Theoretical Population Biology 88 (2013) 57–67

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Correlation between relatives given complete genotypes: From identity by descent to identity by function

Serge Sverdlov *, Elizabeth A. Thompson

Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195, USA

ARTICLE INFO

Article history:
Received 27 December 2012
Available online 11 July 2013

Keywords:
Identity by descent
Identity by function
Whole genome prediction
Functional quantitative genetics
Dominance
Missing Heritability

ABSTRACT

In classical quantitative genetics, the correlation between the phenotypes of individuals with unknown genotypes and a known pedigree relationship is expressed in terms of probabilities of IBD states. In existing approaches to the inverse problem where genotypes are observed but pedigree relationships are not, dependence between phenotypes is either modeled as Bayesian uncertainty or mapped to an IBD model via inferred relatedness parameters. Neither approach yields a relationship between genotypic similarity and phenotypic similarity with a probabilistic interpretation corresponding to a generative model. We introduce a generative model for diploid allele effect based on the classic infinite allele mutation process. This approach motivates the concept of IBF (Identity by Function). The phenotypic covariance between two individuals given their diploid genotypes is expressed in terms of functional identity states. The IBF parameters define a genetic architecture for a trait without reference to specific alleles or population. Given full genome sequences, we treat a gene-scale functional region, rather than a SNP, as a QTL, modeling patterns of dominance for multiple alleles. Applications demonstrated by simulation include phenotype and effect prediction and association, and estimation of heritability and classical variance components. A simulation case study of the Missing Heritability problem illustrates a decomposition of heritability under the IBF framework into Explained and Unexplained components.

1. Introduction

1.1. Correlation between pedigree and genotype relatives

Quantitative genetic models can be described as explaining phenotypic resemblance between individuals on the basis of genetic resemblance, that is, the expected or actual sharing of genes. As we transition from pedigree information to full-sequence genetic information, we change the way we pose a typical question in quantitative genetics, inviting answers with conceptually incompatible probabilistic interpretations:

1. What is the correlation between the heights of an uncle and a nephew?
2. What is the correlation between the heights of individual A_1 with genotype $\mathbf{g}_1$ and individual A_2 with genotype $\mathbf{g}_2$?

The first question has a well-studied interpretation. We can sample n nominally unrelated uncle–nephew pairs from the population and compute their pairwise correlation. The second question only becomes meaningful in the context of models where the set of genotypic effects, or the function mapping genotype to expected phenotype, is treated as random. The meaning of the correlation is not reducible to an ideal sample from the population, and is model dependent. The two major model classes have two distinct interpretations: the purely Bayesian, reflected in the genomic selection (Meuwissen et al., 2001) and reproducing kernel Hilbert spaces (Gianola and van Kaam, 2008) frameworks, and an interpretation based on the inference of the parameters of classical theory from genomic data (Visscher, 2009).

1.2. The three model classes

We can thus identify three distinct classes of quantitative genetic models.

1. Classical theory answers the first question. It predicts the expected fraction of genome shared between two individuals under the principle of Identity by Descent (IBD), based on their pedigree relationship, and maps this measure of genetic similarity to a phenotypic correlation.
2. Several classes of Bayesian, or random effect, models answer the second question on the principle of Identity by State (IBS). The correlation is a function of the similarity between the actual genotypes $\mathbf{g}_1$ and $\mathbf{g}_2$, without reference to pedigrees and descent.

* Corresponding author.
E-mail addresses: serges@uw.edu, sverdlov@gmail.com (S. Sverdlov).

0040-5809/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.tpb.2013.06.004

3. Hybrid models answer the second question by inferring relatedness parameters between A_1 and A_2 from genotype data, and substituting these into the equations of classical theory. This admits neither a pure classical, nor a pure Bayesian interpretation.

1.3. Our alternative model

Our aim is to construct a model within which the second question can be meaningfully asked with a probabilistic interpretation in the non-Bayesian context of a population genetic process based on the classical infinite allele model. The model is generative in the sense of describing a joint distribution of genotypes and effects, given parameters describing a population process and trait architecture.

We adapt the genetic architecture from Fisher (1918), additive across loci but not within a locus, and treat alleles symmetrically, from a neutralist perspective, as draws from an infinite allele process. It is the need to reconcile this symmetry with diploidy and dominance that makes the model mathematically nontrivial. The result is a definition of correlation between two individuals' phenotypes given their genotypes (question 2) with respect to a population process. The thought experiment with respect to which we define our correlation is a replay of evolution. Classically, correlation is defined with respect to a repetition of the same pedigree relationship. For our model, it is the draw of functionally different alleles from the infinite allele process at the same mutation events that gave rise to our observed genotypes.

The model allows for infinitely many alleles at each locus. This is useful for full sequence genomic data, where we routinely encounter novel alleles. The model's genetic architecture parameters specify the relative importance of loci and the pattern of dominance within a locus. These parameters are features of the mutation process that creates alleles, not of individual alleles. They are specific to a trait, but not to a particular population.

