PNAS RESEARCH ARTICLE | GENETICS OPEN ACCESS

Study design and the sampling of deleterious rare variants in biobank-scale datasets

Margaret C. Steiner[a,1], Daniel P. Rice[b,c,1], Arjun Biddanda[d], Mariadaria K. Ianni-Ravn[a], Christian Porras[e], and John Novembre[a,f,2]

Edited by Nils Stenseth, Universitetet i Oslo, Oslo, Norway; received December 3, 2024; accepted April 22, 2025

One key component of study design in population genetics is the "geographic breadth" of a sample (i.e., how broad a region across which individuals are sampled). How the geographic breadth of a sample impacts observations of rare, deleterious variants is unclear, even though such variants are of particular interest for biomedical and evolutionary applications. Here, in order to gain insight into the effects of sample design on ascertained genetic variants, we formulate a stochastic model of dispersal, genetic drift, selection, mutation, and geographically concentrated sampling. We use this model to understand the effects of the geographic breadth of sampling effort on the discovery of negatively selected variants. We find that samples which are more geographically broad will discover a greater number of variants as compared to geographically narrow samples (an effect we label "discovery"); though the variants will be detected at lower average frequency than in narrow samples (e.g., as singletons, an effect we label "dilution"). Importantly, these effects are amplified for larger sample sizes and fitness effects. We validate these results using both population genetic simulations and empirical analyses in the UK Biobank. Our results are particularly important in two contexts: the association of large-effect rare variants with particular phenotypes and the inference of negative selection from allele frequency data. Overall, our findings emphasize the importance of considering geographic breadth when designing and carrying out genetic studies, especially at biobank scale.

population genetics | rare variants | negative selection

Significance

As genetic studies grow, researchers are increasingly seeking to identify rare genetic variants with large impacts on traits. In this paper, we combine theoretical methods and data analysis to show how differences in sampling with respect to geographic location can influence the number and frequency of genetic variants that are found. Our results suggest that geographically broad samples will include more distinct genetic variants, though each variant will be found at a lower frequency, as compared to geographically narrow samples. Our results can help researchers to consider the implications of study design on expected results when constructing new genetic samples.

In recent decades, the size of genetic sequencing cohorts has grown exponentially. Nowhere is this more evident than in human genetics, where the launch of biobanks has transformed the paradigm of data analysis such that sample sizes in the hundreds of thousands are increasingly commonplace (1). Yet, the largest and most commonly utilized biobank-scale genomics datasets are heavily biased toward individuals of European ancestries (2–5), leading to scientific and ethical concerns for precision medicine (6, 7). As a response to this, new biobanks have been launched with specific purposes to diversify available genomics data (8–11). Consequently, not only is the size of human genetics data continuing to increase but the geographic and genetic spaces from which individuals are sampled are growing dramatically.

This development in the field has clear benefits for improving representation in human genetics research and the transferability of results across diverse populations (12–14). What has yet to be addressed is how this change in study design will affect the results of genetic studies at the level of discovered variants. Motivated as such, we ask, as the geographic breadth of a genetic study increases, how should one expect the number and frequency of discovered variants to change? That is, how is the site frequency spectrum (SFS) of observed variants affected by the geographic breadth of a sample? The answer to this question has significant implications for studies in human genetics and genetics more broadly.

For understanding the genetic basis of traits, this question is of interest, as sample design likely impacts the discovery of genetic associations to phenotypes. A key focus of biomedical applications is discovering variants that have large effects on disease susceptibility, as such variants may provide the most biological insight into the etiology of disease and in turn potential therapeutic paths (15, 16). From evolutionary principles, one expects large effect genetic changes most often to be kept at very low population frequencies by the action of natural selection (either due to simple negative selection or via underdominance induced by stabilizing selection; 17). Indeed, rare, deleterious variants have been shown to be enriched in genomic regions of functional interest such as drug target regions (18), have yielded numerous associations with phenotypic outcomes (19, 20), and are argued to be a key component of unexplained heritability in human traits (21). How the geographic breadth of sampling impacts the discovery of

Author affiliations: [a]Department of Human Genetics, University of Chicago, Chicago, IL 60637; [b]Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]SecureBio, Cambridge, MA 02142; [d]Department of Biology, Johns Hopkins University, Baltimore, MD 21218; [e]Department of Genetics and Genomic Sciences, Icahn School of Medicine, New York, NY 10029; and [f]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637

Author contributions: M.C.S., D.P.R., and J.N. designed research; M.C.S., D.P.R., A.B., M.K.I.-R., and C.P. performed research; M.C.S., D.P.R., A.B., M.K.I.-R., and C.P. analyzed data; and M.C.S., D.P.R., and J.N. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]M.C.S. and D.P.R. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jnovembre@uchicago.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2425196122/-/DCSupplemental.

Published June 3, 2025.

these rare, deleterious variants is unknown, yet crucial to the design of studies which aim to characterize such variants.

Understanding how sample design affects the observed SFS of deleterious variants is also important to evolutionary geneticists. In evolutionary genetics, a persistent goal has been to characterize the distribution of fitness effects (DFE)—i.e., the probability with which newly arising mutations are deleterious, advantageous, or selectively neutral—using allele frequency data (22–25), in part because of its implications for genome evolution, mutational load, and conservation efforts (26). Inference of the DFE from population genetic data depends on measurements of the number and frequencies of deleterious variants (equivalently, the SFS). In addition, the SFS in large samples can be used to estimate mutation and demographic parameters (e.g., refs. 27 and 28). Thus, understanding whether and how the geographic breadth of sampling impacts the observed SFS of deleterious variants is also important in order to avoid potential biases in SFS-based inferences of evolutionary parameters.

While the geographic distribution of deleterious variants has relevance in the context of contemporary biobank-scale human and evolutionary genetics, versions of the same question have been explored in the history of population genetics. The earliest interest in this topic was framed in terms of understanding the allele frequencies of recessive lethal variants and the closely related notion of understanding rates of "allelism" of lethal variants (i.e., the rate at which carriers of distinct recessive lethal alleles fail to survive). Notably, Wright and Dobzhansky studied rates of allelism in lethal variants from spatially separated Drosophila populations as early as the 1940s (29, 30). Bruce Wallace later carried out a notable survey to characterize how the rate of allelism of lethals decays as a function of geographic distance (31) and found that the allelism of lethal mutations decayed exponentially with the square root of distance in Drosophila melanogaster samples. In a parallel line of theoretical work developing the stepping-stone model of migration, Weiss and Kimura derived an expression for the spatial covariance of allele frequencies under neutrality in two dimensions (32). Maruyama later extended this to incorporate the effects of negative selection and considered continuous spatial models (33). Motivated by the work of Wallace (31), Yokoyama modeled the rate of allelism of lethals and the closely related covariance in allele frequency as a function of distance in the stepping-stone model (34).

In a second related line of research, motivated by understanding the consequences of spatial structure and sampling on the inference of demographic history, several previous studies have investigated the effect of sample design on neutral variation (35–43). These studies emphasize how in most cases, geographically concentrated (or "narrow") sampling in spatial populations leads to a shift in the neutral SFS with a decrease in observed singletons and enrichment of intermediate and high frequency alleles (i.e. negative Tajima's D; 44).

Thus, the spatial distribution of carriers of deleterious alleles has been a topic of interest in population genetics for over eighty years. Notably, though, previous studies have not considered sample sizes on the scale of modern human biobank cohorts, which reach tens to hundreds of thousands of individuals, nor have they addressed the extent to which distortions of the SFS are amplified or diminished for rare, deleterious variants.

Here, with a focus on the discovery of rare, deleterious variants, we develop a theoretical analysis to obtain the expected counts in the sample SFS from a spatially structured population that is sampled nonuniformly. To do so, we utilize recent mathematical results on the stationary distribution of subcritical measure-valued branching processes with immigration by Friesen et al. (45). The model considers the distribution of carriers of deleterious alleles in continuous geographic space—accounting for dispersal, genetic drift, selection, mutation, and sampling simultaneously—and we derive results that allow the rapid computation of the expected sample SFS across a large range of parameter values

As important background, we note that in the panmictic (or "well-mixed") case, allele frequencies for deleterious variants are well known to follow a two-parameter distribution, such that the probability density $g(x)$ for an allele under negative selection appears at frequency x follows (46, 47):

$$g(x) \propto e^{-\gamma x}[x(1-x)]^{\theta-1}, \quad [1]$$

where γ is the population-scaled selection coefficient ($\gamma = 4N_e s$ with N_e being the effective population size and s is the strength of negative selection acting on heterozygotes, $s \geq 0$) and θ is the population-scaled mutation rate ($\theta = 4N_e\mu$ with mutation rate μ per site per generation). For variants under negative selection ($\gamma > 0$), the exponential term ($e^{-\gamma x}$) induces a reduction in the abundance of observed alleles as a function of the allele frequency x. This mirrors the intuition that alleles under negative selection are less likely to reach higher frequencies.

As we will show, when considering spatially restricted dispersal and geographically concentrated sampling, allele frequencies still follow a two-parameter distribution with scaled "effective" selection (γ_E) and mutation (θ_E) terms. However, these terms are now dependent on the spatial scale of the sampling effort and the offspring dispersal scale, in addition to the usual mutation, selection, and population size parameters. We subscript the parameters with "E" to emphasize that the effects of spatial sampling can be obtained by using these effective parameters in the standard formulas for the SFS of negatively selected alleles. The resulting distributions show that the geographic breadth of a sample has strong effects on the SFS as well as downstream summary statistics, and we assess how these effects change with increasing sample size and selection strength.

