similarities between sequencing chemistry and *in vivo* mutational repair mechanisms, it may also reflect departures from the diffusive model assumed here.[54,56,60] The correlation may also result from an ascertainment issue: in order to detect sequencing error among sites with large mutation rates, the sequencing error rate must itself be large. Altogether, although we find evidence for sequencing error or deviations from diffusive population structure in our dataset, the impact of such effects on our overall results is small.

Our approach allowed us to estimate the effective population size as recently as ten generations ago, a time period that only appreciably affects the site frequency spectrum in datasets with hundreds of thousands of haplotypes. Moreover, we found that, because recurrent mutation is a major determinant of allele frequencies for rare variants in ultra-large sequencing studies, traditional

(B) High mutation rates enable deleterious mutations to rise to high frequency. The horizontal axis shows the heterozygous selection coefficient, and the vertical axis shows the 99th percentile of allele frequencies for segregating alleles. Each subpanel corresponds to a different mutation rate, indicated at the top of the panel, and each line corresponds to a different sample size, indicated by color.
(C) Mutations have multiple origins in large samples. The horizontal axis shows the minor-allele count in a sample of a million haploid genomes, and the vertical axis shows the probability that a mutation has more than one origin. Each subpanel corresponds to a different mutation rate, indicated at the top of the panel, and each line corresponds to a different heterozygous selection coefficient, indicated by color.

population-genetic models that do not account for recurrent mutation result in strongly biased estimates of recent effective population sizes. When applied to humans, we find evidence for super-exponential growth and a large recent effective population size, consistent with known patterns of the census population size. While we found that our empirical results are robust to stratification by the intensity of background selection,[76,85,87] in future work it will be important to use theory and simulation to understand how explosive population growth interacts with background selection to affect patterns of genetic diversity.[90]

More generally, our work can be seen as being broadly in the tradition of site-frequency-spectrum-based approaches to demographic inference.[47,91–93] Although there are existing theories and methods applicable to samples with two mutations in their history, such as tri-allelic data,[93–95] our method allows for an arbitrary number of mutational origins so long as the allele remains rare. Multiple mutational origins are expected for high-mutation-rate alleles even in samples of tens of thousands of individuals.[28,44] As we have shown, recurrent mutation is expected to be the norm in high-mutation-rate classes given current sample sizes. Our model also adds to the literature on branching-process models of rare variants that allow for recurrent mutation. For neutral variants, our model is identical to that of Wakeley et al.,[58] but we derive an efficient sampling formula and generalize their approach by incorporating natural selection and developing a method for estimation from empirical site frequency spectra.

Another domain where our results may be useful is in understanding mutations in highly mutagenic somatic tissues, such as tumors. Branching and birth-death models are commonly used to model these situations, particularly in the case of exponential growth.[96–99] Our results can be used to generalize these approaches to more general models of non-equilibrium tissue growth and may enable more general inference of tissue growth from bulk and single-cell somatic sequencing experiments.

Whereas measures of mutational constraint are critical tools for prioritizing genomic regions in disease,[100–102] drug discovery,[103] and other applications, accurate and interpretable measures of mutational constraint require understanding the demographic and mutational forces that create genetic variation.[104] Using our model, we showed that the interplay among natural selection, mutation, and demography can be complex, supporting the importance of explicit population-genetic modeling for estimating constraint. Moreover, given the possibility of estimating gene trees from tens or hundreds of thousands of samples,[105–108] it may be possible to leverage estimated gene trees to provide enhanced estimates of natural selection. In general, modeling the genealogy of selected variants is difficult,[50,51] but our approach suggests that fitting a simplified birth-death model may substantially augment our ability to learn about natural selection from gene trees of rare variants.[109]

Data and code availability

This study did not generate any new data. The software to run DR EVIL and to reproduce the analyses in this paper is available as R code at https://github.com/Schraiber/drevil.

Acknowledgments

This work is dedicated to the memory of Masatoshi Nei. We thank John Wakeley, Louis Fan, Daniel Balick, Kelley Harris, Alison Feder, Matt Pennell, Molly Przeworski, Guy Sella, Tony Zeng, Hakhamanesh Mostafavi, and Vince Buffalo for helpful discussions during the progress of this work. Funding for J.G.S. and M.D.E. was provided by NIH grant R35GM137758.

Declaration of interests

The authors declare that they have no competing interests.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2025.07.008.

