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], 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] Institute of Ecology and Evolution, School of Biological Sciences, University of Edinburgh, Edinburgh, UK
[8] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK
[9] Laboratory of Genetics, University of Wisconsin-Madison, Madison, US
[10] Natural History Museum, Berlin, DE

[*] corresponding author: jeffrey.d.jensen@asu.edu

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

This paper is dedicated to the memories of Richard Lewontin (1929-2021) and Bill Hill (1940-2021).

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 out-paced 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 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 post-glacial 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

101 been largely validated [18-19], and have provided a means to interpret observed molecular
102 variation and divergence within the context of constantly occurring evolutionary processes
103 including mutation, genetic drift, and purifying selection. While ascribing an important role for
104 positive selection at the level of phenotypic evolution (consistent with Darwin's initial notions),
105 the Neutral Theory hypothesizes that at the genetic level beneficial mutations are rare compared
106 to the much larger input of neutral, nearly neutral, and deleterious mutations that are constantly
107 raining down on the genomes of all species. Accordingly, episodes of positive selection per
108 nucleotide are rare compared to genetic drift and purifying selection. However, the significant
109 effects on evolution at linked sites caused by fitness-altering mutations have been described in
110 detail in the decades since Kimura's initial formulation of the Neutral Theory [20-22].
111 With this framework and the availability of datasets to which it could be applied,
112 statistical approaches for analyzing molecular data began to proliferate, frequently employing
113 some form of neutral expectation as a null model. A wide range of rather sophisticated statistical
114 machinery is now available for reconstructing histories of population size change, population
115 subdivision and migration (*e.g.,* [23-24]), for identifying beneficial mutations based on patterns
116 associated with selective sweeps (*e.g.,* [25-26]), for quantifying the distribution of fitness effects
117 (DFE) of newly arising mutations (*e.g.,* [27-28]), as well as for estimating rates of mutation (*e.g.,*
118 [29-31]) and recombination (*e.g.,* [32-34]). These approaches operate in a variety of statistical
119 frameworks (see [35-37]), and utilize various aspects of the data – including the frequencies of
120 variants in a sample (the site frequency spectrum, SFS), associations between variants (linkage
121 disequilibrium, LD), and/or levels and patterns of between-species divergence at contrasted site
122 classes (*e.g.,* synonymous versus non-synonymous sites).
123

124 *Challenges of model choice and parameter fitting*

125 The growing variety of statistical approaches and associated software implementations presents a
126 dizzying array of choices for any given analysis; although many approaches share the same aims,
127 there also exist important differences. For example, some approaches require a relatively high
128 level of coding ability to implement while others may be applied in easy-to-use software
129 packages; while some are well-tested and justified by population-genetic theory, others are not.
130 Moreover, even the process of translating raw sequencing data into the allele calls and genotypes
131 used as input for these approaches is accompanied by uncertainty that depends on sequencing
132 quality and coverage, availability of a reference genome, and choice of variant calling and
133 filtering strategies [38-39]. Adding to this complexity, it has become increasingly clear that
134 demographic estimation may be highly biased when selection and recombination-associated
135 biased gene conversion are neglected [40-41], whereas estimates of selection intensity and
136 recombination rate may be highly biased when neglecting demographic effects [42-45]. This
137 creates a circular problem when commencing any new analysis: one needs information about the
138 demographic history to estimate parameters of recombination and selection, while at the same
139 time one needs information about recombination and selection to estimate the demographic
140 history. An additional challenge, and a frustration for many, is that there is no single 'best
141 approach'; the correct analysis tools to use, and indeed which questions can be answered at all,
142 depend entirely on the details of the organism under study [46]. Specifically, biological
143 parameters that vary among species – including evolutionary parameters (*e.g.*, effective
144 population size (N_e), mutation rates, recombination rates, and population structure and history),
145 genome structure (*e.g.*, the distribution of functional sites along the genome), and life history

146 traits (*e.g.*, mating system) – must all be considered in order to define addressable hypotheses
147 and optimal approaches.

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

159 Examples abound of enthusiastic promotion of a single preferred model, only to be
160 tempered by subsequent demonstrations of the fit of alternative and often simpler / more
161 biologically realistic models. For example, the view that segregating alleles may be commonly
162 maintained by balancing selection [49] was tempered by the realization that genetic drift is often
163 a sufficient explanation [14], and the view that genome-wide selective sweeps on standing
164 variation are pervasive [50-51] was tempered by the realization that neutral population histories
165 can result in similar patterns [47,52]. While one may readily find such examples of using
166 episodic or hypothesized processes to fit large-scale data patterns by neglecting to define
167 expectations arising from common and certain-to-be-occurring processes, determining which
168 models to evaluate, and how to interpret the fit of a model and its alternatives, are challenges for

169 all researchers. To better illustrate this point, Figure 1 presents three scenarios (constant
170 population size with background selection, constant population size with background selection
171 and selective sweeps, and a population bottleneck with background selection and selective
172 sweeps), and provides the fit of each of those scenarios to two incorrect models (population size
173 change assuming strict neutrality, and recurrent selective sweeps assuming constant population
174 size). As is shown in the figure, each scenario can be well fit by both incorrect models, with
175 selective sweeps and population bottlenecks generally being confounded, as well as background
176 selection and population growth, as has been described several times before (*e.g.*, [40, 53-55]).
177
178
179 **Methodology**
180 In order to provide a series of examples to accompany key points - as with the above Figure 1 -
181 both forward-in-time simulations and coalescent simulations were performed for (1) the
182 inference of demographic history assuming complete neutrality, (2) the inference of positive
183 selection assuming constant population size, and (3) obtaining test datasets representing different
184 evolutionary scenarios. While any statistical framework involving model / parameter exploration
185 and comparison may be consistent with our recommendations, we here utilize approximate
186 Bayesian computation (ABC) for our examples, as it is a particularly useful framework for
187 quantifying uncertainty and for exploring complex models.
188 In all simulations, a chromosomal segment of 99,012 bp was simulated with an intron-
189 exon-intergenic structure resembling the *D. melanogaster* genome. Each gene comprised five
190 exons (of 300 bp each) and four introns (of 100 bp each) separated by intergenic regions of
191 length 1,068 bp. Such a construct resulted in a total of 33 genes across the simulated segment.

192 Population parameters were chosen to resemble those from *D. melanogaster* populations
193 following Campos *et al.* [56], assuming an effective population size (N_e) of 10^6 individuals with
194 a mean mutation rate (μ) of 4.5×10^{-9} per base pair /generation and a mean recombination rate
195 (r) of 1×10^{-8} per base pair /generation. For computational efficiency, all parameters were
196 rescaled by a factor of 200.
197
198 *Modeling and inference of demographic history*
199 A simple demographic history was modeled in which a single population undergoes an
200 instantaneous change from an ancestral size (N_{anc}) to a current size (N_{cur}), τ generations ago.
201 Priors for both N_{anc} and N_{cur} were sampled from a loguniform distribution between 10 and
202 50,000, while priors for the time of change (τ) were sampled from a loguniform distribution
203 between 10 and N_{cur}. One hundred replicates were simulated for each parameter combination.
204 Simulations required for ABC were performed in *msprime* v. 0.7.3 [57] assuming complete
205 neutrality. Mutation and recombination rates were assumed to be constant across the genome and
206 across replicates.
207
208 *Modeling and inference of positive selection*
209 A recurrent selective sweep scenario was modeled in which only neutral and beneficial
210 autosomal mutations were allowed, with simulations performed using *SLiM* v. 3.1 [58]. Introns
211 and intergenic regions were assumed to be neutral, while exons experienced beneficial mutations
212 with fitness effects sampled from an exponential distribution with mean s for homozygotes,
213 assuming semi-dominance. The two parameters varied were the mean population-scaled strength
214 of selection, $\gamma = 2N_{anc}s$, and the proportion of new beneficial mutations, f_{pos}. Priors for these

215 parameters were sampled from a loguniform distribution such that $\gamma \in [0.1, 10000]$ and $f_{pos} \in$

216 $[0.00001, 0.01]$. For all parameter combinations, the true rate of beneficial substitutions per site

217 (d_a) and the true fraction of substitutions due to beneficial mutations (λ, which is related to the α

218 parameter of Eyre-Walker & Keightley [59]) were calculated using the total number of fixations

219 (as provided by *SLiM*), with λ observed to range from 0-0.85 depending on the underlying

220 parameters. Parameter inference was performed for γ and d_a and the corresponding λ was

221 inferred using $\lambda = \frac{d_a}{d_a + ((1-f_{pos}) \times \mu \times num\ of\ generations)}$, where it was assumed that $1 - f_{pos} \sim 1$.