1.4. Quantitative trait locus scale and linkage disequilibrium

Our model treats a functional region, such as a gene, as a quantitative trait locus (QTL). Each (non-synonymous) sequence in that region found in the population is an allele. This contrasts with treating each minimal length unit of genomic variation, such as a SNP or indel, as a QTL, as is the implicit practice in the GWAS or whole exome (e.g. Kiezun et al., 2012) literature, and has analogies to region-based association models such as SKAT (Wu et al., 2011). In earlier QTL mapping methods based on sparse markers and linkage analysis (e.g. Lynch and Walsh, 1998, Chapters 14–16), the QTL was treated as a point along the chromosome and did not need to be defined in sequence terms. Likewise, in the literature related to the genomic relatedness matrix (e.g. Yang et al., 2010), SNP's used to compute kinship are described as markers in linkage disequilibrium with point causal variants. For complete genome or dense SNP (e.g. 1 million SNP's for the human genome, approaching 50 per gene) genotype data, the point abstraction runs into difficulties. Marker SNP's may be inside exons and causal, and multiple markers may occur within the same causal gene. The question of whether these multiple markers should be treated as separate QTL's cannot be avoided. In the GWAS context, it is typical for multiple SNP's in or near a gene to be associated with a trait. Methods extending GWAS approaches to polygenic traits (e.g. Guan and Stephens, 2011) treat this as a variable selection problem, as though only a single SNP per functional region is causal, and the other SNP's in LD with the causal SNP are a nuisance and a source of collinearity. This solution is practical, but restricts the biological effects that can be modeled at gene scale.

We leave outside the scope of this paper the question of defining the boundaries of a biological functional region, and the complication of overlapping genes; we will use the terms gene and functional region interchangeably. From a gene-as-QTL perspective, every new mutation is likely to create a novel gene-scale haplotype, implying an infinite allele population model. From a SNP-as-QTL perspective, every new mutation is likely to target a new SNP at a different base pair location, implying a diallelic, infinite sites model. In SNP-as-QTL, multiple nearby SNP's that occur together in non-random patterns to form functional sequences must be accounted for as linkage disequilibrium (LD), and their action should be expected to exhibit epistasis. The simplest consequence of switching from SNP-as-QTL to gene-as-QTL is reducing the importance of short-range (within-gene) LD and epistasis, introducing instead greater scope for dominance phenomena through the interaction of many alleles. Consider an extreme example, two SNPs within one codon: $[C/A]G[A/T]$. The ancestral sequence *CGA* codes for Arginine, as does either individual SNP mutation, *CGT* or *AGA*. However, the two SNP's together, *AGT*, code for Serine. If we treat the two SNP's as separate QTL's, we would expect both LD and epistasis in any trait affected by this sequence. If, alternatively, we treat the whole codon as a single QTL, there are simply four possible alleles with different frequencies and effects.

The ideal conditions of the classical (Cockerham, 1954; Kempthorne, 1954) decomposition of genetic covariance into variance component terms include an unlinked set of quantitative trait loci, linkage equilibrium, and the absence of mutation and selection. To reconcile quantitative genetic models with genomic data, we must define QTL on a scale which best fits the classical approximation. Unlike the gene-as-QTL approach, the SNP-as-QTL approach conflicts with the ideal conditions by requiring either the violation of linkage equilibrium or the introduction of mutation or selection. When two SNP-scale loci are part of the same functional region, it is biologically plausible that they are in LD. LD between linked, nearby loci decays under random mating due to recombination; but recombination *within* a functional region creates new functional haplotypes (mutation) with varying phenotypic effects (potential selection). In the gene-as-QTL approach, between-gene LD may still occur; but within-gene LD is a phenomenon of the choice of QTL scale, and need not be an issue in full sequence analysis.

2. Existing quantitative genetic models

For all of the three existing model classes we consider, we can write the genotypic value as $G = \mathbf{g} \cdot \mathbf{f}$ for a suitably encoded genotype vector $\mathbf{g}$ and effect vector $\mathbf{f}$. For the commonly used additive model over diallelic SNP's, the genotype vector is a vector of minor allele dosages (0, 1, or 2) for each SNP locus, and the effect the additive contribution to the trait per allele. For a more general model, the genotype vector may contain a 0 or 1 as an indicator for the genotype containing an arbitrarily complex combination of alleles, with the most general possible model including a separate indicator column for each possible genotype.

2.1. Type 1: Fisher's classical polygenic model and identity by descent

In classical quantitative genetics as originated by Fisher (1918) and presented in current form e.g. by Lynch and Walsh (1998), the non-environmental component of randomness is due to the unknown genotypes of the two individuals. Thus genotypes $\mathbf{g}$ are random, and the effects $\mathbf{f}$ are fixed.

The genotypes are modeled as being sampled from the population, with covariance between relatives due to the sharing of genes, as expressed by the probabilities of the identity by descent (IBD)