Web resources

Bmaps, https://github.com/sellalab/HumanLinkedSelectionMaps/tree/master/Bmaps

Received: June 16, 2025
Accepted: July 22, 2025
Published: August 13, 2025

References

1. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
2. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mu tational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
3. The All of Us Research Program Genomics Investigators (2024). Genomic data in the All of Us research program. Nature *627*, 340–346.
4. Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alföldi, J., Watts, N.A., Vittal, C., Gauthier, L. D., et al. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. Nature *625*, 92–100.
5. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature *599*, 628–634.
6. Halldorsson, B.V., Eggertsson, H.P., Moore, K.H.S., Hauswedell, H., Eiriksson, O., Ulfarsson, M.O., Palsson, G., Hardarson, M.T., Oddsson, A., Jensson, B.O., et al. (2022). The sequences of 150,119 genomes in the UK Biobank. Nature *607*, 732–740.

7. Wang, Q., Dhindsa, R.S., Carss, K., Harper, A.R., Nag, A., Tachmazidou, I., Vitsios, D., Deevi, S.V.V., Mackay, A., Muthas, D., et al. (2021). Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature *597*, 527–532.
8. Sun, K.Y., Bai, X., Chen, S., Bao, S., Zhang, C., Kapoor, M., Backman, J., Joseph, T., Maxwell, E., Mitra, G., et al. (2024). A deep catalogue of protein-coding variation in 983,578 individuals. Nature *631*, 583–592.
9. Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740–743.
10. Gazave, E., Ma, L., Chang, D., Coventry, A., Gao, F., Muzny, D., Boerwinkle, E., Gibbs, R.A., Sing, C.F., Clark, A.G., and Keinan, A. (2014). Neutral genomic regions refine models of recent rapid human population growth. Proc. Natl. Acad. Sci. USA *111*, 757–762.
11. Gao, F., and Keinan, A. (2016). Explosive genetic evidence for explosive human population growth. Curr. Opin. Genet. Dev. *41*, 130–139.
12. Gao, F., and Keinan, A. (2016). Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. Genetics *202*, 235–245.
13. Aggarwala, V., and Voight, B.F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat. Genet. *48*, 349–355.
14. Harris, K. (2015). Evidence for recent, population-specific evolution of the human mu tation rate. Proc. Natl. Acad. Sci. USA *112*, 3439–3444.
15. Harris, K., and Pritchard, J.K. (2017). Rapid evolution of the human mutation spectrum. eLife *6*, e24284.
16. DeWitt, W.S., Harris, K.D., Ragsdale, A.P., and Harris, K. (2021). Nonparametric coalescent inference of mutation spectrum history and demography. Proc. Natl. Acad. Sci. USA *118*, e2013798118.
17. Seplyarskiy, V., Koch, E.M., Lee, D.J., Lichtman, J.S., Luan, H.H., and Sunyaev, S.R. (2023). A mutation rate model at the basepair resolution identifies the mutagenic effect of polymerase III transcription. Nat. Genet. *55*, 2235–2242.
18. Agarwal, I., Fuller, Z.L., Myers, S.R., and Przeworski, M. (2023). Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. eLife *12*, e83172.
