Published in final edited form as:

Theor Popul Biol. 2013 September ; 88: . doi:10.1016/j.tpb.2013.06.004.

Correlation between Relatives given Complete Genotypes: from Identity by Descent to Identity by Function

Serge Sverdlov and Elizabeth A. Thompson

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

Serge Sverdlov: serges@uw.edu

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.

Keywords

Identity by Descent; Identity by Function; Whole Genome Prediction; Functional Quantitative Genetics; Dominance; Missing Heritability

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?

© 2013 Elsevier Inc. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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.
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) states. If we were given genotypes of the two individuals in this setting, no source of randomness would remain, and the conditional covariance would be zero. This does not imply that given complete genotype information we know the genotypic values of the two individuals, but rather that we can express them in terms of (fixed) model parameters with no random variables; the number of such parameters in a general polygenic model may well be too large to even consider estimating.

2.2. Type 2: Bayesian, Random Effect, and Identity by State Models

In several classes of models, genotypes $\mathbf{g}$ are fixed, and the effects $\mathbf{f}$ are random.

The reproducing kernel Hilbert space framework described by Gianola and van Kaam (2008) is a general form for a broad class of such models, which may include dominance and epistatic interaction terms. In this framework, the covariance between trait values is explicitly a function of the similarity between genotypes, expressed e.g. as a function of the norm of the difference between the SNP dosage vectors. The description given by de los Campos et al. (2009) argues for judging such covariance structures in terms of their predictive power.

The class of similarity functions is broad, but inherently requires the measurement of similarities between diploid genomes. Thus it does not include e.g. the class of models of the divergence between genome sequences on an evolutionary scale as measured by the time needed for DNA or protein mutation (e.g. Felsenstein, 2004, Chapter 11), or local distance measures used in sequence alignment applications.

In the class of genomic selection models (Meuwissen et al., 2001), a prior on each component of $\mathbf{f}$ is imposed. The prior may place a finite weight on 0, may be heavy tailed, and may be hierarchical, with hyperparameters shared by the components of $\mathbf{f}$. The family of analogous models with different priors and computational methods, recently reviewed by Kärkkäinen and Sillanpää (2012), has become known as the Bayesian alphabet (Gianola et al., 2009).

In such a model, the phenotypes of two individuals given their genotypes have a joint distribution, and therefore a covariance. The interpretation of this distribution is a Bayesian prior, our uncertainty about the effects.

2.3. Type 3: Hybrid and Realized Relationship Models

In these models, genotype data is used to infer relatedness parameters. Then, for the purposes of variance component and heritability estimation or prediction, we treat these as the true parameter values, i.e. probabilities of IBD gene sharing. Such models naturally emerged from a historical context in light of classical theory. “Population and quantitative genetics theory is built with parameters that describe relatedness, and the estimation of these parameters from genetic markers enables progress in fields as disparate as plant breeding, human disease gene mapping and forensic science.” (Weir et al., 2006) With the availability of limited genetic marker data, it became possible to infer relatedness parameters from genotype, when complete pedigrees were unavailable (Thompson, 1975). As we have moved from sparse to dense marker maps, it was noticed that inferred relatedness parameters could outperform pedigree based ones in practice (Hayes et al., 2009), by capturing the actual Mendelian variability within classes of relatives (Hill and Weir, 2011).

The estimated fraction of genome shared by descent is inferred from marker data, either by identifying individual segments passed down without recombination (Browning and Browning, 2012) or from the realized relationship matrix (Hayes et al., 2009). This is a probabilistic inference based, crucially, not only on genotype data, but on population data, such as allele frequencies and assumptions about linkage disequilibrium.

Conditioning on the inference of true relatedness parameters, a type 3 model is a type 1 model. That is, covariance between two individuals’ phenotypes given their genotypes is the classical covariance, given the inferred relatedness. Regardless of the approach used to infer kinship, we can also express a type 3 model as a type 2 model. The type 1 covariance is a function of the relatedness parameters, and the inferred relatedness parameters are a function of genotype vectors. Substituting in the inferred relatedness parameters, we obtain covariances as a function of genotype vectors, corresponding to a type 2 model.

2.4. Population Neutrality

A reference population is needed by classical (type 1) models to specify the probabilities of the random founder genotypes. In a model where genotypes are given and not random, a reference population is not needed. A property of an individual, or a biological property of a single allele in isolation, is *population neutral*. We define population neutrality as the formal invariance of an estimator to the properties of a population, such as allele frequencies and models of linkage disequilibrium. This characterizes the estimator, not its distribution. Its distribution in a population will vary with the distribution of alleles in that population.

The effect of substitution of specific allele *A* for specific allele *B*, in a model without epistasis, is population neutral. The average effect of substituting allele *A* for a random allele (drawn from a population) is not population neutral. The prediction of phenotype from genotype is population neutral; the prediction of a breeding value is not. Under the assumption of the additive model, classical additive effects (and therefore breeding values) do not algebraically depend on the allele frequencies, leading to breeding value estimation methods that do not explicitly refer to the reference population. In the general case, however, the breeding value may be defined as the best predictor of the offspring trait value when mated with a randomly chosen individual from the reference population; such a definition makes the population dependence explicit. Likewise, the heritability of a trait is always dependent on the reference population, and not population neutral.

In the Bayesian (type 2) context, it is difficult to motivate the idea that the prior for the biological effect of an allele would depend on population frequency of that allele. This is clear from a neutralist perspective. Even from a selectionist perspective, such a dependence would only arise from an explicit model of selective forces.

Type 2 models do not have a reference population, but type 3 models do. The type 3 methods explicitly model the fraction of genome shared IBD between the two individuals. IBD is defined with respect to a particular pedigree, or with respect to a founder generation (Thompson, 2013). By choosing extreme pedigrees, the complete population history of the species, or the pedigree containing no relationships, we find IBD fractions of 1 or 0, respectively, for any pair of chromosomes or individuals. The inferred kinship is an estimate of the realized or actual fraction of IBD genome (Cockerham and Weir, 1983; Hill and Weir, 2011) whereas the classical kinship coefficient is the expectation of this fraction. An alternative definition of kinship based on IBD due to Powell et al. (2010) uses the concept of the current reference population, and defines kinship coefficients that may be negative, and thus cannot be interpreted classically as IBD probabilities. In either case, an IBD based inference must refer to a model for a reference population, and is dependent on at least such parameters as allele frequencies.

3. Functional Effect Distribution Model

3.1. Overview

By analogy with genome shared identical by descent (IBD) or by state (IBS) we introduce the concept of *Identity by Function (IBF)* and the *Functional Identity States*, S_0 through S_4 . At a single locus, two individuals are functionally identical if they share a homozygous or heterozygous genotype, AA:AA (S_4) or AB:AB (S_3). They have partial functional identity if they share one identical allele, AA:AB (S_2) or AB:AC (S_1). In all other states (S_0), no alleles are shared between the two individuals. Intuitively we would expect more phenotypic resemblance, the more similarity there is between genotypes, e.g. the more loci are functionally identical.

In practice, we must define locus and identity in genomic sequence terms. A working definition is that a locus is the region surrounding an annotated gene including some fixed upstream and downstream flanking sequences. Identity, given phased full sequence genotypes, is sequence identity excluding known synonymous variation.

As illustrated in Figure 1, the five functional identity states are a different grouping of the 15 partitions of 4 chromosomes. In IBD theory, the 15 IBD states (Thompson, 1974) are partitioned into the 9 Jacquard coefficient (Jacquard, 1972) classes, of which two pairs are mirror images, making 7 classes. Of these, 3 classes do not involve allele sharing between the individuals, and are grouped into one covariance class for our purposes.

