PLOS BIOLOGY

CONSENSUS VIEW

Recommendations for improving statistical inference in population genomics

Parul Johri[1], Charles F. Aquadro[2], Mark Beaumont[3], Brian Charlesworth[4], Laurent Excoffier[5], Adam Eyre-Walker[6], Peter D. Keightley[7], Michael Lynch[1], Gil McVean[8], Bret A. Payseur[9], Susanne P. Pfeifer[1], Wolfgang Stephan[10], Jeffrey D. Jensen[1]*

1 School of Life Sciences, Arizona State University, Tempe, Arizona, United States of America, **2** Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, United States of America, **3** School of Biological Sciences, University of Bristol, Bristol, United Kingdom, **4** Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, **5** Institute of Ecology and Evolution, University of Berne, Berne, Switzerland, **6** School of Life Sciences, University of Sussex, Brighton, United Kingdom, **7** Institute of Ecology and Evolution, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, **8** Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, United Kingdom, **9** Laboratory of Genetics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, **10** Natural History Museum, Berlin, Germany

* jeffrey.d.jensen@asu.edu

Citation: Johri P, Aquadro CF, Beaumont M, Charlesworth B, Excoffier L, Eyre-Walker A, et al. (2022) Recommendations for improving statistical inference in population genomics. PLoS Biol 20(5): e3001669. https://doi.org/10.1371/journal.pbio.3001669

Published: May 31, 2022

Funding: This work was funded by National Institutes of Health grants R01GM135899 and R35GM139383 to JDJ. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: ABC, approximate Bayesian computation; DFE, distribution of fitness effects; LD, linkage disequilibrium; SFS, site frequency spectrum.

Abstract

The field of population genomics has grown rapidly in response to the recent advent of affordable, large-scale sequencing technologies. As opposed to the situation during the majority of the 20th century, in which the development of theoretical and statistical population genetic insights outpaced the generation of data to which they could be applied, genomic data are now being produced at a far greater rate than they can be meaningfully analyzed and interpreted. With this wealth of data has come a tendency to focus on fitting specific (and often rather idiosyncratic) models to data, at the expense of a careful exploration of the range of possible underlying evolutionary processes. For example, the approach of directly investigating models of adaptive evolution in each newly sequenced population or species often neglects the fact that a thorough characterization of ubiquitous nonadaptive processes is a prerequisite for accurate inference. We here describe the perils of these tendencies, present our consensus views on current best practices in population genomic data analysis, and highlight areas of statistical inference and theory that are in need of further attention. Thereby, we argue for the importance of defining a biologically relevant baseline model tuned to the details of each new analysis, of skepticism and scrutiny in interpreting model fitting results, and of carefully defining addressable hypotheses and underlying uncertainties.

Introduction

A brief overview

Population genomic inference—the use of data on molecular variation within species and divergence between species to infer evolutionary processes—has become widely embraced and

highly utilized in fields including evolutionary biology, ecology, anthropology, agriculture, and medicine. The underlying questions may be demographic in nature, be it estimating the timing of the peopling of the world [1] or of viral transmission in a congenitally infected newborn [2]; alternatively, they may concern the selective history of specific populations, be it identifying mutations that confer cryptic coloration in species adapting to major postglacial climatic and geological changes [3] or viral drug resistance to clinical therapeutics [4].

The foundational work allowing for the dissection of these evolutionary processes from levels and patterns of variation and divergence was conducted by Fisher, Wright, and Haldane nearly a century ago (e.g., [5–7]; for a historical overview, see [8]). This work demonstrated the possibility of studying evolution at the genetic level, integrating the revolutionary ideas of Darwin [9] with the turn-of-the-century appreciation of Mendel's [10] research. However, as was famously described by Lewontin [11], this initial theoretical progress during the first half of the 20th century was "like a complex and exquisite machine, designed to process a raw material that no one had succeeded in mining." With the first "mining" of population-level molecular variation in the 1960s (see [12]), this machine was put to work. The next major steps forward were provided by Kimura and Ohta, who offered a comprehensive framework for studying DNA and protein sequence variation based on these fundamental theoretical insights—the Neutral Theory of Molecular Evolution [13–15]—an advance for which molecular biology also provided support [16]. Despite some claims to the contrary [17], Kimura and Ohta's initial postulates have since been largely validated [18,19] and have provided a means to interpret observed molecular variation and divergence within the context of constantly occurring evolutionary processes including mutation, genetic drift, and purifying selection. While ascribing an important role for positive selection at the level of phenotypic evolution (consistent with Darwin's initial notions), the Neutral Theory hypothesizes that at the genetic level, beneficial mutations are rare compared to the much larger input of neutral, nearly neutral, and deleterious mutations that are constantly raining down on the genomes of all species. Accordingly, episodes of positive selection per nucleotide are rare compared to genetic drift and purifying selection. However, the significant effects on evolution at linked sites caused by fitness-altering mutations have been described in detail in the decades since Kimura's initial formulation of the Neutral Theory [20–22].

With this framework and the availability of datasets to which it could be applied, statistical approaches for analyzing molecular data began to proliferate, frequently employing some form of neutral expectation as a null model. A wide range of rather sophisticated statistical machinery is now available for reconstructing histories of population size change, population subdivision, and migration (e.g., [23,24]); for identifying beneficial mutations based on patterns associated with selective sweeps (e.g., [25,26]); and for quantifying the distribution of fitness effects (DFE) of newly arising mutations (e.g., [27,28]), as well as for estimating rates of mutation (e.g., [29–31]) and recombination (e.g., [32–34]). These approaches operate in a variety of statistical frameworks (see [35–37]) and utilize various aspects of the data—including the frequencies of variants in a sample (the site frequency spectrum, SFS), associations between variants (linkage disequilibrium, LD), and/or levels and patterns of between-species divergence at contrasted site classes (e.g., synonymous versus nonsynonymous sites).

Challenges of model choice and parameter fitting

The growing variety of statistical approaches and associated software implementations presents a dizzying array of choices for any given analysis; although many approaches share the same aims, there also exist important differences. For example, some approaches require a relatively high level of coding ability to implement while others may be applied in easy-to-use

software packages; while some are well tested and justified by population genetic theory, others are not. Moreover, even the process of translating raw sequencing data into the allele calls and genotypes used as input for these approaches is accompanied by uncertainty that depends on sequencing quality and coverage, availability of a reference genome, and choice of variant calling and filtering strategies [38,39]. Adding to this complexity, it has become increasingly clear that demographic estimation may be highly biased when selection and recombination-associated biased gene conversion are neglected [40,41], whereas estimates of selection intensity and recombination rate may be highly biased when neglecting demographic effects [42–45]. This creates a circular problem when commencing any new analysis: One needs information about the demographic history to estimate parameters of recombination and selection, while at the same time one needs information about recombination and selection to estimate the demographic history. An additional challenge, and a frustration for many, is that there is no single "best approach"; the correct analysis tools to use, and indeed which questions can be answered at all, depends entirely on the details of the organism under study [46]. Specifically, biological parameters that vary among species—including evolutionary parameters (e.g., effective population size (N_e), mutation rates, recombination rates, and population structure and history), genome structure (e.g., the distribution of functional sites along the genome), and life history traits (e.g., mating system)—must all be considered in order to define addressable hypotheses and optimal approaches.

Beyond these initial considerations, a more difficult issue often emerges. Namely, very different models may be found to provide a good fit to the observed data (e.g., [47]; see [48] for a phylogenetic perspective on the topic). In other words, particular parameter combinations may be found under competing models that are all capable of predicting the observed patterns of variation. For example, assuming neutrality, one may match an empirical observation at a locus by fitting the timing, severity, and duration of a population bottleneck or, alternatively, when assuming a constant population size, by fitting the rate and mean strength of selective sweeps. This fact alone implies a simple truism: The ability to fit the parameters of one's preferred model to data does not alone represent proof of biological reality. Rather, it suggests that this model is one—out of potentially very many—that represents a viable hypothesis, which should be further examined via subsequent analyses or experimentation.

Examples abound of enthusiastic promotion of a single preferred model, only to be tempered by subsequent demonstrations of the fit of alternative and often simpler/more biologically realistic models. For example, the view that segregating alleles may be commonly maintained by balancing selection [49] was tempered by the realization that genetic drift is often a sufficient explanation [14], and the view that genome-wide selective sweeps on standing variation are pervasive [50,51] was tempered by the realization that neutral population histories can result in similar patterns [47,52]. While one may readily find such examples of using episodic or hypothesized processes to fit large-scale data patterns by neglecting to define expectations arising from common and certain-to-be-occurring processes, determining which models to evaluate and how to interpret the fit of a model and its alternatives are challenges for all researchers. To better illustrate this point, Fig 1 presents 3 scenarios (constant population size with background selection, constant population size with background selection and selective sweeps, and a population bottleneck with background selection and selective sweeps) and provides the fit of each of those scenarios to 2 incorrect models (population size change assuming strict neutrality and recurrent selective sweeps assuming constant population size). As is shown in the figure, each scenario can be well fit by both incorrect models, with selective sweeps and population bottlenecks generally being confounded, as well as background selection and population growth, as has been described several times before (e.g., [40,53–55]).