19. Zeng, T., Spence, J.P., Mostafavi, H., and Pritchard, J.K. (2024). Bayesian estimation of gene constraint from an evolutionary model with gene features. Nat. Genet. *56*, 1632–1643.
20. Cassa, C.A., Weghorn, D., Balick, D.J., Jordan, D.M., Nusinow, D., Samocha, K.E., O'Donnell-Luria, A., MacArthur, D.G., Daly, M.J., Beier, D.R., and Sunyaev, S.R. (2017). Estimat ing the selective effects of heterozygous protein-truncating variants from human exome data. Nat. Genet. *49*, 806–810.
21. Fu, W., O'connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.
22. Kimura, M., and Ohta, T. (1973). The age of a neutral mutant persisting in a finite population. Genetics *75*, 199–212.
23. Watterson, G.A. (1977). Reversibility and the age of an allele II. Two-allele models, with selection and mutation. Theor. Popul. Biol. *12*, 179–196.
24. Griffiths, R.C. (2003). The frequency spectrum of a mutation, and its age, in a general diffusion model. Theor. Popul. Biol. *64*, 241–251.
25. Slatkin, M., and Rannala, B. (2000). Estimating allele age. Annu. Rev. Genomics Hum. Genet. *1*, 225–249.
26. Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T., Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A., et al. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature *549*, 519–522.
27. Halldorsson, B.V., Palsson, G., Stefansson, O.A., Jonsson, H., Hardarson, M.T., Eggertsson, H.P., Gunnarsson, B., Oddsson, A., Halldorsson, G.H., Zink, F., et al. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science *363*, eaau1043.
28. Agarwal, I., and Przeworski, M. (2021). Mutation saturation for fitness effects at human CpG sites. eLife *10*, e71513.
29. Carlson, J., Locke, A.E., Flickinger, M., Zawistowski, M., Levy, S., Myers, R.M., Boehnke, M., Kang, H.M., Scott, L.J., Li, J.Z., et al. (2018). Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Nat. Commun. *9*, 3753.
30. Seplyarskiy, V.B., Soldatov, R.A., Koch, E., McGinty, R.J., Goldmann, J.M., Hernandez, R.D., Barnes, K., Correa, A., Burchard, E.G., Ellinor, P.T., et al. (2021). Population sequencing data reveal a compendium of mutational processes in the human germ line. Science *373*, 1030–1035.
31. Agarwal, I., and Przeworski, M. (2019). Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. Proc. Natl. Acad. Sci. USA *116*, 17916–17924.
32. Ehrlich, M., Norris, K.F., Wang, R.Y., Kuo, K.C., and Gehrke, C.W. (1986). DNA cytosine methylation and heat-induced deamination. Biosci. Rep. *6*, 387–393.
33. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.
34. Masson, E., Zou, W.-B., Génin, E., Cooper, D.N., Le Gac, G., Fichou, Y., Pu, N., Rebours, V., Férec, C., Liao, Z., and Chen, J.M. (2022). Expanding ACMG variant classification guidelines into a general framework. Hum. Genomics *16*, 31.
35. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.
36. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067.
37. Gao, H., Hamp, T., Ede, J., Schraiber, J.G., McRae, J., Singer-Berk, M., Yang, Y., Dietrich, A.S.D., Fiziev, P.P., Kuderna, L. F.K., et al. (2023). The landscape of tolerated genetic variation in humans and primates. Science *380*, eabn8153.