3.2. Single Locus, Infinite Allele Model

Consider the standard neutral infinite allele model (e.g., Kimura and Crow, 1964). At each mutation event, we draw a new allele from $\mathcal{A}$, the set of all possible alleles at a locus. Mutations are assumed to be irreversible. When we observe the allele states in the current generation, we can distinguish identical and distinct alleles; but we do not know what mutation event produced a given allele, nor the relationship between allele and phenotype.

Each creation of a novel allele by a new mutation (or, more restrictively, by a non-synonymous mutation) represents an independent draw from the potential allele distribution. This model explicitly rejects the use of information about sequence similarity or mutation distance. There is no sense in which two alleles separated by a single point mutation are

more similar than two alleles that, after many mutations, no longer resemble one another in sequence; two alleles are either identical in function, or different. Alleles are abstract objects, members of set $\mathcal{A}$, and do not correspond directly to real-valued quantities associated with phenotypes. Instead, we introduce a genotypic value mapping function, which turns the allele into a quantitative effect associated with a particular trait.

For a haploid organism with a trait controlled by a single locus, the genotypic trait value depends on a single allele. The genotypic value is then a function of the allele. If alleles are sampled independently from an arbitrary distribution, so are the corresponding genotypic values. We define the haploid genotypic value function $G(\cdot)$ as a mapping $\mathcal{A} \rightarrow \mathbb{R}$. Then the real-valued random variables $G_i = G(A_i)$ are iid. Assuming all required moments exist, $E[G_i] = \mu_G$, $\text{Var}[G_i] = \sigma_G^2$, and $\text{Cov}[G_i, G_j] = 0$. In this simple situation we can think of a genotypic value associated with each allele, and independently generated by each mutation event.

3.3. Diploid Model

For a diploid organism with a trait controlled by a single locus, the genotypic trait value depends on two alleles, in interchangeable order. Define the diploid genotypic value function $G(\cdot, \cdot)$ as a mapping $(\mathcal{A}, \mathcal{A}) \rightarrow \mathbb{R}$, subject to the symmetry constraint $G(A, B) = G(B, A)$. Likewise, define real-valued random variables $G_{ij} = G(A_i, A_j)$. The G_{ij} are no longer all independent; they are only independent in the case $G_{ij} = G_{kl}$ if $\{i, j\} \cap \{k, l\} = \emptyset$. Otherwise there is a number of patterns of dependence we can enumerate. By symmetry, the expectations and variances of all homozygotes and all heterozygotes are separately equal, $E[G_{ii}] = \mu_{AA}$ and $E[G_{ij}] = \mu_{AB}$, and $\text{Var}[G_{ii}] = \sigma_{AA}^2$ and $\text{Var}[G_{ij}] = \sigma_{AB}^2$.

$$\text{Cov}[G_{ii}, G_{ij}] = \rho_3 \sigma_{AA} \sigma_{AB} \quad (1)$$

$$\text{Cov}[G_{ij}, G_{ik}] = \rho_2 \sigma_{AB}^2 \quad (2)$$

$$\text{Cov}[G_{ii}, G_{jk}] = \text{Cov}[G_{ii}, G_{jj}] = \text{Cov}[G_{ij}, G_{kl}] = 0 \quad (3)$$

The five covariances correspond to the five functional identity states, using the compact notation $AB = G(A, B)$, etc.:

S_0 : No shared alleles: $C_0 = \text{Cov}(AB, CD) = \text{Cov}(AB, CC) = \text{Cov}(AA, BB) = 0$

S_1 : Sharing 2 alleles: $C_1 = \text{Cov}(AB, AC) = \rho_2 \sigma_{AB}^2$

S_2 : Sharing 3 alleles: $C_2 = \text{Cov}(AA, AB) = \rho_3 \sigma_{AB} \sigma_{AA}$

S_3 : Full heterozygous functional identity: $C_3 = \text{Cov}(AB, AB) = \sigma_{AB}^2$

S_4 : Full homozygous functional identity: $C_4 = \text{Cov}(AA, AA) = \sigma_{AA}^2$

Not every choice of parameters $\rho_3, \rho_2, \sigma_{AA}, \sigma_{AB}, \mu_{AA}, \mu_{AB}$ corresponds to a valid joint distribution for the effect vector $[G_{ii}, G_{ij}]$. Assuming second moments exist, the requirement is that the covariance matrix for the effects be positive semidefinite. We derive the positive semidefinite constraint

$$\rho_3^2 \leq \rho_2 \leq \frac{1}{2} \quad (4)$$

in Appendix A. The constraint is necessary for the existence of an infinite allele process; that is, for the existence of a joint distribution for the effect vector for an arbitrarily large number k of alleles. It is also sufficient because the proof in Appendix A gives a covariance matrix from which we can construct a multivariate normal distribution for the effect vector for k alleles.

A special case of the diploid model is the *additive model*. We may associate an additive effect with each allele, $C(\cdot) : \mathcal{A} \rightarrow \mathbb{R}$, and define the genotypic value function $G(A_i, A_j) = C(A_i) + C(A_j)$. Then, by the derivation in Appendix B,

$$\mu_{AA} = \mu_{AB} = 2\mu_C \quad (5)$$

$$\rho_2 = 1/2; \quad \rho_3 = 1/\sqrt{2} \quad (6)$$

$$\sigma_{AA} = 2\sigma_C; \quad \sigma_{AB} = \sqrt{2}\sigma_C; \quad \sigma_{AB}/\sigma_{AA} = 1/\sqrt{2} \quad (7)$$

3.4. Generative Models for Genotypic Values

3.4.1. Multivariate Normal—For any choice of μ_{AA} , μ_{AB} , σ_{AA}^2 , σ_{AB}^2 , ρ_2 , and ρ_3 satisfying the positive semidefinite constraint (4), there exists a multivariate normal distribution for the genotypic values. We summarize some of the properties of such distributions.

Using the covariance matrix expression in Appendix A, we can derive the conditional distribution of a heterozygote's genotypic value, conditional on its two homozygotes. The conditional expectation can be decomposed into the mid-homozygote value, a heterozygosity bias, and an effect specific to ρ_3 :

$$E[AB|AA, BB] = \frac{AA+BB}{2} + (\mu_{AB} - \mu_{AA}) + \left(\frac{\sigma_{AB}}{\sigma_{AA}} \rho_3 - \frac{1}{2} \right) (AA - \mu_{AA} + BB - \mu_{AA}) \quad (8)$$

$$\text{Var}[AB|AA, BB] = \sigma_{AB}^2 (1 - 2\rho_3^2) \quad (9)$$

For the purely additive model $AB = (AA + BB)/2$, when $\mu_{AB} = \mu_{AA}$ and when $\sigma_{AB}/\sigma_{AA} = \rho_3 = 1/\sqrt{2}$, the conditional variance for the heterozygote is 0, and the conditional mean is the simple midpoint between the two homozygotes.

Through the derivation in Appendix C, assuming $\mu_{AB} = \mu_{AA}$ we can compute the probability of overdominance, that is, the probability that for two distinct alleles A and B , some ordering of genotypic values other than $AA < AB < BB$ or $AA > AB > BB$ prevails:

$$P_{OD} = \frac{2}{\pi} \tan^{-1} \sqrt{1 - 4\rho_3 \frac{\sigma_{AB}}{\sigma_{AA}} + 2 \frac{\sigma_{AB}^2}{\sigma_{AA}^2}} \quad (10)$$

$P_{OD} = 0$ in the additive case, and positive otherwise.

3.4.2. Effect and Affinity Model—Multivariate normality implies the possibility of overdominance, and requires a marginal normal distribution. A second class of generative models compatible with this framework associates with each allele 1) its homozygous effect, and 2) random values unobservable in the homozygote, that govern how heterozygous effects are formed by merging two homozygous effects. This allows arbitrary marginal distributions to be imposed directly for homozygotes, and indirectly for heterozygotes, and the explicit construction of complete or partial dominance orders.