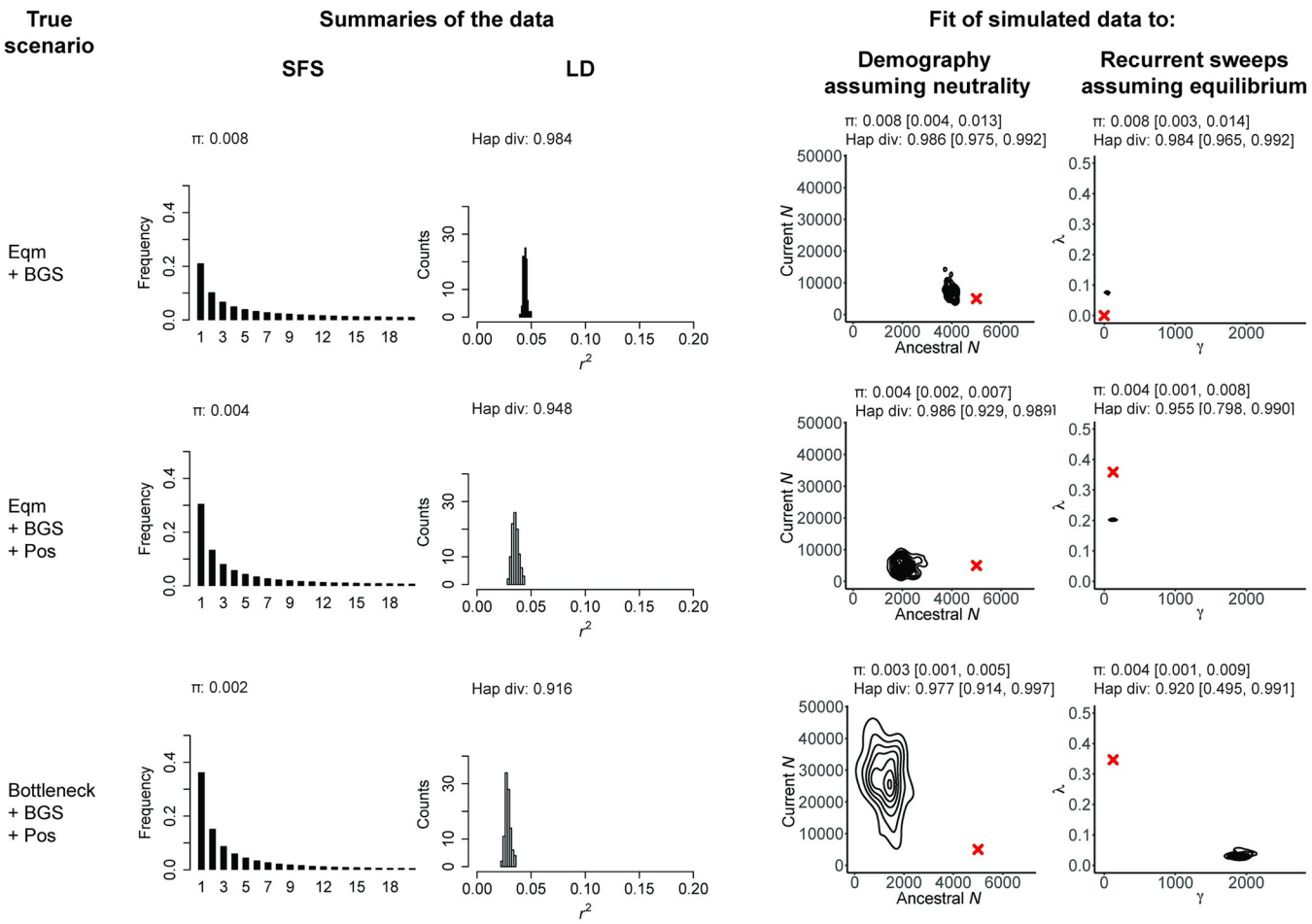

Fig 1. Incorrect models may often readily be fit to a given dataset. Here, we present 3 scenarios varying from simple to more complex: The first row presents a constant-sized population experiencing background selection (denoted by "Eqm + BGS"), the second row is the same scenario with the addition of recurrent selective sweeps (denoted by "Eqm +BGS + Pos"), and the final row adds a population bottleneck (denoted by "Bottleneck + BGS + Pos"). For each scenario, the resulting SFS (truncated to $n = 20$) and LD (r^2) distributions are given, together with mean pairwise (π) and haplotype diversity. To these simulated data, we fit 2 incorrect models: one assuming all sites are neutral but including a change in population size and a second model in which there are recurrent selective sweeps, no change in population size, and all mutations are assumed to be neutral or beneficial (with a population-scaled beneficial selection coefficient (γ) and the fraction of beneficial substitutions (λ) being estimated from the data). For each inference panel, the red cross gives the true value, the distribution presents the joint posterior obtained from the ABC analysis, and the summary statistics given above the posteriors represent the mean values, and the range from the 95% CIs, obtained from posterior checks. In all cases, exonic sites (i.e., directly selected sites) were masked, and the summary statistic calculations as well as inference is based only on neutral regions (see Methodology). As shown, demographic and selection models can be fit to all datasets, often resulting in strong mis-inference when the assumptions underlying the estimation procedure are violated. The scripts underlying this figure may be found at https://github.com/paruljohri/Perspective_Statistical_Inference/tree/main/SimulationsTestSet/Figure1. LD, linkage disequilibrium; SFS, site frequency spectra.

https://doi.org/10.1371/journal.pbio.3001669.g001

Methodology

In order to provide a series of examples to accompany key points—as with the above Fig 1—both forward-in-time simulations and coalescent simulations were performed for (1) the inference of demographic history assuming complete neutrality; (2) the inference of positive selection assuming constant population size; and (3) obtaining test datasets representing different evolutionary scenarios. While any statistical framework involving model/parameter exploration and comparison may be consistent with our recommendations, we here utilize approximate Bayesian computation (ABC) for our examples, as it is a particularly useful framework for quantifying uncertainty and for exploring complex models. All relevant data and scripts

underlying our simulations and analyses may be found at the link below and links pertaining to individual figures may also be found in the respective figure legends:

https://github.com/paruljohri/Perspective_Statistical_Inference.

In all simulations, a chromosomal segment of 99,012 bp was simulated with an intron–exon–intergenic structure resembling the *Drosophila melanogaster* genome. Each gene comprised 5 exons (of 300 bp each) and 4 introns (of 100 bp each) separated by intergenic regions of length 1,068 bp. Such a construct resulted in a total of 33 genes across the simulated segment. Population parameters were chosen to resemble those from *D. melanogaster* populations following Campos and colleagues [56], assuming an effective population size (N_e) of 10^6 individuals with a mean mutation rate (μ) of 4.5×10^{-9} per base pair /generation and a mean recombination rate (r) of 1×10^{-8} per base pair /generation. For computational efficiency, all parameters were rescaled by a factor of 200.

Modeling and inference of demographic history

A simple demographic history was modeled in which a single population undergoes an instantaneous change from an ancestral size (N_{anc}) to a current size (N_{cur}), τ generations ago. Priors for both N_{anc} and N_{cur} were sampled from a loguniform distribution between 10 and 50,000, while priors for the time of change (τ) were sampled from a loguniform distribution between 10 and N_{cur}. A total of 100 replicates were simulated for each parameter combination. Simulations required for ABC were performed in *msprime* v. 0.7.3 [57] assuming complete neutrality. Mutation and recombination rates were assumed to be constant across the genome and across replicates.

Modeling and inference of positive selection

A recurrent selective sweep scenario was modeled in which only neutral and beneficial autosomal mutations were allowed, with simulations performed using *SLiM* v. 3.1 [58]. Introns and intergenic regions were assumed to be neutral, while exons experienced beneficial mutations with fitness effects sampled from an exponential distribution with mean s for homozygotes, assuming semidominance. The 2 parameters varied were the mean population-scaled strength of selection, $\gamma = 2N_{anc}s$, and the proportion of new beneficial mutations, f_{pos}. Priors for these parameters were sampled from a loguniform distribution such that $\gamma\in[0.1, 10{,}000]$ and $f_{pos}\in[0.00001, 0.01]$. For all parameter combinations, the true rate of beneficial substitutions per site (d_a) and the true fraction of substitutions due to beneficial mutations (λ, which is related to the α parameter of Eyre-Walker and Keightley [59]) were calculated using the total number of fixations (as provided by *SLiM*), with λ observed to range from 0 to 0.85 depending on the underlying parameters. Parameter inference was performed for γ and d_a and the corresponding λ was inferred using $\lambda = \frac{d_a}{d_a+((1-f_{pos})\times\mu\times num\ of\ generations)}$, where it was assumed that $1-f_{pos}\sim1$. Populations had a constant size of 5,000 diploid individuals with constant mutation and recombination rates, as specified above. Simulations were run for 100,100 generations (i.e., $20N_e$ + 100 generations).

ABC

The sample size was set to 100 haploid genomes (or 50 diploid individuals). Under the demographic and selection models described above, all exonic regions were masked, and the mean and variance (across replicates) of the following summary statistics were calculated: number of segregating sites, nucleotide site diversity (π), Watterson's theta (θ_W), θ_H, H', Tajima's D, number of singletons, haplotype number and frequency distribution, and statistics summarizing

LD (r^2, D, D'). All statistics were calculated in nonoverlapping sliding windows of 2 kb using *pylibseq* v. 0.2.3 [60]. ABC was performed with the R package "abc" v. 2.1 [61] using all summary statistics, with "neural net" to account for nonlinearity between statistics and parameters. A 100-fold cross-validation was used to identify the optimum tolerance level, which was found to be 0.05 (i.e., 5% of the simulations were accepted during ABC inference to estimate the posterior probability of each parameter). Point estimates of each inferred parameter were calculated as the weighted medians of the posterior estimates.

Simulations of different evolutionary scenarios as "true scenarios"

To consider more biologically realistic models and evaluate model violations, a number of evolutionary scenarios were simulated (using *SLiM*) as follows:

a. Background selection: Exons experienced deleterious mutations modeled by a discrete DFE comprised of 4 nonoverlapping uniform distributions, representing the effectively neutral ($-1<2N_{anc}s\leq 0$), weakly deleterious ($-10<2N_{anc}s\leq -1$), moderately deleterious ($-100<2N_{anc}s\leq -10$), and strongly deleterious ($2N_{anc}s\leq -100$) classes of mutations. All 4 bins were assumed to contribute equally to new mutations (i.e., 25% of all new mutations belonged to each class of mutation).

b. Positive selection: Exons experienced beneficial mutations with $\gamma = 125$ and $f_{pos} = 2.2\times10^{-3}$ (modified from [56]), resulting in $\lambda\approx 0.35$.

c. Population size change: A population decline was simulated such that the population declined from 5,000 to 100 individuals instantaneously 100 generations ago. A population expansion was similarly simulated with parameters $N_{anc} = 5{,}000$ and $N_{cur} = 10{,}000$. A population bottleneck model was also simulated such that $N_{anc} = N_{cur} = 5{,}000$, and a bottleneck occurred 2,000 generations ago with a reduction to 1% of the population size for 100 generations.

d. SNP ascertainment: Genotype error was modeled as an inability to detect the true number of singletons when using low-coverage population-genomic data to call variants [38]. To model this scenario, a random set of singletons, representing a third of all singletons present in the sample, were removed.

e. Progeny skew: A skew in the offspring distribution (ψ) was modeled such that 5% and 10% of the population was replaced by the offspring of a single individual each generation ([62] and see [63,64]).

f. Variation in mutation and recombination rates across the genome (e.g., [65–67]): Every 10 kb of the ≈100 kb genomic region considered was assumed to have a different mutation and recombination rate. For every simulated replicate, these rates were sampled from a Gaussian distribution with the same mean as above, and a coefficient of variation of 0.5. Negative values were truncated to 0.

Posterior checks

For the purposes of illustration, an example of posterior checks are provided in Fig 1 (i.e., showing a simple evaluation of the fit of the inferred posteriors under the incorrect models to the true scenarios under consideration). Specifically, the mean estimates of the inferred parameters were used to simulate the "best-fitting model" in *SLiM* v. 3.1 [58]. Exons were masked and summary statistics were calculated as above in windows of 2 kb using *pylibseq* v.0.2.3 [60]. In order to simulate the inferred models of positive selection, f_{pos} was calculated from λ

assuming a Wright–Fisher diploid population of size N and a total mutation rate of μ_{tot} (which for our purpose is the same as μ). Thus, $\mu_b = f_{pos} \times \mu_{tot}$ and $\mu_{neu} = (1 - f_{pos}) \times \mu_{tot}$, where μ_b and μ_{neu} are the beneficial and neutral mutation rates, respectively. Given a value of λ, and assuming that the DFE of beneficial mutations is exponential (with mean $\bar{s}$), we calculate f_{pos} as follows:

Given that

$$\lambda = \frac{\#of\ beneficial\ subs}{\#of\ beneficial\ subs + \#of\ neutral\ subs}, \quad (1)$$

where

$$\#of\ beneficial\ subs = P_{fix} \times L \times 2N\mu_b \quad (2)$$

and

$$\#of\ neutral\ subs = \mu_{neu} \times L, \quad (3)$$

where L is the length of the region being considered and P_{fix} is the probability of fixation of beneficial mutations, given by

$$P_{fix} = \int_0^\infty \frac{(1 - e^{-x})}{(1 - e^{-2Nx})} \left(\frac{e^{-x/\bar{s}}}{\bar{s}} \right) dx \quad (4)$$

Substituting (2) and (3) in (1), and rearranging, we get

$$f_{pos} = \frac{\lambda}{(1 - \lambda) P_{fix} 2N + \lambda} \quad (5)$$

Integrating (4) in R and substituting it into (5) gives us values of f_{pos}.