We validate our theoretical results using simulations that share our modeling assumptions as well as in a more realistic, individual-based spatial model (42, 48). As continuous-space simulations can be computationally intensive, our development of theoretical approximations allows us to efficiently gain insights across wide parameter ranges without using simulation.

To address the effects of geographically concentrated sampling empirically, we also conduct in silico resampling experiments using the UK Biobank exome sequencing dataset (19) to measure the impact of sampling at different geographic scales. The results broadly confirm our theoretical predictions and yield insights into how sampling design impacts the number of human genetic variants discovered and their frequencies.

Methods

Population Genetic Model. We model how each instance of a rare allele is born, moves in space, reproduces, and dies in a two-dimensional continuous geographic habitat of size $L \times L$. In our model (Fig. 1), deleterious alleles are generated by de novo mutation according to a Poisson point process with rate $\rho_N\mu$, where ρ_N is the population density and μ is the per-generation mutation rate. We note that, in our model, ρ_N and L are constants, which implies an assumption of constant population size. Each de novo allele appears at a random location in the habitat and migrates according to a homogeneous, isotropic diffusion process. In other words, the dispersal process is the same across the habitat and movement is uniform in all directions. With diffusion coefficient σ^2 and time measured in generations, the per-generation displacement in two-dimensional space follows a bivariate Gaussian distribution with SD σ in each

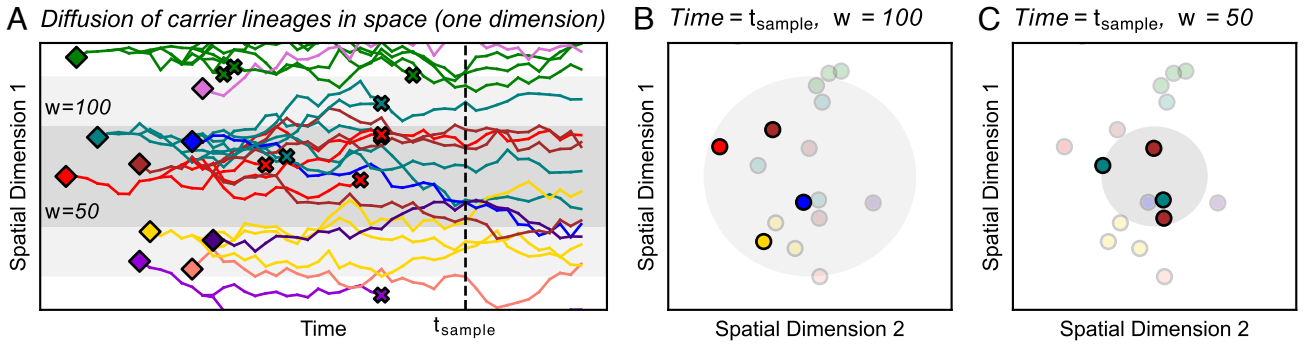

Fig. 1. Illustration of a spatial branching process model with sampling. (*A*) As time progresses, carrier lineages move in space (diffusion), branch into sublineages (reproduction), and die. Diamonds and X's denote lineage birth and death, respectively. For simplicity, we show only one spatial dimension on the vertical axis. Shaded areas represent sampled areas with widths $w = 50$ and $w = 100$. (*B* and *C*) Here, we visualize the locations of carriers from panel (*A*) at a particular time indicated as t_{sample}. Within the sampled area, rare variant carriers can be discovered and included in the sample (opaque points). In this example, sampling from the broader area ($w = 100$) results in a greater number of distinct mutations being observed, all as singletons. The narrow sample ($w = 50$) discovers two distinct mutations with each as a pair of doubletons. This toy example illustrates the potential effects of sampling breadth on entries of the sample SFS (here, the counts of singletons vs. doubletons).

dimension (with zero covariance). Following other classical analyses of spatial models (32, 49, 50), we assume for simplicity that the habitat has periodic boundary conditions (i.e., the habitat is a two-dimensional torus, and so there will be no boundary effects).

Within the habitat, we model reproduction and death as a continuous time branching process, a type of stochastic process which has frequently been applied in theoretical population genetics for rare variants (see, for instance, refs. 51–54). During their lifetime, each instance of a deleterious allele gives rise to offspring alleles with rate $1 - s$ and dies at a rate 1. A larger positive value of s indicates stronger negative selection. We note that our term s is equivalent to the hs parameter used in alternative parameterizations of selection where h is a general dominance parameter and s is a selection coefficient describing the fitness of homozygous carriers (e.g., ref. 55). We assume deleterious variants are rare enough that the fraction of homozygous carriers is negligible and alleles can be approximated in a haploid model with alleles evolving independently. Also, while we model negative selection on individual variants, these dynamics are similar to those of newly arising variants which affect complex traits under stabilizing selection (which experience a form of underdominance, see ref. 17). The use of a branching process model implies that each allele copy evolves independently of one another and of the ancestral alleles (similar to refs. 56–58). In the context of continuous space, the assumption of independence of the branching process and dispersal process are approximate (e.g., see ref. 50), but hold reasonably when every mutation is locally rare.

Modeling Geographically Concentrated Sampling. The spatial model in the previous section describes the process by which rare variants arise and disperse in geographic space. Our next step is to define how sampling of this spatial population occurs. To this end, we model the probability that an individual at a particular position within the habitat is included in the sample. We posit a sampling "center" and assume that the probability of an individual being sampled is determined by the individual's distance from that center using a particular distribution (the "sampling kernel;" *SI Appendix*, Fig. S1). The SD of the sampling kernel, which we denote by w, determines the breadth of sampling effort–or "sampling breadth"–i.e., the extent to which sampling effort is distributed across the habitat. On one extreme, for $w \gtrsim L$, the sampling process converges to "uniform" sampling in which all individuals have an equal probability of being sampled, regardless of spatial position. In the other extreme, as w becomes small, the sampling kernel approaches "point sampling" in which all sampled individuals are located at the same position. In between these two limiting cases, the value of w will determine how spatially "broad" (larger w) or "narrow" (smaller w) a sample will be.

In our implementation, the sampling kernel has the form of a Gaussian distribution with SD w, though we note that our methods are generalizable to other sampling kernels. We employ the Gaussian sampling model to approximate the sampling processes used in constructing real genetic samples, such as sampling centered at field stations for ecological genetics or in biomedical centers for human genetics. As in our population genetic model, the sampling model has periodic boundary conditions (i.e., there are no "edge effects" in our model). This construction is appropriate when the habitat size, L, is sufficiently large compared to the sampling breadth, w, such that we can ignore behavior as the sampling kernel hits the edge of the habitat. Our simulations will show that in cases where w approaches the scale of L, the results converge to those of uniform random sampling.

Theoretical Analysis Methods. In order to solve for the SFS, we frame the problem as special case of a spatial subcritical measure-valued branching process (see refs. 54, 59, and 60). Applying results of Friesen et al. (45), we obtain a nonlinear partial differential equation (PDE) governing the spatially weighted population allele frequency distribution at stationarity. We solve this PDE perturbatively to obtain the mean and variance of the spatially weighted allele frequency. Based on the uniform sampling case and exploratory simulations, we inferred that the full stationary distribution approximately follows a Gamma distribution, and we find the parameters of the Gamma distribution by matching moments.

Having derived the stationary distribution of spatially weighted population allele frequencies, it remains to obtain the SFS for a sample of finite size. For a large sample size n, the number of sampled rare alleles conditional on the population allele frequency approximately follows a Poisson distribution. Then, by a property of Gamma-Poisson mixture distributions, the per-site SFS (i.e. the probability a given site takes on an observed count of 0, 1, . . . , n) follows a Negative Binomial distribution with parameters determined by the shape and rate of the Gamma distribution as well as the sample size n. Last, we use the finite-sample SFS results to derive expectations of several major population genetic summary statistics.

We refer readers to the extended theoretical methods in *SI Appendix* for a detailed description of these methods.

Population Genetic Simulations. We validate our theoretical results with two sets of simulations. First, we simulate a spatial branching process in a two-dimensional continuous habitat and sample according to a Gaussian sampling density, as our theory assumes. These simulations are close to our theory in that they make the same rare-allele approximation. Their role is to check the analytical approximations we make in the course of deriving the SFS.

In addition to the branching process simulations, which align directly to our theoretical model, we carried out more realistic forward-time, individual-based population genetic simulations in SLiM (48) using identical conditions to ref. 42 except that all variants are deleterious with some selection coefficient. In contrast to our other simulations, the SLiM model contains multiple stages of the life cycle, models diploid genomes, and–crucially–does not assume variants are rare and independently evolving.

We refer the reader to *SI Appendix* for additional details on simulation methods. All simulation code and associated scripts are available at: https://doi.org/10.5281/zenodo.15398319 (61).

Analysis of Whole Exome Sequencing Data from the UK Biobank. We perform resampling experiments using the whole exome sequencing dataset ($n = 469{,}835$) in the UK Biobank (UKB; 19) in order to assess the predicted effects of sampling breadth on sample allele frequencies and associated summary statistics. After including only individuals which met quality control and relatedness thresholds used by Bycroft et al. (4), we subset the data to individuals both born within the United Kingdom and having similar genetic ancestry (specifically, individuals within 0.0001 units in Euclidean distance from the centroid in the normalized PC1-PC2 space; applying both filters results in $n = 231{,}073$ individuals). We then use a sampling importance resampling (SIR) method to construct $n = 10{,}000$ samples such that the distribution of birthplace coordinate bins is Gaussian with centers centered at each of three geographic points with SD (w) set to 50 km, 100 km, and 150 km, as well as a uniform sample (see *SI Appendix* for details on the sampling algorithm). We repeat the sampling procedure ten times for each sampling width and center.

