# 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[4], Michael Lynch[1], Gil McVean[7], Bret A. Payseur[8], Susanne P. Pfeifer[1], Wolfgang Stephan[9], and Jeffrey D. Jensen[1]

[1] School of Life Sciences, Arizona State University, Tempe, US
[2] Department of Molecular Biology and Genetics, Cornell University, Ithaca, US
[3] School of Biological Sciences, University of Bristol, Bristol, UK
[4] Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK
[5] Institute of Ecology and Evolution, University of Berne, Berne, CH
[6] School of Life Sciences, University of Sussex, Brighton, UK
[7] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK
[8] Laboratory of Genetics, University of Wisconsin-Madison, Madison, US
[9] Natural History Museum, Berlin, DE

keywords: population genetics, population genomics, statistical inference, model-fitting, positive selection, purifying selection, background selection, selective sweeps, genome scans, demography

## ABSTRACT

The field of population genomics has grown rapidly with 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 non-adaptive processes is a prerequisite for accurate inference. We here describe the perils of these tendencies, present our 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.

*A brief introduction to population genomic inference*

Population genomic inference – the use of molecular variation and divergence data to infer evolutionary processes – has become widely embraced and highly utilized in fields ranging from evolutionary biology, to ecology, to anthropology, and to medicine. The underlying questions may be demographic in nature, be it estimating the timing of the peopling of the world (Nielsen *et al*. 2017) or of viral transmission in a congenitally infected newborn (Renzette *et al*. 2014); alternatively, they may concern the selective history of specific populations, be it identifying mutations that confer cryptic coloration in species adapting to major post-glacial climatic and geological changes (Harris *et al*. 2020) or viral drug-resistance to clinical therapeutics (Irwin *et al*. 2016).

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*., Fisher 1930; Wright 1931; Haldane 1932; for a historical overview, see Provine 1971). This work demonstrated the possibility of studying evolution at the genetic level, integrating the revolutionary ideas of Darwin (1859) with the turn-of-the-century appreciation of Mendel's (1866) research. However, as famously described by Lewontin (1974), 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 Lewontin 1991), 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 (Kimura 1968, 1983; Ohta 1973) – an advance for which molecular biology also provided

support (King & Jukes 1969). Despite some claims to the contrary (Kern & Hahn 2018), Kimura and Ohta's initial postulates have since been largely validated (Walsh & Lynch 2018; Jensen *et al*. 2019), 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 on 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 constantly raining down on the genomes of all species. Accordingly, positive selection per nucleotide is rare compared to genetic drift and purifying selection. 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 (Maynard Smith & Haigh 1974; Charlesworth *et al*. 1993; reviewed in Charlesworth & Jensen 2021).

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.,* Ray & Excoffier 2009; Beichman *et al.* 2018), for identifying beneficial mutations based on patterns associated with selective sweeps (*e.g.,* Booker *et al*. 2017; Stephan 2019), for quantifying the distribution of fitness effects (DFE) of newly arising mutations (*e.g.,* Eyre-Walker & Keightley 2007; Bank *et al*. 2014a), as well as for estimating rates of mutation (*e.g.,* Keightley & Halligan 2009; Keightley 2012; Lynch *et al*. 2016) and recombination (*e.g.,* Stumpf & McVean 2003; Auton & McVean 2012; Spence & Song 2019).

These approaches operate in a variety of statistical frameworks (see Beaumont *et al.* 2002; Beaumont & Rannala 2004; Schraiber & Akey 2015), 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 between-species levels and patterns of divergence at contrasted site classes (*e.g.,* synonymous versus non-synonymous 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 (Han *et al*. 2014; Pfeifer 2021). 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 (Ewing & Jensen 2016; Pouyet *et al.* 2018), whereas estimates of selection intensity and recombination rate may be highly biased when neglecting demographic effects (Dapper & Payseur 2018; Rousselle *et al*. 2018; Johri *et al*. 2020). 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

147 estimate the demographic history. An additional challenge, and a frustration for many, is that
148 there is no single 'best approach'; the correct analysis tools to use, and indeed which questions
149 can be answered at all, depend entirely on the details of the organism under study (Myers *et al*.
150 2008). Specifically, biological parameters that vary among species – including evolutionary
151 parameters (*e.g*., effective population size ($N_e$), mutation rates, recombination rates, and
152 population structure and history), genome structure (*e.g*., the distribution of functional sites along
153 the genome), and life history traits (*e.g*., mating system) – must all be considered in order to
154 define addressable hypotheses and optimal approaches.

155 Beyond these initial considerations, a more difficult issue often emerges. Namely, very
156 different models may be found to provide a good fit to the observed data (*e.g*., Harris *et al*. 2018;
157 and see Louca & Pennell 2020 for a phylogenetic perspective on the topic). In other words,
158 particular parameter combinations may be found under competing models that are all capable of
159 predicting the observed patterns of variation. For example, assuming neutrality, one may match
160 an empirical observation at a locus by fitting the timing, severity, and duration of a population
161 bottleneck; or alternatively, assuming a constant population size, by fitting the rate and mean
162 strength of selective sweeps. This fact alone implies a simple truism: the ability to fit the
163 parameters of one's preferred model to data does not alone represent proof of biological reality.
164 Rather, it suggests that this model is one – out of potentially very many – that represents a viable
165 hypothesis, which should be further examined via subsequent analyses or experimentation.

166 Examples abound of enthusiastic promotion of a single preferred model, only to be
167 tempered by subsequent demonstrations of the fit of alternative and often simpler / more
168 biologically realistic models. For example, the view that segregating alleles may be commonly
169 maintained by balancing selection (Ford 1975) was tempered by the realization that genetic drift

is often a sufficient explanation (Kimura 1983), and the view that genome-wide selective sweeps on standing variation are pervasive (Garud *et al*. 2015; Schrider & Kern 2017) was tempered by the realization that neutral population histories can result in similar patterns (Harris *et al*. 2018). 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, Figure 1 presents three 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 two incorrect models (population size change assuming strict neutrality, and recurrent selective sweeps assuming constant population size). As shown, 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*., Barton 2000; Poh *et al.* 2014; Ewing & Jensen 2016; Johri *et al*. 2021).

*Constructing an appropriate baseline model for population genomic analysis*

The somewhat disheartening exercise depicted in Figure 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 in this regard 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, reassortment, and gene conversion (as applicable), purifying selection acting on functional regions and its effects on linked variants (*i.e.,* background selection: Charlesworth *et al*. 1993, 1995; Charlesworth 2013), as well as genetic drift as modulated by, amongst 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 based on mating system, progeny distributions, ploidy, and so on. 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. Firstly, 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. Secondly, it is not advisable to compare one's empirical observation 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 (Pfeifer 2017). The dataset may also involve a certain ascertainment scheme, either for the variants surveyed (Nielsen 2004), or given some pre-defined criteria for investigating specific genomic regions (*e.g.*, regions representing genomic outliers with respect to a chosen summary statistic; Thornton & Jensen 2007). For the sake of illustration, Figure 2 follows the same format as Figure 1, but considers two scenarios:

population growth with background selection and selective sweeps, and the same scenario together with data ascertainment (in this case, an under-calling of the singleton class). As shown, 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. 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 consequently 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 (Malaspinas *et al.* 2012; Foll *et al*. 2015; Ferrer-Admetlla *et al*. 2016; Lynch & Ho 2020), 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.

With these concerns having 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 (Pfeifer 2020a). However, even for these species, ascertainment issues remain complicating (Smith *et al*. 2018), variation amongst individuals may

be substantial (Ness *et al.* 2015), and estimates only represent a temporal snapshot of rates and patterns that are probably changing over evolutionary time-scales and may be affected by the environment (Lynch *et al*. 2016; Maddamsetti & Grant 2020). 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 crossing rates in cM per Mb have been estimated by comparing genetic maps based on crosses or pedigrees with physical maps (*e.g.,* Kong *et al.* 2002; Cox *et al.* 2009; Comeron *et al.* 2012). In addition, recombination rates scaled by the effective population size have also been estimated from patterns of LD (*e.g*., Auton *et al.* 2012; Pfeifer 2020b) – though this approach typically requires assumptions about evolutionary processes that may be violated (*e.g.,* Dapper & Payseur 2018). As with mutation, the effects on inference of changing the recombination rate – 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 effects. This is a difficult task given the virtually infinite number of potential demographic hypotheses (*e.g*., Chikhi *et al*. 2010); furthermore the interaction of

selection with demography is inherently non-trivial and difficult to treat separately (*e.g.*, Peischl *et al*. 2013, 2015; Johri *et al.* 2021). 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 which is often estimated in both continuous and discrete forms. 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 non-synonymous sites, which presumably experience considerable purifying selection (Keightley & Eyre-Walker 2007; Eyre-Walker & Keightley 2009; Schneider *et al*. 2011). This step-wise approach may become problematic, however, for organisms in which synonymous sites are not themselves neutral (Lynch 2007; Singh *et al*. 2007; Zeng & Charlesworth 2010; Choi & Aquadro 2016; Long *et al*. 2018), 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 non-synonymous sites (Pouyet *et al.* 2018; and see Comeron 2014, 2017).