Statistics were calculated in nonoverlapping windows of 2 kb and intervals (CIs) were calculated as the 0.025 and 0.975 quantiles of the distribution of the statistics.

Recommendations

Constructing an appropriate baseline model for population genomic analysis

The somewhat disheartening exercise of fitting incorrect models to data (as depicted in Fig 1) naturally raises the questions of whether, and if so how, accurate evolutionary inferences can be extracted from DNA sequences sampled from a population. The first point of importance is that the starting point for any genomic analysis should be the construction of a biologically relevant baseline model, which includes the processes that must be occurring and shaping levels and patterns of variation and divergence across the genome. This model should include mutation, recombination, and gene conversion (each as applicable), purifying selection acting on functional regions and its effects on linked variants (i.e., background selection [21,68,69]), as well as genetic drift as modulated by, among other things, the demographic history and geographic structure of the population. Depending on the organism of interest, there may be other significant biological components to include, such as mating system, progeny distributions, ploidy, and so on (although, for certain questions of interest, some of these biological factors may simply be included in the resulting effective population size). It is thus helpful to view this baseline model as being built from the ground up for any new data analysis. Importantly, the point is not that these many parameters need to be fully understood in a given

population in order to perform any evolutionary inference, but rather that they all require consideration, and that the effects of uncertainties in their underlying values on downstream inference can be quantified.

However, even prior to considering any biological processes, it is important to investigate the data themselves. First, there exists an evolutionary variance associated with the myriad of potential realizations of a stochastic process, as well as the statistical variance introduced by finite sampling. Second, it is not advisable to compare one's empirical observations, which may include missing data, variant calling or genotyping uncertainty (e.g., effects of low coverage), masked regions (e.g., regions in which variants were omitted due to low mappability and/or callability) and so on, against either an analytical or simulated expectation that lacks those considerations and thus assumes optimal data resolution [70]. The dataset may also involve a certain ascertainment scheme, either for the variants surveyed [71], or given some predefined criteria for investigating specific genomic regions (e.g., regions representing genomic outliers with respect to a chosen summary statistic [72]). For the sake of illustration, Fig 2 follows the same format as Fig 1, but considers 2 scenarios: population growth with background selection and selective sweeps and the same scenario together with data ascertainment (in this case, an undercalling of the singleton class). As can be seen, due to the changing shape of the frequency spectra, neglecting to account for this ascertainment can greatly affect inference, considerably modifying the fit of both the incorrect demographic and incorrect recurrent selective sweep models to the data.

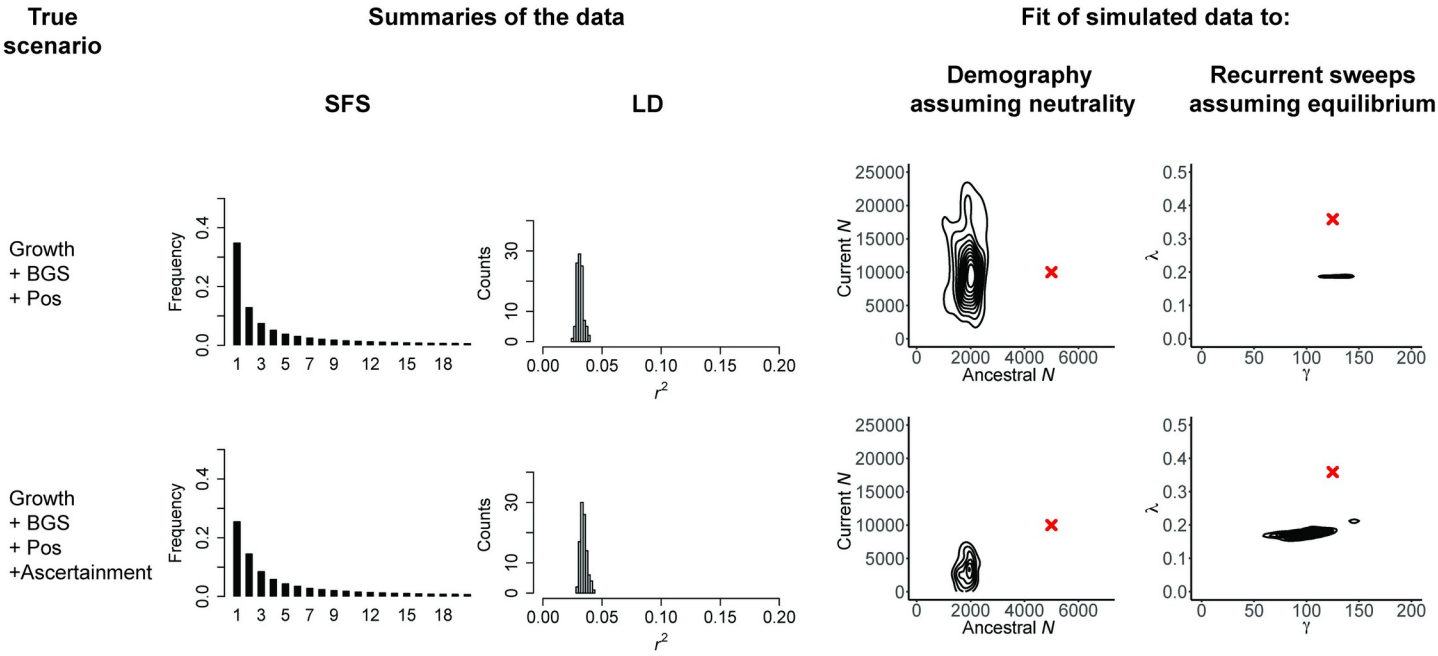

Fig 2. Ascertainment errors may amplify mis-inference, if not corrected. As in Fig 1, the scenarios are given in the first column, here population growth with background selection and recurrent selective sweeps ("Growth + BGS + Pos"), as well as the same scenario in which the imperfections of the variant-calling processes are taken into account—in this case, one-third of singletons are not called ("Growth + BGS + Pos + Ascertainment"). The middle columns present the resulting SFS and LD distributions, and the final columns provide the joint posterior distributions when the data are fit to 2 incorrect models: a demographic model that assumes strict neutrality and a recurrent selective sweep model that assumes a constant population size. All exonic (i.e., directly selected) sites were masked prior to analysis. Red crosses indicate the true values. As shown, unaccounted for ascertainment errors may contribute to mis-inference. The scripts underlying this figure may be found at https://github.com/paruljohri/Perspective_Statistical_Inference/tree/main/SimulationsTestSet/Figure2. LD, linkage disequilibrium; SFS, site frequency spectrum.

https://doi.org/10.1371/journal.pbio.3001669.g002

Hence, if sequencing coverage is such that rare mutations are being excluded from analysis, due to an inability to accurately differentiate genuine variants from sequencing errors, the model used for subsequent testing should also ignore these variants. Similarly, if multiple regions are masked in the empirical analysis due to problems such as alignment difficulties, the expected patterns of LD that are observable under any given model may be affected. Furthermore, while the added temporal dimension of time series data has recently been shown to be helpful for various aspects of population genetic inference [73–76], such data in no way sidestep the need for an appropriate baseline model, but simply requires the development of a baseline that matches the temporal sampling. In sum, as these factors can greatly affect the power of planned analyses and may introduce biases, the precise details of the dataset (e.g., region length, extent and location of masked regions, the number of callable sites, and ascertainment) and study design (e.g., sample size and single time point versus time series data) should be directly matched in the baseline model construction.

Once these concerns have been satisfied, the first biological addition will logically be the mutation rate and mutational spectrum. For a handful of commonly studied species, both the mean of, and genomic heterogeneity in, mutation rates have been quantified via mutation accumulation lines and/or pedigree studies [77]. However, even for these species, ascertainment issues remain complicating [78], variation among individuals may be substantial [79], and estimates only represent a temporal snapshot of rates and patterns that are probably changing over evolutionary timescales and may be affected by the environment [31,80]. In organisms lacking experimental information, often the best available estimates come either from a distantly related species or from molecular clock-based approaches. Apart from stressing the importance of implementing either of the experimental approaches in order to further refine mutation rate estimates for such a species of interest, it is noteworthy that this uncertainty can also be modeled. Namely, if proper estimation has been performed in a closely related species, one may quantify the expected effect on observed levels of variation and divergence of higher and lower rates. The variation in possible data observations induced by this uncertainty is thus now part of the underlying model.

The same logic follows for the next parameter addition(s): crossing over/gene conversion, as applicable for the species in question. For example, for a subset of species, per-generation crossover rates in cM per Mb have been estimated by comparing genetic maps based on crosses or pedigrees with physical maps [81–83]. In addition, recombination rates scaled by the effective population size have also been estimated from patterns of LD (e.g., [84,85])—although this approach typically requires assumptions about evolutionary processes that may be violated (e.g., [42]). As with mutation, the effects on downstream inference arising from the variety of possible recombination rates—whether estimated for the species of interest or a closely related species—can be modeled.

The next additions to the baseline model construction are generally associated with the greatest uncertainty—the demographic history of the population, and the effects of direct and linked purifying selection. This is a difficult task given the virtually infinite number of potential demographic hypotheses (e.g., [86]); furthermore, the interaction of selection with demography is inherently nontrivial and difficult to treat (e.g., [55,87,88]). This realization continues to motivate attempts to jointly estimate the parameters of population history together with the DFE of neutral, nearly neutral, weakly deleterious, and strongly deleterious mutations—a distribution that is often estimated in both continuous and discrete forms [89]. One of the first important advances in this area used putatively neutral synonymous sites to estimate changes in population size based on patterns in the SFS and conditioned on that demography to fit a DFE to nonsynonymous sites, which presumably experience considerable purifying selection [90–92]. This stepwise approach may become problematic, however, for organisms in which

synonymous sites are not themselves neutral [93–95] or when the SFS of synonymous sites is affected by background selection, which is probably the case generally given their close linkage to directly selected nonsynonymous sites ([41] and see [96,97]).

