Theoretical Population Biology 81 (2012) 32–44

Contents lists available at SciVerse ScienceDirect

Theoretical Population Biology

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

The mapping of epistatic effects onto a genealogical tree in haploid populations

Cortland K. Griswold *, Danielle J. Eisner

Department of Integrative Biology, University of Guelph, Guelph Ontario, N1G 2W1, Canada

ARTICLE INFO

Article history:
Received 20 May 2011
Available online 12 November 2011

Keywords:
Coalescent
Epistasis
Neutral
Phenotype
Selection
Genotypic covariance

ABSTRACT

In this paper we present a model that maps epistatic effects onto a genealogical tree for a haploid population. Prior work has demonstrated that genealogical structure causes the genotypic values of individuals to covary. Our results indicate that epistasis can reduce genotypic covariance that is caused by genealogical structure. Genotypic effects (both additive and epistatic) occur along the branches of a genealogical tree, from the base of the tree to its tips. Epistasis reduces genotypic covariance because there is a reweighting of the contribution of branches to the states of genotypes compared to the additive case. Branches near the tips of a genealogical tree contribute proportionally more genetic effects with epistasis than without epistasis. Epistatic effects are most numerous at basal positions in a genealogical tree when a population is constant in size and experiencing no selection, optimizing selection, diversifying selection or directional selection, indicating that epistatic effects are typically old. For a population that is growing in size, epistatic effects are most numerous at midpoints in a genealogical tree, indicating epistatic effects are of moderate age. Our results are important in that they suggest epistatic effects may typically explain deep (old) divergences and broad patterns of divergence that exist in populations, except in growing populations. In a growing population, epistatic effects may cause more within group divergence higher up in a tree and less between group divergence that is deep in a tree. The distribution of the number of epistatic effects and the expected variance and covariance in the number of epistatic effects is also provided assuming neutrality.

1. Introduction

A primary goal of quantitative genetics is to quantify the genetic basis of phenotypic variation in a population. Fisher (1918) initiated the modern statistical approach to quantifying the genetic basis of genetic variation by partitioning phenotypic variation into additive, dominance and epistatic effects. In the same work, Fisher laid the foundation for the more general framework of analysis of variance (ANOVA). A fundamental principle of ANOVA is that a population may have hierarchical structure, which causes predictable structure in the variation of a random variable.

Genetically, reproduction causes a population to be structured hierarchically. For a sample of size n, Kingman (1982) and Tajima (1983) presented the coalescent, which predicts the hierarchical genetic structure of a sample for a single non-recombining locus that evolves neutrally. Subsequent work has incorporated multiple recombining loci (Hudson, 1983; Griffiths and Marjoram, 1997), population structure (Notohara, 1990; Herbots, 1997), and selection (Hudson and Kaplan, 1988; Krone and Neuhauser, 1997).

Coalescent approaches have focused primarily on determining the processes that shape genetic variation. Yet, in principle, the coalescent also provides a framework to understand the genetic basis of phenotypic variation. Some works that illustrate the connection between quantitative genetics and the coalescent are Whitlock (1999) who showed how Q_{ST} is a function of coalescence times and Griswold et al. (2007) who linked the structure of the additive genetic variance covariance matrix (**G**) to the coalescent structure of a population.

In this paper, we seek to link together the coalescent structure of a population and epistasis. In particular, we determine where epistatic effects are mapped onto a gene tree and how the process that shapes the gene tree in turn shapes where epistatic effects map. Our analysis begins with haploid and non-recombining populations. We initially assume neutrality and then consider selection, including optimizing, diversifying and directional forms.

Why is it worthwhile to look at where epistatic effects map onto a gene tree? One reason is that it will help us understand the extent to which differences between individuals are caused by epistatic effects. Secondly, it will provide a more exact expectation for the genetic basis of phenotypic variation in a sample. Thirdly, and perhaps most importantly it will help determine whether epistasis causes haplotypes to covary in their phenotypic states. The hierarchical structuring caused by reproduction causes haplotypes to be non-independent. Here we determine whether epistasis increases or decreases dependence caused by genealogical history.

In Section 2, we define a statistical model of epistasis. In Section 3, we indicate where the effects in the statistical model

* Corresponding author.
E-mail address: cgriswol@uoguelph.ca (C.K. Griswold).