In an attempt to address some of these concerns, Johri *et al.* (2020) recently developed an approximate Bayesian computation (ABC) approach that relaxes the assumption of synonymous site neutrality and accounts 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. While such simulation-

285 based inference (see Cranmer *et al.* 2020), including ABC, provides one promising platform for
286 joint estimation of demographic history and selection, progress on this front has been made using
287 alternative frameworks as well (Williamson *et al.* 2005; Ragsdale *et al.* 2018), and developing
288 analytical expectations under these complex models should remain as the ultimate, if distant,
289 goal. Alternatively, in functionally-sparse genomes with sufficiently high rates of recombination,
290 such that assumptions of strict neutrality are viable for some genomic regions, multiple well-
291 performing approaches have been developed for estimating the parameters of much more
292 complex demographic models (*e.g.*, Gutenkunst *et al*. 2009; Excoffier *et al*. 2013; Kelleher *et al.*
293 2019; Steinrücken *et al*. 2019). In organisms for which such approaches are applicable (*e.g.*,
294 certain large, coding-sparse vertebrate and land-plant genomes), this intergenic demographic
295 estimation assuming strict neutrality may helpfully be compared to estimates derived from data
296 in or near coding regions that account for the effects of direct and linked purifying selection
297 (Pouyet *et al*. 2018; Torres *et al*. 2018; Johri *et al*. 2020). For newly studied species lacking
298 functional annotation and information about coding density, following the joint estimation
299 procedure would remain as the more satisfactory strategy in order to account for possible
300 background selection effects.
301
302 *Quantifying uncertainty in model-choice and parameter estimation, investigating potential model*
303 *violations, and defining answerable questions*
304 One of the useful aspects of these types of analyses is the ability to incorporate uncertainty in
305 underlying parameters under relatively complex models, in order to determine the impact of such
306 uncertainty on downstream inference. The computational burden of incorporating variability in
307 mutation and recombination rate estimates, or drawing from the confidence- or credibility-

intervals of demographic or DFE parameters, may be met with multiple highly-flexible simulation tools (Thornton 2014; Kelleher *et al.* 2018; Haller & Messer 2019). 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 may violate this assumption (*i.e.*, with the large progeny number distributions associated with many plants, viruses, and marine spawners), these distributions may too be modeled in order to quantify potential downstream mis-inference.

To illustrate this point, Figure 3 considers two 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 under-estimated 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 may both generate multiple-merger events (Durrett & Schweinsberg 2004; Hallatschek 2018; Matuszewski *et al.* 2018; Sackman *et al.* 2019). Similarly, one may investigate the assumptions of constant mutation or recombination rates when they are in reality variable. As shown in Figure 4, when these rates are assumed 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

331 modified. Notably, this rate heterogeneity may inflate the inferred strength of selective sweeps.
332 While Figures 3 and 4 serve as examples, the same investigations may be made for cases such as
333 a fixed selective effect when there is in reality a distribution, neutral and unlinked variants when
334 there is in reality linkage disequilibrium, and so on. Simply put, even if a particular biological
335 process / parameter is not being directly estimated, its consequences can nonetheless be
336 explored.

337 As detailed in Box 1, with such a model incorporating both biological and stochastic
338 variance as well as statistical uncertainty in parameter estimates, and with an understanding of
339 the role of likely model violations, one may investigate which additional questions / hypotheses
340 can be addressed with the data at hand. By using a simulation approach starting with the baseline
341 model and adding hypothesized processes on top, it is possible to quantify the extent to which
342 models, and the parameters underlying those models, may be differentiated and which result in
343 overlapping or indistinguishable patterns in the data (*e.g.*, Lapierre *et al*. 2017). For example, if
344 the goal of a given study is to identify recent beneficial fixations in a genome – be they
345 potentially associated with high-altitude adaptation in humans, crypsis in mice, or drug-
346 resistance in a virus – one may begin with the constructed baseline model and simulate selective
347 sweeps under that model. As described in Box 2, by varying the strengths, rates, ages, dominance
348 and epistasis coefficients of beneficial mutations, the patterns in the SFS, LD, and/or divergence
349 that may differentiate the addition of such selective sweep parameters from the baseline
350 expectations can be quantified. Moreover, any intended empirical analyses can be evaluated
351 using simulated data (*i.e.*, the baseline, compared to the baseline + the hypothesis) to define the
352 power and false-positive rates associated. If the differences in resulting patterns cannot be
353 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.*, Poh *et al.* 2014). If the results are favorable, this analysis can further quantify the extent to which the 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 Kim & Stephan 2002; Przeworski 2002), 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.

In this regard, it is worth mentioning two 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 & Wu's (2000) *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 *Drosophila melanogaster*. However, Przeworski (2002) 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 & Stephan (2002) provided a significant improvement by incorporating multiple predictions of a selective sweep model, and was subsequently built upon by Nielsen *et al.* (2005) in proposing the SweepFinder approach. However, Jensen *et al*. (2005) similarly 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 & Nielsen (2004) and Stephan *et al*. (2006) (and see McVean 2007), was also found to be produced under a more limited range of severe neutral population bottlenecks (Jensen *et al*. 2007; Crisci *et al*. 2013). The point here is that the statistics themselves represent important contributions for studying patterns of variation, 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 'bullet-proof' test were to be someday constructed, it would not be possible to establish its utility without appropriate modeling, an examination of model violations, and 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.,* Garud *et al.* 2021). 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 if the observed empirical outliers are of an unexpected severity in the empirical distribution, 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, may one determine if affected loci (*e.g*., subject to recent sweeps) would even be expected to reside in the tails of the chosen statistical distribution, which is far from a given (Teshima *et al.* 2006; Thornton & Jensen 2007). Hence,

400 the approach for which we advocate remains essential for defining expectations, power, and
401 false-positive rates, and thus to interpret the significance of observed empirical outliers. As the
402 appropriate baseline evolutionary model may differ strongly by organism and population, this
403 performance must be carefully defined and quantified for each empirical analysis in order to
404 accurately interpret results.
405
406 *Closing thoughts*
407 When it comes to evolutionary analyses, wanting to answer a question is not necessarily
408 equivalent to being able to answer it. The ability of population genomics to address a hypothesis
409 of interest with a given dataset is something that must be demonstrated, and this may be achieved
410 by constructing a model composed of common biological and evolutionary processes, including
411 the uncertainty in those underlying parameters, as well as the specific features of the dataset at
412 hand. The variation in possible observational outcomes associated with a chosen baseline model,
413 and the ability to distinguish an hypothesized additional evolutionary process from that
414 'background noise', are both quantifiable. Furthermore, even if the model were to be correct,
415 there exists a limit on the precision of estimation imposed by the evolutionary variance in
416 population statistics that requires description, and which no amount of sampling can remove.
417 Demonstrating that multiple models, and/or considerable parameter space within a model,
418 are compatible with the data need not be viewed as a negative or weak finding. Quite the
419 contrary – the honest presentation of such results motivates future theoretical, experimental, and
420 empirical developments and analyses that can further refine the list of competing hypotheses, and
421 this article contains many citations that have succeeded in this vein. At the same time, this
422 analysis can define which degrees of uncertainty are most damaging (*e.g.*, Figures 3 and 4), also

423 highlighting the simple fact that organisms in which basic biological processes have been better
424 characterized are amenable to a wider-range of potential evolutionary analyses. The impact of
425 uncertainty in these parameters in non-model organisms may motivate taking a step back to first
426 better characterize the basic biological processes such as mutation rates and spectra via mutation-
427 accumulation lines or pedigree studies, in order to improve resolution on the primary question of
428 interest.