222 Populations had a constant size of 5000 diploid individuals with constant mutation and

223 recombination rates, as specified above. Simulations were run for 100,100 generations (*i.e.,* $20N_e$

224 + 100 generations).

225

226 *ABC*

227 The sample size was set to 100 haploid genomes (or 50 diploid individuals). Under the

228 demographic and selection models described above, all exonic regions were masked and the

229 mean and variance (across replicates) of the following summary statistics were calculated:

230 number of segregating sites, nucleotide site diversity (π), Watterson's theta (θ_W), θ_H, H',

231 Tajima's *D*, number of singletons, haplotype number and frequency distribution, and statistics

232 summarizing LD (r^2, D, D'). All statistics were calculated in non-overlapping sliding windows of

233 2 kb using *pylibseq* v. 0.2.3 [60]. ABC was performed with the R package "abc" v. 2.1 [61] using

234 all summary statistics, with "neural net" to account for non-linearity between statistics and

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

236 was found to be 0.05 (*i.e.,* 5% of the simulations were accepted during ABC inference to

237 estimate the posterior probability of each parameter). Point estimates of each inferred parameter
238 were calculated as the weighted medians of the posterior estimates.
239
240 *Simulations of different evolutionary scenarios as 'true scenarios'*
241 To consider more biologically realistic models and evaluate model violations, a number of
242 evolutionary scenarios were simulated (using *SLiM*) as follows:
243 a) Background selection: Exons experienced deleterious mutations modeled by a discrete
244 DFE comprised of four non-overlapping uniform distributions, representing the
245 effectively neutral ($-1 < 2N_{anc}s \leq 0$), weakly deleterious ($-10 < 2N_{anc}s \leq -1$),
246 moderately deleterious ($-100 < 2N_{anc}s \leq -10$), and strongly deleterious ($2N_{anc}s \leq$
247 -100) classes of mutations. All four bins were assumed to contribute equally to new
248 mutations (*i.e.*, 25% of all new mutations belonged to each class of mutation).
249 b) Positive selection: Exons experienced beneficial mutations with $\gamma = 125$ and $f_{pos} =$
250 2.2×10^{-3} (modified from [56]), resulting in $\lambda \approx 0.35$.
251 c) Population size change: A population decline was simulated such that the population
252 declined from 5000 to 100 individuals instantaneously 100 generations ago. A population
253 expansion was similarly simulated with parameters $N_{anc} = 5000$ and $N_{cur} = 10000$. A
254 population bottleneck model was also simulated such that $N_{anc} = N_{cur} = 5000$, and a
255 bottleneck occurred 2000 generations ago with a reduction to 1% of the population size
256 for 100 generations.
257 d) SNP ascertainment: Genotype error was modeled as an inability to detect the true number
258 of singletons when using low-coverage population-genomic data to call variants [38]. To

259 model this scenario, a random set of singletons, representing a third of all singletons
260 present in the sample, were removed.

261 e) Progeny skew: A skew in the offspring distribution (ψ) was modeled such that 5% and
262 10% of the population was replaced by the offspring of a single individual each
263 generation ([62], and see [63-64]).

264 f) Variation in mutation and recombination rates across the genome (*e.g.*, [65-67]): Every
265 10 kb of the ≈100 kb genomic region considered was assumed to have a different
266 mutation and recombination rate. For every simulated replicate, these rates were sampled
267 from a Gaussian distribution with the same mean as above, and a coefficient of variation
268 of 0.5. Negative values were truncated to 0.

269

270 *Posterior checks*

271 For the purposes of illustration, an example of posterior checks are provided in Figure 1 (*i.e.,*
272 showing a simple evaluation of the fit of the inferred posteriors under the incorrect models to
273 the true scenarios under consideration). Specifically, the mean estimates of the inferred
274 parameters were used to simulate the "best-fitting model" in *SLiM* v. 3.1 [58]. Exons were
275 masked and summary statistics were calculated as above in windows of 2 kb using *pylibseq*
276 v.0.2.3 [60]. In order to simulate the inferred models of positive selection, f_{pos} was
277 calculated from λ assuming a Wright-Fisher diploid population of size N and a total mutation
278 rate of μ_{tot} (which for our purpose is the same as μ). Thus, $\mu_b = f_{pos} \times \mu_{tot}$ and $\mu_{neu} =$
279 $(1 - f_{pos}) \times \mu_{tot}$ where μ_b and μ_{neu} are the beneficial and neutral mutation rates,
280 respectively. Given a value of λ, and assuming that the distribution of fitness effects of
281 beneficial mutations is exponential (with mean $\bar{s}$), we calculate f_{pos} as follows:

282 given that

283 $$\lambda = \frac{\#of\ beneficial\ subs}{\#of\ beneficial\ subs + \#of\ neutral\ subs} \qquad (1)$$

284

285 where

286 $$\#of\ beneficial\ subs = P_{fix} \times L \times 2N\mu_b \qquad (2)$$

287 and

288 $$\#of\ neutral\ subs = \mu_{neu} \times L \qquad (3)$$

289 where L is the length of the region being considered and P_{fix} is the probability of fixation of

290 beneficial mutations, given by:

291 $$P_{fix} = \int_0^{\infty} \frac{(1 - e^{-x})}{(1 - e^{-2Nx})} \left(\frac{e^{-x/\bar{s}}}{\bar{s}}\right) dx \qquad (4)$$

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

293 $$f_{pos} = \frac{\lambda}{(1 - \lambda)P_{fix}2N + \lambda} \qquad (5)$$

294

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

296

297 Statistics were calculated in non-overlapping windows of 2 kb and intervals (CIs) were

298 calculated as the 0.025 and 0.975 quantiles of the distribution of the statistics.

299

300

301

302

303 **Recommendations**

304 *Constructing an appropriate baseline model for population genomic analysis*

305 The somewhat disheartening exercise of fitting incorrect models to data (as depicted in Figure 1)
306 naturally raises the questions of whether, and if so how, accurate evolutionary inferences can be
307 extracted from DNA sequences sampled from a population. The first point of importance is that
308 the starting point for any genomic analysis should be the construction of a biologically relevant
309 baseline model, which includes the processes that must be occurring and shaping levels and
310 patterns of variation and divergence across the genome. This model should include mutation,
311 recombination, and gene conversion (each as applicable), purifying selection acting on functional
312 regions and its effects on linked variants (*i.e.,* background selection [21, 68-69]), as well as
313 genetic drift as modulated by, amongst other things, the demographic history and geographic
314 structure of the population. Depending on the organism of interest, there may be other significant
315 biological components to include, such as mating system, progeny distributions, ploidy, and so
316 on (though, for certain questions of interest, some of these biological factors may simply be
317 included in the resulting effective population size). It is thus helpful to view this baseline model
318 as being built from the ground up for any new data analysis. Importantly, the point is not that
319 these many parameters need to be fully understood in a given population in order to perform any
320 evolutionary inference, but rather that they all require consideration, and that the effects of
321 uncertainties in their underlying values on downstream inference can be quantified.

322 However, even prior to considering any biological processes, it is important to investigate
323 the data themselves. Firstly, there exists an evolutionary variance associated with the myriad of
324 potential realizations of a stochastic process, as well as the statistical variance introduced by
325 finite sampling. Secondly, it is not advisable to compare one's empirical observations, which

326 may include missing data, variant calling or genotyping uncertainty (*e.g.*, effects of low
327 coverage), masked regions (*e.g.*, regions in which variants were omitted due to low mappability
328 and/or callability) and so on, against either an analytical or simulated expectation that lacks those
329 considerations and thus assumes optimal data resolution [70]. The dataset may also involve a
330 certain ascertainment scheme, either for the variants surveyed [71], or given some pre-defined
331 criteria for investigating specific genomic regions (*e.g.*, regions representing genomic outliers
332 with respect to a chosen summary statistic [72]). For the sake of illustration, Figure 2 follows the
333 same format as Figure 1, but considers two scenarios: population growth with background
334 selection and selective sweeps, and the same scenario together with data ascertainment (in this
335 case, an under-calling of the singleton class). As can be seen, due to the changing shape of the
336 frequency spectra, neglecting to account for this ascertainment can greatly affect inference,
337 considerably modifying the fit of both the incorrect demographic and incorrect recurrent
338 selective sweep models to the data.

