Review

CellPress

Thinking too positive? Revisiting current methods of population genetic selection inference

Claudia Bank[1,2], Gregory B. Ewing[1,2], Anna Ferrer-Admettla[1,2,3], Matthieu Foll[1,2], and Jeffrey D. Jensen[1,2]

[1] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland
[3] Department of Biology and Biochemistry, University of Fribourg, 1700 Fribourg, Switzerland

In the age of next-generation sequencing, the availability of increasing amounts and improved quality of data at decreasing cost ought to allow for a better understanding of how natural selection is shaping the genome than ever before. However, alternative forces, such as demography and background selection (BGS), obscure the footprints of positive selection that we would like to identify. In this review, we illustrate recent developments in this area, and outline a roadmap for improved selection inference. We argue (i) that the development and obligatory use of advanced simulation tools is necessary for improved identification of selected loci, (ii) that genomic information from multiple time points will enhance the power of inference, and (iii) that results from experimental evolution should be utilized to better inform population genomic studies.

Identification of beneficial mutations in the genome: an ongoing quest

The identification of genetic variants that confer an advantage to an organism, and that have spread by forces other than chance, remains an important question in evolutionary biology. Success in this regard will have broad implications, not only for informing our view of the process of evolution itself, but also for evolutionary applications ranging from clinical to ecological. Despite the tremendous quantity of polymorphism data now at our fingertips, which, in principle, ought to allow for a better characterization of such adaptive genetic variants, it remains a challenge to unambiguously identify alleles under selection. This is primarily owing to the difficulty in disentangling the effects of positive selection from those of other factors that shape the composition of genomes, including both demography as well as other selective processes.

Approaches to identify positively selected variants from genomic data can be broadly divided into two categories: those that make use of within-population polymorphism data, and those that make use of between-population/species data. While each approach has its respective merits, in this review, we focus on recent developments in population genetic inference from polymorphism data in both natural and experimental settings (see [1,2] for more general reviews, and [3,4] for recent and specific literature on divergence-based selection inference). For population genetic inference from single time point polymorphism data (as is most commonly the case), this includes not only sophisticated statistical methods, but also simulation programs that enable us to model expected genomic signatures under a wide variety of possible scenarios.

Glossary

Background selection: reduction of genetic diversity due to selection against deleterious mutations at linked sites.
Coalescent simulator: simulation tool that reconstructs the genealogical history of a sample backwards in time. This greatly reduces computational effort, but only models in which mutations are independent of the sample's genealogy can be implemented.
Cost of adaptation: the deleterious effect that a beneficial mutation can have in a different environment. Prominent examples are antibiotic resistance mutations, which have often been observed to cause reduced growth rates (as compared with the wild type) in the absence of antibiotics.
Demographic history: the population history of a sample of individuals, which can include population size changes, differing sex ratios, migration rates, splitting and reconnection of the population, as well as variation over time in these parameters.
Distribution of fitness effects (DFE): the statistical distribution of selection coefficients of all possible new mutations, as compared with a reference genotype.
Epistasis: the interaction of mutational effects, resulting in a dependence of the effect of a mutation on the background it appears on.
Forward simulator: simulation tool that models the evolution of populations forward in time. This allows for implementation of complex models, but also usually results in much longer computation times because all individuals/haplotypes must be tracked.
Nonequilibrium model: any model that incorporates violations of the assumptions of the standard neutral model (see below).
Selection coefficient: a measure of the strength of selection on a selected genotype. Usually, the selection coefficient is measured as the relative difference between the reproductive success of the selected and the ancestral genotypes.
Selective sweep: the process of a beneficial mutation (and its closely linked chromosomal vicinity) being driven ('swept') to high frequency or fixation by natural selection. Selective sweeps result in a genomic signature including a local reduction in genetic variation, and skews in the SFS.
Standard neutral model: under this model [67], the population resides in an equilibrium of allele frequencies determined by the (constant) mutation rate and population size. The model assumptions include random mating, binomial sampling of offspring, and no selection.

Corresponding author: Bank, C. (claudia.bank@epfl.ch).
Keywords: natural selection; background selection; population genetic inference; evolution; computational biology.

0168-9525/
 http://dx.doi.org/10.1016/j.tig.2014.09.010

Alternatively, data from multiple time points – such as those recently afforded by ancient genomic data as well as many clinical and experimental datasets – can be used to greatly improve inference by catching a selective sweep 'in the act'. Finally, recent results from the experimental evolution literature have begun to better illuminate the expected distribution of fitness effects (DFE), the associated costs of adaptation, and the extent of epistasis (see Glossary). In this opinion piece, we present an overview of recent developments for selection inference in the abovementioned areas, and offer a roadmap for future method development.

Selection inference from a single time point

One of the earliest efforts to quantify selection in a natural population was based on multiple time point phenotypic data [5,6]. At the onset of genomics, however, new sequencing methods were both tedious and expensive, limiting data collection to single time points. For this reason, one generally observes only the footprints of the selection process, making it more difficult to distinguish regions shaped by neutral processes from those shaped by selective processes. Over the past number of decades, population genetic theory has predicted the effects of different selection models on molecular variation. These predictions have given rise to test statistics designed to detect selection using polymorphism data, based on patterns of population differentiation (e.g., [7–10]), the shape of the site-frequency spectrum (SFS) (e.g., [11–13]), and haplotype–linkage disequilibrium (LD) structure (e.g., [14–17]).