In an attempt to address some of these concerns, Johri and colleagues [44] recently developed an ABC approach that relaxes the assumption of synonymous site neutrality and corrects for background selection effects by simultaneously estimating parameters of the DFE alongside population history. The posterior distributions of the parameters estimated by this approach in any given data application (i.e., characterizing the uncertainty of inference) represent a logical treatment of population size change and purifying/background selection for the purposes of inclusion within this evolutionarily relevant baseline model. That said, the demographic model in this implementation is highly simplified, and extensions are needed to account for more complex population histories. In particular, estimation biases that may be expected owing to the neglect of cryptic population structure and migration, and indeed the feasibility of co-estimating population size change and the DFE together with population structure and migration within this framework, all remain in need of further investigation. While such simulation-based inference (see [98]), including ABC, provides one promising platform for joint estimation of demographic history and selection, progress on this front has been made using alternative frameworks as well [99,100], and developing analytical expectations under these complex models should remain as the ultimate, if distant, goal. Alternatively, in functionally sparse genomes with sufficiently high rates of recombination, such that assumptions of strict neutrality are viable for some genomic regions, multiple well-performing approaches have been developed for estimating the parameters of much more complex demographic models (e.g., [101–104]). In organisms for which such approaches are applicable (e.g., certain large, coding sequence sparse vertebrate, and land plant genomes), this intergenic demographic estimation assuming strict neutrality may helpfully be compared to estimates derived from data in or near coding regions that account for the effects of direct and linked purifying selection [41,44,105]. For newly studied species lacking functional annotation and information about coding density, following the joint estimation procedure would remain as the more satisfactory strategy in order to account for possible background selection effects.

Quantifying uncertainty in model choice and parameter estimation, investigating potential model violations, and defining answerable questions

One of the useful aspects of these types of analyses is the ability to incorporate uncertainty in underlying parameters under relatively complex models, in order to determine the impact of such uncertainty on downstream inference. The computational burden of incorporating variability in mutation and recombination rate estimates, or drawing from the confidence-or credibility-intervals of demographic or DFE parameters, can be met with multiple highly flexible simulation tools [58,106,107]. These are also useful programs for investigating potential model violations that may be of consequence. For example, if a given analysis for detecting population structure assumes an absence of gene flow, it is possible to begin with one's constructed baseline model, add migration parameters to the model in order to determine the effects of varying rates and directions of migration on the summary statistics being utilized in the empirical analysis, and thereby quantify how a violation of that assumption may affect the subsequent conclusions. Similarly, if an analysis assumes the Kingman coalescent (e.g., a small progeny distribution such that at most one coalescent event occurs per generation), but the organism in question could violate this assumption (i.e., with the large progeny number distributions associated with many plants, viruses, and marine spawners or simply owing to the relatively wide array of evolutionary processes that may similarly lead to multiple

merger coalescent events), these distributions may too be modeled in order to quantify potential downstream mis-inference.

To illustrate this point, Fig 3 considers 2 scenarios of constant population size and strict neutrality but with differing degrees of progeny skew, to demonstrate that a violation of this sort that is not corrected for may result in severely underestimated population sizes as well as the false inference of high rates of strong selective sweeps. In this case, the mis-inference arises from the reduction in contributing ancestors under these models, as well as to the fact that neutral progeny skew and selective sweeps can both generate multiple merger events [63,64,108,109]. Similarly, one may investigate the assumptions of constant mutation or recombination rates when they are in reality variable. As shown in Fig 4, when these rates are assumed to be constant as is common practice, but in reality vary across the genomic region under investigation, the fit of the (incorrect) demographic and selection models considered may again be substantially modified. Notably, this rate heterogeneity may inflate the inferred strength of selective sweeps. While Figs 3 and 4 serve as examples, the same investigations may be made for cases such as a fixed selective effect when there is in reality a distribution, independent neutral variants when there is in reality LD, panmixia when there is in reality population structure, and so on. Simply put, even if a particular biological process/parameter is not being directly estimated, its consequences can nonetheless be explored.

As detailed in Fig 5, with such a model incorporating both biological and stochastic variance as well as statistical uncertainty in parameter estimates, and with an understanding of the role of likely model violations, one may investigate which additional questions/hypotheses can be addressed with the data at hand. By using a simulation approach starting with the baseline model and adding hypothesized processes, it is possible to quantify the extent to which models,

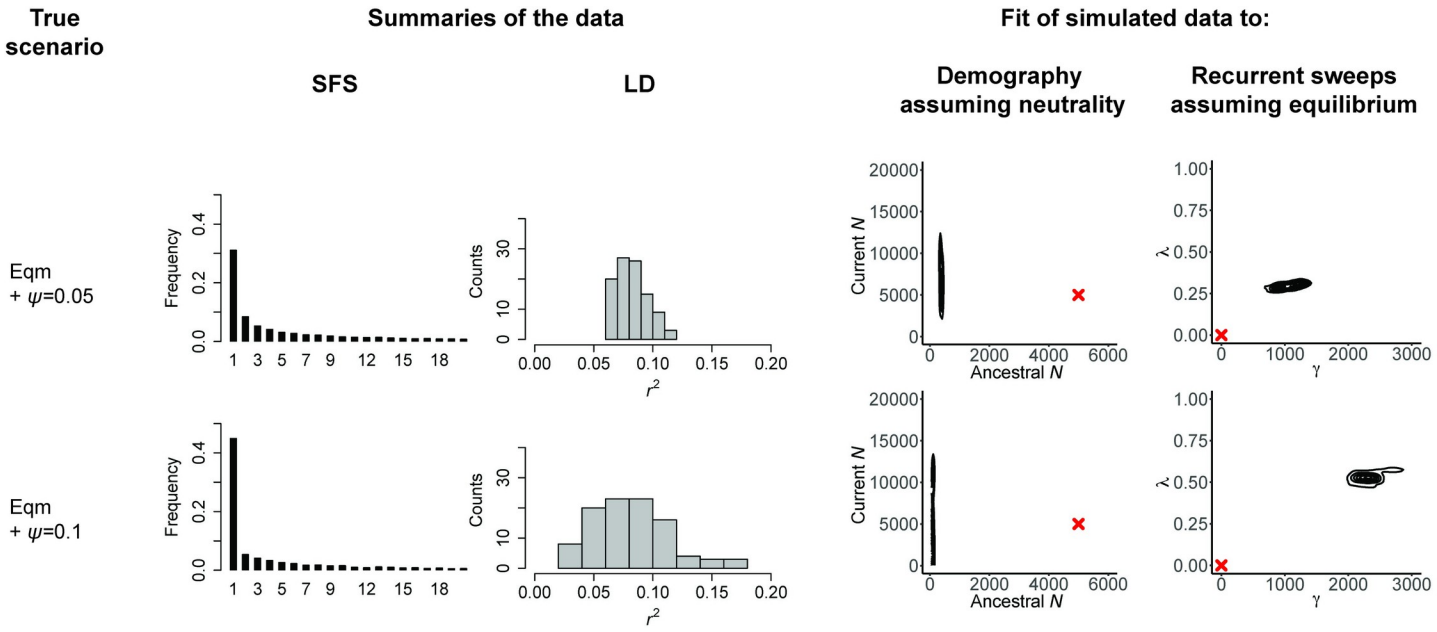

Fig 3. The impact of potential model violations can be quantified. As in Figs 1 and 2, the scenarios are given in the first column, here, equilibrium population size together with a moderate degree of progeny skew ("Eqm + $\psi = 0.05$") as well as with a high degree of progeny skew ("Eqm + $\psi = 0.1$") (see Methodology); the middle columns present the resulting SFS and LD distributions, and the final columns provide the joint posterior distributions when the data are fit to 2 incorrect models: a demographic model assuming neutrality and a recurrent selective sweep model assuming equilibrium population size. Red crosses indicate the true values. As shown, this violation of Kingman coalescent assumptions can lead to drastic mis-inference, but the biases resulting from such potential model violations can readily be described. The scripts underlying this figure may be found at https://github.com/paruljohri/Perspective_Statistical_Inference/tree/main/SimulationsTestSet/Figure3. LD, linkage disequilibrium; SFS, site frequency spectrum.

https://doi.org/10.1371/journal.pbio.3001669.g003

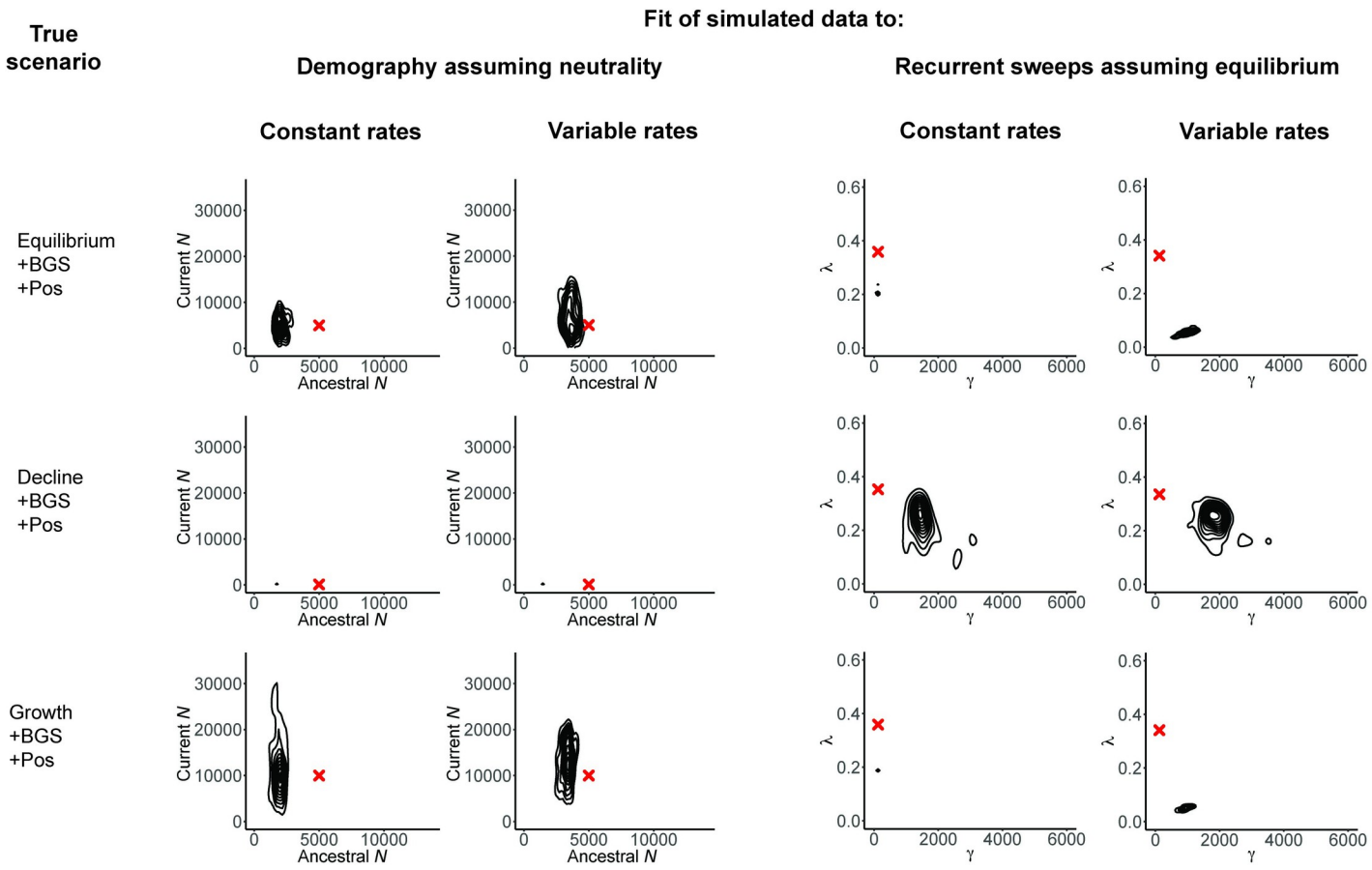

Fig 4. The effects of not correcting for mutation and recombination rate heterogeneity. Three scenarios are here considered: equilibrium population size with background selection and recurrent selective sweeps ("Eqm +BGS + Pos"), declining population size together with background selection and recurrent selective sweeps ("Decline + BGS + Pos"), and growing population size together with background selection and recurrent selective sweeps ("Growth + BGS + Pos"). Inference is again made under an incorrect demographic model assuming neutrality, as well as an incorrect recurrent selective sweep model assuming equilibrium population size. However, within each category, inference is performed under 2 settings: mutation and recombination rates are constant and known and mutation and recombination rates are variable across the region but assumed to be constant (see Methodology). Red crosses indicate the true values, and all exonic (i.e., directly selected) sites were masked prior to analysis. As shown, neglecting mutation and recombination rate heterogeneity across the genomic region in question can have an important impact on inference, particularly with regard to selection models. The scripts underlying this figure may be found at https://github.com/paruljohri/Perspective_Statistical_Inference/tree/main/SimulationsTestSet/Figure4.