339 Hence, if sequencing coverage is such that rare mutations are being excluded from
340 analysis, due to an inability to accurately differentiate genuine variants from sequencing errors,
341 the model used for subsequent testing should also ignore these variants. Similarly, if multiple
342 regions are masked in the empirical analysis due to problems such as alignment difficulties, the
343 expected patterns of LD that are observable under any given model may be affected.
344 Furthermore, while the added temporal dimension of time-series data has recently been shown to
345 be helpful for various aspects of population genetic inference [73-76], such data in no way
346 sidestep the need for an appropriate baseline model, but simply requires the development of a
347 baseline that matches the temporal sampling. In sum, as these factors can greatly affect the
348 power of planned analyses and may introduce biases, the precise details of the dataset (*e.g.*,

349 region length, extent and location of masked regions, the number of callable sites, and
350 ascertainment) and study design (*e.g.,* sample size and single time-point versus time-series data)
351 should be directly matched in the baseline model construction.

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

367 The same logic follows for the next parameter addition(s): crossing over / gene
368 conversion, as applicable for the species in question. For example, for a subset of species, per-
369 generation crossover rates in cM per Mb have been estimated by comparing genetic maps based
370 on crosses or pedigrees with physical maps [81-83]. In addition, recombination rates scaled by
371 the effective population size have also been estimated from patterns of LD (*e.g.*, [84-85]) –

372 though this approach typically requires assumptions about evolutionary processes that may be
373 violated (*e.g.,* [42]). As with mutation, the effects on downstream inference arising from the
374 variety of possible recombination rates – whether estimated for the species of interest or a
375 closely related species – can be modeled.

376 The next additions to the baseline model construction are generally associated with the
377 greatest uncertainty – the demographic history of the population, and the effects of direct and
378 linked purifying selection. This is a difficult task given the virtually infinite number of potential
379 demographic hypotheses (*e.g*., [86]); furthermore, the interaction of selection with demography
380 is inherently non-trivial and difficult to treat (*e.g*., [55, 87-88]). This realization continues to
381 motivate attempts to jointly estimate the parameters of population history together with the DFE
382 of neutral, nearly neutral, weakly deleterious and strongly deleterious mutations – a distribution
383 which is often estimated in both continuous and discrete forms [89]. One of the first important
384 advances in this area used putatively neutral synonymous sites to estimate changes in population
385 size based on patterns in the SFS and conditioned on that demography to fit a DFE to non-
386 synonymous sites, which presumably experience considerable purifying selection [90-92]. This
387 stepwise approach may become problematic, however, for organisms in which synonymous sites
388 are not themselves neutral [93-95], or when the SFS of synonymous sites is affected by
389 background selection, which is probably the case generally given their close linkage to directly
390 selected non-synonymous sites ([41], and see [96-97]).

391 In an attempt to address some of these concerns, Johri *et al.* [44] recently developed an
392 ABC approach that relaxes the assumption of synonymous site neutrality and corrects for
393 background selection effects by simultaneously estimating parameters of the DFE alongside
394 population history. The posterior distributions of the parameters estimated by this approach in

395 any given data application (*i.e.*, characterizing the uncertainty of inference) represent a logical
396 treatment of population size change and purifying / background selection for the purposes of
397 inclusion within this evolutionarily relevant baseline model. That said, the demographic model in
398 this implementation is highly simplified, and extensions are needed to account for more complex
399 population histories. In particular, estimation biases that may be expected owing to the neglect of
400 cryptic population structure and migration, and indeed the feasibility of co-estimating population
401 size change and the DFE together with population structure and migration within this framework,
402 all remain in need of further investigation. While such simulation-based inference (see [98]),
403 including ABC, provides one promising platform for joint estimation of demographic history and
404 selection, progress on this front has been made using alternative frameworks as well [99-100],
405 and developing analytical expectations under these complex models should remain as the
406 ultimate, if distant, goal. Alternatively, in functionally sparse genomes with sufficiently high
407 rates of recombination, such that assumptions of strict neutrality are viable for some genomic
408 regions, multiple well-performing approaches have been developed for estimating the parameters
409 of much more complex demographic models (*e.g.*, [101-104]). In organisms for which such
410 approaches are applicable (*e.g.*, certain large, coding sequence sparse vertebrate and land-plant
411 genomes), this intergenic demographic estimation assuming strict neutrality may helpfully be
412 compared to estimates derived from data in or near coding regions that account for the effects of
413 direct and linked purifying selection [41, 44, 105]. For newly studied species lacking functional
414 annotation and information about coding density, following the joint estimation procedure would
415 remain as the more satisfactory strategy in order to account for possible background selection
416 effects.
417

418 *Quantifying uncertainty in model choice and parameter estimation, investigating potential model*
419 *violations, and defining answerable questions*
420 One of the useful aspects of these types of analyses is the ability to incorporate uncertainty in
421 underlying parameters under relatively complex models, in order to determine the impact of such
422 uncertainty on downstream inference. The computational burden of incorporating variability in
423 mutation and recombination rate estimates, or drawing from the confidence- or credibility-
424 intervals of demographic or DFE parameters, can be met with multiple highly flexible simulation
425 tools [58, 106-107]. These are also useful programs for investigating potential model violations
426 that may be of consequence. For example, if a given analysis for detecting population structure
427 assumes an absence of gene flow, it is possible to begin with one's constructed baseline model,
428 add migration parameters to the model in order to determine the effects of varying rates and
429 directions of migration on the summary statistics being utilized in the empirical analysis, and
430 thereby quantify how a violation of that assumption may affect the subsequent conclusions.
431 Similarly, if an analysis assumes the Kingman coalescent (*e.g.*, a small progeny distribution such
432 that at most one coalescent event occurs per generation), but the organism in question could
433 violate this assumption (*i.e.*, with the large progeny number distributions associated with many
434 plants, viruses, and marine spawners, or simply owing to the relatively wide array of
435 evolutionary processes that may similarly lead to multiple merger coalescent events), these
436 distributions may too be modeled in order to quantify potential downstream mis-inference.
437 To illustrate this point, Figure 3 considers two scenarios of constant population size and
438 strict neutrality but with differing degrees of progeny skew, to demonstrate that a violation of
439 this sort that is not corrected for may result in severely under-estimated population sizes as well
440 as the false inference of high rates of strong selective sweeps. In this case, the mis-inference

441 arises from the reduction in contributing ancestors under these models, as well as to the fact that
442 neutral progeny skew and selective sweeps can both generate multiple-merger events [63-64,
443 108-109]. Similarly, one may investigate the assumptions of constant mutation or recombination
444 rates when they are in reality variable. As shown in Figure 4, when these rates are assumed to be
445 constant as is common practice, but in reality vary across the genomic region under
446 investigation, the fit of the (incorrect) demographic and selection models considered may again
447 be substantially modified. Notably, this rate heterogeneity may inflate the inferred strength of
448 selective sweeps. While Figures 3 and 4 serve as examples, the same investigations may be made
449 for cases such as a fixed selective effect when there is in reality a distribution, independent
450 neutral variants when there is in reality linkage disequilibrium, panmixia when there is in reality
451 population structure, and so on. Simply put, even if a particular biological process / parameter is
452 not being directly estimated, its consequences can nonetheless be explored.