For each weighted subsample, we compute the SFS for putative LoF sites on chromosome 1 (32,320 variants) as well as equal-sized random subsets of synonymous and missense variants (subsets generated using PLINK v1.90b6.26; variant annotation provided by UKB). We then calculate summary statistics (number of variant sites, number of singletons, and heterozygosity). We refer the reader to the *SI Appendix* for additional details.

Results

The Finite Sample SFS Depends on Ratios Between Spatial Scales as Well as Sample Size. In our model, a key emergent feature is the length scale across which an allele spreads during the time from the initial mutation to the extinction of all its descendant copies, which we denote as ℓ_c, the characteristic length scale for this problem. As the lineages diffuse with coefficient σ^2 and alleles go extinct on a time-scale of $1/s$ generations, $\ell_c = \sqrt{\sigma^2/s}$. We note that the same ratio arises in related but different models in the spatial population genetics literature (for instance, refs. 62, 63). In our model, alleles that spread more quickly (large σ) will move farther during the lifespan of the allele (ℓ_c is large). Conversely, alleles which are under stronger negative selection (large s) die more quickly and thus the descendant alleles will have spread over shorter lengths (ℓ_c is small). As we will see later in our results, how the spatial scale of the sampling effort (w) compares to the spatial spread of the allele (ℓ_c) will be an important factor in the behavior of the SFS.

In order to derive the form of the SFS for a finite sample of size n with sampling breadth w, we first consider the distribution of allele frequencies across the entire spatially extended population with a weighting on each position provided by the sampling kernel. In a panmictic population, the population SFS of rare, deleterious alleles approximately follows a Gamma density (by ignoring the $x \to 1$ tail of Eq. **1**). We show analytically that this approximation holds for spatially uniform samples under our model (*SI Appendix*) and confirm via simulations that allele frequencies of spatially concentrated samples are also well-approximated by a Gamma distribution (*SI Appendix*, Figs. S2 and S3). Intuitively, the Gamma distribution captures two important effects: power-law behavior at low frequencies due to mutation-drift balance and an exponentially decaying tail at high frequencies due to selection.

We analyze our model in order to obtain the two parameters of this Gamma distribution, which we refer to as the effective mutation supply, θ_E, and the effective selection intensity, γ_E. Then, we derive an expression for the expected SFS of a finite sample with size n in terms of these parameters. First, let the random variable K denote the count of derived alleles at a single site in a sample of size n. Given a Gamma density for the allele frequencies, and approximating the binomial sampling process for large n as a Poisson distribution, K follows a Negative Binomial distribution with number of successes θ_E and success probability $\gamma_E/(\gamma_E + n)$:

$$K \sim \text{NegBin}\left(\theta_E, \frac{\gamma_E}{\gamma_E + n}\right). \quad [2]$$

The k-th entry of the normalized sample SFS is then given by $\xi_k^{(n)} \equiv \Pr(K = k)$. In the limit that θ_E is small, this becomes approximately:

$$\xi_k^{(n)} = \frac{\theta_E}{k}\left(\frac{n}{\gamma_E + n}\right)^k. \quad [3]$$

In the case of spatially concentrated sampling ($w \ll L$), we show (*SI Appendix*) that the effective mutation and selection terms are given by

$$\theta_E \equiv \mu\rho_N\ell_c^2\lambda, \quad [4]$$

and

$$\gamma_E \equiv s\rho_N\ell_c^2\lambda, \quad [5]$$

where $\rho_N\ell_c^2$ is akin to a population size. We denote by λ a term we refer to as the sampling effect scalar, which is a function of w/ℓ_c:

$$\lambda \equiv \frac{4\pi}{\exp((w/\ell_c)^2)E_1((w/\ell_c)^2)}, \quad [6]$$

where $E_1(x) = \int_1^\infty (e^{-tx}dt)/t$ is the exponential integral function.

We note that in Eq. **5**, the leading selection coefficient s cancels with the $1/s$ term in ℓ_c^2, meaning that the only dependence of either effective parameter on selection is via λ.

The value of λ increases as w increases (Fig. 2*A*) and depends on the ratio between the sampling breadth and characteristic length scale: w/ℓ_c (Fig. 2*B*). When w/ℓ_c is small ($\ll 1$), λ is approximately $-2\pi/\ln(w/\ell_c)$ and thus is only logarithmically dependent on w and s in this regime. Conversely, as w/ℓ_c becomes large, λ becomes approximately $4\pi(w/\ell_c)^2$ (Fig. 2*B*), in which case the effective parameters simplify to $\theta_E = \mu\rho_N 4\pi w^2$ and $\gamma_E = s\rho_N 4\pi w^2$ (Fig. 2 *C* and *D*). In this regime, the effective parameters scale quadratically with w, and are the equivalent to what one would expect for a panmictic population of size $\rho_N\pi(2w)^2$, which is interpretable as the size of a population of density ρ_N found in a circle whose radius is given by two times the sampling breadth ($2w$).

As w approaches L, both effective parameters eventually converge to their values in the uniform sampling limit ($N\mu$ and Ns, respectively, where N is the total population size, similarly to Eq. **1**; Fig. 2 *C* and *D*). To understand this result, one can think of the term $\rho_N\ell_c^2\lambda$ as approximating the size of the population effectively being sampled, which grows as $\rho_N\pi(2w)^2$ before converging to $\rho_N L^2 = N$ as sampling approaches the uniform case.

Selection and Sampling Induce a Trade-Off Between Discovery and Dilution. Having derived an expression for the sample SFS, we now consider its behavior with respect to sampling breadth (Fig. 3 *A* and *B*). As expected from the effective parameters, we find that broader sampling effort (larger w) induces an upward shift in the intercept of the SFS on the vertical axis. We also observe a decrease in the frequency of segregating variants for broader samples. These results arise from what we call a

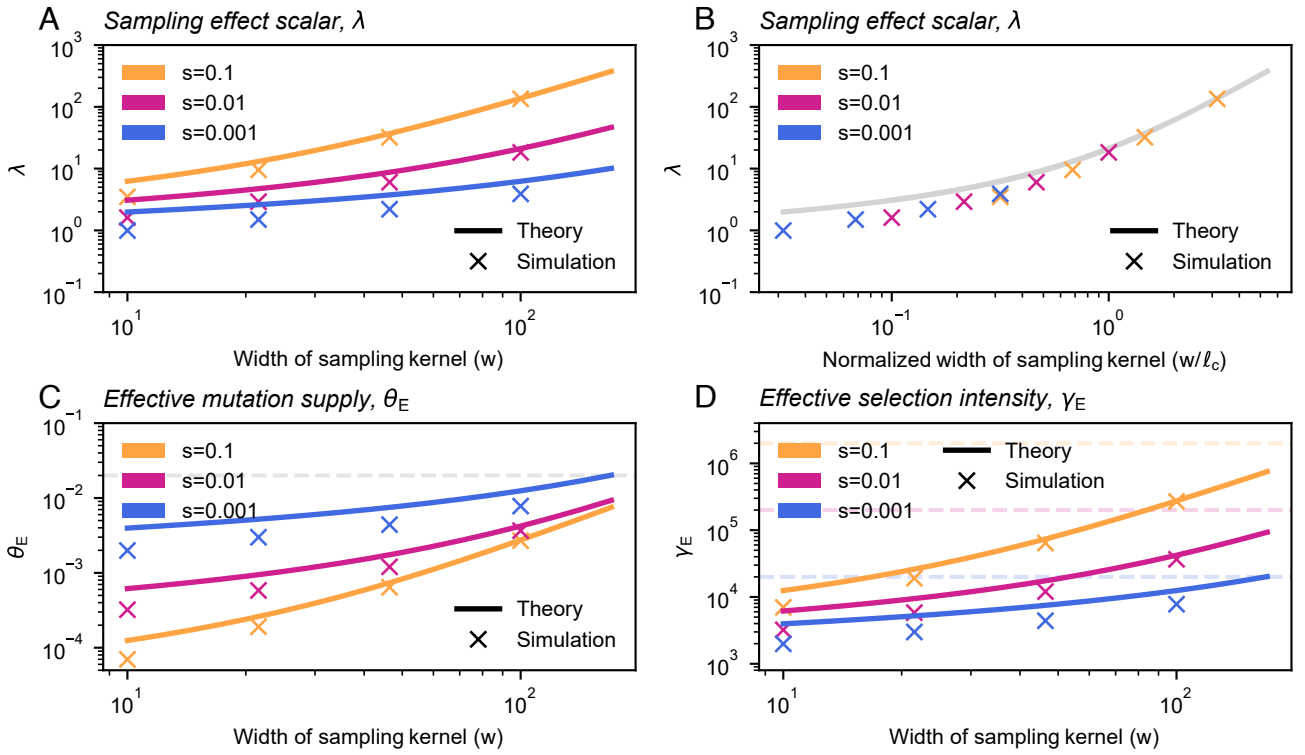

Fig. 2. The sampling effect scalar and effective parameters of the SFS. (*A*) As the breadth of the sampling kernel (w) increases, the sampling effect scalar λ also increases. (*B*) When plotted as a function of w/ℓ_c, the relationship is identical across a range of selection intensities. (*C* and *D*) Both the effective mutation supply, θ_E, and the effective selection intensity, γ_E, depend on the selection coefficient (via ℓ_c) and the breadth of the sampling kernel. Dashed lines show values of θ_E and γ_E for the uniform case in panels (*C* and *D*), respectively. Other parameters are $\sigma = 10$, $\rho = 20$, and $\mu = 10^{-9}$. Branching process simulations shown are from the Gillespie algorithm run with a habitat size of $L = 1{,}000$.