https://doi.org/10.1371/journal.pbio.3001669.g004

and the parameters underlying those models, may be differentiated and which result in overlapping or indistinguishable patterns in the data (e.g., [110]). For example, if the goal of a given study is to identify recent beneficial fixations in a genome—be they potentially associated with high-altitude adaptation in humans, crypsis in mice, or drug resistance in a virus—one may begin with the baseline model and simulate selective sweeps under that model. As illustrated in Fig 6, by varying the strengths, rates, ages, dominance and epistasis coefficients of beneficial mutations, the patterns in the SFS, LD, and/or divergence that may differentiate the addition of such selective sweep parameters from the baseline expectations can be quantified. Moreover, any intended empirical analyses can be evaluated using simulated data (i.e., the baseline, compared to the baseline + the hypothesis) to define the power and false positive rates associated. If the differences in resulting patterns cannot be distinguished from the expected variance under the baseline model (in other words, if the power and false positive rate of the analyses are not favorable), the hypothesis is not addressable with the data at hand (e.g., [54]). If the results are favorable, this analysis can further quantify the extent to which the

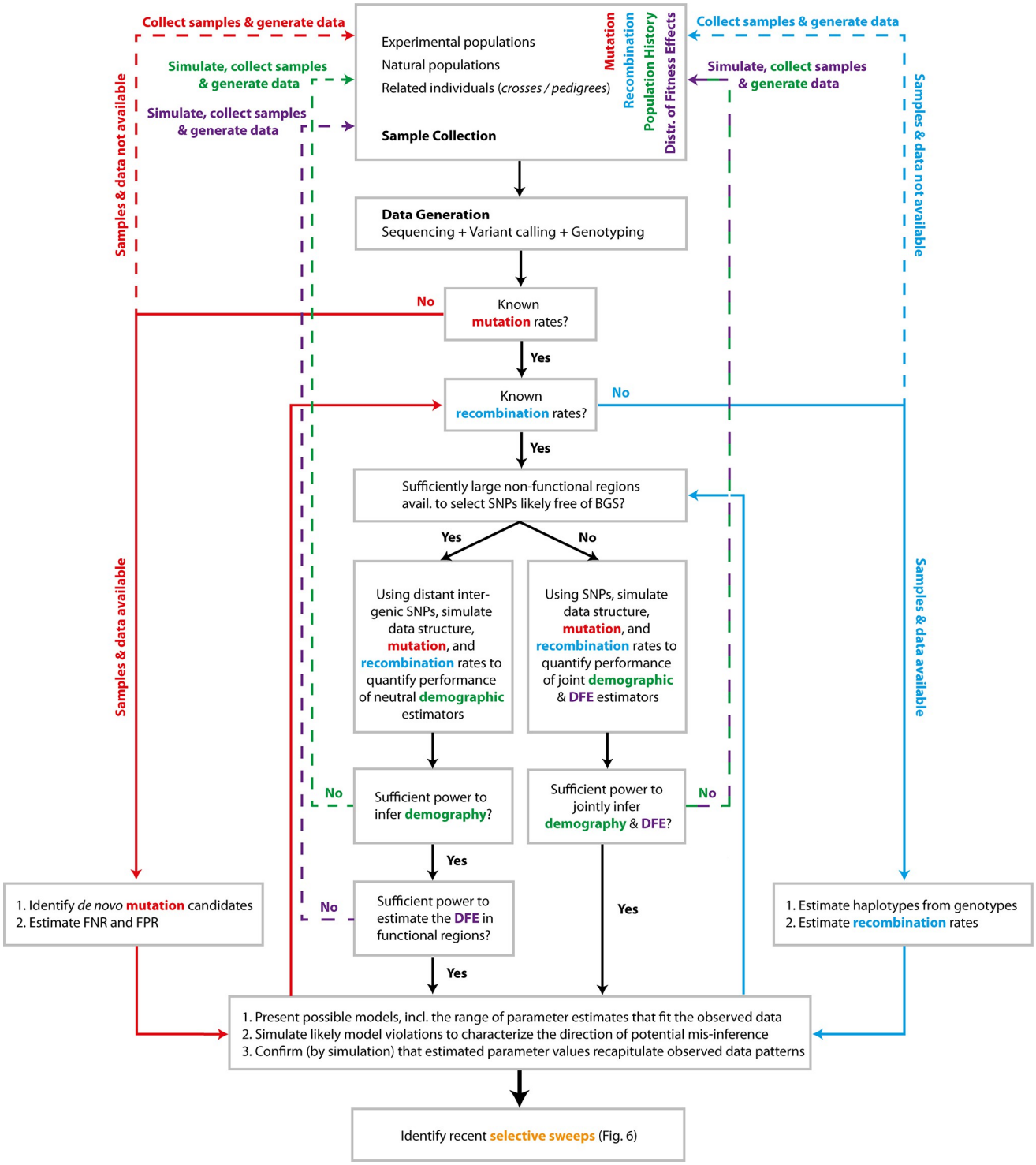

Fig 5. Diagram of important considerations in constructing a baseline model for genomic analysis. Considerations related to mutation rate are coded in red, recombination rate in blue, demographic history in green, and the DFE in purple—as well as combinations thereof. Beginning from the top with the source of data collected, the arrows suggest a path that is needed to be considered. Dotted lines indicate a return to the starting point. DFE, distribution of fitness effects; FNR, false negative rate; FPR, false positive rate.

https://doi.org/10.1371/journal.pbio.3001669.g005

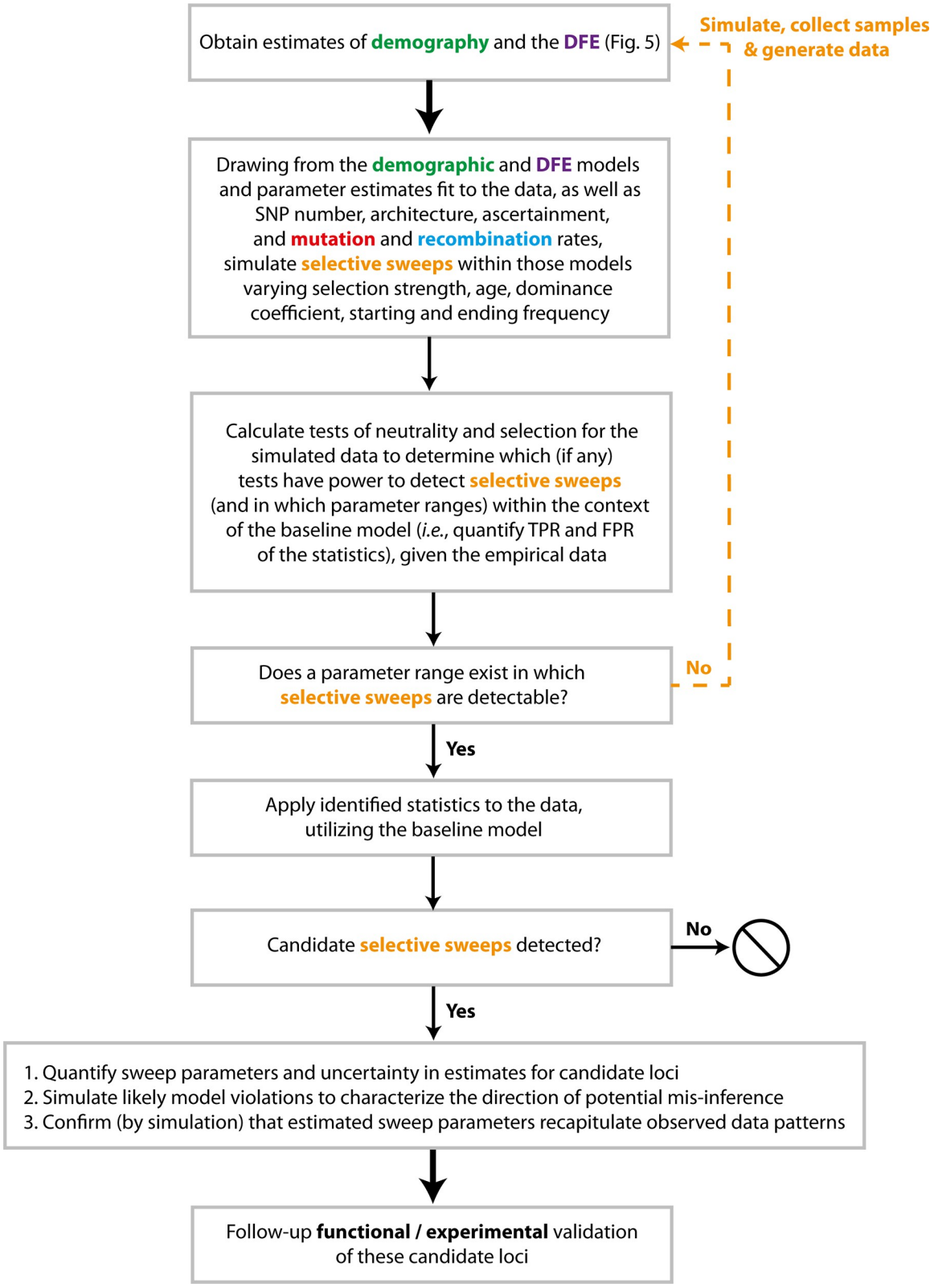

Fig 6. Diagram of important considerations in detecting selective sweeps. The color scheme matches that in Fig 5, with "selective sweeps" coded in orange. DFE, distribution of fitness effects; FPR, false positive rate; TPR, true positive rate.

https://doi.org/10.1371/journal.pbio.3001669.g006

hypothesis may be tested; perhaps only selective sweeps from rare mutations with selective effects greater than 1% and that have fixed within the last 0.1 N_e generations are detectable (see [111,112]), and any others could not be statistically distinguished from expected patterns under the baseline model. Hence, such an exercise provides a critically essential key for interpreting the resulting data analysis.