For each allele A_i , we construct two random variables, the homozygote value G_{ii} and the strength or affinity S_i . These are independent for different i , but G_{ii} and S_i need not be independent. We define the heterozygote value as a weighted blend of the homozygote values,

$$G_{ij} = G_{ii} w(S_i, S_j) + G_{jj} w(S_j, S_i) \quad (11)$$

There is no possibility of overdominance in this model unless w is allowed to exceed 1. A partial dominance model can be constructed by a “soft cutoff” weight function,

$$w_\sigma(A, B) = \frac{1}{1 + e^{\frac{B-A}{\sigma}}} \quad \text{or} \quad w_0(A, B) = \begin{cases} 1, & A > B \\ 1/2, & A = B \\ 0, & B < A \end{cases} \quad (12)$$

In the limit as $\sigma \rightarrow 0$, this becomes a hard cutoff corresponding to strict dominance.

The simplest model of dominance is dominance with respect to the trait itself; that is, $G_{ij} = S_i$ in which case a heterozygote has the genotypic value of the higher of its homozygotes. Another choice is the random dominance order, where S_i is a continuous random variable independent of G_{ii} . An intermediate model would let S_i be correlated with G_{ii} but not be deterministically related.

The random dominance order case has a closed form solution regardless of distribution. By the derivation in Appendix D,

$$\mu_{AA} = \mu_{AB}; \quad \sigma_{AB} = \sigma_{AA} \quad (13)$$

$$\rho_2 = 1/3; \quad \rho_3 = 1/2 \quad (14)$$

Other cases do not yield closed form solutions, but we can compute ρ_2 and ρ_3 by simulation. In Figure 2, we simulate the strict dominance model with trait dominance order for several distributions. Location and scale are irrelevant for the calculation of these correlations; the distributions shown are the uniform, normal, exponential, the double exponential (Laplace), and the Bernoulli ($p = 0.5$). For the normal distribution, we simulate different values of the partial dominance parameter α generate S_i and G_{ii} as jointly normal with different choices of correlation α . The choice of $\alpha = 0$ and $\alpha = 1$ is the random strict dominance order model, and increasing the parameter α from strict to partial dominance attracts toward the additive model. Other parameter choices lie approximately on the line between those two models.

4. Multiple Loci

4.1. Multiple loci and covariance between two individuals

Suppose the genome consists of L loci. The trait's genotypic value is a sum of contributions from the L loci, each following an infinite allele process with its own μ_{AA} , μ_{AB} , σ_{AA}^2 , σ_{AB}^2 , ρ_2 , and ρ_3 , designated $\mu_{AA(l)}$ etc. This is a genetic architecture additive across loci, but not within a locus.

We introduce the strong assumption of independence of effects across loci. This is not an assumption about linkage equilibrium, but about the mutation processes that generate alleles at each locus. For the sake of model completeness, we may assume unlinked Mendelian segregation. Extending the model to a linked pattern of inheritance, so long as inheritance is independent of trait values, does not affect the problem, since in the end we are conditioning on final allele state.

At a given locus, two individuals are in one of the identity states defined above, S_0 through S_4 . We can express the relationship between the two individuals through a vector $S(l)$, where $S(l) \in \{0, 1, 2, 3, 4\}$ gives the identity state at locus l . The covariance between phenotypes P_A and P_B of the two individuals can be written in terms of $S(l)$, without reference to specific alleles:

$$\text{Cov}(P_A, P_B) = \sum_{l=1}^L C_{S(l)} \quad (15)$$

Two individuals that share an identical genotype, whether it is homozygote or heterozygote (S_4 or S_3) experience the same contribution from that locus to the trait, and a correlation of 1. Individuals sharing no alleles (S_0) have independent contributions to the trait from the locus, and a correlation of 0. It is also possible to share two (S_1) or three (S_2) alleles, with corresponding effect correlations of ρ_2 and ρ_3 respectively.

4.2. Variance Component Representation

For a single locus, assuming an environmental error term with variance σ_E^2 , we can write the covariance of a vector of individuals' trait values,

$$\text{Cov}(\mathbf{P}) = C_4 \mathbf{I}_4 + C_3 \mathbf{I}_3 + C_2 \mathbf{I}_2 + C_1 \mathbf{I}_1 + \sigma_E^2 \mathbf{I} \quad (16)$$

where $C_4 = \sigma_{AA}^2$, $C_3 = \sigma_{AB}^2$, $C_2 = \rho_3 \sigma_{AB} \sigma_{AA}$, $C_1 = \rho_2 \sigma_{AB}^2$; $\mathbf{I}_k$ is a matrix whose entry ij is an indicator, 1 when individuals i and j are in identity state k and 0 otherwise. We can write the positive semidefinite constraint (4) in terms of the variance components,

$$\rho_2 = \frac{C_1}{C_3}; \quad \rho_3 = \frac{C_2}{\sqrt{C_4 C_3}} \quad (17)$$

$$2C_1 \leq C_3; \quad 2C_2^2 \leq 2C_4 C_1 \quad (18)$$

Both the additive model (5) and the strict dominance random ordering model (13) constrain the relationship between the four variance components down to one parameter. In the additive model we can solve

$$\frac{C_1}{C_3} = \rho_2 = \frac{1}{2}; \quad \frac{C_2}{\sqrt{C_4 C_3}} = \rho_3 = \frac{1}{\sqrt{2}}; \quad \frac{C_3}{C_4} = \frac{\sigma_{AB}^2}{\sigma_{AA}^2} = \frac{1}{2} \quad (19)$$

$$[C_1, C_2, C_3, C_4] \propto \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{2}, 1 \right] \quad (20)$$

and equivalently, in the strict dominance random ordering model:

$$\frac{C_1}{C_3} = \rho_2 = \frac{1}{3}; \quad \frac{C_2}{\sqrt{C_4 C_3}} = \rho_3 = \frac{1}{2}; \quad \frac{C_3}{C_4} = \frac{\sigma_{AB}^2}{\sigma_{AA}^2} = 1 \quad (21)$$

$$[C_1, C_2, C_3, C_4] \propto \left[\frac{1}{2}, \frac{1}{\sqrt{2}}, 1, 1 \right] \quad (22)$$

If the distribution of the effects G_{ij} is multivariate normal, so is that of $\mathbf{P}$. The posterior distributions of G_{ij} conditional on observed $\mathbf{P}$ is then also normal and expressible in matrix terms. Any such inferences are, of course, also conditional on known IBF parameters μ_{AA} , μ_{AB} , σ_{AA}^2 , σ_{AB}^2 , 2, and 3.

In the multi-locus case, the number of variance component parameters required to specify the model is very large:

$$\text{Cov}(\mathbf{P}) = \sigma_E^2 \mathbf{I} + \sum_{l=1}^L C_{4(l)} \mathbf{I}_{4(l)} + C_{3(l)} \mathbf{I}_{3(l)} + C_{2(l)} \mathbf{I}_{2(l)} + C_{1(l)} \mathbf{I}_{1(l)} \quad (23)$$

at each locus, the variances $C_{4(l)}$ and $C_{3(l)}$ must be nonnegative, and the positive semidefinite constraint must be satisfied. Estimation of these parameters from data poses special challenges. Under the restrictive assumption of equal IBF parameters for each locus, we can aggregate the identity state matrices over loci, and thus estimate only four variance component (plus σ_E^2). In this case the inference problem is a conventional mixed model, though with a nonlinear positive semidefinite constraint (4).

The most general inference problem is that in which IBF parameters are allowed to vary by locus. This is a mixed model with a number of variance components proportional to the number of loci, and quite possibly larger than the number of observations in $\mathbf{P}$. In the genomic selection literature surveyed by Kärkkäinen and Sillanpää (2012), analogous sparse regression problems are addressed by imposing single or hierarchical priors on weight parameters and finding point estimates via MAP (maximum a posteriori) methods or posterior distribution estimates via MCMC methods. In the machine learning literature, the work of Tipping (2001) and Faul and Tipping (2001) on what in statistical terminology would be called random coefficient regression proposes an algorithm for large scale local optimization of models with large numbers of variance components.