"discovery" effect and a "dilution" effect. As the geographic breadth of a sample increases, the number of potential localized mutations one can find grows, and this is reflected in the increase in mutation supply (θ_E) and the expected number of variants discovered (the discovery effect). At the same time, for a broader sample, each sampled deleterious variant is found at "diluted" frequencies because sampling broadly inadvertently captures many carriers of the alternative allele at each site (dilution effect). This is reflected in the larger γ_E term and concomitant faster rate of decay of the number of observed sites of count k as k increases. As sampling effort broadens (w increases), the SFS converges to the result under uniform sampling. Conversely, geographically narrow samples will capture fewer variants, but they are "concentrated" in the sample, meaning that they are observed at higher sampled allele frequencies on average than they would be found otherwise in a uniform sample.

Besides displaying the SFS directly, we visualize these results by comparing each entry of the SFS of a sample with breadth w to the SFS in the uniform limit (Fig. 3 *C* and *D*). The results show how relative to a uniform sample, as one makes the sample increasingly narrow in geographic breadth (w decreases), more deleterious alleles are found at higher allele counts and fewer are found at the lowest allele counts. At some value of allele count, the frequency spectra for a sample with breadth w and a uniform value are equal (which we see as the intersection between curves in Fig. 3 *A* and *B*), and the allele count at which the frequency spectra intersect is dependent on the sample size, n: For smaller samples, the intersection point is low, perhaps at the level of singletons; for larger samples, the intersection point is higher and the magnitude of effect in the low allele count range is also much larger (Fig. 3).

As expected, the changes to the observed SFS with increasing sampling breadth have varying effects on downstream summary statistics (Fig. 4). Under our model, the expected allele frequency in the sample is equal to μ/s regardless of sampling strategy, including sample breadth and sample size (Fig. 4 *A* and *E*). This indicates that the discovery and dilution effects effectively cancel each other out, such that the average frequency (and in turn the expected heterozygosity) remains the same regardless of sampling (*SI Appendix*, Fig. S10).

However, other quantities of interest are indeed sensitive to the sampling breadth. First, we see that broader samples are expected to have a greater proportion of variant sites (Fig. 4 *C* and *G*). Second, the variant sites in broader samples are expected to have lower allele frequency (Fig. 4 *B* and *F*), as indicated by a higher fraction of singletons (Fig. 4 *D* and *H*) as compared to narrower samples. Together, these two results imply that broader samples will include more variants, but each variant will segregate at lower frequency on average. This result is consistent with intuition following the discovery and dilution effects described previously.

The exact behavior of these statistics with respect to w depends on the strength of selection, albeit weakly (Fig. 4 *A*–*D*). With stronger selection, the observations converge to those expected under uniform sampling more rapidly as w is increased. When instead considering the values of expected summary statistics over the ratio w/ℓ_c, we see that the rate and magnitude of change are consistent across selection coefficients (*SI Appendix*, Fig. S11). This is a result of the ratio w/ℓ_c being the key length scale in our model (Fig. 2*B*): For a fixed w, stronger selection reduces ℓ_c because allele carriers are more tightly clustered in space, and as a result, the sample is in effect more broad relative to the spatial dispersion of the carriers. Conversely, with the strength of

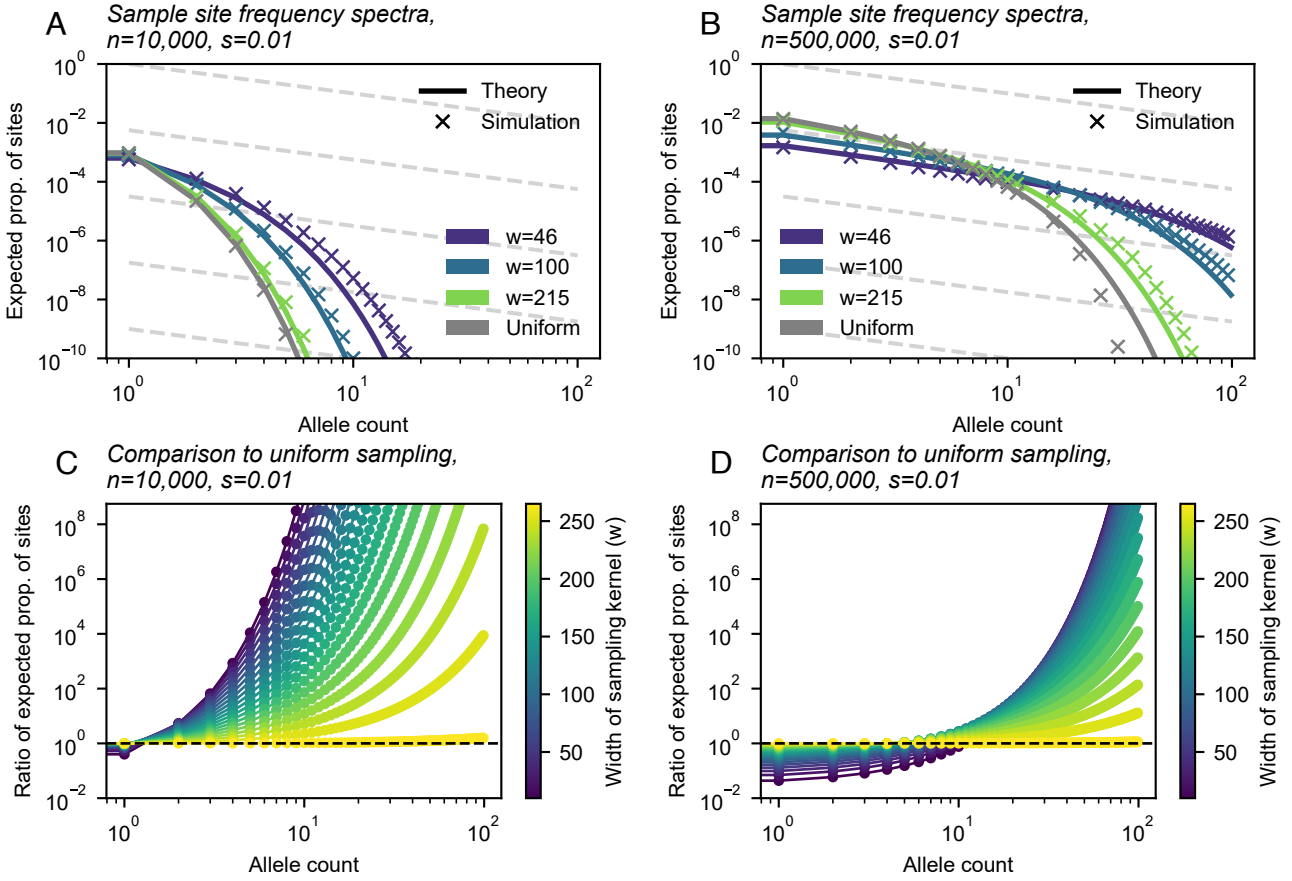

Fig. 3. Site frequency spectra for a range of sampling widths. (*A* and *B*) Sample site frequency spectra for $n = 10{,}000$ and $n = 500{,}000$ sample sizes, respectively, shown for three sampling breadths (wrapped Gaussian sampling) and uniform sampling. (*C* and *D*) Ratio between frequency spectra elements for a range of w values (Gaussian sampling), relative to those of the SFS under uniform sampling, for parameter regimes shown in (*A* and *B*). For all panels, $\sigma = 10$, $\rho = 20$, $\mu = 10^{-9}$, and $s = 0.01$. Branching process simulations shown were run with a habitat size of $L = 1{,}000$.

selection held constant, increasing w results in the sample being more broad relative to the spatial dispersion of carriers.

Holding selection constant, we also see that the magnitude of effect as w increases becomes larger as n increases (Fig. 4 *E*–*H*). These effects span several orders of magnitude.

In- and Out-of-Model Simulations Validate the Results of Theoretical Analysis. To validate our theoretical results, we performed two sets of population genetic simulations. First, we ran branching process simulations which correspond directly to our model. Inspecting the results, we see that the simulations and theoretical computations generally align for key outputs: the first two moments of the allele frequency (*SI Appendix*, Fig. S4), the joint parameters λ, θ_E, and γ_E (Fig. 2), as well as the SFS itself (Fig. 3 and *SI Appendix*, Figs. S2 and S3). We observe some deviation from simulation results and theoretical expectations in two regimes (*SI Appendix*, Fig. S4): weaker selection regimes, in which we expect violations of the modeling assumptions regarding the rareness of the alleles and absence of homozygous carriers, and in regimes of very small sampling widths (w close to σ), which approach spatial scales where our measure-valued process may become a poor approximation to the reality of discrete individuals.