A consideration of alternative strategies

In this regard, it is worth mentioning 2 common approaches that may be viewed as alternatives to the strategy that we recommend. The first tactic concerns identifying patterns of variation that are uniquely and exclusively associated with one particular process, the presence of which could support that model regardless of the various underlying processes and details composing the baseline. For example, Fay and Wu's [113] *H*-statistic, capturing an expected pattern of high-frequency derived alleles generated by a selective sweep with recombination, was initially proposed as a powerful statistic for differentiating selective sweep effects from alternative models. Results from the initial application of the *H*-statistic were interpreted as evidence of widespread positive selection in the genome of *D. melanogaster*. However, Przeworski [112] subsequently demonstrated that the statistic was characterized by low power to detect positive selection, and that significant values could readily be generated under multiple neutral demographic models. The composite likelihood framework of Kim and Stephan [111] provided a significant improvement by incorporating multiple predictions of a selective sweep model and was subsequently built upon by Nielsen and colleagues [114] in proposing the SweepFinder approach. However, Jensen and colleagues [115] described low power and high false positive rates under certain neutral demographic models. The particular pattern of LD generated by a beneficial fixation with recombination described by Kim and Nielsen [116] and Stephan and colleagues [117] (and see [118]) was also found to be produced under an (albeit more limited) range of severe neutral population bottlenecks [119,120].

The point here is that the statistics themselves represent important tools for studying patterns of variation and are useful for visualizing multiple aspects of the data, but in any given empirical application, they are impossible to interpret without the definition of an appropriate baseline model and related power and false positive rates. Thus, the search for a pattern unique to a single evolutionary process is not a work-around, and, historically, such patterns rarely turn out to be process specific after further investigation. Even if a "bulletproof" test were to be someday constructed, it would not be possible to establish its utility without appropriate modeling, an examination of model violations, and extensive power/sensitivity–specificity analyses. But in reality, the simple fact is that some test statistics and estimation procedures perform well under certain scenarios, but not under others.

The second common strategy involves summarizing empirical distributions of a given statistic, and assuming that outliers of that distribution represent the action of a process of interest, such as positive selection (e.g., [121]). However, such an approach is problematic. To begin with, any distribution has outliers, and there will always exist a 5% or 1% tail for a chosen statistic under a given model. Consequently, a fit baseline model remains necessary to determine whether the observed empirical outliers are of an unexpected severity, and if the baseline model together with the hypothesized process has, for example, a significantly improved likelihood. Moreover, only by considering the hypothesized process within the context of the baseline model can one determine whether affected loci (e.g., those subject to recent sweeps) would even be expected to reside in the tails of the chosen statistical distribution, which is far from a given [72,122]. As such, approaches which may not necessarily require a defined baseline model in order to perform the initial analyses (e.g., [114]),

nonetheless require such modeling to accurately define expectations, power and false positive rates, and thus to interpret the significance of observed empirical outliers. For these reasons, the approach for which we advocate remains essential. As the appropriate baseline evolutionary model may differ strongly by organism and population, this performance must be carefully defined and quantified for each empirical analysis in order to accurately interpret results.

Conclusions

When it comes to evolutionary analyses, wanting to answer a question is not necessarily equivalent to being able to answer it. The ability of population genomics to address a hypothesis of interest with a given dataset is something that must be demonstrated, and this may be achieved by constructing a model composed of common biological and evolutionary processes, including the uncertainty in those underlying parameters, as well as the specific features of the dataset at hand. The variation in possible observational outcomes associated with a chosen baseline model and the ability to distinguish a hypothesized additional evolutionary process from such "background noise" are both quantifiable. Furthermore, even if the model were correct, there exists a limit on the precision of estimation imposed by the evolutionary variance in population statistics that requires description, and which no amount of sampling can remove.

Demonstrating that multiple models, and/or considerable parameter space within a model, are compatible with the data need not be viewed as a negative or weak finding. Quite the contrary—the honest presentation of such results motivates future theoretical, experimental, and empirical developments and analyses, which can further refine the list of competing hypotheses, and this article contains many citations that have succeeded in doing this. At the same time, this analysis can define which degrees of uncertainty are most damaging (e.g., Figs 3 and 4), also highlighting the simple fact that organisms in which basic biological processes have been better characterized are amenable to a wider range of potential evolutionary analyses. The impact of uncertainty in these parameters in nonmodel organisms may motivate taking a step back to first better characterize the basic biological processes such as mutation rates and spectra via mutation accumulation lines or pedigree studies, in order to improve resolution on the primary question of interest.

Importantly, the framework we describe will also generally identify many models and parameter realizations that are in fact inconsistent with the observed data. This "ruling out" process can often be just as useful as model fitting, and rejecting possible hypotheses is frequently the more robust exercise of the 2. The value of this narrowing down, rather than the enthusiastic promotion of individual scenarios, is worthy of heightened appreciation. Nevertheless, all models should not be viewed equally. Decades of work supporting the central tenets of the Neutral Theory [19], high-quality experimental and computational work quantifying mutation and recombination rates [77–79,83,84,123], constantly improving experimental and theoretical approaches to quantify the neutral and deleterious DFE from natural population, mutation accumulation, or directed mutagenesis data [44,90,124–126], and historical knowledge (e.g., anthropological, ecological, and clinical) of population size change or structure—combined with the fact that all of these factors may strongly shape observed levels and patterns of variation and divergence—justify their role in comprising the appropriate baseline model for genomic analysis.

Given this, and particularly after accounting for the inflation of variance contributed by uncertainty in relevant parameters, potential model violations, as well as the quantity and quality of data available in any given analysis, it will often be the case that many hypotheses of interest may not be addressable with the dataset and knowledge at hand. However, recognizing

that a question cannot be accurately answered, and defining the conditions under which it could become answerable, should be preferred over making unfounded and thus misleading claims. Consistent with this call for caution, however, it should equally be emphasized that the fit of a baseline model to data is certainly not inherent evidence that the model encompasses all relevant processes shaping the population. In reality, it is virtually guaranteed not to be all encompassing, and building these models involves simplifying more complex processes (for a helpful and more general perspective, see [127]). When an additional process cannot be satisfactorily detected, that may rather be viewed as a statement about statistical identifiability—the inability to distinguish a hypothesized process from other processes that are known to be acting—and in such scenarios, absence of evidence need not be taken as evidence of absence.

While the many considerations we describe may appear daunting, it is our hope that these recommendations may serve as a useful roadmap for future data analyses in population genomics, one that may inform not only the perspectives of authors, but also that of reviewers and editors as well. Helpfully, these strategies can save considerable time, money, and effort prior to the start of empirical data handling, by determining which questions are accessible to the researcher. If a question is addressable, this preliminary analysis can additionally define what types of data are needed, for example, the number of variants or sample size necessary to obtain sufficient power or how alternative data collections (e.g., temporal samples) could improve resolution. This further highlights the value of defining specific hypotheses and of studying specific patterns as opposed to running a general suite of software on each new dataset in the hopes of identifying something of interest—namely, one cannot define the power of a study to address an unformulated question. Such hypothesis-driven population genomics has resulted in a number of success stories over the past decade: systems in which specific hypotheses were formed, data were collected for the purpose, detailed population genomic analyses were designed, and, ultimately, important insights were gained about the evolutionary history of the population in question (e.g., the study of cryptic coloration has proven fruitful in this regard [3]). One feature common to these studies is interdisciplinarity: the utilization of population genetic theory and inference as described here, combined with classical genetic crosses, large-scale field studies, and genetic manipulation in order to connect genotype to phenotype to fitness and to validate statistical inference. Importantly, however, without a population genetic framework for defining hypotheses, quantifying processes contributing to observed variation and divergence, evaluating and distinguishing among competing models, and defining uncertainty and potential biases, the observations remain merely descriptive.

Acknowledgments

This paper is dedicated to the memories of Richard Lewontin (1929–2021) and Bill Hill (1940–2021). We would like to thank Nick Barton, Matt Dean, Fabian Freund, Ryan Gutenkunst, Mark Kirkpatrick, Sarah Marion, Mohamed Noor, Sally Otto, Kevin Thornton, and John Wakeley for helpful comments and suggestions.

References

1. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017; 541(7637):302–10. https://doi.org/10.1038/nature21347 PMID: 28102248
2. Renzette N, Gibson L, Jensen JD, Kowalik TF. Human cytomegalovirus intrahost evolution—a new avenue for understanding and controlling herpesvirus infections. Curr Opin Virol. 2014; 8:109–15. https://doi.org/10.1016/j.coviro.2014.08.001 PMID: 25154343
3. Harris RB, Irwin K, Jones MR, Laurent S, Barrett RDH, Nachman MW, et al. The population genetics of crypsis in vertebrates: recent insights from mice, hares, and lizards. Heredity. 2020; 124(1):1–14. https://doi.org/10.1038/s41437-019-0257-4 PMID: 31399719