38. Jurgens, S.J., Wang, X., Choi, S.H., Weng, L.-C., Koyama, S., Pirruccello, J.P., Nguyen, T., Smadbeck, P., Jang, D., Chaffin, M., et al. (2024). Rare coding variant analysis for human diseases across biobanks and ancestries. Nat. Genet. *56*, 1811–1820.
39. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. USA *111*, E455–E464.
40. Auer, P.L., and Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. Genome Med. *7*. 16–11.
41. Sawyer, S.A., and Hartl, D.L. (1992). Population genetics of polymorphism and divergence. Genetics *132*, 1161–1176.
42. Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics *61*, 893–903.
43. Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics *105*, 437–460.
44. Harpak, A., Bhaskar, A., and Pritchard, J.K. (2016). Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. PLoS Genet. *12*, e1006489.
45. Polanski, A., and Kimmel, M. (2003). New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics *165*, 427–436.
46. Polanski, A., Bobrowski, A., and Kimmel, M. (2003). A note on distributions of times to coalescence, under time-dependent population size. Theor. Popul. Biol. *63*, 33–40.
47. Bhaskar, A., Wang, Y.X.R., and Song, Y.S. (2015). Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Res. *25*, 268–279.
48. Kamm, J.A., Terhorst, J., and Song, Y.S. (2017). Efficient computation of the joint sample frequency spectra for multiple populations. J. Comput. Graph Stat. *26*, 182–194.
49. Evans, S.N., Shvets, Y., and Slatkin, M. (2007). Non-equilibrium theory of the allele frequency spectrum. Theor. Popul. Biol. *71*, 109–119.
50. Krone, S.M., and Neuhauser, C. (1997). Ancestral processes with selection. Theor. Popul. Biol. *51*, 210–237.
51. Neuhauser, C., and Krone, S.M. (1997). The genealogy of samples in models with selection. Genetics *145*, 519–534.
52. Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., and François, O. (2010). Approximate Bayesian computation (ABC) in practice. Trends Ecol. Evol. *25*, 410–418.
53. Schrider, D.R., and Kern, A.D. (2018). Supervised machine learning for population genetics: a new paradigm. Trends Genet. *34*, 301–312.
54. Wakeley, J., and Takahashi, T. (2003). Gene genealogies when the sample size exceeds the effective size of the population. Mol. Biol. Evol. *20*, 208–213.
55. Fu, Y.-X. (2006). Exact coalescent for the Wright–Fisher model. Theor. Popul. Biol. *69*, 385–394.
56. Bhaskar, A., Clark, A.G., and Song, Y.S. (2014). Distortion of genealogical properties when the sample is very large. Proc. Natl. Acad. Sci. USA *111*, 2385–2390.
57. Melfi, A., and Viswanath, D. (2018). The Wright–Fisher site frequency spectrum as a perturbation of the coalescent's. Theor. Popul. Biol. *124*, 81–92.
58. Wakeley, J., Fan, W.T.L., Koch, E., and Sunyaev, S. (2023). Recurrent mutation in the ancestry of a rare variant. Genetics *224*, iyad049.
59. Fan, W.-T.L., and Wakeley, J. (2024). Latent mutations in the ancestries of alleles under selection. Theor. Popul. Biol. *158*, 1–20.
60. Spence, J.P., Zeng, T., Mostafavi, H., and Pritchard, J.K. (2023). Scaling the discrete-time Wright–Fisher model to biobank-scale datasets. Genetics *225*, iyad168.
61. Feller, W. (1951). Diffusion processes in genetics. In Proc. Second Berkeley Symp. on Math. Stat. and Prob., pp. 227–246.
62. Haldane, J.B.S. (1927). A mathematical theory of natural and artificial selection, part V: selection and mutation. Math. Proc. Camb. Phil. Soc. *23*, 838–844. Cambridge University Press.
63. Nei, M. (1968). The frequency distribution of lethal chromosomes in finite populations. Proc. Natl. Acad. Sci. USA *60*, 517–524.
64. Charlesworth, B., and Hill, W.G. (2019). Selective effects of heterozygous protein-truncating variants. Nat. Genet. *51*, 2.
65. Kendall, D.G. (1949). Stochastic processes and population growth. J. Roy. Stat. Soc. B *11*, 230–264.
66. Giorno, V., and Nobile, A.G. (2020). Bell Polynomial Approach for Time-Inhomogeneous Linear Birth–Death Process with Immigration. Mathematics *8*, 1123.
67. Bell, E.T. (1927). Partition polynomials. Ann. Math. *29*, 38–46.
68. Bell, E.T. (1934). Exponential polynomials. Ann. Math. *35*, 258–277.
69. Turkmen, A., and Lin, S. (2017). Are rare variants really independent? Genet. Epidemiol. *41*, 363–371.
70. Brent, R.P. (1971). An algorithm with guaranteed convergence for finding a zero of a function. Comput. J. *14*, 422–425.
71. Liu, D.C., and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. Math. Program. *45*, 503–528.
72. Rosen, Z., Bhaskar, A., Roch, S., and Song, Y.S. (2018). Geometry of the sample frequency spectrum and the perils of demographic inference. Genetics *210*, 665–682.
73. Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statist. *9*, 60–62.
74. Nassar, L.R., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., et al. (2023). The UCSC genome browser database: 2023 update. Nucleic Acids Res. *51*, D1188–D1195.
75. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.
76. Murphy, D.A., Elyashiv, E., Amster, G., and Sella, G. (2023). Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and noncoding elements. eLife *12*, e76065.
77. Sargsyan, O., and Wakeley, J. (2008). A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theor. Popul. Biol. *74*, 104–114.