As a stronger test of the theory, we performed simulations in SLiM (48) using a modified version of the model of Battey et al. (42). These simulations are individual-based and do not make a number of the simplifying approximations used in our theoretical analysis. Fig. 5 shows the alignment between these simulations and our theoretical results across several parameter values. We generally see quite close alignment, with the exception of higher allele frequencies for the narrow sampling kernel ($w = 1.22$) in stronger selection regimes ($s = 0.1$) in which case our theory overestimates the number of variants relative to SLiM simulations.

These comparisons also reveal how computational efficiency varies greatly across the different approaches. SLiM simulation time ranged between 7.58 and 11.89 h (average: 8.64 h) per replicate for a habitat of length 75 units (50 replicates run per sample size and selection coefficient pair). On average, the branching process simulations completed in 18.59 min, 1.22 h, and 2.85 h per one million time steps for $s = 0.1$, 0.01, and 0.001, respectively for a habitat size of 10,000 units. In contrast, the time to generate theoretical frequency spectra shown in Fig. 3 ranged from 6.88 to 9.91 milliseconds per curve.

Resampling Experiments in the UK Biobank Reveal Evidence of Discovery and Dilution Effects. Having identified relationships between the spatial breadth of sampling effort and observed variant frequencies in our theoretical work, we now consider to what extent these patterns would arise in human genetic studies.

To do so, we artificially mimic sampling designs that vary in geographic sampling breadth via in silico subsampling ($n = 10{,}000$) of individuals from the large ($n = 469{,}835$) exome sequencing dataset of the UKB, using sequencing data from

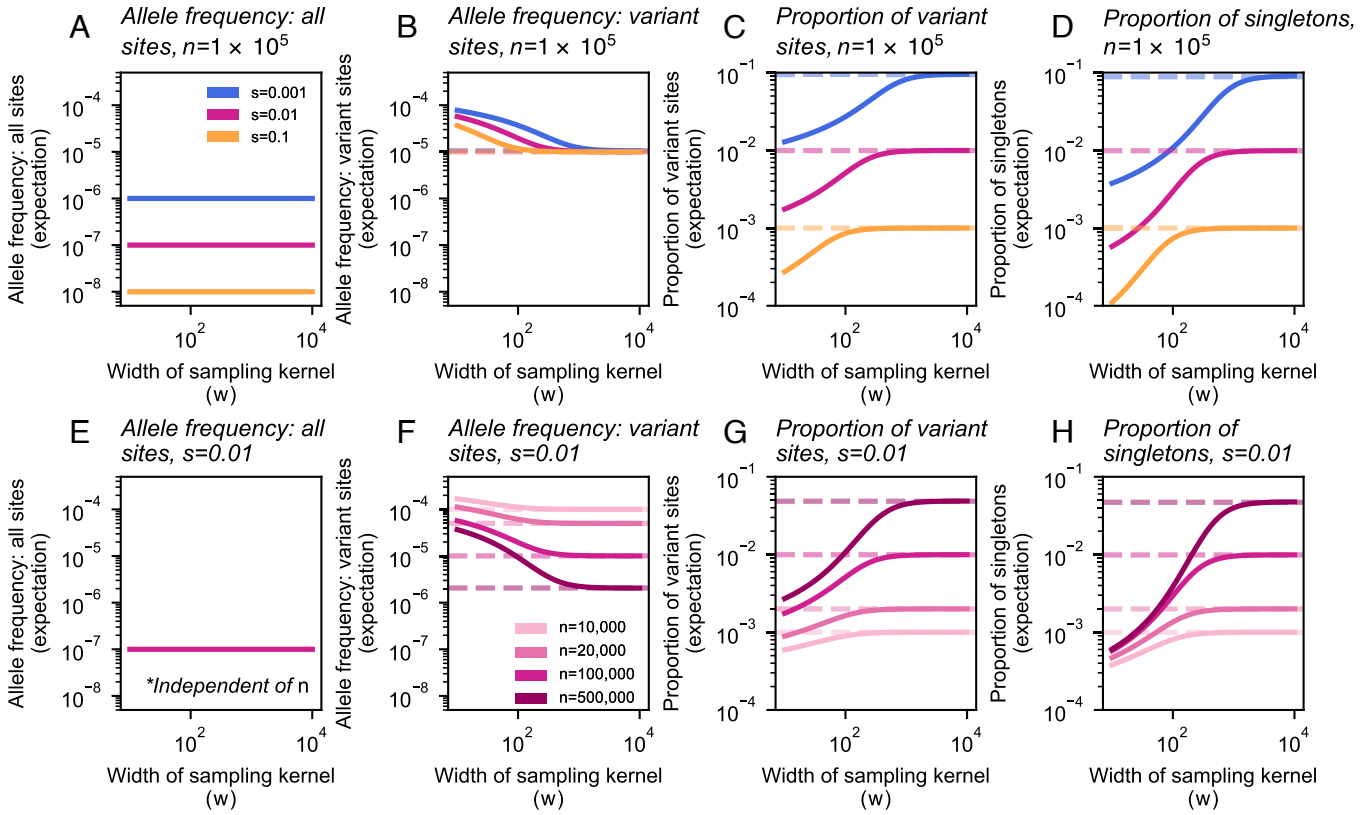

Fig. 4. Expected values of summary statistics as a function of the width of the sampling kernel (w). (*A–D*) As the breadth of the sampling kernel increases, our model implies that the expected frequency across all sites will remain constant, though the expected frequency of variant sites will decrease. The expected proportion of variant sites and the expected proportion of singletons will both increase. Values for these statistics (excluding expected frequency across all sites, *A*) converge to the theoretical expectation under uniform sampling (dashed lines) as w increases, with convergence occurring more quickly for stronger selection coefficients. For more deleterious variants (larger s), the expected proportions of variant sites and singletons, as well as expected frequency across all sites, are shifted downward across the range of w. (*E–H*) Fixing s and instead varying sample size (n), we see that the magnitude of change as w increases grows with sample size (n), with the exception of expected allele frequency across all sites, which is independent of n. In plots shown, $\sigma = 10$, $\rho_N = 20$, and $\mu = 10^{-9}$.

Chromosome 1 (19). More specifically, we constructed samples spanning geographic scales of $w = 50$ km to uniform sampling based on birthplace data within the island of Britain for individuals with high genetic similarity to the centroid of the UK Biobank cohort (Fig. 6 *A–D* and *SI Appendix*, Fig. S14). In constructing the samples, we adjusted for the underlying spatial heterogeneity in sampling density in the UKB using sampling importance resampling (*SI Appendix*, Fig. S13; 64).

Visualizing frequency spectra from these samples, we clearly observe differences in the SFS according to sampling breadth (Fig. 6 *E* and *F*). For instance, we observe an intersection between the SFS for uniform vs. narrow samples similar to that in our theoretical results (Fig. 6*E* and *SI Appendix*, Figs. S16 and S17). Additionally, when we compute the ratio of SFS values for the narrow sample ($w = 50$ km) vs. uniform sample, we identify a pattern similar to that seen in theory for rare alleles (counts less than 10; Fig. 6*F* and *SI Appendix*, Fig. S20).

Consistent with our theoretical predictions, as sampling scale becomes broader, one generally observes more variant sites, more singletons, and lower mean heterozygosity at variant sites, while mean heterozygosity across all sites is largely insensitive to changing sampling breadth (Fig. 6 *G–J* and Table 1). The shifts in these summary statistics as sampling breadth increases are larger for the more putatively deleterious variant categories, though the differences are relatively small (Table 1). This is consistent with the theoretical results, which convey a dependence of the effects of sampling on the strength of selection that is relatively weak. The scale of the effects is such that for samples of size $n = 10{,}000$, moving from $w = 50$ km to uniform sampling leads to on average 72.3% more discovered LoF variants per kb with a 36.75% reduction in heterozygosity at variant LoF sites (Table 1).

A notable deviation from our theoretical predictions is that we observe a convergence of the larger allele counts in the empirical SFS across different sampling strategies (Fig. 6 *E* and *SI Appendix*, Figs. S16 and S17). We speculate this is due to how variants with larger allele counts in the narrow sample are old enough to have been broadly shared across the scale of Britain, as they possibly predate the genetic structure present in Britain and are thus plausibly less affected by sampling breadth.

As an additional in silico experiment, to emulate sampling on scales greater than the spatial scale of the island of Britain, we also construct subsamples across the UK Biobank at different scales of sampling breadth based on positions in PC1-PC2 space. Though the shift in the SFS is more subtle, we again find that as sampling breadth scale increases, the total number of variants increases, with a concomitant decrease in the heterozygosity of variant sites, such that the expected heterozygosity over all sites remains constant (*SI Appendix*, Fig. S15).