4. Irwin KK, Renzette N, Kowalik TF, Jensen JD. Antiviral drug resistance as an adaptive process. Virus Evol. 2016; 2(1):vew014. https://doi.org/10.1093/ve/vew014 PMID: 28694997
5. Fisher RA. The genetical theory of natural selection. Clarendon Press, Oxford, UK; 1930.
6. Wright S. Evolution in Mendelian populations. Genetics. 1931; 16(2):97–159. https://doi.org/10.1093/genetics/16.2.97 PMID: 17246615
7. Haldane JBS. The causes of evolution. Longmans, London, UK; 1932.
8. Provine WB. The origins of theoretical population genetics. University of Chicago Press; 1971.
9. Darwin C. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London, UK; 1859.
10. Mendel G. Versuche über Pflanzenhybriden. Verh Naturforsch Ver Brünn. 1866; 4:3–47.
11. Lewontin RC. The genetic basis of evolutionary change. Columbia Univ. Press, New York; 1974.
12. Lewontin RC. Twenty-five years ago in Genetics: electrophoresis in the development of evolutionary genetics: milestone or millstone? Genetics. 1991; 128(4):657–62. https://doi.org/10.1093/genetics/128.4.657 PMID: 1916239
13. Kimura M. Evolutionary rate at the molecular level. Nature. 1968; 217(5129):624–6. https://doi.org/10.1038/217624a0 PMID: 5637732
14. Kimura M. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge; 1983.
15. Ohta T. Slightly deleterious mutant substitutions in evolution. Nature. 1973; 246(5428):96–8. https://doi.org/10.1038/246096a0 PMID: 4585855
16. King JL, Jukes TH. Non-Darwinian evolution. Science. 1969; 164(3881):788–98. https://doi.org/10.1126/science.164.3881.788 PMID: 5767777
17. Kern AD, Hahn MW. The neutral theory in light of natural selection. Mol Biol Evol. 2018; 35(6):1366–71. https://doi.org/10.1093/molbev/msy092 PMID: 29722831
18. Walsh B, Lynch M. Evolution and selection of quantitative traits. Oxford University Press, Oxford; 2018.
19. Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, et al. The importance of the Neutral Theory in 1968 and 50 years on: a response to Kern & Hahn 2018. Evolution. 2019; 73(1):111–4. https://doi.org/10.1111/evo.13650 PMID: 30460993
20. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974; 23(1):23–5. PMID: 4407212
21. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993; 134(4):1289–303. https://doi.org/10.1093/genetics/134.4.1289 PMID: 8375663
22. Charlesworth B, Jensen JD. The effects of selection at linked sites on patterns of genetic variability. Annu Rev Ecol Evol Syst. 2021; 52:177–97.
23. Ray N, Excoffier L. Inferring past demography using spatially explicit population genetic models. Hum Biol. 2009; 81(2–3):141–57. https://doi.org/10.3378/027.081.0303 PMID: 19943741
24. Beichman AC, Huerta-Sanchez E, Lohmueller KE. Using genomic data to infer historic population dynamics of non-model organisms. Annu Rev Ecol Evol Syst. 2018; 49:433–56.
25. Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. BMC Biol. 2017; 15(1):98. https://doi.org/10.1186/s12915-017-0434-y PMID: 29084517
26. Stephan W. Selective sweeps. Genetics. 2019; 211(1):5–13. https://doi.org/10.1534/genetics.118.301319 PMID: 30626638
27. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007; 8(8):610–8. https://doi.org/10.1038/nrg2146 PMID: 17637733
28. Bank C, Foll M, Ferrer-Admetlla A, Ewing G, Jensen JD. Thinking too positive? Revisiting current methods in population genetic selection inference. Trends Genet. 2014; 30(12):540–6. https://doi.org/10.1016/j.tig.2014.09.010 PMID: 25438719
29. Keightley PD, Halligan DL. Analysis and implications of mutational variation. Genetica. 2009; 136(2):359–69. https://doi.org/10.1007/s10709-008-9304-4 PMID: 18663587
30. Keightley PD. Rates and fitness consequences of new mutations in humans. Genetics. 2012; 190(2):295–304. https://doi.org/10.1534/genetics.111.134668 PMID: 22345605
31. Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. 2016; 17(11):704–14. https://doi.org/10.1038/nrg.2016.104 PMID: 27739533

32. Stumpf MP, McVean GA. Estimating recombination rates from population-genetic data. Nat Rev Genet. 2003; 4(12):959–68. https://doi.org/10.1038/nrg1227 PMID: 14631356

33. Auton A, Fledel-Alon A, Pfeifer SP, Venn O, Ségurel L, Street T, et al. A fine-scale chimpanzee genetic map from population sequencing. Science. 2012; 336(6078):193–8. https://doi.org/10.1126/science.1216872 PMID: 22422862

34. Spence JP, Song YS. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. Sci Adv. 2019; 5(10):eaaw9206. https://doi.org/10.1126/sciadv.aaw9206 PMID: 31681842

35. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002; 162(4):2025–35. https://doi.org/10.1093/genetics/162.4.2025 PMID: 12524368

36. Beaumont MA, Rannala B. The Bayesian revolution in genetics. Nat Rev Genet. 2004; 5(4):251–61. https://doi.org/10.1038/nrg1318 PMID: 15131649

37. Schraiber JG, Akey JM. Methods and models for unravelling human evolutionary history. Nat Rev Genet. 2015; 16(12):727–40. https://doi.org/10.1038/nrg4005 PMID: 26553329

38. Han E, Sinsheimer JS, Novembre J. Characterizing bias in population genetic inferences from low-coverage sequencing data. Mol Biol Evol. 2014; 31(3):723–35. https://doi.org/10.1093/molbev/mst229 PMID: 24288159

39. Pfeifer SP. Studying mutation rate evolution in primates–the effects of computational pipeline and parameter choices. GigaScience. 2021; 10(10):giab069. https://doi.org/10.1093/gigascience/giab069 PMID: 34673929

40. Ewing G, Jensen JD. The consequences of not accounting for background selection in demographic inference. Mol Ecol. 2016; 25(1):135–41. https://doi.org/10.1111/mec.13390 PMID: 26394805

41. Pouyet F, Aeschbacher S, Thiery A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Elife. 2018; 7:e36317. https://doi.org/10.7554/eLife.36317 PMID: 30125248

42. Dapper AL, Payseur BA. Effects of demographic history on the detection of recombination hotspots from linkage disequilibrium. Mol Biol Evol. 2018; 35(2):335–53. https://doi.org/10.1093/molbev/msx272 PMID: 29045724

43. Rousselle M, Maeva M, Nabholz B, Bataillon T, Galtier N. Overestimation of the adaptive substitution rate in fluctuating populations. Biol Lett. 2018; 14(5):20180055. https://doi.org/10.1098/rsbl.2018.0055 PMID: 29743267

44. Johri P, Charlesworth B, Jensen JD. Towards an evolutionarily appropriate null model: jointly inferring demography and purifying selection. Genetics. 2020; 215(1):173–92. https://doi.org/10.1534/genetics.119.303002 PMID: 32152045

45. Samuk K, Noor MAF. Gene flow biases population genetic inference of recombination rate. biorxiv 2021. https://www.biorxiv.org/content/10.1101/2021.09.26.461846v1.full.pdf

46. Myers S, Fefferman C, Patterson N. Can one learn history from the allelic spectrum? Theor Popul Biol. 2008; 73(3):342–8. https://doi.org/10.1016/j.tpb.2008.01.001 PMID: 18321552

47. Harris RB, Sackman A, Jensen JD. On the unfounded enthusiasm for soft selective sweeps II: examining recent evidence from humans, flies, and viruses. PLoS Genet. 2018; 14(12):e1007859. https://doi.org/10.1371/journal.pgen.1007859 PMID: 30592709

48. Louca S, Pennell MW. Extant timetrees are consistent with a myriad of diversification histories. Nature. 2020; 580(7804):502–5. https://doi.org/10.1038/s41586-020-2176-1 PMID: 32322065

49. Ford EB. Ecological genetics. Chapman and Hall, London, UK; 1975.

50. Garud N, Messer P, Buzbas E, Petrov D. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLoS Genet. 2015; 11(2):e1005004. https://doi.org/10.1371/journal.pgen.1005004 PMID: 25706129

51. Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the human genome. Mol Biol Evol. 2017; 34(8):1863–77. https://doi.org/10.1093/molbev/msx154 PMID: 28482049

52. Johri P, Stephan W, Jensen JD. Soft selective sweeps: addressing new definitions, evaluating competing models, and interpreting empirical outliers. PLoS Genet. 2022; 18(2):e1010022. https://doi.org/10.1371/journal.pgen.1010022 PMID: 35202407

53. Barton NH. Genetic hitchhiking. Philos Trans R Soc B. 2000; 355(1403):1553–62. https://doi.org/10.1098/rstb.2000.0716 PMID: 11127900

54. Poh YP, Domingues V, Hoekstra HE, Jensen JD. On the prospect of identifying adaptive loci in recently bottlenecked populations. PLoS ONE. 2014; 9(11):e110579. https://doi.org/10.1371/journal.pone.0110579 PMID: 25383711

55. Johri P, Riall K, Becher H, Excoffier L, Charlesworth B, Jensen JD. The impact of purifying and background selection on the inference of population history: problems and prospects. Mol Biol Evol. 2021; 38(7):2986–3003. https://doi.org/10.1093/molbev/msab050 PMID: 33591322

56. Campos JL, Charlesworth B. The effects on neutral variability of recurrent selective sweeps and background selection. Genetics. 2019; 212(1):287–303. https://doi.org/10.1534/genetics.119.301951 PMID: 30923166

57. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol. 2016; 12(5):e1004842. https://doi.org/10.1371/journal.pcbi.1004842 PMID: 27145223

58. Haller BC, Messer PW. SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. Mol Biol Evol. 2019; 36(3):632–7. https://doi.org/10.1093/molbev/msy228 PMID: 30517680

59. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 2009; 26(9):2097–108. https://doi.org/10.1093/molbev/msp119 PMID: 19535738

60. Thornton K. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics. 2003; 19(17):2325–7. https://doi.org/10.1093/bioinformatics/btg316 PMID: 14630667

61. Csilléry K, François O, Blum M. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 2012; 3:475–9.

62. Eldon B, Wakeley J. Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics. 2006; 172(4):2621–33. https://doi.org/10.1534/genetics.105.052175 PMID: 16452141

63. Matuszewski M, Hildebrandt ME, Achaz G, Jensen JD. Coalescent processes with skewed offspring distributions and non-equilibrium demography. Genetics. 2018; 208(1):323–38. https://doi.org/10.1534/genetics.117.300499 PMID: 29127263

64. Sackman A, Harris RB, Jensen JD. Inferring demography and selection in organisms characterized by skewed offspring distributions. Genetics. 2019; 211(3):1019–28. https://doi.org/10.1534/genetics.118.301684 PMID: 30651284

65. McVean G, Myers S, Hunt S, Deloukas P, Bentley D, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004; 304(5670):581–4. https://doi.org/10.1126/science.1092500 PMID: 15105499

66. Chan AH, Jenkins P, Song Y. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. PLoS Genet. 2012; 8(12):e1003090. https://doi.org/10.1371/journal.pgen.1003090 PMID: 23284288

67. Penalba JV, Wolf JB. From molecules to populations: appreciating and estimating recombination rate variation. Nat Rev Genet. 2020; 21(8):476–92. https://doi.org/10.1038/s41576-020-0240-1 PMID: 32472059

68. Charlesworth D, Charlesworth B, Morgan MT. The pattern of neutral molecular variation under the background selection model. Genetics. 1995; 141(4):1619–32. https://doi.org/10.1093/genetics/141.4.1619 PMID: 8601499

69. Charlesworth B. Background selection 20 years on. The Wilhelmine E. Key 2012 invitational lecture. J Hered. 2013; 104(2):161–71. https://doi.org/10.1093/jhered/ess136 PMID: 23303522

70. Pfeifer SP. From next-generation resequencing reads to a high quality variant data set. Heredity. 2017; 118(2):111–24. https://doi.org/10.1038/hdy.2016.102 PMID: 27759079

71. Nielsen R. Population genetic analysis of ascertained SNP data. Hum Genomics. 2004; 1(3):218–24. https://doi.org/10.1186/1479-7364-1-3-218 PMID: 15588481

72. Thornton KR, Jensen JD. Controlling the false positive rate in multi-locus genome scans for selection. Genetics. 2007; 175(2):737–50. https://doi.org/10.1534/genetics.106.064642 PMID: 17110489

73. Malaspinas AS, Malaspinas O, Evans SN, Slatkin M. Estimating allele age and selection coefficient from time-serial data. Genetics. 2012; 192(2):599–607. https://doi.org/10.1534/genetics.112.140939 PMID: 22851647

74. Foll M, Shim H, Jensen JD. A Wright-Fisher ABC-based approach for inferring per-site effective population sizes and selection coefficients from time-sampled data. Mol Ecol Resour. 2015; 15(1):87–98. https://doi.org/10.1111/1755-0998.12280 PMID: 24834845

75. Ferrer-Admetlla A, Leuenberger C, Jensen JD, Wegmann D. An approximate Markov model for the Wright-Fisher diffusion and its application to time series data. Genetics. 2016; 203(2):831–46. https://doi.org/10.1534/genetics.115.184598 PMID: 27038112

76. Lynch M, Ho WC. The limits to estimating population-genetic parameters with temporal data. Genome Biol Evol. 2020; 12(4):443–55. https://doi.org/10.1093/gbe/evaa056 PMID: 32181820

77. Pfeifer SP. Spontaneous mutation rates. *InThe Molecular Evolutionary Clock. Theory and Practice*. Springer Nature; 2020.