4.3. Locus Weight and Identity State Contribution

Our method expresses the covariance between two individuals, given the IBF parameters, as sum of contributions by identity state, weighted over loci. Some, but not all, inferred relationship IBD (type 3) and IBS (type 2) covariance structures can also be expressed in this way. For a simple diallelic SNP model to which all three approaches are applicable, we

can express all three models in terms that allow direct comparison. Consider two individuals X and Y . Each pair of SNP dosages at (X_l, Y_l) , is an identity state for the pair of individuals at locus l . Each model defines a covariance, relatedness, or similarity index r_{XY} . We decompose r_{XY} into a locus weight and an identity state contribution $C_l(X_l, Y_l)$ where we fix $C_l(2, 2) = 1$ and $C_l(2, 0) = C_l(0, 2) = 0$. Extracting a constant k , we can write

$$r_{XY} = k + \sum_{l=1}^L w_l C_l(X_l, Y_l) \quad (24)$$

4.3.1. Representative IBD method—The realized relationship coefficient, justified in IBD terms by Powell et al. (2010), and here considered representative of a type 3, hybrid, model, is computed for genotypes of individuals X and Y in terms of p_l , the population frequency of the A allele at locus l , and X_l and Y_l , the respective dosages of the A alleles at locus l in the two individuals:

$$r_{XY} = \frac{1}{2L} \sum_{l=1}^L \frac{(X_l - 2p_l)(Y_l - 2p_l)}{p_l(1 - p_l)} \quad (25)$$

The classical phenotypic covariance, in the additive model, is proportional to this coefficient. For this parametrization, $w_l = 2/Lp_l$, and the normalized contributions $C_l(X_l, Y_l)$ are given in Table 1.

4.3.2. Representative IBS method—An early estimator of relatedness on the basis of IBS genome sharing (Nejati-Javaremi et al., 1997) counts the number of the pairs of alleles IBS between the two individuals at each locus (0, 2 or 4) and averages this count over polymorphic loci. The arbitrariness of this definition is in the choice of polymorphic loci to be equally weighted; for example, adding loci presumed polymorphic, but actually not polymorphic, would skew the score by adding 4's into the average. This is equivalent to the similarity index:

“For a given single locus, a similarity index S_{XY} between 2 individuals x and y is calculated, where $S_{XY} = 1$ when genotype $x = kk$ (i.e., both alleles at locus l are identical) and genotype $y = kk$. $S_{XY} = 0.5$ when $x = kk$ and $y = kl$, or vice versa, or when $x = kl$ and $y = kl$, $S_{XY} = 0.25$ when $x = kl$ and $y = km$, and $S_{XY} = 0$ when the 2 individuals have no alleles in common at the locus (Eding and Meuwissen, 2001). The similarity index was averaged over loci.” (Hayes and Goddard, 2008)

We adapt the similarity index directly as our identity state contribution $C_l(X_l, Y_l) = S_{XY}$, and choose weights $w_l = 1/L$ for the L arbitrarily chosen loci.

4.3.3. The IBF method—Following the IBF definition of the covariance between individuals (Eq. 15) we choose the identity state contribution $C_l(X_l, Y_l) = C_{S(l)} / \sigma_{AA(l)}^2$. We scale by choosing weights $w_l = \sigma_{AA(l)}^2$ and introducing the parameter $v = AB(l) / AA(l)$.

4.3.4. Comparison—To allow comparison with the IBD method, we consider only diallelic combinations, excluding e.g. $AB : AC$. The results are demonstrated in Table 1. We can observe that the IBD method treats the two alleles asymmetrically, unless the allele frequency of $p = 0.5$ imposes symmetry. This offsets the asymmetry of the weighting scheme, $w_l = 2/Lp_l$, of course when $p_l = 0.5$, $w_l = 1/L$.

The key observation is that the same identity state contribution function is derived in the simplest and most symmetric case from each approach. The IBS similarity index is constructed from arbitrary symmetry considerations. The IBD-based realized relationship method gives the same function when the allele frequency $p = 0.5$. The IBF parametrization $v = \rho_3 = \frac{\sqrt{2}}{2}$ corresponds to the additive model.

We can interpret the difference thus: the IBF framework should deviate from the IBS method when there is dominance in the genetic architecture of the specific trait. The IBD method should deviate when major and minor allele frequencies differ.

5. Applications

For a set of known IBF parameters, and given data in the form of genotype-phenotype pairs, there exists a conditional distribution of the effects G_{ij} from which other inferences can be made about the effects themselves, classical variance components and heritabilities, and phenotypes and breeding values.

Under the assumption of normality and known IBF parameters, most operations involved in these applications are reducible to linear algebra. Section 4.2 describes the separate, and greater, computational challenge of estimating the IBF parameters from data.

5.1. Association

The generative model produces genotypic values G_{ij} for each allele pair at each locus. By classical methods (Lynch and Walsh, 1998, Chapter 4), these can be mapped to additive effects μ , dominance deviations d_{ij} and the additive variance σ_A^2 , given allele frequencies.

For fixed IBF parameters the distribution of effects G_{ij} if assumed multivariate normal as in section 3.4.1, is jointly multivariate normal with the phenotype distribution in section 4.2. Then the effect distribution conditional on observed phenotype is also normal and has a matrix form.

Formally, this implies the ability to perform a GWAS style test for the significance of individual loci and alleles. The power to do so is severely limited by the large number of parameters, one for each allele pair at each locus. Such testing is practical when IBF parameters are either assumed known or inferred from outside the sample. Significance for an individual allele is easiest to achieve when IBF parameters induce sparsity or concentration on particular loci.

5.2. Heritability

An important application of type 3 models (Yang et al., 2010) has been the estimation of heritability explainable by common variants. It has been pointed out (Visscher, 2010; Makowsky et al., 2011) that such estimates of heritability are not the same as predictive R^2 . With reasonable sample size, heritability can be estimated with limited error, but the large number of individual effects needed to build an accurate predictive model cannot. The narrow-sense heritability estimation problem, ignoring complications, is reducible to the estimation of the additive variance, σ_A^2 , which in a type 3 context (e.g. Yang et al., 2010) is a natural variance component parameter of the model. Lacking such a parametrization, we must construct the additive variance bottom-up from additive effects and dominance deviations. The additive variance, σ_A^2 is a quadratic form in the G_{ij} and a random variable under our model, for fixed allele frequencies. There are two ways to calculate an estimate of σ_A^2 and therefore heritability, conditional on observed data:

1. Substitute the expectations of individual G_{ij} into the quadratic form.
2. Compute the expectation of the quadratic form over the uncertainty about the G_{ij} .

By convexity the second, expected, σ_A^2 is greater than the first σ_A^2 . We interpret the second as an estimate of the true additive variance; and the first as the estimate of the additive variance explained by the model as inferred given finite data.

We illustrate heritability estimation in the Simulation and Missing Heritability Case Study section.

5.3. Prediction

Given the IBF parameters, a covariance between each pair of individuals, and therefore a covariance matrix among n individuals, can be computed. The IBF covariance can be interpreted as another kernel in the framework of Gianola and van Kaam (2008), and applied to the prediction of phenotype for an individual with known genotype, given a set of genotype-phenotype pairs. Such prediction has many analogies, and is algebraically equivalent to prediction in Henderson's BLUP, Kriging, or Gaussian Processes (Robinson, 1991). A prediction variance is computed together with the prediction, and is generally lower when data on "neighbors" of the individual to be predicted is available. A neighbor conducive to such prediction is an individual that shares genotype at important loci; the importance weight of a locus roughly corresponds to high variance in its IBF parameters.