453 As detailed in Figure 5, with such a model incorporating both biological and stochastic
454 variance as well as statistical uncertainty in parameter estimates, and with an understanding of
455 the role of likely model violations, one may investigate which additional questions / hypotheses
456 can be addressed with the data at hand. By using a simulation approach starting with the baseline
457 model and adding hypothesized processes, it is possible to quantify the extent to which models,
458 and the parameters underlying those models, may be differentiated and which result in
459 overlapping or indistinguishable patterns in the data (*e.g.*, [110]). For example, if the goal of a
460 given study is to identify recent beneficial fixations in a genome – be they potentially associated
461 with high-altitude adaptation in humans, crypsis in mice, or drug-resistance in a virus – one may
462 begin with the baseline model and simulate selective sweeps under that model. As illustrated in
463 Figure 6, by varying the strengths, rates, ages, dominance and epistasis coefficients of beneficial

464 mutations, the patterns in the SFS, LD, and/or divergence that may differentiate the addition of
465 such selective sweep parameters from the baseline expectations can be quantified. Moreover, any
466 intended empirical analyses can be evaluated using simulated data (*i.e.*, the baseline, compared
467 to the baseline + the hypothesis) to define the power and false-positive rates associated. If the
468 differences in resulting patterns cannot be distinguished from the expected variance under the
469 baseline model (in other words, if the power and false-positive rate of the analyses are not
470 favorable), the hypothesis is not addressable with the data at hand (*e.g.*, [54]). If the results are
471 favorable, this analysis can further quantify the extent to which the hypothesis may be tested;
472 perhaps only selective sweeps from rare mutations with selective effects greater than 1% and that
473 have fixed within the last 0.1 N_e generations are detectable (see [111-112]), and any others could
474 not be statistically distinguished from expected patterns under the baseline model. Hence, such
475 an exercise provides a critically essential key for interpreting the resulting data analysis.
476
477 *A consideration of alternative strategies*
478 In this regard, it is worth mentioning two common approaches that may be viewed as alternatives
479 to the strategy that we recommend. The first tactic concerns identifying patterns of variation that
480 are uniquely and exclusively associated with one particular process, the presence of which could
481 support that model regardless of the various underlying processes and details composing the
482 baseline. For example, Fay & Wu's [113] *H*-statistic, capturing an expected pattern of high-
483 frequency derived alleles generated by a selective sweep with recombination, was initially
484 proposed as a powerful statistic for differentiating selective sweep effects from alternative
485 models. Results from the initial application of the *H*-statistic were interpreted as evidence of
486 widespread positive selection in the genome of *Drosophila melanogaster*. However, Przeworski

487 [112] subsequently demonstrated that the statistic was characterized by low power to detect
488 positive selection, and that significant values could readily be generated under multiple neutral
489 demographic models. The composite likelihood framework of Kim & Stephan [111] provided a
490 significant improvement by incorporating multiple predictions of a selective sweep model, and
491 was subsequently built upon by Nielsen *et al.* [114] in proposing the SweepFinder approach.
492 However, Jensen *et al*. [115] described low power and high false-positive rates under certain
493 neutral demographic models. The particular pattern of LD generated by a beneficial fixation with
494 recombination described by Kim & Nielsen [116] and Stephan *et al*. [117] (and see [118]), was
495 also found to be produced under an (albeit more limited) range of severe neutral population
496 bottlenecks [119-120].

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

507 The second common strategy involves summarizing empirical distributions of a given
508 statistic, and assuming that outliers of that distribution represent the action of a process of
509 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 an hypothesized additional evolutionary process from such

533 'background noise', are both quantifiable. Furthermore, even if the model were correct, there
534 exists a limit on the precision of estimation imposed by the evolutionary variance in population
535 statistics that requires description, and which no amount of sampling can remove.

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

548 Importantly, the framework we describe will also generally identify many models and
549 parameter realizations that are in fact inconsistent with the observed data. This 'ruling-out'
550 process can often be just as useful as model-fitting, and rejecting possible hypotheses is
551 frequently the more robust exercise of the two. The value of this narrowing down, rather than the
552 enthusiastic promotion of individual scenarios, is worthy of heightened appreciation.
553 Nevertheless, all models should not be viewed equally. Decades of work supporting the central
554 tenets of the Neutral Theory [19], high-quality experimental and computational work quantifying
555 mutation and recombination rates [77-79, 83-84, 123], constantly improving experimental and

556 theoretical approaches to quantify the neutral and deleterious DFE from natural population,
557 mutation-accumulation, or directed mutagenesis data [44, 90, 124-126], and historical knowledge
558 (*e.g.*, anthropological, ecological, clinical) of population size change or structure – combined
559 with the fact that all of these factors may strongly shape observed levels and patterns of variation
560 and divergence – justify their role in comprising the appropriate baseline model for genomic
561 analysis.

562 Given this, and particularly after accounting for the inflation of variance contributed by
563 uncertainty in relevant parameters, potential model violations, as well as the quantity and quality
564 of data available in any given analysis, it will often be the case that many hypotheses of interest
565 may not be addressable with the dataset and knowledge at hand. However, recognizing that a
566 question cannot be accurately answered, and defining the conditions under which it could
567 become answerable, should be preferred over making unfounded and thus misleading claims.
568 Consistent with this call for caution however, it should equally be emphasized that the fit of a
569 baseline model to data is certainly not inherent evidence that the model encompasses all relevant
570 processes shaping the population. In reality, it is virtually guaranteed not to be all-encompassing,
571 and building these models involves simplifying more complex processes (for a helpful and more
572 general perspective, see [127]). When an additional process cannot be satisfactorily detected, that
573 may rather be viewed as a statement about statistical identifiability – the inability to distinguish a
574 hypothesized process from other processes that are known to be acting – and in such scenarios,
575 absence of evidence need not be taken as evidence of absence.

576 While the many considerations we describe may appear daunting, it is our hope that these
577 recommendations may serve as a useful roadmap for future data analyses in population
578 genomics, one that may inform not only the perspectives of authors, but also that of reviewers

579 and editors as well. Helpfully, these strategies can save considerable time, money, and effort
580 prior to the start of empirical data handling, by determining which questions are accessible to the
581 researcher. If a question is addressable, this preliminary analysis can additionally define what
582 types of data are needed; for example, the number of variants or sample size necessary to obtain
583 sufficient power, or how alternative data collections (*e.g.*, temporal samples) could improve
584 resolution. This further highlights the value of defining specific hypotheses and of studying
585 specific patterns as opposed to running a general suite of software on each new dataset in the
586 hopes of identifying something of interest – namely, one cannot define the power of a study to
587 address an unformulated question. Such hypothesis-driven population genomics has resulted in a
588 number of success stories over the past decade; systems in which specific hypotheses were
589 formed, data was collected for the purpose, detailed population genomic analyses were designed,
590 and ultimately important insights were gained about the evolutionary history of the population in
591 question (*e.g.*, the study of cryptic coloration has proven fruitful in this regard [3]). One feature
592 common to these studies is interdisciplinarity: the utilization of population genetic theory and
593 inference as described here, combined with classical genetic crosses, large-scale field studies,
594 and genetic manipulation in order to connect genotype to phenotype to fitness and to validate
595 statistical inference. Importantly however, without a population genetic framework for defining
596 hypotheses, quantifying processes contributing to observed variation and divergence, evaluating
597 and distinguishing amongst competing models, and defining uncertainty and potential biases, the
598 observations remain merely descriptive.
599
600
601

ACKNOWLEDGEMENTS

This work was funded by National Institutes of Health grants R01GM135899 and R35GM139383 to JDJ. 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.

610 **CITATIONS**

611

612 [1] Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E.
613 Tracing the peopling of the world through genomics. Nature 2017;541(7637): 302-310.
614

615 [2] Renzette N, Gibson L, Jensen JD, Kowalik TF. Human cytomegalovirus intrahost
616 evolution - a new avenue for understanding and controlling herpesvirus infections. Curr
617 Opin Virol 2014;8: 109-115.
618

619 [3] Harris RB, Irwin K, Jones MR, Laurent S, Barrett RDH, Nachman MW, et al. The
620 population genetics of crypsis in vertebrates: recent insights from mice, hares, and
621 lizards. Heredity 2020;124(1): 1-14.
622

623 [4] Irwin KK, Renzette N, Kowalik TF, Jensen JD. Antiviral drug resistance as an adaptive
624 process. Virus Evol 2016;2(1): vew014.

625 [5] Fisher RA. The genetical theory of natural selection. Clarendon Press, Oxford, UK; 1930.

626 [6] Wright S. Evolution in Mendelian populations. Genetics 1931;16(2): 97-159.

627 [7] Haldane JBS. The causes of evolution. Longmans, London, UK; 1932.
628

629 [8] Provine WB. The origins of theoretical population genetics. University of Chicago Press;
630 1971.
631

632 [9] Darwin C. On the origin of species by means of natural selection, or the preservation of
633 favoured races in the struggle for life. John Murray, London, UK; 1859.
634

635 [10] Mendel G. Versuche über Pflanzenhybriden. Verh Naturforsch Ver Brünn 1866;4: 3-47.

636 [11] Lewontin RC. The genetic basis of evolutionary change. Columbia Univ. Press, New
637 York; 1974.

638 [12] Lewontin RC. Twenty-five years ago in Genetics: electrophoresis in the development of
639 evolutionary genetics: milestone or millstone? Genetics 1991;128(4): 657-662.