429 Importantly, the framework we describe will also generally identify many models and
430 parameter realizations that are in fact inconsistent with the observed data. This 'ruling-out'
431 process can often be just as useful as model-fitting, and rejecting possible hypotheses is
432 frequently the more robust exercise of the two. The value of this narrowing down, rather than the
433 enthusiastic promotion of individual scenarios, is worthy of heightened appreciation.
434 Nevertheless, all models should not be viewed equally. Decades of work supporting the central
435 tenets of the Neutral Theory (Jensen *et al*. 2019), high-quality experimental and computational
436 work quantifying mutation rate and recombination rate (*e.g.,* Lynch *et al*. 2008; Auton &
437 McVean 2012; Comeron *et al.* 2012; Ness *et al*. 2015; Smith *et al*. 2018; Pfeifer 2020a),
438 constantly improving experimental and theoretical approaches to quantify the neutral and
439 deleterious DFE from natural population, mutation-accumulation, or directed mutagenesis data
440 (*e.g*., Keightley & Eyre-Walker 2007; Bank *et al*. 2014b; Foll *et al*. 2014; Böndel *et al*. 2019;
441 Johri *et al*. 2020), and often historical knowledge (*e.g*., anthropological, ecological, clinical) of
442 population size change or structure – combined with the fact that all of these processes may
443 strongly shape observed levels and patterns of variation and divergence – justify their status in
444 comprising the appropriate baseline model for genomic analysis. Given this, and particularly
445 once accounting for the inflation of variance contributed by uncertainty in these parameters,

446 potential model violations, as well as the quantity and quality of data available in any given
447 analysis, it will often be the case that many hypotheses of interest may not be addressable with
448 the dataset and knowledge at hand. However, recognizing that a question cannot be accurately
449 answered, and defining the conditions under which it could become answerable, must be the
450 preference over making unfounded and thus misleading claims. Consistent with this call for
451 caution however, it should equally be emphasized that the fit of a baseline model to data is
452 certainly not inherent evidence that the model encompasses all relevant processes shaping the
453 population. In reality, it is virtually guaranteed not to be all-encompassing, and building these
454 models involves simplifying more complex processes (for a helpful and more general
455 perspective, see Gelman & Shalizi 2013). When an additional process on top of this model
456 cannot be satisfactorily detected, that may rather be viewed as a statement about statistical
457 identifiability – the inability to distinguish a hypothesized process from other processes that are
458 known to be acting – and in such scenarios, absence of evidence need not be taken as evidence of
459 absence.

460 While the many considerations described may appear daunting, it is our hope that this
461 may serve as a useful roadmap for future data analyses in population genomics, one that may
462 inform not only the perspectives of authors, but also that of reviewers and editors as well.
463 Helpfully, these strategies can save considerable time, money, and effort prior to the start of
464 empirical data handling, by determining which questions are accessible to the researcher. If a
465 question is addressable, this preliminary analysis can additionally define what types of data are
466 needed; for example, the number of variants or sample size necessary to obtain sufficient power,
467 or how alternative data collections (*e.g.*, temporal samples) could improve resolution. This
468 further highlights the value of defining specific hypotheses and of studying specific patterns as

469 opposed to running a general suite of software on each new dataset in the hopes of identifying
470 something of interest – namely, one cannot define power to address an unformulated question.
471 Such hypothesis-driven population genomics has resulted in a number of success stories over the
472 past decade; systems in which specific hypotheses were formed, data was collected for the
473 purpose, detailed population genomic analyses were designed, and ultimately important insights
474 were gained about the evolutionary history of the population in question (*e.g.*, the study of
475 cryptic coloration has proven particularly fruitful in this regard, Harris *et al.* 2020). One feature
476 common amongst these studies is interdisciplinarity: the utilization of population genetic theory
477 and inference as described here, combined with classical genetic crosses, large-scale field
478 studies, and genetic manipulation in order to connect genotype to phenotype to fitness and to
479 validate statistical inference. Importantly however, without the population genetic framework to
480 define hypotheses, quantify processes contributing to observed variation and divergence,
481 evaluate and distinguish amongst competing models, and define uncertainty and potential biases,
482 these empirical observations alone remain merely descriptive.
483
484
485 **METHODS**
486 Both forward- and coalescent-simulations were performed (see below for details) for (1) the
487 inference of demographic history using ABC assuming complete neutrality, (2) the inference of
488 parameters of positive selection using ABC assuming constant population size, and (3) to obtain
489 test datasets representing different evolutionary scenarios. In all cases, a chromosomal segment
490 of 99,012 bp was simulated with an intron-exon-intergenic structure resembling the *D.*
491 *melanogaster* genome. Each gene comprised five exons (of 300 bp each) and four 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 *et al.* (2019), assuming an effective population size ($N_e$) of $10^6$ individuals with a mean mutation rate ($\mu$) of $4.5 \times 10^{-9}$ per bp/gen and a mean recombination rate ($r$) of $1 \times 10^{-8}$ per bp/gen. For computational efficiency, all parameters were re-scaled 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 ($N_{cur}$), 100 generations ago. Priors for both $N_{anc}$ and $N_{cur}$ were sampled from a loguniform distribution between 10 and 50,000, while priors for $\tau$ were sampled from a loguniform distribution between 10 and $N_{cur}$. One hundred replicates were simulated for each parameter combination. Simulations required for ABC were performed in *msprime* v. 0.7.3 (Kelleher *et al*. 2016) 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 mutations were assumed to be present, with simulations performed using SLiM v. 3.1 (Haller & Messer 2019). Introns and intergenic regions were assumed to be neutral, while exons experienced beneficial mutations with fitness effects sampled from an exponential distribution with mean $s$. The two 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 $\gamma \in [0.1, 10000]$ 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 ($\lambda$, which is related to the $\alpha$ parameter of Eyre-Walker & Keightley) was calculated using the total number of fixations (as provided by SLiM), which was found to range from 0-0.85 depending on the underlying parameters. Parameter inference was performed for $\gamma$ and $d_a$ and the corresponding $\lambda$ was inferred using $\lambda = \frac{d_a}{d_a + ((1-f_{pos}) \times \mu \times num\ of\ generations)}$, where it was assumed that $1 - f_{pos} \sim 1$. Populations were assumed to have a constant size and comprised of 5000 diploid individuals with constant mutation and recombination rates as specified above. Selection coefficients were re-scaled by the factor 200 and 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 both 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 ($\pi$), Watterson's theta ($\theta_W$), $\theta_H$, $H'$, Tajima's *D*, number of singletons, haplotype diversity, and statistics summarizing LD ($r^2, D, D'$). All statistics were calculated in non-overlapping sliding windows of 2 kb using *pylibseq* v. 0.2.3 (Thornton 2003). ABC was performed using the R package "abc" v. 2.1 (Csilléry *et al*. 2012) using all summary statistics, with "neural net" to account for non-linearity between statistics and

parameters. A 100-fold cross-validation was used to identify the optimum tolerance level, which was found to be 0.05.

**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 four non-overlapping 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 four 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 $f_{pos} = 2.2 \times 10^{-3}$ (modified from Campos & Charlesworth 2019), resulting in $\lambda$~0.35.

c) Population size change: A population decline was simulated such that the population declined from 5000 to 100 individuals instantaneously 100 generations ago. A population expansion was similarly simulated with parameters $N_{anc} = 5000$ and $N_{cur} = 10000$. A population bottleneck model was also simulated such that $N_{anc} = N_{cur} = 5000$ and a bottleneck occurred 2000 generations ago with a severity of 1% and a duration of 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 (Han *et*

559 *al*. 2014). To account for this scenario, a random set of singletons, representing a third of
560 all singletons present in the sample, were removed.

561 e) Progeny skew: A skew in the offspring distribution was modeled as a $\psi$-coalescent
562 (Eldon & Wakeley 2006; and see Matuszewski *et al.* 2018; Sackman *et al.* 2019), such
563 that 5% and 10% of the population was replaced by the offspring of a single individual
564 each generation.

565 f) Variation in mutation and recombination rates across the genome (*e.g.*, McVean *et al.*
566 2004; Chan *et al.* 2012; Penalba & Wolf 2020): Every 10 kb of the ~100 kb genomic
567 region considered was assumed to have a different mutation and recombination rate. For
568 every simulated replicate, these rates were sampled from a Gaussian distribution with the
569 same mean as above, and a standard deviation of 0.5×mean value. Negative values were
570 truncated to 0.

571

572 **Posterior checks**

573 For the purposes of illustration, an example of posterior checks are provided in Figure 1 (*i.e.,*
574 showing a simple evaluation of the fit of the inferred posteriors under the incorrect models to
575 the true scenarios under consideration). Specifically, the mean estimates of the inferred
576 parameters were used to simulate the "best-fitting model" in SLiM v. 3.1 (Haller & Messer
577 2019). Exons were masked and summary statistics were calculated as above in windows of 2
578 kb using *pylibseq* v.0.2.3 (Thornton 2003). In order to simulate the inferred models of
579 positive selection, $f_{pos}$ was calculated from $\lambda$ assuming a Wright-Fisher diploid population
580 of size $N$ and a total mutation rate of $\mu_{tot}$ (which for our purpose is the same as $\mu$). Thus,
581 $\mu_b = f_{pos} \times \mu_{tot}$ and $\mu_{neu} = (1 - f_{pos}) \times \mu_{tot}$ where $\mu_b$ $and$ $\mu_{neu}$ are the beneficial and

582 neutral mutation rates, respectively. Given a value of $\lambda$, and assuming that the distribution of
583 fitness effects of beneficial mutations is exponential (with mean $\bar{s}$), we calculate $f_{pos}$ as
584 follows:
585 given that

586 $$\lambda = \frac{\#of\ beneficial\ subs}{\#of\ beneficial\ subs + \#of\ neutral\ subs} \qquad (1)$$

587

588 where,

589 $$\#of\ beneficial\ subs = P_{fix} \times L \times 2N\mu_b \qquad (2)$$

590 and,

591 $$\#of\ neutral\ subs = \mu_{neu} \times L \qquad (3)$$

592 where $L$ is the length of the region being considered and $P_{fix}$ is the probability of fixation of
593 beneficial mutations, such that

594 $$P_{fix} = \int_0^{\infty} \frac{(1 - e^{-x})}{(1 - e^{-2Nx})} \left(\frac{e^{-x/\bar{s}}}{\bar{s}}\right) dx \qquad (4)$$

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

596 $$f = \frac{\lambda}{(1 - \lambda)P_{fix}2N + \lambda} \qquad (5)$$

597

598 Integrating (4) in $R$ and substituting it in (5) gives us values of $f$.

599

600 Statistics were calculated in non-overlapping windows of 2 kb and confidence intervals (CIs)
601 were calculated as the 0.025 and 0.975 quantiles of the distribution of the statistics.