Finally, we carried out an exercise using maximum likelihood estimation to fit the ρ_N, σ, and s parameters to see whether the simple model we assume can plausibly fit the data and what scales the fitted parameters would be (*SI Appendix*, Figs. S22 and S23). The fitted parameters $\hat{\sigma} \approx 21.37$ to 51.92km, $\hat{\rho}_N \approx 0.31$ to 1.53 per km^2, $\hat{s_M} \approx 0.00702$ for missense variants and $\hat{s_L} \approx 0.01$ to 0.0119 for LoF variants are arguably within plausible ranges given

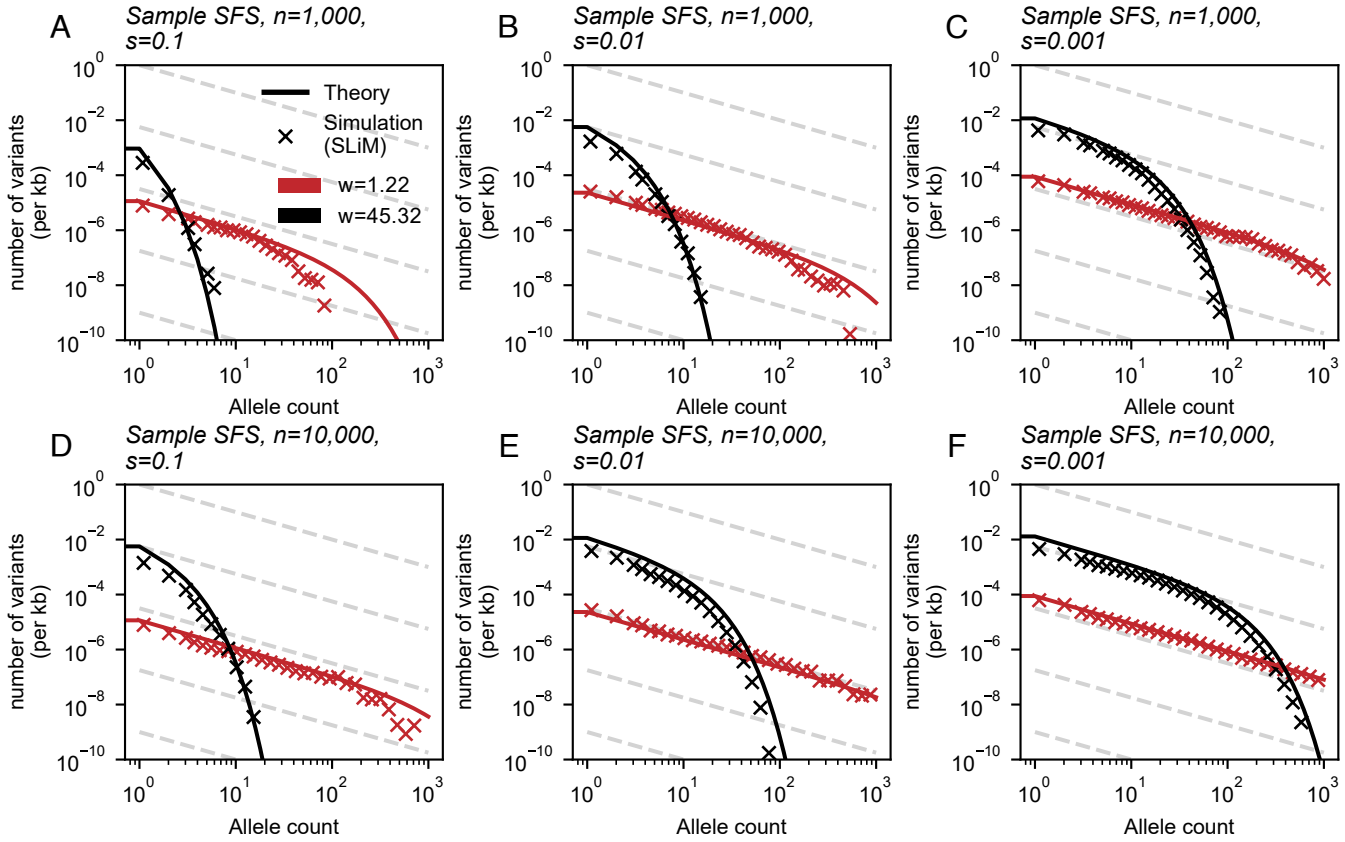

Fig. 5. Simulations in SLiM compared to expected results under the model. SLiM simulations were performed using a modified version of the model of Battey et al. (42). After simulation, samples of size $n = 1{,}000$ (*A–C*) and $n = 10{,}000$ (*D–F*) were taken at varying sampling widths. Simulations shown were run using a habitat of width 75 units, population density of 5 individuals/unit squared, and deleterious mutation rate of 10^{-10} per basepair per generation. Frequency spectra shown are averaged over 100 sampling iterations. Parameters for theory results are directly matched to those of the simulations.

historical estimates (see *SI Appendix*; 65, 66). However, $\hat{\rho}_N$ and $\hat{\sigma}$ are sensitive to the inclusion of singletons and doubletons, and the expected SFS using the best fit values still deviates from the observed SFS, suggesting the simple model should be extended for fitting the data more fully (e.g. to accommodate nonstationary population dynamics).

Discussion

Here, we have addressed the question of how the geographic "breadth" of sampling effort in genetic studies impacts the discovery of rare, deleterious genetic variants using a theoretical approach. Our analysis shows how sampling affects the expected site frequency spectrum via both discovery and dilution effects: Geographically broad samples will discover a greater number of variants but each will be observed at lower allele frequencies than in narrow samples. In contrast, geographically narrow samples will include fewer variants, and these variants will appear concentrated in the sample, often at frequencies above what would be found in uniform samples.

In several ways, our results echo the impacts of sampling on neutral variation described in previous studies: Spatially broader samples tend to discover more variant sites overall; however, these alleles tend to be singletons and other low-frequency alleles (36–38, 42, 43, 67). Using our approach, we can directly account for and vary the strength of negative selection, and observe its effects on the frequency spectra (Figs. 3 and 5 and *SI Appendix*, Figs. S6 and S7) and summary statistics (Fig. 4) of selected alleles. The more deleterious a class of variants is, the smaller the spatial scale of their spread (ℓ_c) will be, and in turn, the effects of increasing sample breadth saturate most quickly for more deleterious variants. That is, for more deleterious sites, the discovered alleles will be as diluted as they would be in a fully uniform sample at a comparatively smaller scale of sampling.

A notable outcome in the model is that expected allele frequency (and in turn expected heterozygosity) is constant with respect to sampling breadth. The result also holds empirically across several annotation categories in our analysis of the UK Biobank. For instance, with a sample size of 10,000, our experiments show broad resamples of the UK Biobank have on average 72.3% more variant sites (and 146.7% more singletons, for LoF variants), but 36.75% lower heterozygosity at variant sites than when samples are narrowly concentrated (Table 1 and Fig. 6), yet either sampling design will see an average heterozygosity of 0.013%. While perhaps not immediately obvious from observing how the SFS changes with sampling breadth, the result can be understood as an expected outcome of our model (assuming birth/death/mutation rates are constant and independent of geography) and other models of deleterious allele frequency evolution in the presence of variable demographic histories (68). Indeed, recent work by Stolyarova et al. (69) identifies similar patterns across sampling groups in the gnomAD dataset (5), and explains how the result is expected from considerations of how under mutation–selection balance, the average frequency of deleterious alleles is expected to be similar across populations.

Our results have important implications for two major areas of research that use observations of rare, deleterious variants: i) genetic association studies and ii) evolutionary genetic inferences