640 [13] Kimura M. Evolutionary rate at the molecular level. Nature 1968;217(5129): 624-626.

641 [14] Kimura M. The neutral theory of molecular evolution. Cambridge Univ. Press,
642 Cambridge; 1983.

643 [15] Ohta T. Slightly deleterious mutant substitutions in evolution. Nature 1973;246(5428):
644 96-98.

645

[16] King JL, Jukes TH. Non-Darwinian evolution. Science 1969;164(3881): 788-798.

[17] Kern AD, Hahn MW. The neutral theory in light of natural selection. Mol Biol Evol 2018;35(6): 1366-1371.

[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): 111-114.

[20] Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res 1974;23(1): 23-25.

[21] Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics 1993;134(4): 1289-1303.

[22] Charlesworth B, Jensen JD. The effects of selection at linked sites on patterns of genetic variability. Annu Rev Ecol Evol Syst 2021;52: 177-197.

[23] Ray N, Excoffier L. Inferring past demography using spatially explicit population genetic models. Hum Biol 2009;81(2-3): 141-157.

[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: 433-456.

[25] Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. BMC Biol 2017;15(1): 98.

[26] Stephan W. Selective sweeps. Genetics 2019;211(1): 5-13.

[27] Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet 2007;8(8): 610-618.

[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): 540-546.

[29] Keightley PD, Halligan DL. Analysis and implications of mutational variation. Genetica 2009;136(2): 359-369.

[30] Keightley PD. Rates and fitness consequences of new mutations in humans. Genetics 2012;190(2): 295-304.

682 [31] Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic drift,
683 selection and the evolution of the mutation rate. Nat Rev Genet 2016;17(11): 704-714.
684
685 [32] Stumpf MP, McVean GA. Estimating recombination rates from population-genetic data.
686 Nat Rev Genet 2003;4(12): 959-968.
687
688 [33] Auton A, Fledel-Alon A, Pfeifer SP, Venn O, Ségurel L, Street T, et al. A fine-scale
689 chimpanzee genetic map from population sequencing. Science 2012;336(6078): 193-198.
690
691 [34] Spence JP, Song YS. Inference and analysis of population-specific fine-scale
692 recombination maps across 26 diverse human populations. Sci Adv 2019;5(10):
693 eaaw9206.

694 [35] Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population
695 genetics. Genetics 2002;162(4): 2025-2035.

696 [36] Beaumont MA, Rannala B. The Bayesian revolution in genetics. Nat Rev Genet
697 2004;5(4): 251-261.
698
699 [37] Schraiber JG, Akey JM. Methods and models for unravelling human evolutionary history.
700 Nat Rev Genet 2015;16(12): 727-740.
701
702 [38] Han E, Sinsheimer JS, Novembre J. Characterizing bias in population genetic inferences
703 from low-coverage sequencing data. Mol Biol Evol 2014;31(3): 723-735.
704
705 [39] Pfeifer SP. Studying mutation rate evolution in primates – the effects of computational
706 pipeline and parameter choices. GigaScience 2021;10(10): giab069.
707
708 [40] Ewing G, Jensen JD. The consequences of not accounting for background selection in
709 demographic inference. Mol Ecol 2016;25(1): 135-141.
710
711 [41] Pouyet F, Aeschbacher S, Thiery A, Excoffier L. Background selection and biased gene
712 conversion affect more than 95% of the human genome and bias demographic inferences.
713 eLife 2018;7: e36317.
714
715 [42] Dapper AL, Payseur BA. Effects of demographic history on the detection of
716 recombination hotspots from linkage disequilibrium. Mol Biol Evol 2018;35(2): 335-353.
717
718 [43] Rousselle M, Maeva M, Nabholz B, Bataillon T, Galtier N. Overestimation of the
719 adaptive substitution rate in fluctuating populations. Biol Lett 2018;14(5): 20180055.
720
721 [44] Johri P, Charlesworth B, Jensen JD. Towards an evolutionarily appropriate null model:
722 jointly inferring demography and purifying selection. Genetics 2020;215(1): 173-192.
723

[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): 342-348.

[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.

[48] Louca S, Pennell MW. Extant timetrees are consistent with a myriad of diversification histories. Nature 2020;580(7804): 502-505.

[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.

[51] Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the human genome. Mol Biol Evol 2017;34(8): 1863-1877.

[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.

[53] Barton NH. Genetic hitchhiking. Phil Trans R Soc B 2000;355(1403): 1553-1562.

[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.

[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.

[56] Campos JL, Charlesworth B. The effects on neutral variability of recurrent selective sweeps and background selection. Genetics 2019;212(1): 287-303.

[57] Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol 2016;12(5): e1004842.

[58] Haller BC, Messer PW. SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. Mol Biol Evol 2019;36(3): 632-637.

763 [59] Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the
764 presence of slightly deleterious mutations and population size change. Mol Biol Evol
765 2009;26(9): 2097-2108.

766 [60] Thornton K. Libsequence: a C++ class library for evolutionary genetic analysis.
767 Bioinformatics 2003;19(17): 2325–2327.
768
769 [61] Csilléry K, François O, Blum M. abc: an R package for approximate Bayesian
770 computation (ABC). Methods Ecol Evol 2012;3: 475-479.
771
772 [62] Eldon B, Wakeley J. Coalescent processes when the distribution of offspring
773 number among individuals is highly skewed. Genetics 2006;172(4): 2621-2633.
774
775 [63] Matuszewski M, Hildebrandt ME, Achaz G, Jensen JD. Coalescent processes with
776 skewed offspring distributions and non-equilibrium demography. Genetics 2018;208(1):
777 323-338.
778
779 [64] Sackman A, Harris RB, Jensen JD. Inferring demography and selection in organisms
780 characterized by skewed offspring distributions. Genetics 2019;211(3): 1019-1028.
781
782 [65] McVean G, Myers S, Hunt S, Deloukas P, Bentley D, Donnelly P. The fine-scale
783 structure of recombination rate variation in the human genome. Science 2004;304(5670):
784 581-584.
785
786 [66] Chan AH, Jenkins P, Song Y. Genome-wide fine-scale recombination rate variation in
787 *Drosophila melanogaster*. PLoS Genet 2012;8(12): e1003090.
788
789 [67] Penalba JV, Wolf JB. From molecules to populations: appreciating and estimating
790 recombination rate variation. Nat Rev Genet 2020;21(8): 476-492.

791 [68] Charlesworth D, Charlesworth B, Morgan MT. The pattern of neutral molecular variation
792 under the background selection model. Genetics 1995;141(4): 1619-1632.

793 [69] Charlesworth B. Background selection 20 years on. The Wilhelmine E. Key 2012
794 invitational lecture. J Hered 2013;104(2): 161-171.

795 [70] Pfeifer SP. From next-generation resequencing reads to a high quality variant data
796 set. Heredity 2017;118(2): 111-124.
797
798 [71] Nielsen R. Population genetic analysis of ascertained SNP data. Hum Genomics
799 2004;1(3): 218-224.
800
801 [72] Thornton KR, Jensen JD. Controlling the false positive rate in multi-locus genome scans
802 for selection. Genetics 2007;175(2): 737-750.
803

[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.

[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 Res 2015;15(1): 87-98.

[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): 831-846.

[76] Lynch M, Ho WC. The limits to estimating population-genetic parameters with temporal data. Gen Biol Evol 2020;12(4): 443-455.

[77] Pfeifer SP. Spontaneous mutation rates. In *The 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.

[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): 1739-1749.

[80] Maddamsetti R, Grant NA. Divergent evolution of mutation rates and biases in the long-term evolution experiment with *Escherichia coli*. Gen Biol Evol 2020;12(9): 1591-1603.

[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): 241-247.

[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): 1335-1344.

[83] Comeron JM, Ratnappan R, Bailin S. The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet 2012;8(10): e1002905.

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

[85] Pfeifer SP. A fine-scale genetic map for vervet monkeys. Mol Biol Evol 2020;37(7): 1855-1865.

842 [86] Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. The confounding effects of
843 population structure, genetic diversity and the sampling scheme on the detection and
844 quantification of population size change. Genetics 2010;186(3): 983-995.

845 [87] Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L. On the accumulation of
846 deleterious mutations during range expansions. Mol Ecol 2013;22(24): 5972-5982.
847
848 [88] Peischl S, Kirkpatrick M, Excoffier L. Expansion load and the evolutionary
849 dynamics of a species range. Am Nat 2015;185(4): E81-E93.
850
851 [89] Johri P, Eyre-Walker A, Gutenkunst RN, Lohmueller KE, Jensen JD. On the prospect of
852 achieving accurate joint estimation of selection with population history. In revision, Gen
853 Biol Evol.