5.4. Prediction, Populations, and Population Neutrality

Population neutral applications of the IBF model, in particular phenotype and individual genotypic effect prediction, do not require us to specify a reference population in the form of a set of allele frequencies. Unlike phenotype prediction, breeding value prediction is not population neutral in a model with dominance; indeed the breeding value can be defined through regression of the offspring trait onto a parent crossed with a mate chosen at random from a given population. Classically, we can predict the breeding value by inferring, and summing, the additive effects μ_j at each locus.

The bottom-up reconstruction of μ_j , breeding value, and heritability thus take a set of allele frequencies as an input. These may be directly estimated from the same population as the effects; alternatively, we may transpose our effect estimates onto a different population, or use an adjusted allele frequency estimation method (e.g. Sillanpää, 2011).

6. Simulation and Missing Heritability Case Study

6.1. Simulation and Prediction

A toolkit for simulation and prediction using the IBF model was built in the R language and is included in the supplementary materials section. We use this simulation to demonstrate the implementation details of the model, and to illustrate our approach to prediction and heritability estimation.

The infinite allele model is simulated using the Chinese Restaurant Process (e.g. Hoppe, 1984) with a fixed population size N and Ewens parameter θ (e.g. Ewens, 1972). This gives population allele frequencies. Treating this population as infinite, i.e. sampling with replacement, we draw pairs of alleles with these frequencies to construct genotypes. We assume Hardy-Weinberg equilibrium for the purpose of the demonstration, but the code as written will accept arbitrary pairwise genotype frequencies. For a single locus, we fix the full set of IBF parameters as a midpoint in covariance space between the Additive and

Random Order, Strict Dominance Models with $\mu_{AA} = \mu_{AB} = 0$ and $\sigma_{AA}^2 = 1$. We simulate genotypic values using the multivariate normal model with these IBF parameters.

Given genotypic values and allele frequencies, we can compute classical Fisher additive effects α_i and dominance deviations d_{ij} . These are simulated “true” values, not estimates from observed data. Likewise, we can compute breeding values by removing dominance deviations from centered genotypic effects.

We add an iid normal environmental contribution with $\sigma_E^2 = 1$ to the true genotypic values for a sample of 1000 individuals randomly selected genotypes to construct simulated phenotypes. Assuming we do not know the true simulated genotypic values, but do know the IBF parameters and σ_E^2 , we compute the mean and covariance matrix of genotypic values conditional on observed genotype-phenotype pairs by a procedure algebraically similar to BLUP. We then use these predicted genotypic values to construct predicted additive effects α_i and dominance deviations d_{ij} , and compare these with the true simulated values. Each of these predictions is linear in the genotypic values, holding allele frequencies fixed.

6.2. Simulation Case Study: Interpreting Missing Heritability

Classically, the contributions of each locus to the variance components σ_A^2 and σ_D^2 are functions of the allele frequencies p_i and additive effects α_i and dominance residuals d_{ij} . The α_i and d_{ij} in turn, coefficients and residuals of a weighted regression of the genotypic effects G_{ij} onto an allele dosage matrix, with the weighting a function of the p_i . Generally, σ_A^2 and σ_D^2 , and for our non-epistatic model $\sigma_G^2 = \sigma_A^2 + \sigma_D^2$, are quadratic in G_{ij} ; that is, each has the form $\mathbf{G} \mathbf{Q} \mathbf{G}$ where $\mathbf{G}$ is the vector of the G_{ij} and $\mathbf{Q}$ depends on p_i .

In our model, both the G_{ij} and the p_i are random variables; thus σ_A^2 and σ_D^2 are random. Our model has variance components, $C_1 \dots C_4$, which determine the IBF parameters at each locus. These do not translate deterministically to the classical variance components σ_A^2 and σ_D^2 , but to *expected* σ_A^2 and σ_D^2 , either under given allele frequencies, or under a distribution of allele frequencies.

Consider a fixed population with known allele frequencies p_i and a single locus trait with known IBF parameters and environmental variance σ_E^2 . If we have complete knowledge of the G_{ij} , the variance components σ_A^2 and σ_D^2 are given deterministically by the classical calculations. If we have no knowledge of G_{ij} , we can determine the expectations over possible realizations of genotypic values, $E[\sigma_A^2]$ and $E[\sigma_D^2]$. From a Bayesian perspective, the former case corresponds to a degenerate complete-data posterior, and the latter to a prior, on genotypic effects. If we observe some pairs of genotypes and phenotypes, and condition on this information, we obtain a posterior over genotypic effects, and the Bayes estimator of the variance components is the expectation under that posterior.

Under the normal model, taking for simplicity $\mu_{AA} = \mu_{AB} = 0$,

1. The prior is the IBF normal genotypic value model and has the form $\mathbf{G} \sim N(\mathbf{0}, \mathbf{I})$
2. The posterior is obtained by conditioning the prior on observed phenotypes, and has the form $\mathbf{G} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$
3. Complete knowledge of $\mathbf{G}$ is equivalent to $\mathbf{G} \sim N(\boldsymbol{\mu}_3, \mathbf{0})$

The expectation of a quadratic in a random vector is a sum of mean and covariance terms. Under the assumption of normality, so is the variance:

$$E[\mathbf{G}'\mathbf{Q}\mathbf{G}] = \mu' \mathbf{Q} \mu + \text{tr}[\mathbf{Q} \Sigma] \quad (26)$$

$$\text{Var}[\mathbf{G}'\mathbf{Q}\mathbf{G}] = 4\mu' \mathbf{Q} \Sigma \mathbf{Q} \mu + 2\text{tr}[\mathbf{Q} \Sigma \mathbf{Q} \Sigma] \quad (27)$$

In the methodology of Yang et al. (2010) addressing the “Missing Heritability” problem, variance component estimates imply a heritability that may be significantly higher than the R^2 achievable by prediction via the corresponding covariance model or attributable to individual markers. Here, we decompose the *Total* heritability or variance component estimate into an *Explained* and an *Unexplained* component. The mean term is the variance component, or equivalently heritability, *Explained* by inferred effects; it is the variance component we would get by substituting the predicted (i.e. expected, conditional on observed data) genotypic values into the classical expression for the variance component. The covariance term then corrects for the bias due to the genotypic values being predicted from limited data. In the extreme cases, when we observe no phenotype data (case 1), we have only the covariance term (unless $\mu_{AA} = \mu_{AB}$). Given infinite phenotype data (case 3), the mean term converges to the true variance component value; both the covariance term and the error term (SD) vanish. In the simulation, we obtain Table 2 by conditioning on 1000 genotype-phenotype pairs.

7. Discussion

7.1. Model Capabilities

Our IBF method defines a novel probability interpretation for the phenotypic covariance between genotyped individuals on the basis of a population process. It is an adaptation of the Fisher (1918) linear model, additive across loci, but not within a locus, with many alleles at each locus. It allows, in principle, the direct estimation of all individual i and ij terms of the Fisher model.

The IBF parameters define a genetic architecture, giving the relative importances of the different loci, and the pattern of dominance at each locus. This architecture is separate from the effects of individual alleles at each locus, and of the population frequencies of those alleles. This architecture allows us to express covariance between the traits of individuals in terms of alleles that need not have previously been observed. Inferences may be made about individuals across populations with different allele frequencies, including alleles which occur in one population but not the other. This is crucial for full sequence genomic analysis, where the number of individual alleles at a locus may be large, and the genotype of any given individual may include alleles at multiple loci not previously encountered in the population.