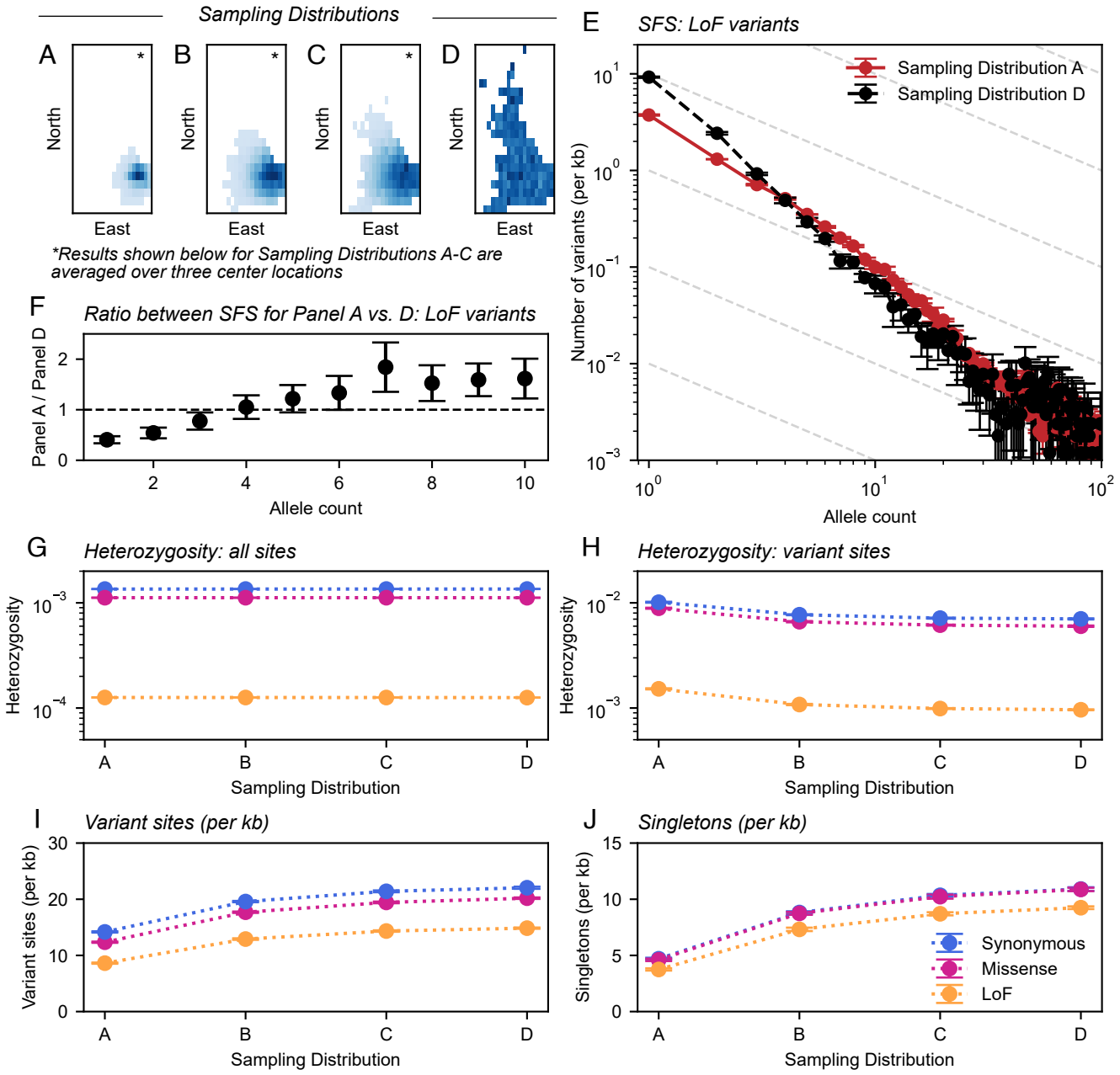

Fig. 6. In silico resampling experiments in the UK Biobank using exome sequencing data from chromosome 1. Panels (*A–C*) depict outputs of the SIR procedure, where the distribution of birthplace coordinates follows a discretized Gaussian distribution. Sampling SDs (w), in order, are 50 km, 100 km, and 150 km. Gaussian sampling was repeated over three center locations (*SI Appendix*, Fig. S14) with averages across centers shown in other panels. Panel (*D*) depicts uniform sampling in this scheme. (*E*) Average folded sample SFS for putative LoF sites on chromosome 1 across sampling distributions, truncated at allele count of 100. (*F*) Ratio between SFS values between panels (*A* and *D*) for singletons through ten-tons. (*G–J*) Variation in summary statistics across sampling distributions and variant annotations. All results are averaged over ten sampling replicates with error bars representing two SEs.

of fitness effects. In the next paragraphs, we discuss our results within these two contexts.

In genetic association studies of disease phenotypes and complex traits, observed frequencies are intrinsically tied to statistical power. GWAS power is roughly linear in allele frequency for low frequency alleles, as $x(1-x) \approx x$ for small x, and the cumulative MAF that impacts power in burden tests is also directly dependent on the average allele frequency. While many studies have considered the impact of increasing sample size on power, our results suggest interesting trade-offs related to geographic sampling breadth. Discovering a greater number of variants in broader samples will expand the space of potentially identifiable associations. Yet, the dilution of allele frequency seen in broader samples will hinder power to detect an association to each particular variant using single-variant GWAS designs. These effects perfectly compensate for one another in terms of their effects on average allele frequency observed at deleterious sites, suggesting negligible impact overall on the rate of detecting phenotypic associations.

Such implications are tentative though—more in-depth analyses of the impacts on GWAS and burden test power are needed which consider factors not addressed by our model. Such factors include linkage disequilibrium patterns, corrections for population stratification, the increased rate of cryptic relatedness in narrow samples, increased variance due to nongenetic factors in broad samples, and the effects of recent human population

Table 1. Relative change in summary statistics between narrow ($w = 50$ km) and uniform sampling distributions in birthplace (panels *A* and *D*, Fig. 6) based on resampling experiments across variant types

Variant type	$w = 50$ km	Uniform sampling	Relative change
Heterozygosity (all sites)			
Synonymous	1.352×10^{-3}	1.354×10^{-3}	0.15%
Missense	1.120×10^{-3}	1.119×10^{-3}	−0.09%
LoF	1.260×10^{-4}	1.256×10^{-4}	−0.32%
Heterozygosity (variant sites)			
Synonymous	1.014×10^{-2}	7.038×10^{-3}	−30.59%
Missense	8.889×10^{-3}	5.999×10^{-3}	−32.51%
LoF	1.522×10^{-3}	9.626×10^{-4}	−36.75%
Variant sites (per kb)			
Synonymous	14.18	22.05	55.5%
Missense	12.35	20.19	63.5%
LoF	8.62	14.85	72.3%
Singletons (per kb)			
Synonymous	4.70	10.91	132.1%
Missense	4.60	10.87	136.3%
LoF	3.75	9.25	146.7%

Results shown for sampling distribution $w = 50$ km are averaged over three center locations.

growth. Furthermore, the question of how best to construct samples for human genetics research is intrinsically linked to discussions of representation in biomedical research (2, 3, 7). So, while the work here contributes insights into the impact of sampling on the SFS of discovered variants, we emphasize that sampling is only one element of the multifaceted challenge of study design.

A second area of research for which our results have key implications is the inference of fitness effects in evolutionary genetics. Many studies aim to infer the DFE from the frequencies of observed variants of different classes (22–25). Such studies often focus on the population-scaled selection coefficient (commonly, Ns) as the parameter of interest. Empirically, population genetic samples are typically taken from one or a few distinct locations, yet are modeled as a random sample from the total population. Our results imply that this practice will lead to biases in the inference of selection coefficients which will tend toward underestimating the strength of negative selection.

Specifically, we expect that sampling narrowly from a particular location will lead to artificially high (or "concentrated") frequencies of deleterious variants. This can be readily seen in how the shape of the SFS becomes flatter for more narrow samples (Fig. 3 *A* and *B*). In terms of our theory, this corresponds to our result that for spatially concentrated sampling, the effective selection intensity, γ_E, can be substantially less than Ns. In fact, values of γ_E are orders of magnitude lower than Ns within our test settings (Fig. 2). Thus, the frequencies used in the inference framework will be higher than expected under random mating, leading to biased estimates of s. This bias is likely to be most prominent for alleles under stronger selection, as the deviation of γ_E from Ns will be larger (Fig. 2). We also expect to see a downward bias in the inferred variance of the DFE for spatially concentrated samples: Estimates for variants with stronger fitness effects will be biased more strongly than those with weak effects, leading to an overall reduction in variability among inferred effects.

For both of these downstream applications, another relevant finding from our model is that the magnitude of effect of sampling width on allele frequencies and summary statistics is highly dependent on the sample size (Figs. 3 and 4*B* and *SI Appendix*, Fig. S5). Thus, as sample sizes in genetics continue to grow toward millions of individuals, we may expect the impact of sampling breadth to become more evident.

In relationship to other theoretical approaches to this problem, we note that previous analyses of spatial population genetic models have derived the spatial covariance between allele frequencies sampled at two locations (32–34, see *SI Appendix*). In retrospect, given our finding that second-order moments of the allele frequency distribution provide useful approximations of the SFS, an alternative route to obtain the results we obtain would be to use weighted averages of previously derived two-point spatial covariance functions to adjust for uneven sampling. The approach taken here has the advantage of being more quickly generalizable to higher-order moments (which can be used to refine the approximation), as well as to alternative spatial models of dispersal and sampling.

Indeed, the most important caveat of our work is that we analyze a highly abstract model of a spatial population and sampling effort. The deviations between the observed SFS and fitted SFS observed when we attempted to fit the model directly indicates deviations in the abundance of singletons and in how the SFS shifts more modestly with w than predicted (*SI Appendix*). We have also not considered various departures from equilibrium such as recent population growth, recent admixture from diverged lineages (e.g. archaic hominids), and recent origins from a shared ancestral population (e.g. shared African origins of humans). Also, our simple model of migration via local diffusion does not account for repeated layers of long-range dispersal events which are plausibly frequent in human and nonhuman populations (though such long-range dispersal can be approximated in the model using as an additional global homogenizing force similar to mutation, as done in Kimura and Weiss; 70). As a result, a geographically "narrow" sample in real data (e.g. sampling a city like London) are not truly "narrow" in the sense of our model. Also, our model also assumes that there are no boundaries on where carriers can disperse and as a result, no "boundary effects" are present. Thus, especially for settings beyond the UK Biobank, the relevance of these more complex factors should be kept in mind.

Overall, in real studies of populations of humans or other organisms, the patterns of movement and of sampling may greatly deviate from what we investigated here. Nonetheless, the general alignment of our empirical and theoretical results suggest the real-world importance of sampling breadth for interpreting the outcomes of existing studies and designing future ones.

Data, Materials, and Software Availability. Some study data are available: All simulation data and associated scripts are available at: https://doi.org/10.5281/zenodo.15398319 (61). Empirical analyses used the published exome sequence data from the UK Biobank resource (19, https://www.ukbiobank.ac.uk/).

ACKNOWLEDGMENTS. We thank Luis Barreiro, Jeremy Berg, Jennifer Blanc, Maryn Carlson, Graham Coop, Castedo Ellerman, Molly Przeworski, Yuval Simons, Matthias Steinrücken, and Anastasia Stolyarova, and the anonymous reviewers for helpful comments and discussion. This work has been supported by the NSF (DGE1746045 to M.C.S.) and the NIH (R01 GM132383 and R35 GM149521 to J.N.). This research has been conducted using the UK Biobank Resource under Application Number 88057.