854 [90] Keightley PD, Eyre-Walker A. Joint inference of the distribution of fitness effects of
855 deleterious mutations and population demography based on nucleotide polymorphism
856 frequencies. Genetics 2007;177(4): 2251-2261.

857 [91] Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. A method for
858 inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics
859 2011;189(4): 1427-1437.

860 [92] Lynch M. The origins of genome architecture. Sinauer Associates, Sunderland, MA;
861 2007.

862 [93] Singh ND, Bauer DuMont VL, Hubisz MJ, Nielsen R, Aquadro CF. Patterns of mutation
863 and selection at synonymous sites in *Drosophila*. Mol Biol Evol 2007;24(12): 2687-
864 2697.

865 [94] Zeng K, Charlesworth B. Studying patterns of recent evolution at synonymous sites and
866 intronic sites in *Drosophila melanogaster*. J Mol Evol 2010;70(1): 116-128.

867 [95] Choi JY, Aquadro CF. Recent and long term selection across synonymous sites in
868 *Drosophila ananassae*. J Mol Evol 2016;83(1-2): 50-60.

869 [96] Comeron JM. Background selection as baseline for nucleotide variation across the
870 *Drosophila* genome. PLoS Genet. 2014;10(6): e1004434.

871 [97] Comeron JM. Background selection as a null hypothesis in population genomics:
872 insights and challenges from *Drosophila* studies. Phil Trans R Soc B 2017;372(1736):
873 20160471.

874 [98] Cranmer K, Brehmer J, Louppe G. The frontier of simulation-based inference. Proc Natl
875 Acad Sci USA 2020;117(48): 30055-30062.
876

[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 USA 2005;102(22): 7882-7887.

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

[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.

[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.

[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): 1330-1338.

[104] Steinrücken M, Kamm J, Spence J, Song YS. Inference of complexpopulation histories using whole-genome sequences from multiple populations. Proc Natl Acad Sci USA 2019;116(34): 17115-17120.

[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.

[106] Thornton KR. A C++ template library for efficient forward-time population genetic simulation of large populations. Genetics 2014;198(1): 157-166.

[107] Kelleher J, Thornton K, Ashander J, Ralph P. Efficient pedigree recording for fast population genetics simulation. PLoS Comput. Biol 2018;14(11): e1006581.

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

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

[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): 439-449.

[111] Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 2002;160(2): 765-777.

[112] Przeworski M. The signature of positive selection at randomly chosen loci. Genetics 2002;160(3): 1179-1189.

914 [113] Fay J, Wu CI. Hitchhiking under positive Darwinian selection. Genetics 2000;155(3):
915 1405-1413.

916 [114] Nielsen R., Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante CD. Genomic
917 scans for selective sweeps using SNP data. Genome Res 2005;15(11): 1566-1575.

918 [115] Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD. Distinguishing between
919 selective sweeps and demography using DNA polymorphism data. Genetics 2005;170(3):
920 1401-1410.

921 [116] Kim Y, Nielsen R. Linkage disequilibrium as a signature of selective sweeps. Genetics
922 2004;167(3): 1513-1524.

923 [117] Stephan W, Song YS, Langley CH. Hitchhiking effect on linkage disequilibrium between
924 linked neutral loci. Genetics 2006;172(4): 2647-2663.

925 [118] McVean G. The structure of linkage disequilibrium around a selective sweep.
926 Genetics 2007;175(3): 1395-1406.
927
928 [119] Jensen JD, Thornton KR, Bustamante CD, Aquadro CF. On the utility of linkage
929 disequilibrium as a statistic for identifying targets of positive selection in non-equilibrium
930 populations. Genetics 2007;176(4): 2371-2379.
931
932 [120] Crisci J, Poh YP, Mahajan S, Jensen JD. The impact of equilibrium assumptions on tests
933 of selection. Front Genet 2013; 4: 235.

934 [121] Garud N, Messer P, Petrov D. Detection of hard and soft selective sweeps from
935 *Drosophila melanogaster* population genomic data. PLoS Genet. 2021;17(2): e1009373.

936 [122] Teshima K, Coop G, Przeworski M. How reliable are empirical genome scans for
937 selective sweeps? Genome Res 2006;16(6): 702-712.
938
939 [123] Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, et al. A genome-wide
940 view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci USA
941 2008;105(27): 9272-9277.
942
943 [124] Bank C, Hietpas RT, Wong A, Bolon DNA, Jensen JD. A Bayesian MCMC approach to
944 assess the complete distribution of fitness effects of new mutations: uncovering the
945 potential for adaptive walks in challenging environments. Genetics 2014;196(3): 841-
946 852.
947
948 [125] Foll M, Poh YP, Renzette N, Ferrer-Admetlla A, Shim H, Malaspinas AS, et al. Influenza
949 virus drug resistance: a time-sampled population genetics perspective. PLoS
950 Genet 2014;10(2): e1004185.
951

[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.

[127] Gelman A, Shalizi CR. Philosophy and the practice of Bayesian statistics. Br J Math Stat Psychol 2013;66(1): 8-38.

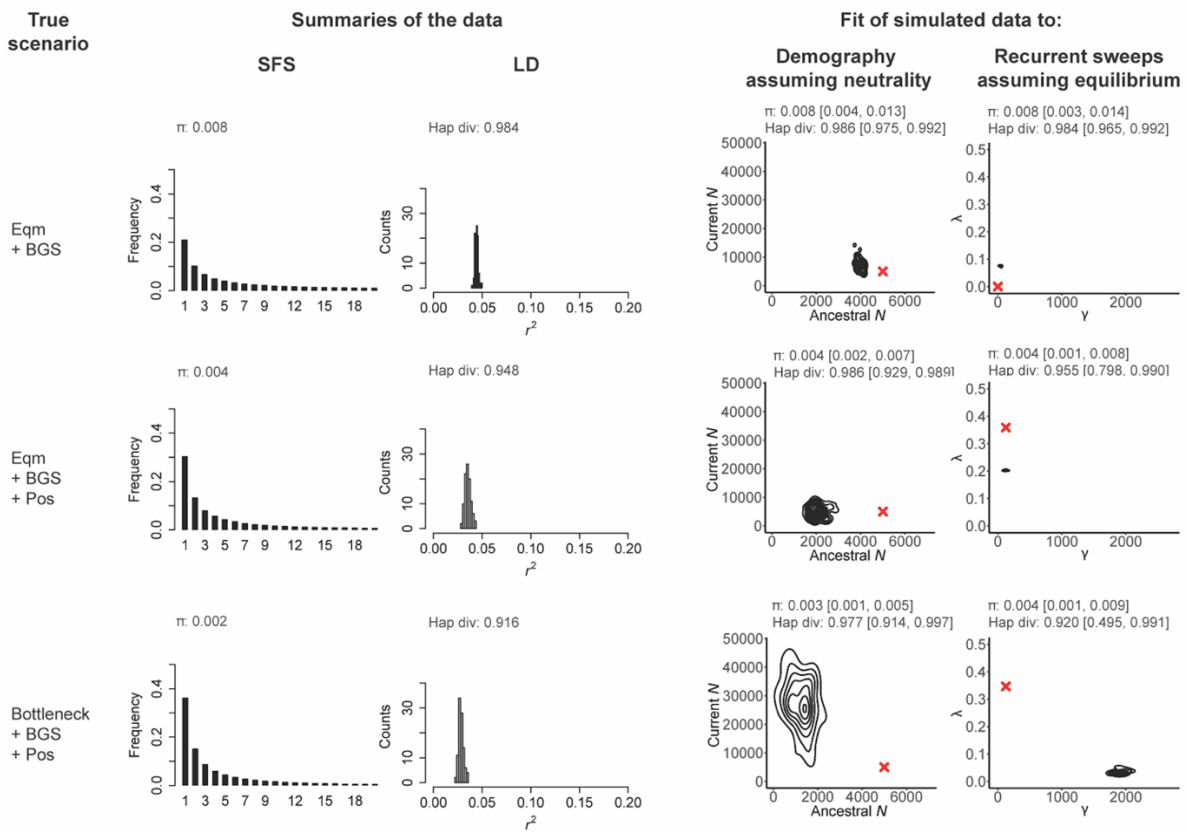

Figure 1. Incorrect models may often readily be fit to a given dataset
Here we present three scenarios varying from simple to more complex: the first row presents a constant-sized population experiencing background selection (denoted by 'Eqm + BGS'), the second row is the same scenario with the addition 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 (π) 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, 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.