7.2. Diploidy and Dominance

The inherently diploid nature of the model, and the ability to treat the phenomenon of dominance coherently for a large number of alleles, allows for a reexamination of the relative importance of dominance effects. Existing models can be parametrized to incorporate dominance. Type 2 models through kernels designed to include allele interaction terms. In type 1 or type 3 models, the dominance genetic relationship matrix, or the matrix of coefficients of fraternity, coupled with the dominance variance component, can be incorporated into the covariance model together with the kinship matrix and the additive variance component. In our model, individual dominance effects have a covariance structure

implied by the IBF parameters. This allows for an arbitrary allele count and estimation of both the individual dominance deviations and variance components.

Dominance has not received much attention in the animal breeding literature, which is focused on breeding value estimation, or in the GWAS literature, where the linear SNP association models predominate. But dominance at the functional locus or gene level, as a distinct biological phenomenon, may be more detectable than nonlinearity in SNP dosage. Furthermore, even if additive, rather than dominance, variance predominates at the population level, (e.g. Hill et al., 2008), dominance could be important for individual phenotype prediction, especially where rare alleles are collectively common.

7.3. Extensions

The model as presented is strictly neutralist, and as such invites selectionist extensions. In particular, the model recognizes only a single class of alleles, and treats mutations between such alleles symmetrically. If we partition alleles into multiple functional classes (e.g. *normal* and *detrimental*) based on bioinformatics and population genetics, we can construct an extended set of IBF states and a set of covariances between such states, at the cost of a larger number of parameters. The set of homozygotes and heterozygotes within each class (e.g. two normal alleles, or two detrimental alleles) requires one set of IBF parameters μ_{AA} , μ_{AB} , σ_{AA}^2 , σ_{AB}^2 , σ_{AB}^2 for each class. To these would be added cross-class heterozygotes (normal-detrimental), requiring a mean and variance parameter for each class pairing, as well as a set of correlations (among cross-class heterozygotes sharing alleles, and between cross-class heterozygotes and within-class homozygotes and heterozygotes). To use this kind of model for inference given genotype, we need not explicitly model the population process giving rise to the different mutations. Breaking the symmetry between designated allele classes would be sufficient. Further extensions of the state space can integrate sex chromosomes and copy number variants into the polygenic framework.

The integration of bioinformatic allele classification with a quantitative genetic model is a step toward *Functional Quantitative Genetics*. Plant and animal breeding applications have repeatedly demonstrated the predictive power of quantitative genetics over a variety of complex traits, but a quantitative genetic model does not elucidate the causal mechanisms of gene action. Sequence bioinformatics aids the understanding of individual gene function at the molecular level, but not quantitative effects at organism scale. The relationship between genotype and phenotype for complex polygenic traits must be sought by the reconciliation of the two approaches.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by NIH training grant GM081062, and grants GM046255 and GM099568. The population neutrality concept benefited from discussions with, and an unpublished manuscript by, Joseph Felsenstein. We are grateful for valuable commentary and suggestions from three anonymous reviewers.

References

Browning SR, Browning BL. Identity by Descent Between Distant Relatives: Detection and Applications. *Annual Review of Genetics*. 2012; 46:617–633.

- de los Campos G, Gianola D, Rosa GJM. Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation. *Journal of Animal Science*. 2009; 87:1883–1887. [PubMed: 19213705]
- Cockerham CC. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*. 1954; 39:859. [PubMed: 17247525]
- Cockerham CC, Weir BS. Variance of actual inbreeding. *Theoretical Population Biology*. 1983; 23:85–109. [PubMed: 6857551]
- Eding JH, Meuwissen TH. Marker based estimates of between and within population kinships for the conservation of genetic diversity. *Journal of Animal Breeding Genetics*. 2001; 118:141–159.
- Ewens WJ. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*. 1972; 3:87–112. [PubMed: 4667078]
- Faul, AC.; Tipping, ME. *Advances in Neural Information Processing Systems 14*. MIT Press; 2001. Analysis of Sparse Bayesian Learning; p. 383–389.
- Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates; 2004.
- Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*. 1918; 52:399–433.
- Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R. Additive Genetic Variability and the Bayesian Alphabet. *Genetics*. 2009; 183:347–363. [PubMed: 19620397]
- Gianola D, van Kaam JB. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*. 2008; 178:2289–2303. [PubMed: 18430950]
- Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*. 2011; 5:1780–1815.
- Hayes BJ, Goddard ME. Technical note: Prediction of breeding values using marker-derived relationship matrices. *Journal of Animal Science*. 2008; 86:2089–2092. [PubMed: 18407982]
- Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetical Research*. 2009; 91
- Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*. 2008; 4:e1000008. [PubMed: 18454194]
- Hill WG, Weir BS. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*. 2011; 93:47–64. [PubMed: 21226974]
- Hoppe FM. Pólya-like urns and the ewens' sampling formula. *Journal of Mathematical Biology*. 1984; 20:91–94.
- Jacquard A. Genetic Information Given by a Relative. *Biometrics*. 1972; 28:1101–1114. [PubMed: 4648793]
- Kärkkäinen HP, Sillanpää MJ. Back to Basics for Bayesian Model Building in Genomic Selection. *Genetics*. 2012; 191:969–987. [PubMed: 22554888]
- Kempthorne O. The correlation between relatives in a random mating population. *Proceedings of the Royal Society of London Series B-Biological Sciences*. 1954; 143:103–113.
- Kiezun A, Garimella K, Do R, Stitzel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, et al. Exome sequencing and the genetic basis of complex traits. *Nature Genetics*. 2012; 44:623–630. [PubMed: 22641211]
- Kimura M, Crow JF. The Number of Alleles That Can Be Maintained in a Finite Population. *Genetics*. 1964; 49:725–738. [PubMed: 14156929]
- Lynch, M.; Walsh, B. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates; 1998.
- Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los Campos G. Beyond Missing Heritability: Prediction of Complex Traits. *PLoS Genetics*. 2011; 7:e1002051. [PubMed: 21552331]
- Meuwissen THE, Hayes BJ, Goddard ME. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*. 2001; 157:1819–1829. [PubMed: 11290733]
- Nejati-Javaremi A, Smith C, Gibson JP. Effect of total allelic relationship on accuracy of evaluation and response to selection. *Journal of Animal Science*. 1997; 75:1738–1745. [PubMed: 9222829]

- Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*. 2010; 11:800–805.
- Robinson GK. That BLUP is a good thing: The estimation of random effects. *Statistical Science*. 1991; 6:15–32.
- Sillanpää MJ. On statistical methods for estimating heritability in wild populations. *Molecular Ecology*. 2011; 20:1324–1332. [PubMed: 21426431]
- Thompson EA. Gene identities and multiple relationships. *Biometrics*. 1974; 30:667–680. [PubMed: 4429760]
- Thompson EA. The estimation of pairwise relationships. *Annals of Human Genetics*. 1975; 39:173–188. [PubMed: 1052764]
- Thompson EA. Identity by descent: Variability in meiosis, across genomes and in populations. *Genetics*. 2013; 194:301–326. [PubMed: 23733848]
- Tipping M. Sparse Bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*. 2001; 1:211–244.
- Visscher PM. Whole genome approaches to quantitative genetics. *Genetica*. 2009; 136:351–358. [PubMed: 18668208]
- Visscher PM. A Commentary on ‘Common SNPs Explain a Large Proportion of the Heritability for Human Height’ by Yang et al.(2010). *Twin Research and Human Genetics*. 2010; 13:517–524. [PubMed: 21142928]
- Weir BS, Anderson AD, Hepler AB. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*. 2006; 7:771–780.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*. 2011; 89:82–93.
- Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*. 2010; 42:565–569. [PubMed: 20562875]