1. C. S. Gallagher, G. S. Ginsburg, A. Musick, Biobanking with genetics shapes precision medicine and global health. *Nat. Rev. Genet.* **26**, 191–202 (2024).
2. C. D. Bustamante, F. M. De La Vega, E. G. Burchard, Genomics for the world. *Nature* **475**, 163–165 (2011).
3. A. B. Popejoy, S. M. Fullerton, Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
4. C. Bycroft *et al*., The UK biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
5. K. J. Karczewski *et al*., The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
6. A. R. Martin *et al*., Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
7. D. D. Dolan, M. K. Cho, S. S. J. Lee, Innovating for a just and equitable future in genomic and precision medicine research. *Am. J. Bioeth.* **23**, 1–4 (2023).
8. M. Sohail *et al*., Mexican biobank advances population and medical genomics of diverse ancestries. *Nature* **622**, 775–783 (2023).
9. All of Us Research Program Genomics Investigators, Genomic data in the all of us research program. *Nature* **627**, 340–346 (2024).
10. A. Verma *et al*., Diversity and scale: Genetic architecture of 2068 traits in the VA million veteran program. *Science* **385**, eadj1182 (2024).
11. A. Elfatih, C. Saad, Qatar Genome Program Research Consortium, B. Mifsud, H. Mbarek, Analysis of 14,392 whole genomes reveals 3.5% of Qataris carry medically actionable variants. *Eur. J. Hum. Genet.* **32**, 1465–1473 (2024).
12. G. Sirugo, S. M. Williams, S. A. Tishkoff, The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
13. A. Durvasula, K. E. Lohmueller, Negative selection on complex traits limits phenotype prediction accuracy between populations. *Am. J. Hum. Genet.* **108**, 620–631 (2021).
14. Y. Ding *et al*., Polygenic Scoring Accuracy Varies across the Genetic Ancestry Continuum. *Nature* **618**, 774–781 (2023).
15. J. D. Szustakowski *et al*., Advancing human genetics research and drug discovery through exome sequencing of the UK biobank. *Nat. Genet.* **53**, 942–948 (2021).
16. M. Ghoussaini, M. R. Nelson, I. Dunham, Future prospects for human genetics and genomics in drug discovery. *Curr. Opin. Struct. Biol.* **80**, 102568 (2023).
17. G. Sella, N. H. Barton, Thinking about the evolution of complex traits in the era of Genome-Wide association studies. *Annu. Rev. Genomics Hum. Genet.* **20**, 461–493 (2019).
18. D. J. Weiner *et al*., Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614**, 1–8 (2023).
19. J. D. Backman *et al*., Exome sequencing and analysis of 454,787 UK biobank participants. *Nature* **599**, 1–10 (2021).
20. B. B. Sun *et al*., Genetic associations of protein-coding variants in human disease. *Nature* **603**, 95–102 (2022).
21. P. Wainschtein *et al*., Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* **54**, 263–273 (2022).
22. S. H. Williamson *et al*., Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7882–7887 (2005).
23. A. R. Boyko *et al*., Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4**, e1000083 (2008).
24. R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
25. B. Y. Kim, C. D. Huber, K. E. Lohmueller, Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics* **206**, 345–361 (2017).
26. J. Robinson, C. C. Kyriazis, S. C. Yuan, K. E. Lohmueller, Deleterious variation in natural populations and implications for conservation genetics. *Annu. Rev. Anim. Biosci.* **11**, 93–114 (2023).
27. M. R. Nelson *et al*., An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
28. J. G. Schraiber, J. P. Spence, M. D. Edge, Estimation of demography and mutation rates from one million haploid genomes. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.09.18.613708 (Accessed 27 March 2025).
29. S. Wright, T. Dobzhansky, W. Hovanitz, Genetics of natural populations. VII. The Allelism of Lethals in the third chromosome of *Drosophila pseudoobscura*. *Genetics* **27**, 363–394 (1942).
30. S. Wright, Isolation by distance. *Genetics* **28**, 114–138 (1943).
31. B. Wallace, Distance and the Allelism of Lethals in a tropical population of *Drosophila melanogaster*. *Am. Nat.* **100**, 565–578 (1966).
32. G. H. Weiss, M. Kimura, A mathematical analysis of the stepping stone model of genetic correlation. *J. Appl. Probab.* **2**, 129–149 (1965).
33. T. Maruyama, Distribution of gene frequencies in a geographically structured population. 3. Distribution of deleterious genes and genetic correlation between different localities. *Ann. Hum. Genet.* **36**, 99–108 (1972).
34. S. Yokoyama, The rate of Allelism of lethal genes in a geographically structured population. *Genetics* **93**, 245–262 (1979).
35. J. Wakeley, Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871 (1999).
36. S. E. Ptak, M. Przeworski, Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet.* **18**, 559–563 (2002).
37. U. Arunyawat, W. Stephan, T. Städler, Using multilocus sequence data to assess population structure, natural selection, and linkage disequilibrium in wild tomatoes. *Mol. Biol. Evol.* **24**, 2310–2322 (2007).
38. T. Städler, B. Haubold, C. Merino, W. Stephan, P. Pfaffelhuber, The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**, 205–216 (2009).
39. E. Quéméré *et al*., Spatial variation in density and total size estimates in fragmented primate populations: The golden-crowned sifaka (*Propithecus tattersalli*). *Am. J. Primatol.* **72**, 72–80 (2010).
40. K. R. St Onge, A. E. Palmé, S. I. Wright, M. Lascoux, Impact of sampling schemes on demographic inference: An empirical study in two species with different mating systems and demographic histories. *G3* **2**, 803–814 (2012).
41. O. Mazet, W. Rodríguez, L. Chikhi, Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theor. Popul. Biol.* **104**, 46–58 (2015).
42. C. J. Battey, P. L. Ralph, A. D. Kern, Space is the place: Effects of continuous spatial structure on analysis of population genetic data. *Genetics* **215**, 193–214 (2020).
43. A. D. Gloss *et al*., Genome-wide association mapping within a local *Arabidopsis* thaliana population more fully reveals the genetic architecture for defensive metabolite diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20200512 (2022).
44. F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
45. M. Friesen, Long-time behavior for subcritical measure-valued branching processes with immigration. *Potential Anal.* **59**, 705–730 (2023).
46. S. Wright, The distribution of gene frequencies in populations. *Proc. Natl. Acad. Sci. U.S.A.* **23**, 307–320 (1937).
47. J. F. Crow, M. Kimura, *An Introduction to Population Genetics Theory* (Harper and Row, 1970).
48. B. C. Haller, P. W. Messer, SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Mol. Biol. Evol.* **36**, 632–637 (2019).
49. G. Malecot, *Les Mathematiques de l'heredite* (Masson & Cie, 1948).
50. J. Felsenstein, A pain in the torus: Some difficulties with models of isolation by distance. *Am. Nat.* **109**, 359–368 (1975).
51. W. J. Ewens, Some applications of multiple-type branching processes in population genetics. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **30**, 164–175 (1968).
52. S. Tavaré, Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* **26**, 119–164 (1984).
53. B. M. Peter, M. Slatkin, The effective founder effect in a spatially expanding population. *Evolution* **69**, 721–734 (2015).
54. A. Etheridge, N. Freeman, S. Penington, D. Straulino, Branching Brownian motion and selection in the spatial *Lambda*-Fleming-Viot process. *Ann. Appl. Probab.* **27**, 2605–2645 (2017).
55. Z. L. Fuller, J. J. Berg, H. Mostafavi, G. Sella, M. Przeworski, Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
56. J. B. S. Haldane, A mathematical theory of natural and artificial selection, Part V: Selection and mutation. *Math. Proc. Camb. Philos. Soc.* **23**, 838–844 (1927).
57. M. Slatkin, B. Rannala, The sampling distribution of disease-associated alleles. *Genetics* **147**, 1855–1861 (1997).
58. J. Novembre, M. Slatkin, Likelihood-based inference in isolation-by-distance models using the spatial distribution of low-frequency alleles. *Evolution* **63**, 2914–2925 (2009).
59. A. Etheridge, *An Introduction to Superprocesses* (American Mathematical Soc., 2000).
60. A. M. Etheridge, Survival and extinction in a locally regulated population. *Ann. Appl. Probab.* **14**, 188–214 (2004).
61. M. Steiner, M. K. Ianni-Ravn, D. P. Rice, Spatial Rare Alleles. Zenodo. https://doi.org/10.5281/zenodo.15398319. Deposited 13 May 2025.
62. R. A. Fisher, The wave of advance of advantageous genes. *Ann. Eugen.* **7**, 355–369 (1937).
63. M. Slatkin, Gene flow and selection in a cline. *Genetics* **75**, 733–756 (1973).
64. D. B. Rubin, The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *J. Am. Stat. Assoc.* **82**, 543 (1987).
65. E. M. Wijsman, L. L. Cavalli-Sforza, Migration and genetic population structure with special reference to humans. *Annu. Rev. Ecol. Syst.* **15**, 279–301 (1984).
66. C. McEvedy, R. Jones, *Atlas of World Population History* (Factos on File, New York, 1978).
67. A. De, R. Durrett, Stepping-stone spatial structure causes slow decay of linkage disequilibrium and shifts the site frequency spectrum. *Genetics* **176**, 969–981 (2007).
68. Y. B. Simons, M. C. Turchin, J. K. Pritchard, G. Sella, The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46**, 220–224 (2014).
69. A. Stolyarova, G. Coop, M. Przeworski, The distribution of highly deleterious variants across human ancestry groups. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.01.31.635988 (Accessed 27 March 2025).
70. M. Kimura, G. H. Weiss, The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561–576 (1964).