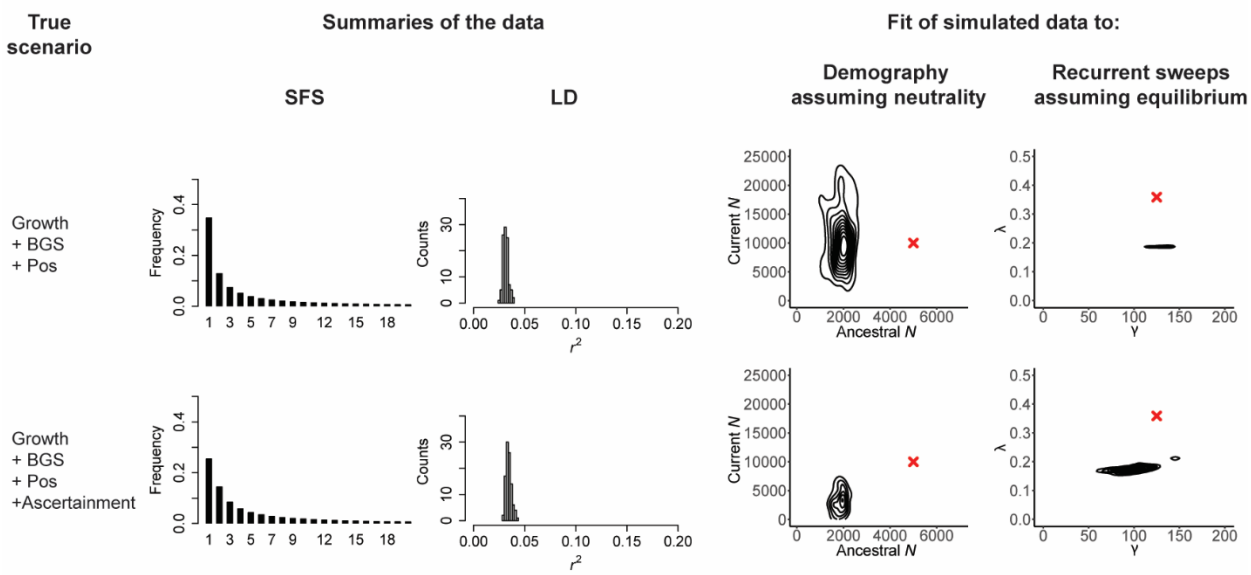

Figure 2. Ascertainment errors may amplify mis-inference, if not corrected
As in Figure 1, the scenarios are given in the first column, here population growth with background selection and recurrent selective sweeps ('Growth + BGS + Pos'), as well as the same scenario in which the imperfections of the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

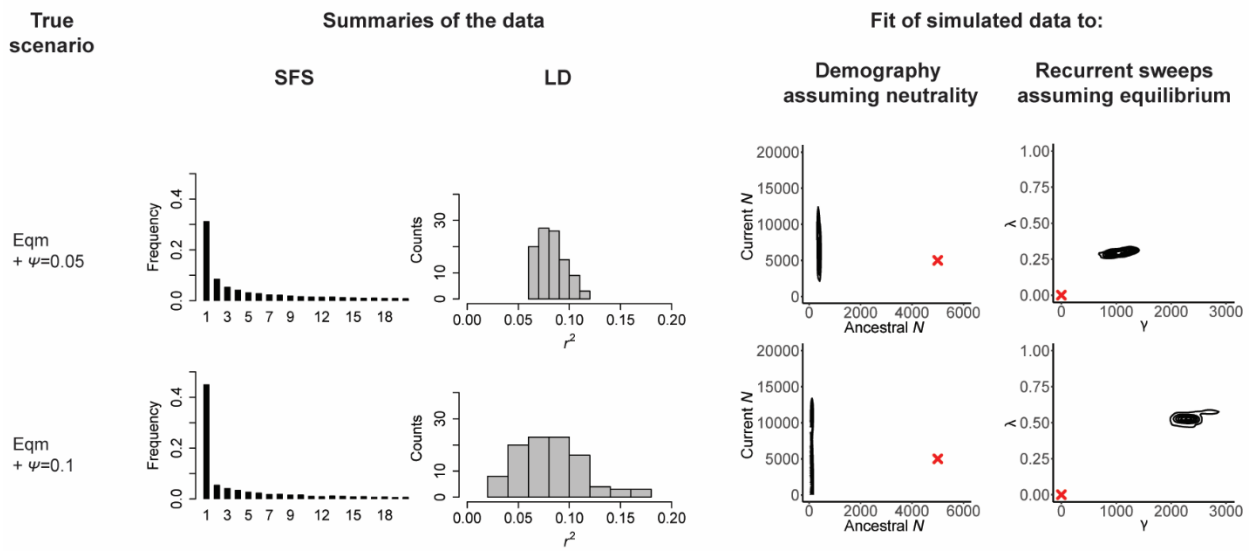

Figure 3. The impact of potential model violations can be quantified
As in Figures 1 and 2, the scenarios are given in the first column, here equilibrium population size together with a moderate degree of progeny skew ('Eqm + $\psi = 0.05$') as well as with a high degree of progeny skew ('Eqm + $\psi = 0.1$') (see 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 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.

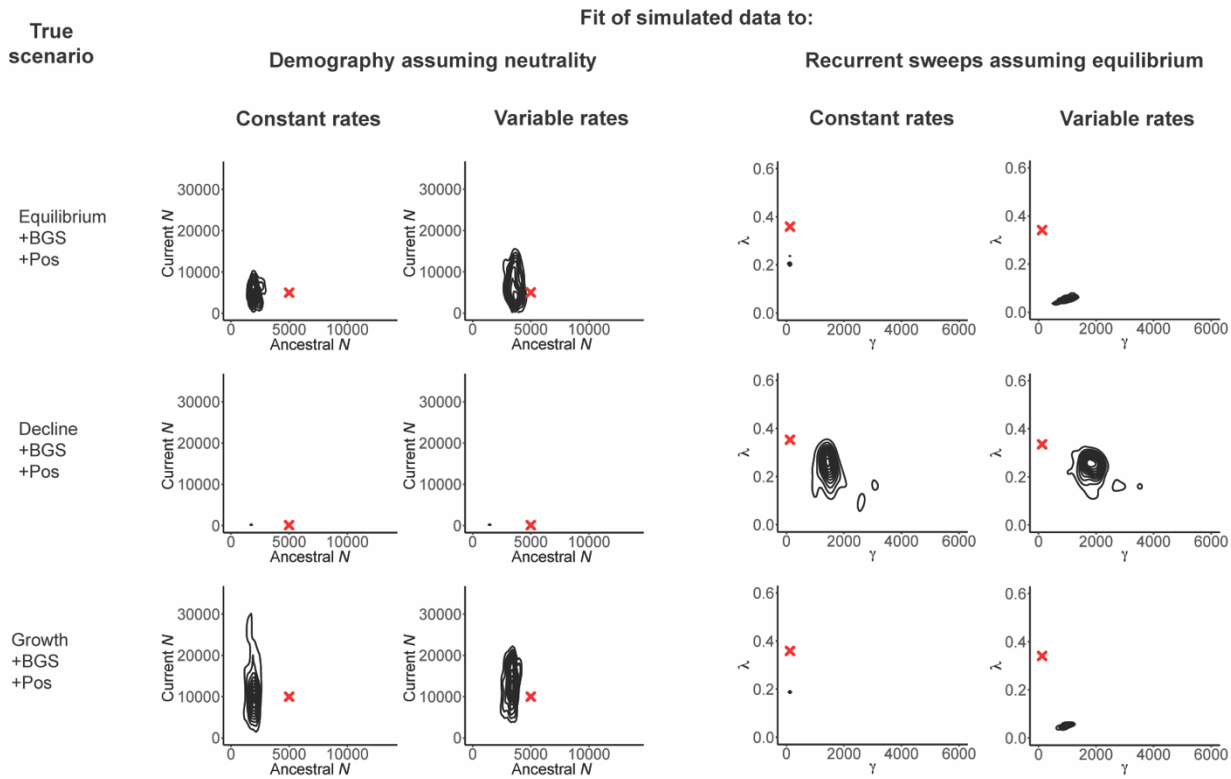

Figure 4. The effects of not correcting for mutation and recombination rate heterogeneity Three scenarios are here considered: equilibrium population size with background selection and recurrent selective sweeps ('Eqm +BGS + Pos'), declining population size together with background selection and recurrent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