Appendix A. Positive Definite Conditions

In order for the joint distribution to exist, the two correlations ρ_3 and ρ_2 must be such that the covariance matrix of all homozygote and heterozygote values $[G_{ij}, G_{ij}]$ is positive semidefinite. The joint covariance matrix has the form

$$= \text{Cov} \left(([G_{11} \ G_{22} \ G_{33} \ G_{44}]; [G_{12} \ G_{13} \ G_{14} \ G_{23} \ G_{24} \ G_{34}])^T \right) \quad (A.1)$$

$$= \begin{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \sigma_{AA}^2 & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \rho_3 \sigma_{AB} \sigma_{AA} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \rho_3 \sigma_{AB} \sigma_{AA} & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \sigma_{AB}^2 + \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} \rho_2 \sigma_{AB}^2 \end{bmatrix}$$

Or, generally, using analogous (homozygotes, then row major) ordering of the G_{ij} ,

$$\text{Cov} \left(\begin{bmatrix} G_{ii} \\ G_{jk} \end{bmatrix} \right) = A = \begin{pmatrix} I\sigma_{AA}^2 & M_{i,jk}^{(3)}\rho_3\sigma_{AB}\sigma_{AA} \\ M_{i,jk}^{(3)T}\rho_3\sigma_{AB}\sigma_{AA} & I\sigma_{AB}^2 + M_{jk;lm}^{(2)}\rho_2\sigma_{AB}^2 \end{pmatrix} \quad (\text{A.2})$$

Here $M_{i,jk}^{(3)}$ and $M_{jk;lm}^{(2)}$ are binary matrices, containing 1 when exactly one index matches between the column and row identifiers, and 0 otherwise.

To find the values of ρ_3 and ρ_2 for which the matrix is positive semidefinite, we observe that $I\sigma_{AA}^2$ is positive definite. Then A is positive semidefinite if and only if the Schur complement of $I\sigma_{AB}^2 + M_{jk;lm}^{(2)}\rho_2\sigma_{AB}^2$ in A is positive semidefinite. The Schur complement is

$$S = I\sigma_{AB}^2 + M_{jk;lm}^{(2)}\rho_2\sigma_{AB}^2 - \begin{bmatrix} M_{i,jk}^{(3)}\rho_3\sigma_{AB}\sigma_{AA} \end{bmatrix} \begin{bmatrix} I\sigma_{AA}^2 \end{bmatrix}^{-1} \begin{bmatrix} M_{i,jk}^{(3)T}\rho_3\sigma_{AB}\sigma_{AA} \end{bmatrix} \quad (\text{A.3})$$

$$= \sigma_{AB}^2 \begin{bmatrix} I + M_{jk;lm}^{(2)}\rho_2 - M_{i,jk}^{(3)}M_{i,jk}^{(3)T}\rho_3^2 \end{bmatrix}$$

We observe that

$$M_{i,jk}^{(3)}M_{i,jk}^{(3)T} = M_{jk;lm}^{(2)} + 2I \quad (\text{A.4})$$

Then

$$S = \sigma_{AB}^2 \begin{bmatrix} I + (M_{i,jk}^{(3)}M_{i,jk}^{(3)T} - 2I)\rho_2 - M_{i,jk}^{(3)}M_{i,jk}^{(3)T}\rho_3^2 \end{bmatrix} \quad (\text{A.5})$$

$$= \sigma_{AB}^2 \begin{bmatrix} (1 - 2\rho_2)I + M_{i,jk}^{(3)}M_{i,jk}^{(3)T}(\rho_2 - \rho_3^2) \end{bmatrix}$$

$M_{i,jk}^{(3)}M_{i,jk}^{(3)T}$ is positive semidefinite; it is not of full rank for nontrivial numbers of alleles; thus it has an eigenvalue of 0 and no negative eigen-values. The eigenvalues of $M_{i,jk}^{(3)}M_{i,jk}^{(3)T}$ grow without bound with the size of the matrix since the vector 1 is always an eigenvector with an eigenvalue that grows with the count of heterozygotes. The eigenvalues of S are simply $\lambda' = (1 - 2\rho_2) + (\rho_2 - \rho_3^2)\lambda$ for all eigenvalues λ of $M_{i,jk}^{(3)}M_{i,jk}^{(3)T}$. We want to constrain the values of ρ_2 and ρ_3^2 so that any number of heterozygotes, S is positive semidefinite. We must choose $\rho_2 - \rho_3^2 \geq 0$ because otherwise we can find a large enough λ for a large number of heterozygotes to make $\lambda' < 0$. Likewise, we must choose $1 - 2\rho_2 \geq 0$, since otherwise $\lambda' < 0$ for $\lambda = 0$. Together these create the constraint:

$$\rho_3^2 \leq \rho_2 \leq \frac{1}{2} \quad (\text{A.6})$$

Appendix B. Diploid Additive Model

For the diploid additive model we associate an additive effect with each allele, $C(\cdot) : \mathcal{A} \rightarrow \mathbb{R}$, and define the genotypic value function $G(A_i A_j) = C(A_i) + C(A_j)$; then, for distinct i, j, k, l ,

$$\begin{aligned} E[G_{ii}] &= E[C(A_i)] + E[C(A_i)] = 2\mu_C = \\ E[G_{ij}] &= E[C(A_i)] + E[C(A_j)] = 2\mu_C \end{aligned} \quad (\text{B.1})$$

This implies $\mu_{AA} = \mu_{AB} = 2\mu_C$.

$$\text{Var}[G_{ii}] = \text{Var}[2C(A_i)] = 4 \text{Var}[C(A_i)] = 4\sigma_C^2 \quad (\text{B.2})$$

$$\text{Var}[G_{ij}] = \text{Var}[C(A_i)] + \text{Var}[C(A_j)] = 2\sigma_C^2 \quad (\text{B.3})$$

Substituting into the general model,

$$\text{Var}[G_{ii}] = \sigma_{AA}^2 = 4\sigma_C^2 \quad (\text{B.4})$$

$$\text{Var}[G_{ij}] = \sigma_{AB}^2 = 2\sigma_C^2 \quad (\text{B.5})$$

This implies $\sigma_{AA} = 2\sigma_C$ and $\sigma_{AB} = \sqrt{2}\sigma_C$, or $\sigma_{AB}/\sigma_{AA} = 1/\sqrt{2}$.

$$\text{Cov}[G_{ii}, G_{ij}] = \text{Cov}[2C(A_i), C(A_i)] + \text{Cov}[2C(A_i), C(A_j)] = 2\sigma_C^2 \quad (\text{B.6})$$

$$\text{Cov}[G_{ij}, G_{ik}] = \text{Var}[C(A_i)] + \text{Cov}[C(A_i), C(A_j)] + \text{Cov}[C(A_i), C(A_k)] + \text{Cov}[C(A_j), C(A_k)] = \sigma_C^2 \quad (\text{B.7})$$

Substituting into the general model,

$$\text{Cov}[G_{ii}, G_{ij}] = \rho_3 \sigma_{AA} \sigma_{AB} = \rho_3 2\sqrt{2}\sigma_C^2 = 2\sigma_C^2 \quad (\text{B.8})$$

$$\text{Cov}[G_{ij}, G_{ik}] = \rho_2 \sigma_{AB}^2 = \rho_2 2\sigma_C^2 = \sigma_C^2 \quad (\text{B.9})$$

This implies $\rho_2 = 1/2$ and $\rho_3 = 1/\sqrt{2}$.

Appendix C. Overdominance in the Multivariate Normal Model

For two distinct alleles F and G , we define the dominance coefficient Y in terms of the genotypic values FF , FG , and GG :

$$Y = \frac{(FG - FF) + (FG - GG)}{FF - GG} \quad (\text{C.1})$$

A value of $|Y| > 1$ indicates overdominance, an ordering other than $GG > FG > FF$ or $FF > FG > GG$. We will derive the probability of overdominance, P_{OD} , for the multivariate normal model with $\mu_{AB} = \mu_{AA}$.