78. Smith TCA, Arndt PF, Eyre-Walker A. Large scale variation in the rate of germ-line *de novo* mutations, base composition, divergence and diversity in humans. PLoS Genet. 2018; 14(3):e1007254. https://doi.org/10.1371/journal.pgen.1007254 PMID: 29590096

79. Ness RW, Morgan AD, Radhakrishnan V, Colegrave N, Keightley PD. Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. Genome Res. 2015; 25(11):1739–49. https://doi.org/10.1101/gr.191494.115 PMID: 26260971

80. Maddamsetti R, Grant NA. Divergent evolution of mutation rates and biases in the long-term evolution experiment with *Escherichia coli*. Genome Biol Evol. 2020; 12(9):1591–603. https://doi.org/10.1093/gbe/evaa178 PMID: 32853353

81. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir G, Gudjonsson S, Richardsson B, et al. 2002. A high-resolution recombination map of the human genome. Nat Genet. 2002; 31(3):241–7. https://doi.org/10.1038/ng917 PMID: 12053178

82. Cox A, Ackert-Bicknell C, Dumont B, Ding Y, Tzenova Bell J, Brockmann G, et al. A new standard genetic map for the laboratory mouse. Genetics. 2009; 182(4):1335–44. https://doi.org/10.1534/genetics.109.105486 PMID: 19535546

83. Comeron JM, Ratnappan R, Bailin S. The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. 2012; 8(10):e1002905. https://doi.org/10.1371/journal.pgen.1002905 PMID: 23071443

84. Auton A, McVean G. Estimating recombination rates from genetic variation in humans. Methods Mol Biol. 2012; 856:217–37. https://doi.org/10.1007/978-1-61779-585-5_9 PMID: 22399461

85. Pfeifer SP. A fine-scale genetic map for vervet monkeys. Mol Biol Evol. 2020; 37(7):1855–65. https://doi.org/10.1093/molbev/msaa079 PMID: 32211856

86. Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size change. Genetics. 2010; 186(3):983–95. https://doi.org/10.1534/genetics.110.118661 PMID: 20739713

87. Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L. On the accumulation of deleterious mutations during range expansions. Mol Ecol. 2013; 22(24):5972–82. https://doi.org/10.1111/mec.12524 PMID: 24102784

88. Peischl S, Kirkpatrick M, Excoffier L. Expansion load and the evolutionary dynamics of a species range. Am Nat. 2015; 185(4):E81–93. https://doi.org/10.1086/680220 PMID: 25811091

89. Johri P, Eyre-Walker A, Gutenkunst RN, Lohmueller KE, Jensen JD. On the prospect of achieving accurate joint estimation of selection with population history. In revision. Genome Biol Evol.

90. Keightley PD, Eyre-Walker A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics. 2007; 177(4):2251–61. https://doi.org/10.1534/genetics.107.080663 PMID: 18073430

91. Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics. 2011; 189(4):1427–37. https://doi.org/10.1534/genetics.111.131730 PMID: 21954160

92. Lynch M. The origins of genome architecture. Sinauer Associates, Sunderland, MA; 2007.

93. Singh ND, Bauer DuMont VL, Hubisz MJ, Nielsen R, Aquadro CF. Patterns of mutation and selection at synonymous sites in *Drosophila*. Mol Biol Evol. 2007; 24(12):2687–97. https://doi.org/10.1093/molbev/msm196 PMID: 18000010

94. Zeng K, Charlesworth B. Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. J Mol Evol. 2010; 70(1):116–28. https://doi.org/10.1007/s00239-009-9314-6 PMID: 20041239

95. Choi JY, Aquadro CF. Recent and long term selection across synonymous sites in *Drosophila ananassae*. J Mol Evol. 2016; 83(1–2):50–60. https://doi.org/10.1007/s00239-016-9753-9 PMID: 27481397

96. Comeron JM. Background selection as baseline for nucleotide variation across the *Drosophila* genome. PLoS Genet. 2014; 10(6):e1004434. https://doi.org/10.1371/journal.pgen.1004434 PMID: 24968283

97. Comeron JM. Background selection as a null hypothesis in population genomics: insights and challenges from *Drosophila* studies. Philos Trans R Soc B. 2017; 372(1736):20160471. https://doi.org/10.1098/rstb.2016.0471 PMID: 29109230

98. Cranmer K, Brehmer J, Louppe G. The frontier of simulation-based inference. Proc Natl Acad Sci U S A. 2020; 117(48):30055–62. https://doi.org/10.1073/pnas.1912789117 PMID: 32471948

99. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. 2005; 102(22):7882–7. https://doi.org/10.1073/pnas.0502300102 PMID: 15905331

100. Ragsdale A, Moreau C, Gravel S. Genomic inference using diffusion models and the allele frequency spectrum. Curr Opin Gen Deve. 2018; 53:140–7. https://doi.org/10.1016/j.gde.2018.10.001 PMID: 30366252

101. Gutenkunst R, Hernandez R, Williamson S, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP data. PLoS Genet. 2009; 5(10):e1000695. https://doi.org/10.1371/journal.pgen.1000695 PMID: 19851460

102. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013; 9(10):e1003905. https://doi.org/10.1371/journal.pgen.1003905 PMID: 24204310

103. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. Nat Genet. 2019; 51(9):1330–8. https://doi.org/10.1038/s41588-019-0483-y PMID: 31477934

104. Steinrücken M, Kamm J, Spence J, Song YS. Inference of complex population histories using whole-genome sequences from multiple populations. Proc Natl Acad Sci U S A. 2019; 116(34):17115–20. https://doi.org/10.1073/pnas.1905060116 PMID: 31387977

105. Torres R, Szpiech Z, Hernandez RD. Human demographic history has amplified the effects of background selection across the genome. PLoS Genet. 2018; 14(6):e1007387. https://doi.org/10.1371/journal.pgen.1007387 PMID: 29912945

106. Thornton KR. A C++ template library for efficient forward-time population genetic simulation of large populations. Genetics. 2014; 198(1):157–66. https://doi.org/10.1534/genetics.114.165019 PMID: 24950894

107. Kelleher J, Thornton K, Ashander J, Ralph P. Efficient pedigree recording for fast population genetics simulation. PLoS Comput Biol. 2018; 14(11):e1006581. https://doi.org/10.1371/journal.pcbi.1006581 PMID: 30383757

108. Durrett R, Schweinsberg J. Approximating selective sweeps. Theor Popul Biol. 2004; 66(2):129–38. https://doi.org/10.1016/j.tpb.2004.04.002 PMID: 15302222

109. Hallatschek O. Selection-like biases emerge in population models with recurrent jackpot events. Genetics. 2018; 210(3):1053–73. https://doi.org/10.1534/genetics.118.301516 PMID: 30171032

110. Lapierre M, Lambert A, Achaz G. Accuracy of demographic inference from the site frequency spectrum: the case of the Yoruba population. Genetics. 2017; 206(1):439–49. https://doi.org/10.1534/genetics.116.192708 PMID: 28341655

111. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 2002; 160(2):765–77. https://doi.org/10.1093/genetics/160.2.765 PMID: 11861577

112. Przeworski M. The signature of positive selection at randomly chosen loci. Genetics. 2002; 160(3):1179–89. https://doi.org/10.1093/genetics/160.3.1179 PMID: 11901132

113. Fay J, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000; 155(3):1405–13. https://doi.org/10.1093/genetics/155.3.1405 PMID: 10880498

114. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante CD. Genomic scans for selective sweeps using SNP data. Genome Res. 2005; 15(11):1566–75. https://doi.org/10.1101/gr.4252305 PMID: 16251466

115. Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics. 2005; 170(3):1401–10.

116. Kim Y, Nielsen R. Linkage disequilibrium as a signature of selective sweeps. Genetics. 2004; 167(3):1513–24. https://doi.org/10.1534/genetics.103.025387 PMID: 15280259

117. Stephan W, Song YS, Langley CH. Hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics. 2006; 172(4):2647–63. https://doi.org/10.1534/genetics.105.050179 PMID: 16452153

118. McVean G. The structure of linkage disequilibrium around a selective sweep. Genetics. 2007; 175(3):1395–406. https://doi.org/10.1534/genetics.106.062828 PMID: 17194788

119. Jensen JD, Thornton KR, Bustamante CD, Aquadro CF. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in non-equilibrium populations. Genetics. 2007; 176(4):2371–9. https://doi.org/10.1534/genetics.106.069450 PMID: 17565955

120. Crisci J, Poh YP, Mahajan S, Jensen JD. The impact of equilibrium assumptions on tests of selection. Front Genet. 2013; 4:235. https://doi.org/10.3389/fgene.2013.00235 PMID: 24273554

121. Garud N, Messer P, Petrov D. Detection of hard and soft selective sweeps from *Drosophila melanogaster* population genomic data. PLoS Genet. 2021; 17(2):e1009373. https://doi.org/10.1371/journal.pgen.1009373 PMID: 33635910

122. Teshima K, Coop G, Przeworski M. How reliable are empirical genome scans for selective sweeps? Genome Res. 2006; 16(6):702–12. https://doi.org/10.1101/gr.5105206 PMID: 16687733

123. Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, et al. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A. 2008; 105(27):9272–7. https://doi.org/10.1073/pnas.0803466105 PMID: 18583475

124. Bank C, Hietpas RT, Wong A, Bolon DNA, Jensen JD. A Bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. Genetics. 2014; 196(3):841–52. https://doi.org/10.1534/genetics.113.156190 PMID: 24398421

125. Foll M, Poh YP, Renzette N, Ferrer-Admetlla A, Shim H, Malaspinas AS, et al. Influenza virus drug resistance: a time-sampled population genetics perspective. PLoS Genet. 2014; 10(2):e1004185.

126. Böndel KB, Kraemer SA, Samuels TS, McClean D, Lachapelle J, Ness RW, et al. Inferring the distribution of fitness effects of spontaneous mutations in *Chlamydomonas reinhardtii*. PLoS Biol. 2019; 17(6):e3000192. https://doi.org/10.1371/journal.pbio.3000192 PMID: 31242179

127. Gelman A, Shalizi CR. Philosophy and the practice of Bayesian statistics. Br J Math Stat Psychol. 2013; 66(1):8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x PMID: 22364575