602

603

604
605
606 **ACKNOWLEDGEMENTS**

607 We would like to thank Mark Kirkpatrick and Kevin Thornton for helpful comments and critical

608 feedback. This work was funded by National Institutes of Health grants R01GM135899 and

609 R35GM139383 to JDJ.

610
611
612

## CITATIONS

Auton, A., A. Fledel-Alon, S.P. Pfeifer, O. Venn, L. Ségurel, T. Street, E.M. Leffler, R. Bowden, I. Aneas, J. Broxholme, P. Humburg, Z. Iqbal, G. Lunter, J. Maller, R.D. Hernandez, C. Melton, A. Venkat, M.A. Nobrega, R. Bontrop, S. Myers, P. Donnelly, M. Przeworski, and G. McVean, 2012. A fine-scale chimpanzee genetic map from population sequencing. Science 336(6078): 193-198.

Auton, A., and G. McVean, 2012. Estimating recombination rates from genetic variation in humans. Methods Mol. Biol. 856: 217-237.

Bank, C., M. Foll, A. Ferrer-Admetlla, G. Ewing, and J.D. Jensen, 2014a. Thinking too positive? Revisiting current methods in population genetic selection inference. Trends Genet. 30(12): 540-546.

Bank, C., R.T. Hietpas, A. Wong, D.N.A. Bolon, and J.D. Jensen, 2014b. 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 196(3): 841-852.

Barton, N.H., 2000. Genetic hitchhiking. Phil. Trans. R. Soc. B 355(1403): 1553-1562.

Beaumont, M.A., and B. Rannala, 2004. The Bayesian revolution in genetics. Nat. Rev. Genet. 5(4): 251-261.

Beaumont, M.A., W. Zhang, and D.J. Balding, 2002. Approximate Bayesian computation in population genetics. Genetics 162(4): 2025-2035.

Beichman, A.C., E. Huerta-Sanchez, and K.E. Lohmueller, 2018. Using genomic data to infer historic population dynamics of non-model organisms. Annu. Rev. Ecol. Evol. Syst. 49: 433-456.

Böndel, K.B., S.A. Kraemer, T.S. Samuels, D. McClean, J. Lachapelle, R.W. Ness, N. Colegrave, and P.D. Keightley, 2019. Inferring the distribution of fitness effects of spontaneous mutations in *Chlamydomonas reinhardtii*. PLoS Biol. 17(6): e3000192.

Booker, T.R., B.C. Jackson, and P.D. Keightley, 2017. Detecting positive selection in the genome. BMC Biol. 15(1):98.

Campos, J.L., and B. Charlesworth, 2019. The effects on neutral variability of recurrent selective sweeps and background selection. Genetics 212(1): 287-303.

Chan, A.H., P. Jenkins, and Y.S. Song, 2012. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. PLoS Genet. 8(12): e1003090.

Charlesworth, B., 2013. Background selection 20 years on. The Wilhelmine E. Key 2012 invitational lecture. J. Hered. 104(2): 161-171.

Charlesworth, D., B. Charlesworth, and M.T. Morgan, 1995. The pattern of neutral molecular variation under the background selection model. Genetics 141(4): 1619-1632.

Charlesworth, B., and J.D. Jensen, 2021. The effects of selection at linked sites on patterns of genetic variability. Annu. Rev. Ecol. Evol. Syst., in press.

Charlesworth, B., M.T. Morgan, and D. Charlesworth, 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134(4): 1289-1303.

Chikhi, L., V.C. Sousa, P. Luisi, B. Goossens, and M.A. Beaumont, 2010. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size change. Genetics 186(3): 983-995.

Choi, J.Y., and C.F. Aquadro, 2016. Recent and long term selection across synonymous sites in *Drosophila ananassae*. J. Mol. Evol. 83(1-2): 50-60.

Comeron, J.M., 2014. Background selection as baseline for nucleotide variation across the Drosophila genome. PLoS Genet. 10(6): e1004434.

Comeron, J.M., 2017. Background selection as a null hypothesis in population genomics: insights and challenges from Drosophila studies. Phil. Trans. R. Soc. B 372(1736): 20160471.

Comeron, J.M., R. Ratnappan, and S. Bailin, 2012. The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. 8(10): e1002905.

Cox, A., C. Ackert-Bicknell, B. Dumont, Y. Ding, J. Tzenova Bell, G. Brockmann, J. Wergedal, C. Bult, B. Paigen, J. Flint, S-W. Tsaih, G. Churchill, and K. Broman, 2009. A new standard genetic map for the laboratory mouse. Genetics 182(4): 1335-1344.

Cranmer, K., J. Brehmer, and G. Louppe, 2020. The frontier of simulation-based inference. Proc. Natl. Acad. Sci. USA 117(48): 30055-30062.

Crisci, J., Y.-P. Poh, S. Mahajan, and J.D. Jensen, 2013. The impact of equilibrium assumptions on tests of selection. Front. Genet. 4: 235.

Csilléry K., O. François, and M.G.B. Blum, 2012. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3: 475-479.

Dapper, A.L., and B.A. Payseur, 2018. Effects of demographic history on the detection of recombination hotspots from linkage disequilibrium. Mol. Biol. Evol. 35(2): 335-353.

Darwin, C., 1859. 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.

Durrett, R. and J. Schweinsberg, 2004. Approximating selective sweeps. Theor. Popul. Biol. 66(2): 129-138.

Eldon, B., and J. Wakeley, 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172(4): 2621-2633.

Ewing, G., and J.D. Jensen, 2016. The consequences of not accounting for background selection in demographic inference. Mol. Ecol. 25(1): 135-141.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V.C. Sousa, and M. Foll, 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9(10): e1003905.

Eyre-Walker, A., and P.D. Keightley, 2007. The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8(8): 610-618.

Eyre-Walker, A., and P.D. Keightley, 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26(9): 2097-2108.

Fay, J., and C.-I Wu, 2000. Hitchhiking under positive Darwinian selection. Genetics 155(3): 1405-1413.

Ferrer-Admetlla, A., C. Leuenberger, J.D. Jensen, and D. Wegmann, 2016. An approximate Markov model for the Wright-Fisher diffusion and its application to time series data. Genetics 203(2): 831-846.

Fisher, R.A., 1930. The genetical theory of natural selection. Clarendon Press, Oxford, UK.

Foll, M., Y.-P. Poh, N. Renzette, A. Ferrer-Admetlla, H. Shim, A.-S. Malaspinas, G. Ewing, C. Bank, P. Liu, D. Wegmann, D.R. Caffrey, K.B. Zeldovich, D.N.A. Bolon, J. Wang, T.F. Kowalik, C.A. Schiffer, R.W. Finberg, and J.D. Jensen, 2014. Influenza virus drug resistance: a time-sampled population genetics perspective. PLoS Genet. 10(2): e1004185.

Foll, M., H. Shim, and J.D. Jensen, 2015. A Wright-Fisher ABC-based approach for inferring per-site effective population sizes and selection coefficients from time-sampled data. Mol. Ecol. Res. 15(1): 87-98.

Ford, E.B., 1975. Ecological genetics. Chapman and Hall, London, UK.

Garud, N., P. Messer, E. Buzbas, and D. Petrov, 2015. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLoS Genet. 11(2): e1005004.

Garud, N., P. Messer, and D. Petrov, 2021. Detection of hard and soft selective sweeps from *Drosophila melanogaster* population genomic data. PLoS Genet. 17(2): e1009373.

Gelman, A., and C.R. Shalizi, 2013. Philosophy and the practice of Bayesian statistics. Br J Math Stat Psychol. 66(1): 8-38.

Gutenkunst, R., R. Hernandez, S. Williamson, and C. Bustamante, 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP data. PLoS Genet. 5(10): e1000695.

Haldane, J.B.S., 1932. The causes of evolution. Longmans, London, UK.

Hallatschek, O., 2018. Selection-like biases emerge in population models with recurrent jackpot events. Genetics 210(3): 1053-1073.

Haller, B.C., and P.W. Messer, 2019. SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. Mol. Biol. Evol. 36(3): 632-637.

