Theoretical Population Biology 102 (2015) 85–93

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Sensitivity of quantitative traits to mutational effects and number of loci

Joshua G. Schraiber [a,*], Michael J. Landis [b]

[a] Department of Genome Sciences, University of Washington, Seattle, WA 98195, United States
[b] Department of Integrative Biology, University of California, Berkeley, CA 94720, United States

ARTICLE INFO

Article history:
Received 19 January 2015
Available online 31 March 2015

Keywords:
Quantitative genetics
Coalescent theory
Characteristic function
Neutral theory

ABSTRACT

When models of quantitative genetic variation are built from population genetic first principles, several assumptions are often made. One of the most important assumptions is that traits are controlled by many genes of small effect. This leads to a prediction of a Gaussian trait distribution in the population, via the Central Limit Theorem. Since these biological assumptions are often unknown or untrue, we characterized how finite numbers of loci or large mutational effects can impact the sampling distribution of a quantitative trait. To do so, we developed a neutral coalescent-based framework, allowing us to gain a detailed understanding of how number of loci and the underlying mutational model impacts the distribution of a quantitative trait. Through both analytical theory and simulation we found the normality assumption was highly sensitive to the details of the mutational process, with the greatest discrepancies arising when the number of loci was small or the mutational kernel was heavy-tailed. In particular, skewed mutational effects will produce skewed trait distributions and fat-tailed mutational kernels result in multimodal sampling distributions, even for traits controlled by a large number of loci. Since selection models and robust neutral models may produce qualitatively similar sampling distributions, we advise extra caution should be taken when interpreting model-based results for poorly understood systems of quantitative traits.

1. Introduction

Questions about the distribution of traits that vary continuously in populations were critical in motivating early evolutionary biologists. The earliest studies of quantitative trait variation relied on phenomenological models, because the underlying nature of heritable variation was not yet well understood (Galton, 1883, 1889; Pearson, 1894, 1895). Despite the rediscovery of the work of Mendel (1866), researchers studying continuous variation in natural populations were initially skeptical that Mendel's laws could explain what they observed (Weldon, 1902; Pearson, 1904). These views were reconciled when Fisher (1918) showed that the observations of correlation and variation between phenotypes in natural populations could be explained by a model in which many genes made small contributions to the phenotype of an individual.

The insights of Fisher (1918) made it possible to build models of quantitative trait evolution from population genetic first principles. Early work focused primarily on the interplay between mutation and natural selection in the maintenance of quantitative genetic variation in natural populations, while typically ignoring the effects of genetic drift (Fisher, 1930; Haldane, 1954; Latter, 1960; Kimura, 1965).

However, genetic drift plays an important role in shaping variation in natural populations. While earlier work assumed that a finite number of alleles control quantitative genetic variation (e.g. Latter (1970)), Lande (1976) used the continuum-of-alleles model proposed by Kimura (1965) to model the impact of genetic drift on differentiation within and between populations. A key assumption of Lande's models is that the additive genetic variance in a trait is constant over time. In fact, in finite populations the genetic variance itself is random; at equilibrium, there are still stochastic fluctuations around the deterministic value assumed by Lande, even if none of the underlying genetic architecture changes (Bürger and Lande, 1994).

Several later papers explored more detailed models to understand how genetic variance changes through time due to the joint effects of mutation and drift (e.g. Chakraborty and Nei (1982)). Lynch and Hill (1986) undertook an extremely thorough analysis

* Corresponding author.
E-mail address: schraib@uw.edu (J.G. Schraiber).

http://dx.doi.org/10.1016/j.tpb.2015.03.005
0040-5809/

of the evolution of neutral quantitative traits. They analyzed the moments (e.g. mean and variance) of trait distributions that arise due to mutation and genetic drift and provided several quantities that can be used to interpret variation within and between species and analyze mutation accumulation experiments.

Much of this earlier work has made several simplifying assumptions about the distribution of mutational effects and the genetic architecture of the traits in question. For instance, Lynch and Hill (1986), despite analyzing quite general models of dominance and epistasis, ignored the impact of heavy tailed or skewed mutational effects. While, in many cases, such properties of the mutational effect distribution are not expected to have an impact if a large number of genes determine the phenotype in question, it is unknown what impact they may have when only a small number of genes determine the genetic architecture of the trait. Moreover, when mutational effects display "power-law" or "fat-tailed" behavior, the impact of the details of the mutational effects may persist even in the so-called infinitesimal limit of a large number of loci with small effects. Finally, mutation accumulation experiments have produced skewed and/or leptokurtic samples of quantitative traits (Mackay et al., 1992), which is a direct motivation to relax assumptions on the mutational effects distribution.

Such deviations that stem from the violations of common modeling assumptions have the potential to influence our understanding of variation in natural populations. For instance, leptokurtic trait distributions may be a signal of some kind of diversifying selection (Kopp and Hermisson, 2006) but are also possible under neutrality when the number of loci governing a trait is small. Similarly, multimodal trait distributions may reflect some kind of underlying selective process (Doebeli et al., 2007) but may also be due to rare mutations of large effect.