Despite efforts to create statistics robust to demography, all currently available methods to detect selection are prone to misinference under nonequilibrium models [18,19]. Therefore, in parallel to the production of statistics for inferring selection, a separate class of methods has been developed to estimate the demographic history of populations utilizing the same patterns of variation [20,21]. This lends itself to a two-step approach when analyzing population genetic data: first, demography is inferred using a putatively neutral class of sites, and that model is then used to test for selection among a putatively selected class of sites [13,16]. However, the assumptions enabling this inference are highly problematic because they rely on the correct identification of a class of sites that is both neutral and untouched by linked selection. BGS, for example, may indeed influence a large fraction of the genome in many species [22]. Thus, the misidentification of such a class of sites in the genome may not only result in the misinference of the underlying demographic history, but also in misinference of selection owing to the incorrect estimation of the demographic null model – leading to both false positives and false negatives (Figure 1).

In order to circumvent this problem, it is, therefore, necessary to develop methods that can jointly infer the demographic and selective history of the population simultaneously, recognizing that both processes are likely to shape the majority of the genome in concert [23]. The development of simulation software that can model both positive and negative selection in nonequilibrium populations (see below) is a step towards this goal, as it allows for the generation of expected patterns of variation under such scenarios. Recently, the introduction of likelihood-free inference frameworks such as Approximate Bayesian Computation (ABC) [24] has made it computationally feasible to combine demographic and selective inference. Recent

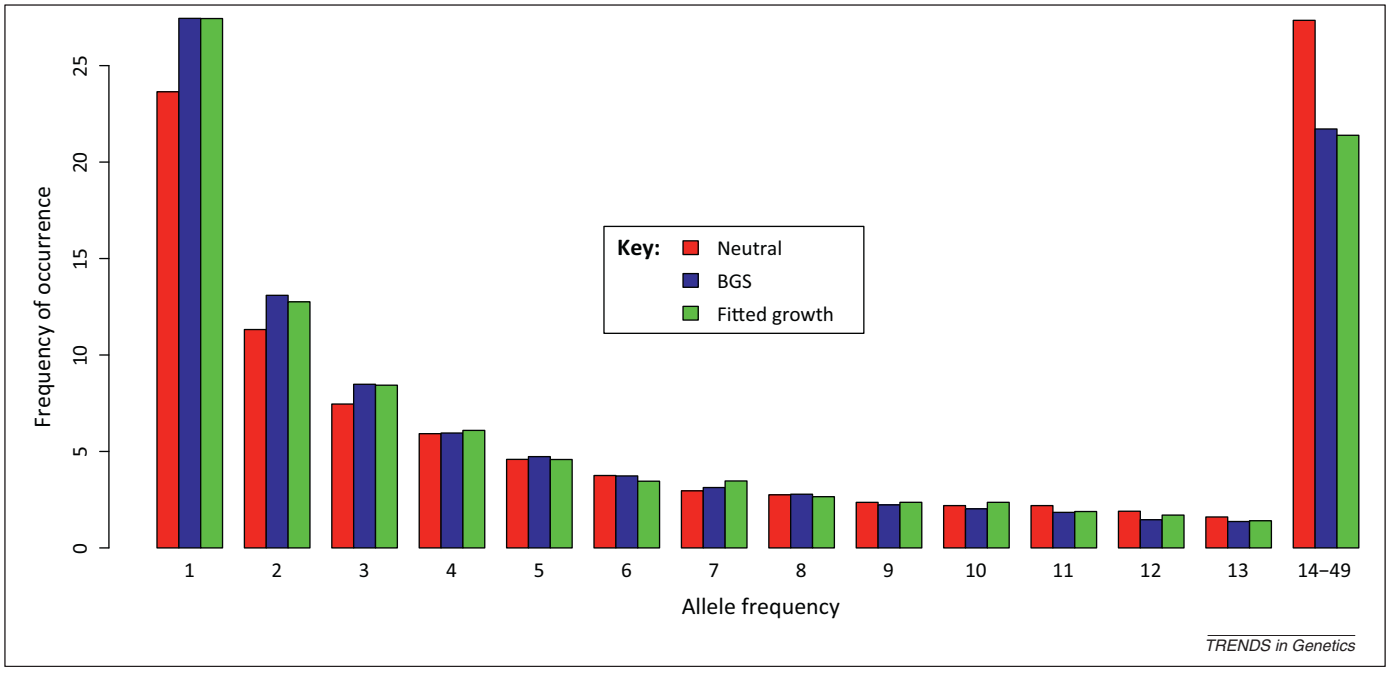

Figure 1. An example of the ability of background selection to mimic neutral nonequilibrium (here, exponential growth) site-frequency spectrum (SFS) based patterns. Assuming an absence of background selection (BGS) effects when they are present (as is common in demographic inference), may result in substantial misinference of the underlying demographic model – here inferring population growth when the population is, in fact, at equilibrium. Simulations were performed using SFSCode [40], with 50 samples from a single time point, conditioned on 100 single nucleotide polymorphisms per locus. The BGS coefficient is $\alpha = 2Ns = -4$ and the probability of a deleterious mutation is 0.1, with recombination rate $\rho = 2Nr = 50$ between chromosome ends. The single nucleotide polymorphisms (SFS) shows the frequency of occurrence (y-axis) of a number of derived alleles (x-axis) in the simulated sample. Site classes 14–49 were binned for illustrative purposes.