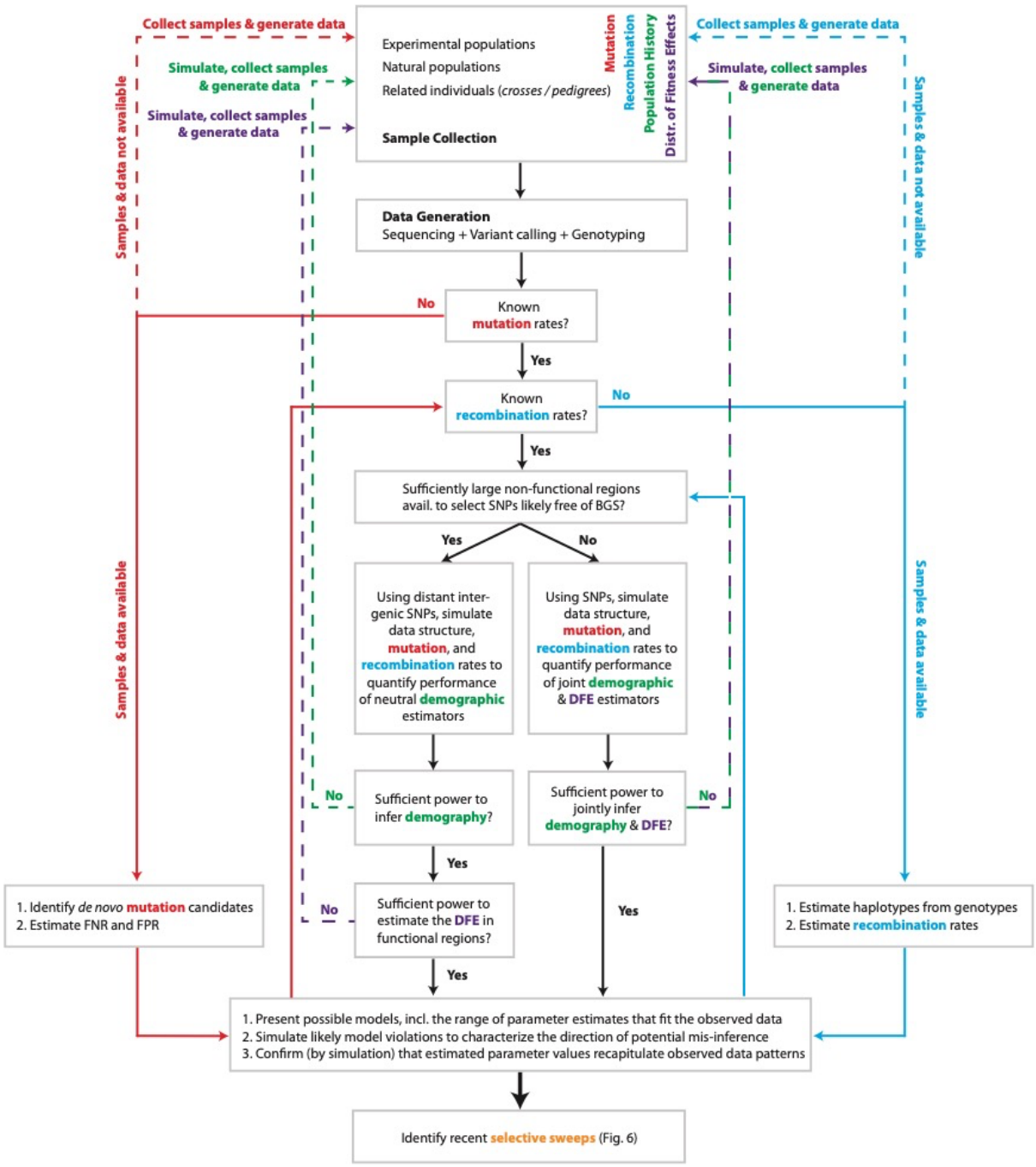

Figure 5: **Diagram of important considerations in constructing a baseline model for genomic analysis**

Considerations related to mutation rate are coded in red, recombination rate in blue, demographic history in green, and the 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. Dotted lines indicate a return to the starting point.

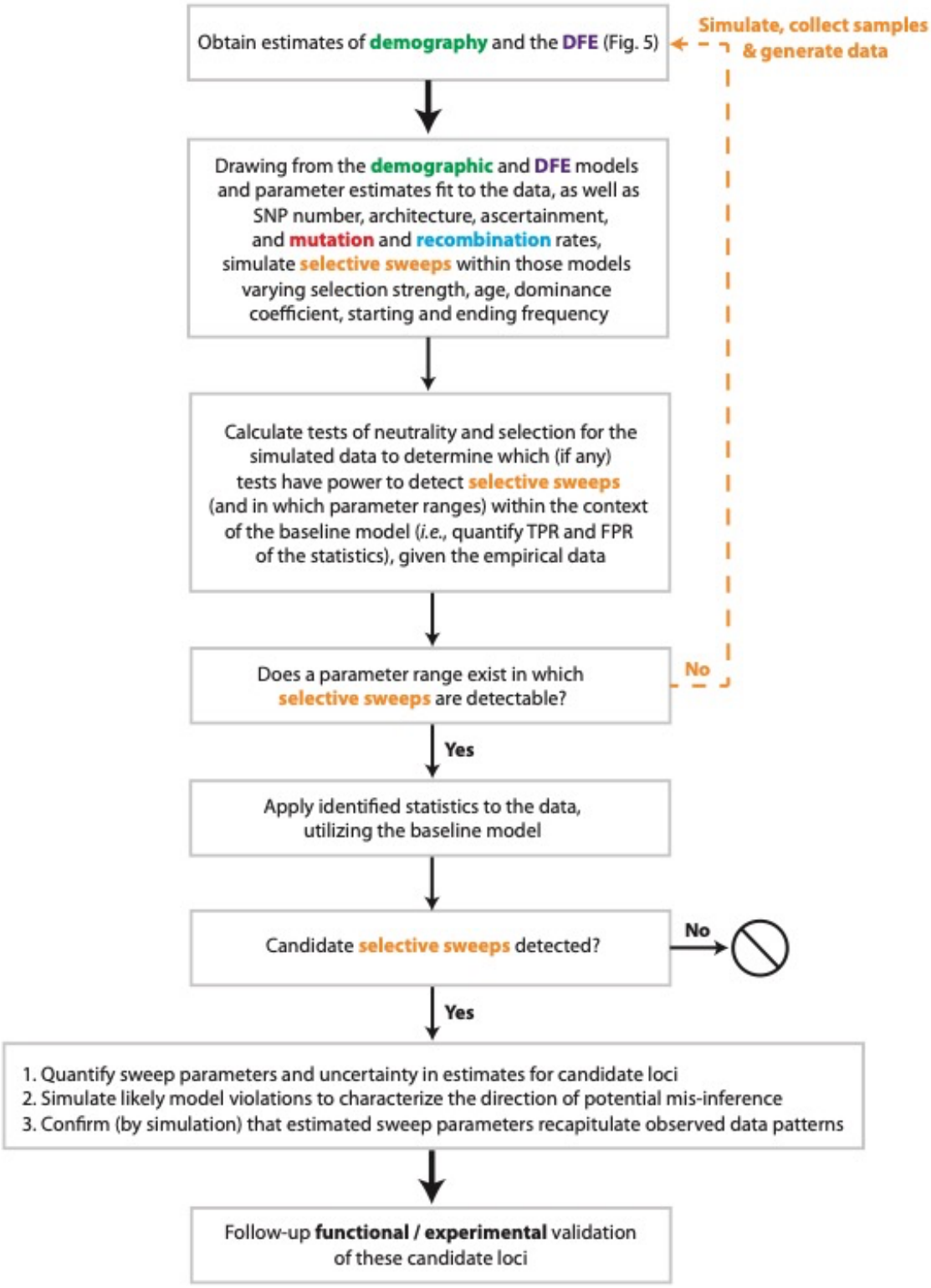

Figure 6: **Diagram of important considerations in detecting selective sweeps**
The color scheme matches that in Figure 5, with 'selective sweeps' coded in orange.