78. Ségurel, L., Wyman, M.J., and Przeworski, M. (2014). Determinants of mutation rate variation in the human germline. Annu. Rev. Genomics Hum. Genet. *15*, 47–70.
79. Schraiber, J.G., and Akey, J.M. (2015). Methods and models for unravelling human evolutionary history. Nat. Rev. Genet. *16*, 727–740.
80. Bergeron, L.A., Besenbacher, S., Turner, T., Versoza, C.J., Wang, R.J., Price, A.L., Armstrong, E., Riera, M., Carlson, J., Chen, H.-y., et al. (2022). The Mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. eLife *11*, e73577.
81. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.
82. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. *46*, 919–925.
83. Spence, J.P., Steinrücken, M., Terhorst, J., and Song, Y.S. (2018). Inference of population history using coalescent HMMs: review and outlook. Curr. Opin. Genet. Dev. *53*, 70–76.
84. Ewing, G.B., and Jensen, J.D. (2016). The consequences of not accounting for background selection in demographic inference. Mol. Ecol. *25*, 135–141.
85. Johri, P., Charlesworth, B., and Jensen, J.D. (2020). Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. Genetics *215*, 173–192.
86. Johri, P., Riall, K., Becher, H., Excoffier, L., Charlesworth, B., and Jensen, J.D. (2021). The impact of purifying and background selection on the inference of population history: problems and prospects. Mol. Biol. Evol. *38*, 2986–3003.
87. Soni, V., Pfeifer, S.P., and Jensen, J.D. (2024). The effects of mutation and recombination rate heterogeneity on the inference of demography and the distribution of fitness effects. Genome Biol. Evol. *16*, evae004.
88. Han, E., Sinsheimer, J.S., and Novembre, J. (2014). Characterizing bias in population genetic inferences from low-coverage sequencing data. Mol. Biol. Evol. *31*, 723–735.
89. Johri, P., Aquadro, C.F., Beaumont, M., Charlesworth, B., Excoffier, L., Eyre-Walker, A., Keightley, P.D., Lynch, M., McVean, G., Payseur, B.A., et al. (2022). Recommendations for improving statistical inference in population genomics. PLoS Biol. *20*, e3001669.
90. Barroso, G.V., and Ragsdale, A.P. (2025). A model for background selection in nonequilibrium populations. Preprint at bioRxiv. https://doi.org/10.1101/2025.02.19.639084.
91. Williamson, S.H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C.D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA *102*, 7882–7887.
92. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. *5*, e1000695.
93. Jouganous, J., Long, W., Ragsdale, A.P., and Gravel, S. (2017). Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. Genetics *206*, 1549–1567.
94. Jenkins, P.A., and Song, Y.S. (2011). The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. Theor. Popul. Biol. *80*, 158–173.
95. Desai, M.M., and Plotkin, J.B. (2008). The polymorphism frequency spectrum of finitely many sites under selection. Genetics *180*, 2175–2191.
96. Durrett, R. (2013). Population genetics of neutral mutations in exponentially growing cancer cell populations. Ann. Appl. Probab. *23*, 230–250.
97. Ohtsuki, H., and Innan, H. (2017). Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. Theor. Popul. Biol. *117*, 43–50.
98. Cheek, D., and Antal, T. (2018). Mutation frequencies in a birth–death branching process. Ann. Appl. Probab. *28*, 3922–3947.
99. Cheek, D., and Antal, T. (2020). Genetic composition of an exponentially growing cell population. Stoch. Process. their Appl. *130*, 6580–6624.
100. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Aslibekyan, S., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat. Genet. *52*, 969–983.
101. Fu, J.M., Satterstrom, F.K., Peng, M., Brand, H., Collins, R.L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S.P., et al. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat. Genet. *54*, 1320–1331.
102. Gazal, S., Weissbrod, O., Hormozdiari, F., Dey, K.K., Nasser, J., Jagadeesh, K.A., Weiner, D.J., Shi, H., Fulco, C.P., O'Connor, L.J., et al. (2022). Combining SNP-togene linking strategies to identify disease genes and assess disease omnigenicity. Nat. Genet. *54*, 827–836.
103. Szustakowski, J.D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P.G., Sasson, A., Wong, E., Liu, D., Wade Davis, J., Haefliger, C., et al. (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat. Genet. *53*, 942–948.
104. Fuller, Z.L., Berg, J.J., Mostafavi, H., Sella, G., and Przeworski, M. (2019). Measuring intolerance to mutation in human genetics. Nat. Genet. *51*, 772–776.
105. Kelleher, J., Wong, Y., Wohns, A.W., Fadil, C., Albers, P.K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. Nat. Genet. *51*, 1330–1338.
106. Speidel, L., Forest, M., Shi, S., and Myers, S.R. (2019). A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet. *51*, 1321–1329.
107. Zhang, B.C., Biddanda, A., Gunnarsson, Á.F., Cooper, F., and Palamara, P.F. (2023). Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. Nat. Genet. *55*, 768–776.
108. Deng, Y., Nielsen, R., and Song, Y.S. (2024). Robust and accurate bayesian inference of genome-wide genealogies for large samples. Preprint at bioRxiv. https://doi.org/10.1101/2024.03.16.585351.
109. Rannala, B. (1997). On the genealogy of a rare allele. Theor. Popul. Biol. *52*, 216–223.