Han E., J.S. Sinsheimer, and J. Novembre, 2014. Characterizing bias in population genetic inferences from low-coverage sequencing data. Mol. Biol. Evol. 31(3): 723-735.

Harris, R.B., K. Irwin, M.R. Jones, S. Laurent, R.D.H. Barrett, M.W. Nachman, J.M. Good, C.R. Linnen, J.D. Jensen, and S.P. Pfeifer, 2020. The population genetics of crypsis in vertebrates: recent insights from mice, hares, and lizards. Heredity 124(1): 1-14.

Harris, R.B., A. Sackman, and J.D. Jensen, 2018. On the unfounded enthusiasm for soft selective sweeps II: examining recent evidence from humans, flies, and viruses. PLoS Genet. 14(12): e1007859.

Irwin, K.K., N. Renzette, T.F. Kowalik, and J.D. Jensen, 2016. Antiviral drug resistance as an adaptive process. Virus Evol. 2(1): vew014.

Jensen, J.D., Y. Kim, V.B. DuMont, C.F. Aquadro, and C.D. Bustamante, 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170(3): 1401-1410.

Jensen, J.D., B.A. Payseur, W. Stephan, C.F. Aquadro, M. Lynch, D. Charlesworth, and B. Charlesworth, 2019. The importance of the Neutral Theory in 1968 and 50 years on: a response to Kern & Hahn 2018. Evolution 73(1): 111-114.

Jensen, J.D., K.R. Thornton, C.D. Bustamante, and C.F. Aquadro, 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in non-equilibrium populations. Genetics 176(4): 2371-2379.

Johri, P., B. Charlesworth, and J.D. Jensen, 2020. Towards an evolutionarily appropriate null model: jointly inferring demography and purifying selection. Genetics 215(1): 173-192.

769 Johri, P., K. Riall, H. Becher, L. Excoffier, B. Charlesworth, and J.D. Jensen, 2021. The impact
770 of purifying and background selection on the inference of population history: problems and
771 prospects. Mol. Biol. Evol. 38(7): 2986-3003.

772

773 Keightley, P.D., 2012. Rates and fitness consequences of new mutations in humans. Genetics
774 190(2): 295-304.

775 Keightley, P.D., and A. Eyre-Walker, 2007. Joint inference of the distribution of fitness effects
776 of deleterious mutations and population demography based on nucleotide polymorphism
777 frequencies. Genetics 177(4): 2251-2261.

778 Keightley, P.D., and D.L. Halligan, 2009. Analysis and implications of mutational variation.
779 Genetica 136(2): 359-369.
780
781 Kelleher J., A.M. Etheridge, and G. McVean, 2016. Efficient coalescent simulation and
782 genealogical analysis for large sample sizes. PLoS Comput. Biol. 12(5): e1004842.
783
784 Kelleher, J., K. Thornton, J. Ashander, and P. Ralph, 2018. Efficient pedigree recording for fast
785 population genetics simulation. PLoS Comput. Biol. 14(11): e1006581.

786 Kelleher, J., Y. Wong, A.W. Wohns, C. Fadil, P.K. Albers, and G. McVean, 2019. Inferring
787 whole-genome histories in large population datasets. Nat. Genet. 51(9): 1330-1338.

788 Kern, A.D., and M.W. Hahn, 2018. The neutral theory in light of natural selection. Mol. Biol.
789 Evol. 35(6): 1366-1371.
790
791 Kim, Y., and R. Nielsen, 2004. Linkage disequilibrium as a signature of selective sweeps.
792 Genetics 167(3): 1513-1524.

793 Kim, Y., and W. Stephan, 2002. Detecting a local signature of genetic hitchhiking along a
794 recombining chromosome. Genetics 160(2): 765-777.

795 Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217(5129): 624-626.

796 Kimura, M., 1983. The neutral theory of molecular evolution. Cambridge Univ. Press,
797 Cambridge.

798 King, J.L., and T.H. Jukes, 1969. Non-Darwinian evolution. Science 164(3881): 788-798.

799 Kong, A., D.F. Gudbjartsson, J. Sainz, G. Jonsdottir, S. Gudjonsson, B. Richardsson, S.
800 Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. Palsson, M. Frigge, T.
801 Thorgeirsson, J. Gulcher, and K. Stefansson, 2002. A high-resolution recombination map of the
802 human genome. Nat. Genet. 31(3): 241-247.

803 Lapierre, M., A. Lambert, and G. Achaz, 2017. Accuracy of demographic inference from the site
804 frequency spectrum: the case of the Yoruba population. Genetics 206(1): 439-449.

805 Lewontin, R.C., 1974. The genetic basis of evolutionary change. Columbia Univ. Press, New
806 York.

807 Lewontin, R.C., 1991. Twenty-five years ago in Genetics: electrophoresis in the development of
808 evolutionary genetics: milestone or millstone? Genetics 128(4): 657-662.

809 Long, H., W. Sung, S. Kucukyildirim, E. Williams, S. Miller, W. Guo, C. Patterson, C. Gregory,
810 C. Strauss, C. Stone, C. Berne, D. Kysela, W. Shoemaker, M. Muscarella, H. Luo, J. Lennon, Y.
811 Brun, and M. Lynch, 2018. Evolutionary determinants of genome-wide nucleotide composition.
812 Nat. Ecol. Evol. 2(2): 237-240.

813 Louca, S., and M.W. Pennell, 2020. Extant timetrees are consistent with a myriad of
814 diversification histories. Nature 580(7804): 502-505.

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

816 Lynch, M., M.S. Ackerman, J.F. Gout, H. Long, W. Sung, W.K. Thomas, and P.L. Foster, 2016.
817 Genetic drift, selection and the evolution of the mutation rate. Nat. Rev. Genet. 17(11): 704-714.
818
819 Lynch, M., and W.-C. Ho, 2020. The limits to estimating population-genetic parameters with
820 temporal data. Gen. Biol. Evol. 12(4): 443-455.
821
822 Lynch, M., W. Sung, K. Morris, N. Coffey, C.R. Landry, E.B. Dopman, W.J. Dickinson, K.
823 Okamoto, S. Kulkarni, D.L. Hartl, and W.K. Thomas, 2008. A genome-wide view of the
824 spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. Sci. USA 105(27): 9272-9277.
825
826 Maddamsetti, R., and N.A. Grant, 2020. Divergent evolution of mutation rates and biases in the
827 long-term evolution experiment with *Escherichia coli*. Gen. Biol. Evol. 12(9): 1591-1603.
828
829 Malaspinas, A-S., O. Malaspinas, S.N. Evans, and M. Slatkin, 2012. Estimating allele age and
830 selection coefficient from time-serial data. Genetics 192(2): 599-607.
831
832 Matuszewski, M., M.E. Hildebrandt, G. Achaz, and J.D. Jensen, 2018. Coalescent processes with
833 skewed offspring distributions and non-equilibrium demography. Genetics 208(1): 323-338.
834
835 Maynard Smith, J. and J. Haigh, 1974. The hitch-hiking effect of a favourable gene. Genet. Res.
836 23(1): 23-25.
837
838 McVean, G., 2007. The structure of linkage disequilibrium around a selective sweep. Genetics
839 175(3): 1395-1406.
840
841 McVean, G., S. Myers, S. Hunt, P. Deloukas, D. Bentley, and P. Donnelly, 2004. The fine-scale
842 structure of recombination rate variation in the human genome. Science 304(5670): 581-584.

Mendel, G., 1866. Versuche über Pflanzenhybriden. Ver. Naturforsch. Ver. Brünn 4: 3-47.

Myers, S., C. Fefferman, and N. Patterson, 2008. Can one learn history form the allelic spectrum? Theor. Popul. Biol. 73(3): 342-348.

Ness, R.W., A.D. Morgan, V. Radhakrishnan, N. Colegrave, and P.D. Keightley, 2015. Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. Genome Res. 25(11): 1739-1749.

Nielsen, R., 2004. Population genetic analysis of ascertained SNP data. Hum. Genomics 1(3): 218-224.

Nielsen, R., J.M. Akey, M. Jakobsson, J.K. Pritchard, S. Tishkoff, and E. Willerslev, 2017. Tracing the peopling of the world through genomics. Nature 541(7637): 302-310.

Nielsen, R., S. Williamson, Y. Kim, M.J. Hubisz, A.G. Clark, and C.D. Bustamante, 2005. Genomic scans for selective sweeps using SNP data. Genome Res. 15(11): 1566-1575.

Ohta, T., 1973. Slightly deleterious mutant substitutions in evolution. Nature 246(5428): 96-98.

Peischl, S., I. Dupanloup, M. Kirkpatrick, and L. Excoffier, 2013. On the accumulation of deleterious mutations during range expansions. Mol. Ecol. 22(24): 5972-5982.

Peischl, S., M. Kirkpatrick, and L. Excoffier, 2015. Expansion load and the evolutionary dynamics of a species range. Am. Nat. 185(4): E81-E93.