We have two main goals in this work. Primarily, we want to assess the impact of violations of common assumptions on properties of the sampling distribution of a quantitative trait (e.g. variance, kurtosis, modality). Secondly, we believe that the formalism that we present here can be useful in a variety of situations in quantitative trait evolution, particularly in the development of robust null models for detecting selection at microevolutionary time scales. To this end, we introduce a novel framework for computing sampling distributions of quantitative traits. Our framework builds upon the coalescent approach of Whitlock (1999), but allows us to recover the full sampling distribution, instead of merely its moments.

First, we outline the biological model and explain how we can compute quantities of interest using a formalism based on characteristic functions. We then use this approach to compute the sample central moments. While much previous work focuses on only the first two central moments (mean and variance), we are able to compute arbitrarily high central moments, which are related to properties such as skewness and kurtosis. By doing so, we are able to determine the regime in which the details of the mutational effect distribution are visible in a sample from a natural population. Additionally, we explore the convergence to the infinitesimal limit and find that when "fat-tailed" effects are present, traditional theory based on the assumption of normality can lead to misleading predictions about phenotypic variation.

2. Model

The mechanistic model we construct has two components: a coalescent process, and a genetic mutational process that acts upon the controlling quantitative trait loci by sampling effect sizes from a mutational kernel. Together, these processes generate the values of quantitative traits sampled from the study population while explicitly modeling their shared genetic ancestry. Although we opt for simple model components during this exposition, the model generally supports more realistic and complex extensions, such as population structure and epistasis.

We assume that we sample n haploid individuals from a randomly mating population of size N. Initially, we consider a trait governed by a single locus and we will later extend the theory to traits governed by multiple loci. Let μ be the mutation rate per generation at the locus, and $\theta = 2N\mu$ be the coalescent-scaled mutation rate. We model mutation as a process by which a new mutant adds an independent and identically distributed random effect to the ancestral state. Note that when the distribution of random effects is continuous, this corresponds to the Kimura (1965) continuum-of-alleles model. However, it is also possible for the effect distribution to be discrete, similar to the discrete model of Chakraborty and Nei (1982). While this model does not capture the impact of a biallelic locus with exactly two effects, the following theory could easily be modified to analyze that case.

Fig. 1 shows one realization of both the coalescent and mutational processes for a sample of size 5. Given the phenotype at the root of the tree and the locations and effects of each mutation on the tree, the phenotypes at the tips are determined by adding mutant effects from the root to tip. To specify the root, we can assume without loss of generality that the ancestral phenotype for the entire population has a value 0 (this is similar to the common assumption in quantitative genetics literature that the ancestral state at each locus can be assigned a value of 0).

This mutational process can be described as a compound Poisson process (see also Khaitovich et al. (2005b); Chaix et al. (2008); Landis et al. (2013) for compound Poisson processes in a phylogenetic context). To ensure that this paper is self contained, we briefly review relevant facts about compound Poisson processes in Appendix A.1.

In the following, we ignore the impact of non-genetic variation and focus on the breeding value of individuals, i.e. the average phenotype of an individual harboring a given set of mutations.

3. Results

3.1. Computing the characteristic function of a sample

In many analyses, the object of interest is the joint probability of the data. If we let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be the vector representing the values of the quantitative trait observed in a sample of n individuals, we denote the joint probability of the data as $p(x_1, x_2, \ldots, x_n)$. Note that, in general, X_i and X_j are correlated due to shared ancestry, and that p must be computed by integrating over all mutational histories consistent with the data. Hence, computing p directly is extremely difficult.

Instead, we compute the characteristic function of $\mathbf{X}$. For a one-dimensional random variable, X, the characteristic function is defined as $\mathbb{E}(e^{ikX})$ where i is the imaginary unit, and k is a dummy variable. Generalizing this definition to an n-dimensional random variable, we are interested in computing

$$\begin{aligned}\lambda_n(\mathbf{k}) &= \mathbb{E}(e^{i\mathbf{k}^T\mathbf{X}}) \\ &= \mathbb{E}(e^{i(k_1X_1+k_2X_2+\cdots+k_nX_n)})\end{aligned}$$

where $\mathbf{k} = (k_1, k_2, \ldots, k_n)$ is a vector of dummy variables. Like a probability density function, the characteristic function of $\mathbf{X}$ contains all the information about the distribution of $\mathbf{X}$. Moreover, computing moments of $\mathbf{X}$ is reduced to calculating derivatives of the characteristic function, which will prove useful in the following.

We calculate the characteristic function of $\mathbf{X}$ in two parts. First, we compute a recursive formula for ϕ_n, the characteristic function given that ancestral phenotype of the *sample* is equal to 0. Then, we compute ρ_n, the characteristic function of the ancestral