0040-5809/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2011.11.001

map genealogically, count the number of effects in a tree and define the genotypic variance–covariance matrix in the context of epistasis. In Section 4 we present results that provide expectations for the number of epistatic effects that are expected to occur in particular parts of a genealogical tree and expectations for the eigenstructure of the genotypic covariance matrix.

2. Statistical model of epistasis

We assume that individuals are haploid and that a haplotype consists of L sites. We assume that a site has two possible states. A random sample of n haplotypes has a most recent common ancestor (MRCA) designated with the empty set {}. A haplotype that differs from the MRCA is designated by set of numbers that indicate the sites that differ from the MRCA. For instance, a haplotype (H) with mutations at sites 1, 10 and 23 relative to the MRCA is represented by the set $H = \{1, 10, 23\}$.

In principle, different haplotypes express different phenotypes. The genotypic value of a haplotype is determined by the genetic effects of mutations in the haplotype. In our model, genetic effects are defined relative to a reference haplotype, as in the work of Hansen and Wagner (2001). In the context of the coalescent, a natural reference haplotype is the MRCA of a sample. For simplicity, we assume no environmental effects and that there is no genotype-by-environment interaction. Under these assumptions, the genotypic value of a haplotype (H) is

$$G_H = \sum_{i \in W(H)} g_i \tag{1}$$

where g_i is a genetic effect. The set of genetic effects of a haplotype H is determined by its power set $W(H)$. The power set of a haplotype gives all possible combinations of mutations relative to the MRCA, including the ancestral state of the MRCA. For instance, the power set of the haplotype $\{1, 10, 23\}$ is $W(\{1, 10, 23\}) = \langle\langle\rangle, \langle 1\rangle, \langle 10\rangle, \langle 23\rangle, \langle 1, 10\rangle, \langle 1, 23\rangle, \langle 10, 23\rangle, \langle 1, 10, 23\rangle\rangle$.
These genetic effects can involve mutations that occur within the same gene or between different genes.

The genotypic value of the MRCA is $G_{\{\}} = g_{\langle\rangle}$. A haplotype $\{1\}$ has genotypic value $G_{\{1\}} = g_{\langle\rangle} + g_{\langle 1\rangle}$. The effect $g_{\langle 1\rangle}$ is given by $g_{\langle 1\rangle} = G_{\{1\}} - g_{\langle\rangle}$. For haplotype $\{1, 10\}$, its genotypic value is $G_{\{1,10\}} = g_{\langle\rangle} + g_{\langle 1\rangle} + g_{\langle 10\rangle} + g_{\langle 1,10\rangle}$, where $g_{\langle 1,10\rangle} = G_{\{1,10\}} - g_{\langle\rangle} - g_{\langle 1\rangle} - g_{\langle 10\rangle}$. Genetic effects that involve more than one mutation are epistatic. Genetics effects involving a single mutation are additive. The order of an epistatic effect corresponds to the number of mutations involved in the expressing the effect. For instance, the effect $g_{\langle 1,10,23\rangle}$ is third-order and the effect $g_{\langle 1\rangle}$ is first-order (or additive).

The number of effects g_i of order r that differ between two haplotypes (X, Y) is

$$Q_r(X, Y) = \binom{|X|}{r} + \binom{|Y|}{r} - 2\binom{|X \cap Y|}{r}. \tag{2}$$

The total number of effects g_i that differ between haplotypes is $Q(X, Y) = \sum_r Q_r(X, Y)$. It is important to recognize that $Q(X, Y)$ and $Q_r(X, Y)$ are defined with respect to the MRCA of a sample of size n. To illustrate the importance of this point, consider two haplotypes, $\{1, 3, 5\}$ and $\{1, 3\}$ that are present in a sample of size $n > 2$. These haplotypes differ by a single site, which may lead to the conclusion that only the additive effect of mutation 5 will cause a difference between haplotypes. But, in context of a sample of size $n > 2$ in which the MRCA has haplotype {}, the epistatic effects $\langle 1, 5\rangle, \langle 3, 5\rangle$ and $\langle 1, 3, 5\rangle$ may also cause differences between the two haplotypes. Overall, the distances between haplotypes $\{1, 3, 5\}$ and $\{1, 3\}$ are $Q_1 = 1, Q_2 = 2, Q_3 = 1$ and $Q = 4$. Although it may seem odd that epistatic effects are defined relative to a sample, this in fact the norm in quantitative genetics. For instance, in the classic equation for a genotypic value (e.g. Eq (5.7) in Lynch and Walsh (1998)) a locus must be polymorphic — in the sample — for it to be included as a distinct effect in the model.