Penalba, J.V., and J.B. Wolf, 2020. From molecules to populations: appreciating and estimating recombination rate variation. Nat. Rev. Genet. 21(8): 476-492.

Pfeifer, S.P., 2017. From next-generation resequencing reads to a high quality variant data set. Heredity 118(2): 111-124.

Pfeifer, S.P., 2020a. Spontaneous mutation rates. In *The Molecular Evolutionary Clock. Theory and Practice.* Springer Nature.

Pfeifer, S.P., 2020b. A fine-scale genetic map for vervet monkeys. Mol. Biol. Evol. 37(7): 1855-1865.

Pfeifer, S.P., 2021. Studying mutation rate evolution in primates – the impacts of computational pipeline and parameter choices. GigaScience, in press.

Poh, Y.-P., V. Domingues, H.E. Hoekstra, and J.D. Jensen, 2014. On the prospect of identifying adaptive loci in recently bottlenecked populations. PLoS One 9(11): e110579.

885 Pouyet, F., S. Aeschbacher, A. Thiery, and L. Excoffier, 2018. Background selection and biased
886 gene conversion affect more than 95% of the human genome and bias demographic inferences.
887 eLife 7: e36317.
888
889 Provine, W.B., 1971. The origins of theoretical population genetics. University of Chicago Press.

890 Przeworski, M., 2002. The signature of positive selection at randomly chosen loci. Genetics
891 160(3): 1179-1189.

892 Ragsdale, A., C. Moreau, and S. Gravel, 2018. Genomic inference using diffusion models and
893 the allele frequency spectrum. Curr. Opin. Gen. Deve. 53: 140-147.

894 Ray, N., and L. Excoffier, 2009. Inferring past demography using spatially explicit population
895 genetic models. Hum. Biol. 81(2-3): 141-157.

896 Renzette, N., L. Gibson, J.D. Jensen, and T.F. Kowalik, 2014. Human cytomegalovirus intrahost
897 evolution - a new avenue for understanding and controlling herpesvirus infections. Curr. Opin.
898 Virol. 8: 109-115.
899
900 Rousselle, M., M. Maeva, B. Nabholz, T. Bataillon, and N. Galtier, 2018. Overestimation of the
901 adaptive substitution rate in fluctuating populations. Biol. Lett. 14(5): 20180055.
902
903 Sackman, A., R.B. Harris, and J.D. Jensen, 2019. Inferring demography and selection in
904 organisms characterized by skewed offspring distributions. Genetics 211(3): 1019-1028.
905
906 Schneider, A., B. Charlesworth, A. Eyre-Walker, and P.D. Keightley, 2011. A method for
907 inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189(4):
908 1427-1437.
909
910 Schraiber, J.G., and J.M. Akey, 2015. Methods and models for unravelling human evolutionary
911 history. Nat. Rev. Genet. 16(12): 727-740.

912 Schrider, D.R., and A.D. Kern, 2017. Soft sweeps are the dominant mode of adaptation in the
913 human genome. Mol. Biol. Evol. 34(8): 1863-1877.

914 Singh, N.D., V.L. Bauer DuMont, M.J. Hubisz, R. Nielsen, and C.F. Aquadro, 2007. Patterns of
915 mutation and selection at synonymous sites in Drosophila. Mol. Biol. Evol. 24(12): 2687-2697.
916
917 Smith, T.C.A., P.F. Arndt, and A. Eyre-Walker, 2018. Large scale variation in the rate of germ-
918 line *de novo* mutations, base composition, divergence and diversity in humans. PLoS Genet.
919 14(3): e1007254.
920
921 Spence, J.P. and Y.S. Song, 2019. Inference and analysis of population-specific fine-scale
922 recombination maps across 26 diverse human populations. Sci. Adv. 5(10): eaaw9206.
923

924 Steinrücken, M., J. Kamm, J. Spence, and Y. Song, 2019. Inference of complex population
925 histories using whole-genome sequences from multiple populations. Proc. Natl. Acad. Sci. USA
926 116(34): 17115-17120.

927 Stephan, W., 2019. Selective sweeps. Genetics 211(1): 5-13.

928 Stephan, W., Y.S. Song and C.H. Langley, 2006. Hitchhiking effect on linkage disequilibrium
929 between linked neutral loci. Genetics 172(4): 2647-2663.

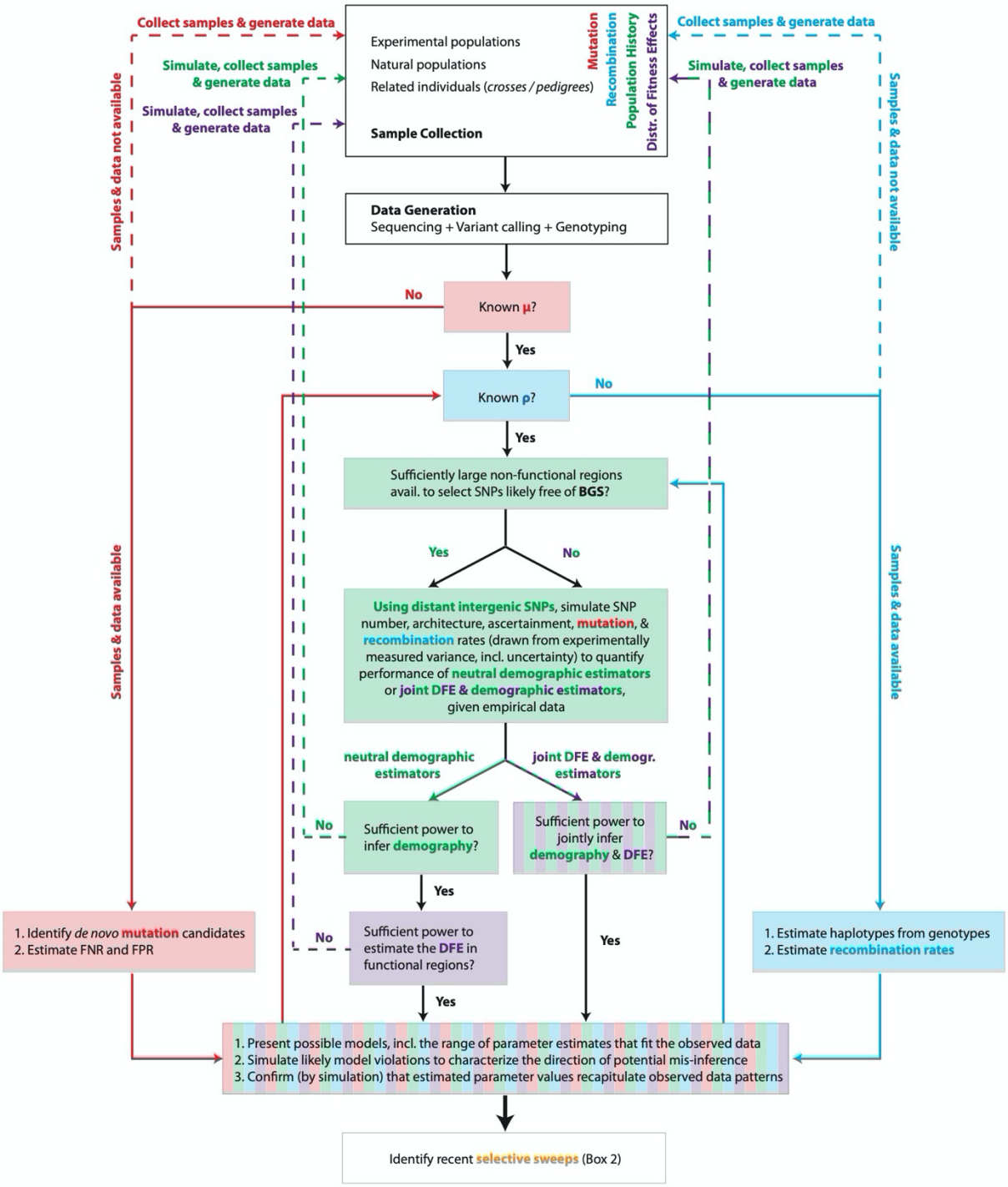
930 Stumpf, M.P., and G.A. McVean, 2003. Estimating recombination rates from population-genetic
931 data. Nat. Rev. Genet. 4(12): 959-968.
932
933 Teshima, K., G. Coop, and M. Przeworski, 2006. How reliable are empirical genome scans for
934 selective sweeps? Genome Res. 16(6): 702-712.
935
936 Thornton K., 2003. Libsequence: a C++ class library for evolutionary genetic analysis.
937 Bioinformatics 19(17): 2325–2327.
938
939 Thornton, K.R., 2014. A C++ template library for efficient forward-time population genetic
940 simulation of large populations. Genetics 198(1): 157-166.
941
942 Thornton, K.R., and J.D. Jensen, 2007. Controlling the false positive rate in multi-locus genome
943 scans for selection. Genetics 175(2): 737-750.
944
945 Torres, R., Z. Szpiech, and R.D. Hernandez, 2018. Human demographic history has amplified
946 the effects of background selection across the genome. PLoS Genet. 14(6): e1007387.