Consider a general multivariate normal distribution of the form:

$$\begin{bmatrix} A_0 \\ B \\ A_1 \end{bmatrix} \sim N \left(\begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_A^2 & \sigma_A \sigma_B \rho_{AB} & \sigma_A^2 \rho_A \\ \sigma_A \sigma_B \rho_{AB} & \sigma_B^2 & \sigma_A \sigma_B \rho_{AB} \\ \sigma_A^2 \rho_A & \sigma_A \sigma_B \rho_{AB} & \sigma_A^2 \end{bmatrix} \right) \quad (\text{C.2})$$

$$\begin{aligned} & \text{Var}(-A_1 + 2B - A_0) \\ &= \sigma_A^2 + 4\sigma_B^2 + \sigma_A^2 + 2(-2\sigma_A \sigma_B \rho_{AB} + 1\sigma_A^2 \rho_A - 2\sigma_A \sigma_B \rho_{AB}) \\ &= 2\sigma_A^2 + 4\sigma_B^2 - 8\sigma_A \sigma_B \rho_{AB} + 2\sigma_A^2 \rho_A \\ &= 2(1 + \rho_A) \sigma_A^2 - 8\rho_{AB} \sigma_A \sigma_B + 4\sigma_B^2 \\ & \text{Cov}(-A_1 + 2B - A_0, A_1 - A_0) \\ &= -\sigma_A^2 + 2\sigma_A^2 \rho_{AB} - \sigma_A^2 \rho_A - (-\sigma_A^2 \rho_A + 2\sigma_A^2 \rho_{AB} - \sigma_A^2) \\ &= 0 \\ & \text{Var}(A_1 - A_0) \\ &= 2\sigma_A^2 - 2\sigma_A^2 \rho_A \\ &= 2(1 - \rho_A) \sigma_A^2 \end{aligned}$$

The distribution of a ratio of zero mean uncorrelated normals $Y = N/D$ with respective variances σ_N^2 and σ_D^2 and correlation ρ is a Cauchy with pdf $p(y) = \frac{1}{\pi} \beta / (y^2 + \beta^2)$ where $\beta = \frac{\sigma_N}{\sigma_D}$.

$$\sigma_N^2 = 2(1 + \rho_A) \sigma_A^2 - 8\rho_{AB} \sigma_A \sigma_B + 4\sigma_B^2 \quad (\text{C.3})$$

$$\sigma_D^2 = 2(1 - \rho_A) \sigma_A^2 \quad (\text{C.4})$$

In the overdominance case, $A_1 = FF$; $A_2 = GG$; $B = FG$; $\rho_{AB} = \frac{1}{2}$; $\rho_A = 0$; the variances of the heterozygote and homozygote are σ_{AB}^2 and σ_{AA}^2 respectively.

$$\beta = \frac{\sigma_N}{\sigma_D} = \sqrt{\frac{2(1 + \rho_A) \sigma_A^2 - 8\rho_{AB} \sigma_A \sigma_B + 4\sigma_B^2}{2(1 - \rho_A) \sigma_A^2}} \quad (\text{C.5})$$

$$= \sqrt{\frac{2 + 2\rho_A - 8\rho_{AB} \frac{\sigma_B}{\sigma_A} + 4\frac{\sigma_B^2}{\sigma_A^2}}{2 - 2\rho_A}} \quad (\text{C.6})$$

$$= \sqrt{\frac{2 - 4\rho_{AB} \frac{\sigma_B}{\sigma_A} + 2\frac{\sigma_B^2}{\sigma_A^2}}{1 - \rho_A}} - 1 \quad (\text{C.7})$$

The Cauchy CDF gives

$$\Pr(|Y| > 1) = \frac{2}{\pi} \tan^{-1} \beta \quad (\text{C.8})$$

$$\beta = \sqrt{1 - 4\rho_3 \frac{\sigma_{AB}}{\sigma_{AA}} + 2 \frac{\sigma_{AB}^2}{\sigma_{AA}^2}} \quad (\text{C.9})$$

This gives the probability of overdominance P_{OD} :

$$P_{OD} = \Pr(|Y| > 1) = \frac{2}{\pi} \tan^{-1} \beta = \frac{2}{\pi} \tan^{-1} \sqrt{1 - 4\rho_3 \frac{\sigma_{AB}}{\sigma_{AA}} + 2 \frac{\sigma_{AB}^2}{\sigma_{AA}^2}} \quad (\text{C.10})$$

In the additive case, $\rho_3 = \sigma_{AB} / \sigma_{AA} = 1 / \sqrt{2}$; then $P_{OD} = \frac{2}{\pi} \tan^{-1} \sqrt{1 - 2 + 1} = 0$ as expected.

Appendix D. Strict Dominance, Random Ordering

The marginal distribution of the heterozygotes is the same as that of the homozygotes, since under random ordering a homozygote's distribution is an equiprobable mixture of its two homozygotes' equal distributions. Table D.3 lists the 6 equally probable orderings of the affinity variable S_i . Conditional on each ordering, each heterozygote is equal to one of the three homozygotes, as tabulated. Taking wlog $\sigma_{AB} = \sigma_{AA} = 1$, the ρ_2 and ρ_3 columns give the conditional covariances $\text{Cov}(G_{AB}, G_{AC})$ and $\text{Cov}(G_{AB}, G_{AA})$ respectively, which, when averaged over the 6 random orderings, give the unconditional covariances and therefore correlations ρ_2 and ρ_3 .

Table D.3

Affinity orderings, heterozygote-homozygote equivalences, and conditional correlations for the Strict Dominance, Random Ordering model.

Ordering:	G_{AB}	G_{AC}	G_{BC}	ρ_2	ρ_3
$S_A > S_B > S_C$	G_{AA}	G_{AA}	G_{BB}	1	1
$S_A > S_C > S_B$	G_{AA}	G_{AA}	G_{CC}	1	1
$S_B > S_A > S_C$	G_{BB}	G_{AA}	G_{BB}	0	0
$S_B > S_C > S_A$	G_{BB}	G_{CC}	G_{BB}	0	0
$S_C > S_A > S_B$	G_{AA}	G_{CC}	G_{CC}	1	0
$S_C > S_B > S_A$	G_{BB}	G_{CC}	G_{CC}	0	0

Figure 1.

Functional Identity States. The 15 partitions of four chromosomes in two diploid individuals correspond to the 9 Jacquard coefficients (left) and map to the 5 Functional Identity States (right).

Figure 2.

Feasible region $\rho_3^2 \leq \rho_2 \leq \frac{1}{2}$ and (ρ_2, ρ_3) for various simulated distributions and dominance models. Numbers correspond to the Strict Dominance, Trait Order model under different distributional assumptions, as described in the text. Symbols correspond to the Partial Dominance model for different values of σ . For each symbol, full color is Random Order ($\rho = 0$), and black is Trait Order ($\rho = 1$).

Table 1

Comparison of identity state contributions under IBD, IBS, and IBF methods.

	IBD	IBD with $p = 1/2$	IBS	IBF	IBF with $v = \rho_3 = \frac{\sqrt{2}}{2}$
AA:AA	1	1	1	1	1
AA:AB	1/2	1/2	1/2	3^v	1/2
AA:BB	0	0	0	0	0
AB:AB	$\frac{1}{4} \left(1 + \frac{p}{1-p} \right)$	1/2	1/2	v^2	1/2
AB:BB	$\frac{1}{2} \frac{p}{1-p}$	1/2	1/2	3^v	1/2
BB:BB	$\frac{p}{1-p}$	1	1	1	1

Table 2

Decomposition of the expected variance components into Explained and Unexplained components; numerical values are from a simulation described in the Simulation Supplement

	σ_A^2	σ_D^2	$\sigma_G^2 = \sigma_A^2 + \sigma_D^2$
Explained	0.41	0.05	0.46
Unexplained	0.01	0.02	0.04
Total	0.43	0.07	0.50
SD	0.04	0.01	0.04