2.1. Transformation to a new reference

In certain situations, it may be helpful to change the reference haplotype (Hansen and Wagner, 2001). For instance, in a model of optimizing selection a haplotype needs to be chosen as the reference to assign fitnesses to other haplotypes, but this haplotype may not be the MRCA at the time of sampling. This section illustrates two important principles related to changing the reference haplotype. First, genetic effects using the new reference are functions of genetic effects using the old reference (e.g. Hansen and Wagner, 2001). Second, we prove that the scaling of genetic effects is preserved upon a change of reference.

The genotypic values of haplotypes are independent of a reference haplotype. Consequently, the difference in genotypic values between haplotypes is also independent of a reference haplotype (Hansen and Wagner, 2001). This invariance property is helpful when understanding how genetic effects in one reference are transformed to another reference. For two haplotypes, H_1 and H_2 the difference in genotypic values is

$$G_{H_1} - G_{H_2} = \sum_{i \in W(H_1)} g_i - \sum_{i \in W(H_2)} g_i. \tag{3}$$

In the expression above, both $\sum_{i \in W(H_1)} g_i$ and $\sum_{i \in W(H_2)} g_i$ have the term $g_{\langle\rangle}$, which is canceled out with subtraction. Denote $W(H_i) - g_{\langle\rangle}$ to be the power set of H_i minus the element $g_{\langle\rangle}$ then expression (3) is equivalent to

$$G_{H_1} = G_{H_2} + \sum_{i \in W(H_1) - g_{\langle\rangle}} g_i - \sum_{i \in W(H_2) - g_{\langle\rangle}} g_i. \tag{4}$$

Expression (4) has the same form as Eq. (1) with haplotype H_2 being the new reference haplotype, but genetic effects measured relative to the original reference. Next, match up terms g_i' in the expression

$$G_{H_1} = \sum_{i \in W_{H_2}(H_1) - B(H_1, H_2)} g_i', \tag{5}$$

with terms g_i from Eq. (4), where the genetic effects g_i' are based on using haplotype H_2 as the new reference, noting that $g_{\langle\rangle}' = G_{H_2}$. In Eq. (5), $W_{H_2}(H_1)$ indicates the power set of the haplotype H_1 using H_2 as a reference and $B(H_1, H_2)$ is the set of interactions between back mutations and interactions between back mutations and forward mutations that occur between haplotypes H_1 and H_2.

Forward and back mutations in this context can be confusing because there are two layers of reference. The base reference is the original haplotype that was used as a reference. For instance, haplotypes $H_1 = \{3, 5, 6\}$ and $H_2 = \{1, 3, 5\}$ are measured relative to the original reference such that there are mutations at sites 3, 5, and 6 in H_1 and mutations at sites 1, 3, and 5 in H_2, relative to the original reference. Switching to H_2 as the new reference, there is a back mutation at site 1 in H_1 because that site now takes on the state of the original reference and a forward mutation at site 6 in H_1 because that site takes on a different state relative to the original reference. A forward mutation is a change in state relative to the original reference genotype, and a back mutation is a change in state of a haplotype that brings it back to the state of the original reference. For haplotypes $H_1 = \{3, 5, 6\}$ and $H_2 = \{1, 3, 5\}$, $B(H_1, H_2)$ involves the interaction of the back mutation at site 1 and the forward mutation at site 6 ($B(H_1, H_2) = \langle\langle 1, 6\rangle\rangle$).

If A is the set of mutations separating haplotype H_2 from the original reference then the appropriate matching is

$$\kappa_i g_i' = \sum_{j \in \{i, i \times A\}} g_j,$$

where $\{i, i \times A\}$ is a set that consists of all unique pairings of element i with elements in A ($i \times A$) unioned with the element i