947 Walsh, B., and M. Lynch, 2018. Evolution and selection of quantitative traits. Oxford University
948 Press, Oxford.

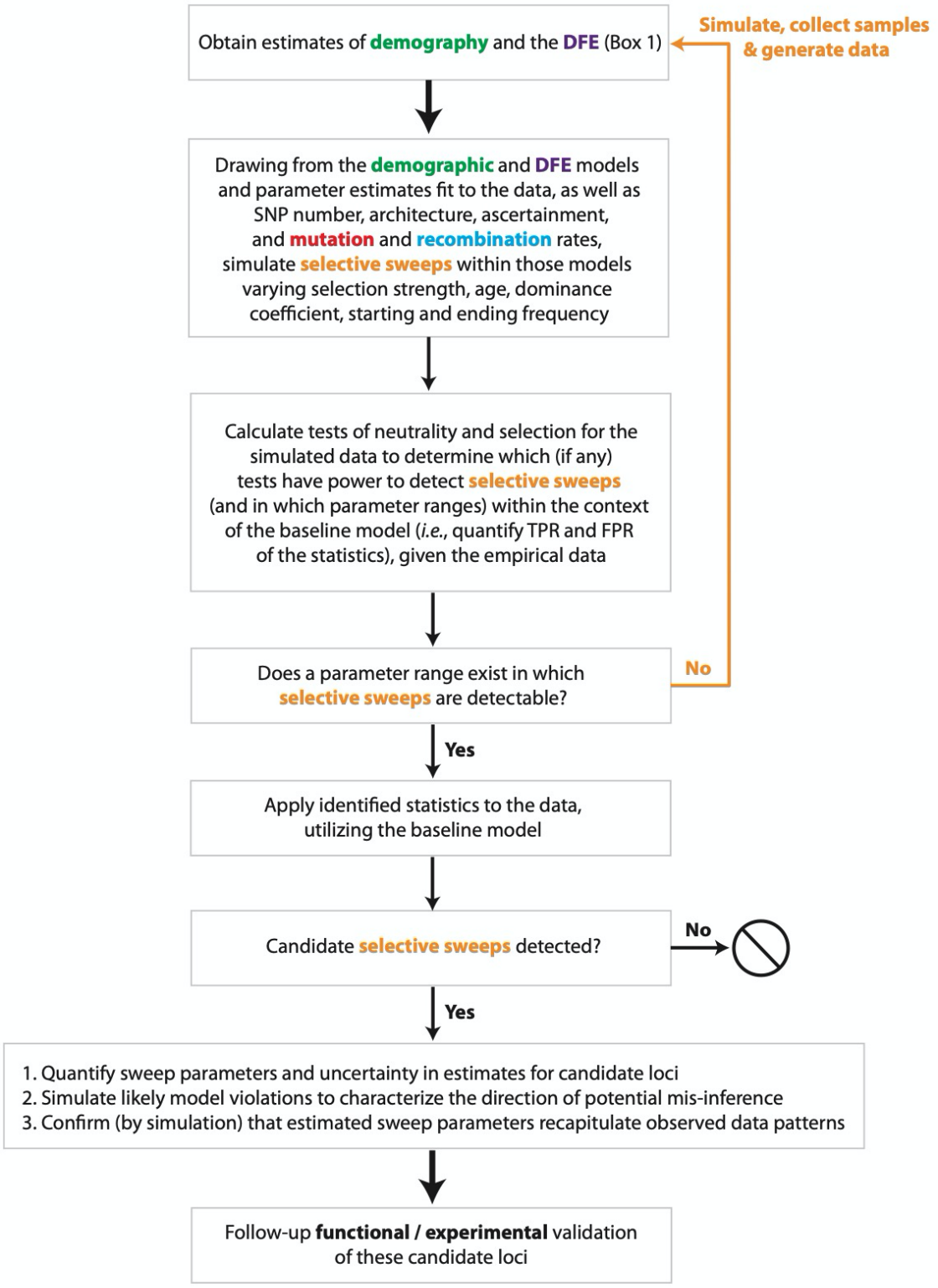
949 Williamson, S.H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen, and C.D. Bustamante,
950 2005. Simultaneous inference of selection and population growth from patterns of variation in
951 the human genome. Proc. Natl. Acad. Sci. USA 102(22): 7882-7887.

952 Wright, S., 1931. Evolution in Mendelian populations. Genetics 16(2): 97-159.

953 Zeng, K., and B. Charlesworth, 2010. Studying patterns of recent evolution at synonymous sites
954 and intronic sites in *Drosophila melanogaster*. J. Mol. Evol. 70(1): 116-128.




**Box 1**: 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 distribution of fitness effects 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.




**Box 2**: Diagram of important considerations in detecting selective sweeps. The color scheme matches that in Box 1, with 'selective sweeps' coded in orange.

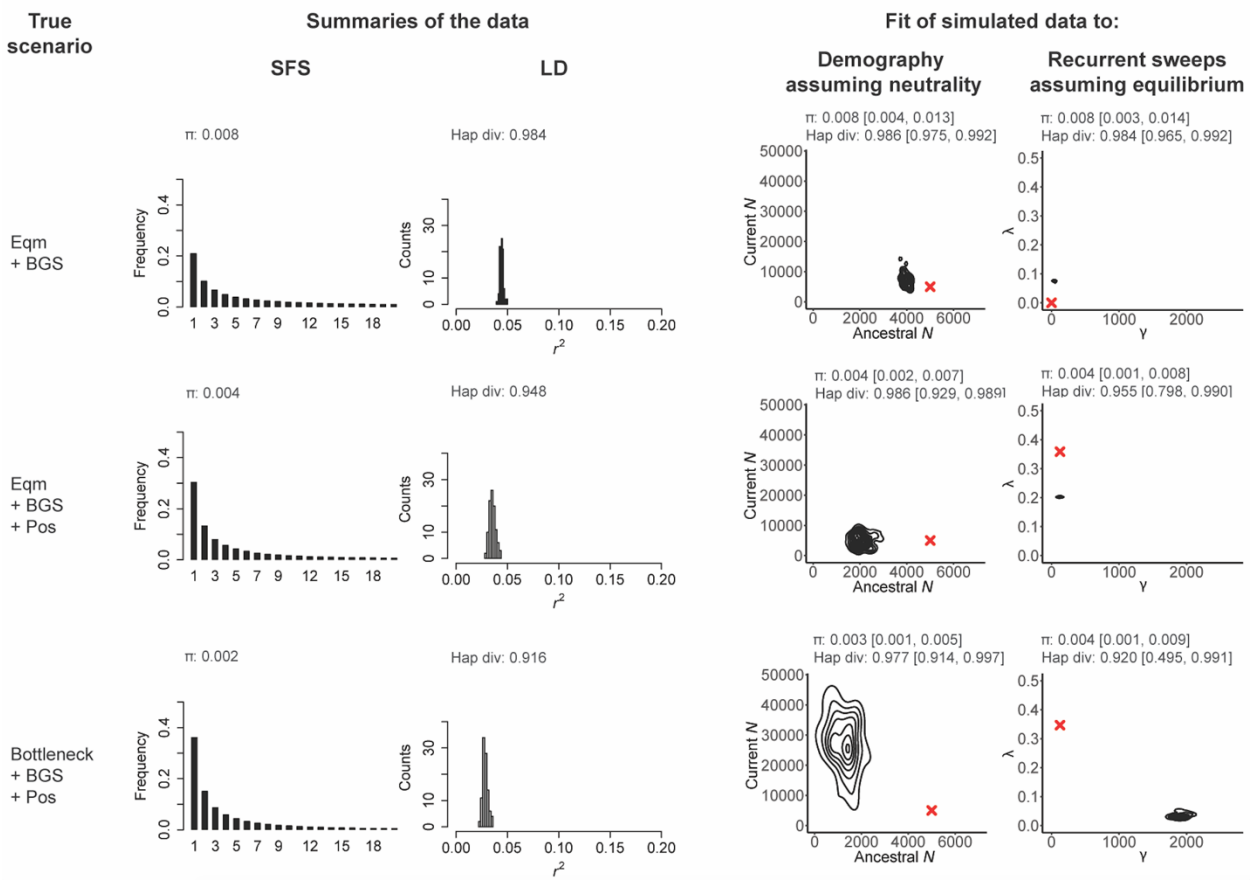



**Figure 1. Incorrect models may often readily be fit to a given dataset**
Here we present three scenarios varying from simple to more complex: the first row presents a constant-sized population experiencing background selection (denoted by 'Eqm + BGS'), the second row is the same scenario with the addition 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 site frequency spectra (SFS, truncated to $n = 20$) and linkage disequilibrium ($r^2$) distributions are given, together with mean pairwise ($\pi$) and haplotype diversity. To these simulated data we fit two incorrect models; one assuming all sites are neutral but including a change in population size (with the current size, ancestral size, and time of change being estimated from the data), 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 ($\gamma$) and the fraction of beneficial substitutions ($\lambda$) 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 Methods). 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.

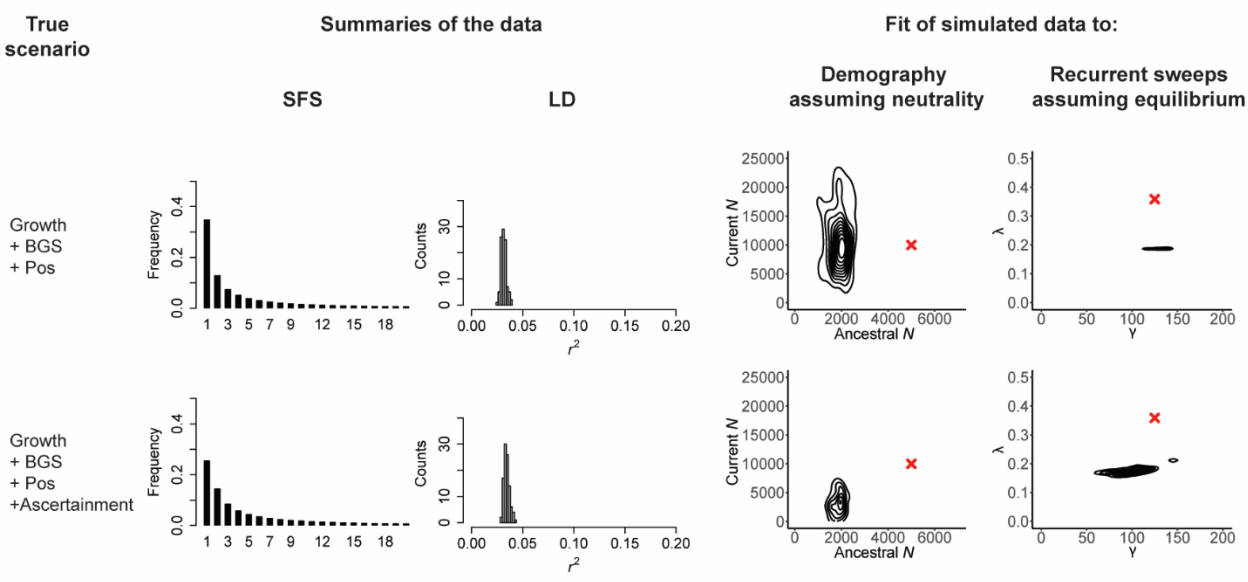



**Figure 2. Ascertainment errors may amplify mis-inference, if not corrected**
As in Figure 1, the scenarios are given in the first column, here population growth with background selection and recurrent selective sweeps ('Growth + BGS + Pos'), as well as the same scenario in which the imperfections of 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 panels provide the joint posterior distributions when the data are fit to two 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 may contribute to mis-inference.

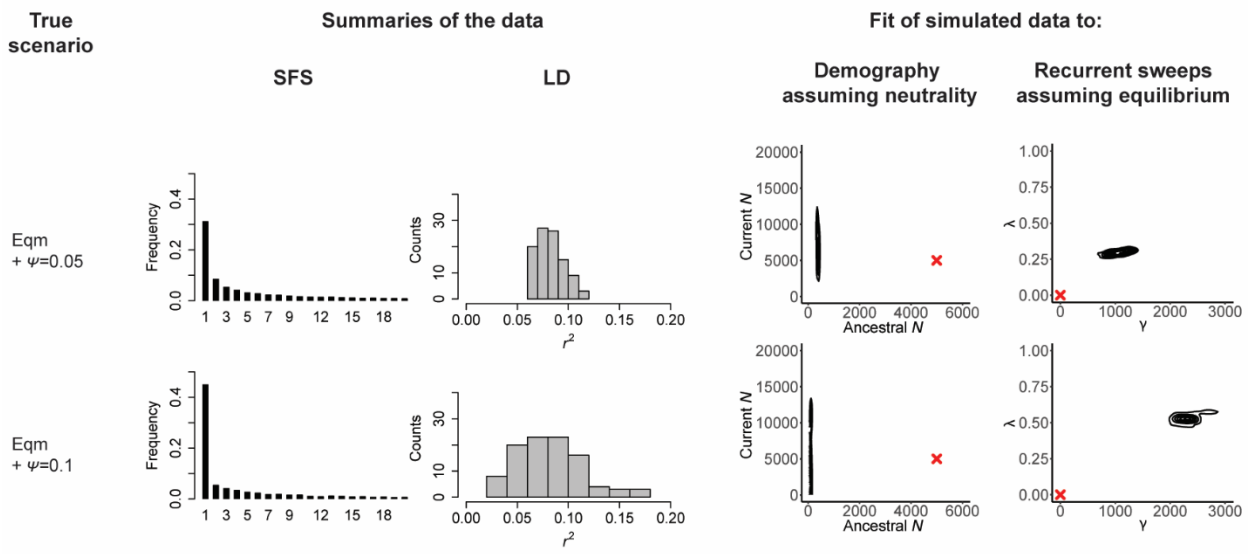



**Figure 3. The impact of potential model violations can be quantified**

As in Figures 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 Methods); the middle columns present the resulting SFS and LD distributions, and the final panels provide the joint posterior distributions when the data are fit to two 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.

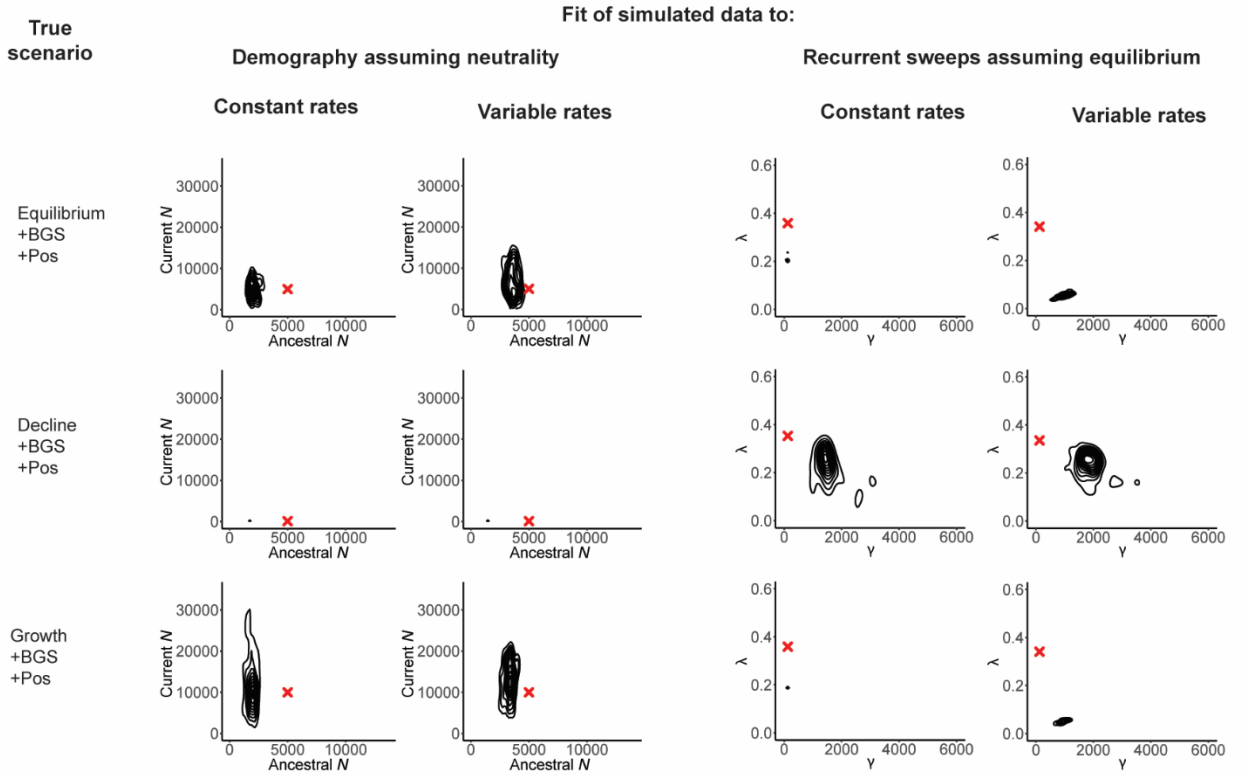



**Figure 4. The effects of not correcting for mutation and recombination rate heterogeneity** Three scenarios are here considered, equilibrium population size with background selection and recurrent selective sweeps ('Eqm +BGS + Pos'), declining population size together with background selection and 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 two 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 Methods). 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 regards to selection models.