Received: 20 November 2018 | Revised: 4 March 2019 | Accepted: 25 March 2019

DOI: 10.1002/gepi.22206

REVIEW ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Correlations between relatives: From Mendelian theory to complete genome sequence

Elizabeth A. Thompson

Department of Statistics, University of Washington, Seattle, Washington

Correspondence
Elizabeth A. Thompson, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322.
Email: eathomp@uw.edu

Funding information
National Institutes of Health, Grant/Award Number: R37-GM046255

Abstract

It is 100 years since R. A. Fisher proposed that a Mendelian model of genetic variant effects, additive over loci, could explain the patterns of observed phenotypic correlations between relatives. His loci were hypothetical and his model theoretical. It is only about 50 years since the first genetic markers allowed the detection of even variants with major effects on phenotype, and only 20 years since the development of single-nucleotide polymorphism technology provided dense markers over the genome. Then both mappings in defined pedigrees and population-based genome-wide association studies samples allowed the detection of multiple contributing variants of smaller effect. Finally, with methods based on genotypic correlations between individuals, or on allelic associations between loci, the additive heritability contributions of the genome can be estimated from large population samples. In this review we trace, from 1918 to 2018, the analysis of observed phenotypic correlations between relatives to estimate underlying genetic components of traits in human populations. As with studies from 1918 onward, we use height as the example trait where not only data are readily available, but where Fisher's model of large numbers of variants of infinitesimal effect appears to provide a good approximation to reality. However, we also trace the use of phenotypic and genotypic correlations between relatives in mapping causal variants and resolving genetic contributions to more complex human traits. With the availability of DNA sequence data, we can hope to not only estimate the total genetic contribution to a trait, but to resolve effects of individual genetic variants on biological function.

KEYWORDS

additive models, assortative mating, genotypic correlation matrix, heritability, identity by function, population structure, realized descent

1 | INTRODUCTION

In the 100 years since Fisher (1918) (hereafter *F18*) was published, knowledge of genomes has changed out of all recognition, but explaining quantitative trait variation remains a challenge. Fisher's critical contribution was in showing that, under a Mendelian model, genetic variance would be maintained over generations. Fisher's approach was explicit modeling of multiple causal variants, which, in principle, leads to the identification of these variants and estimation of their effects on a trait. The model also leads to expressions for the genotypic correlations

between individuals in terms of their realized or expected pedigree or population relationships. Working directly with these expected correlations is the basis of the method of path coefficients initiated by Wright (1921), also almost a century ago.

1.1 | Fisher's Mendelian model for a quantitative trait

Fisher's model is of multiple variants each of small effect, distributed at loci across the genome. The effects are assumed additive over loci, but with general values for the three genotypes at any locus. Thus the phenotype y_i of individual i with genotypic effect $G_{i\ell}$ at locus ℓ is

$$y_i = \sum_{\ell} G_{i\ell} + e_i, \tag{1}$$

where e_i is an independent environmental effect, with variance V_e. Using more modern terminology and notation, the genotype of an individual i at a diallelic locus ℓ may be coded as $x_{il} = 0, 1$ or 2, the number copies of the reference allele. The phenotypic means G_{il} for the three genotypes are $(\mu - a_\ell)$, $(\mu + d_\ell)$and $(\mu + a_\ell)$. A linear regression of the genetic contribution of locus ℓ is

$$G_{i\ell} = \mu + \beta_\ell x_{i\ell} + \delta_{i\ell}, \tag{2}$$

where $\delta_{i\ell}$ is a zero-mean residual effect.

In a random-mating population under Hardy–Weinberg equilibrium, in which the reference allele has frequency p_ℓ, $x_{il} = 0, 1, 2$ with probabilities $(1 - p_l)^2$, $2p_\ell(1 - p_\ell)$, p_ℓ^2, and the variance of x_{il} is $2p_\ell(1 - p_\ell)$. From the linear regression (2), the mean and regression slope for the genetic effect of locus ℓ are

$$\begin{aligned} &\mu = a_\ell(2p_\ell - 1) + 2d_\ell p_\ell(1 - p_\ell) \quad \text{and} \\ &\beta_\ell = (a_\ell + (1 - 2p_\ell)d_\ell). \end{aligned} \tag{3}$$

Here β_ℓ is the expected increase in phenotype for unit increase in x_{il}, or the substitution of a reference allele for an alternate. This additive component of the genotypic values of Equation (2) was referred to by *F18* as the "*essential genotype*." It is also known as the *genic value* or *breeding value*.

The variance of the genetic contribution of locus ℓ to phenotype is the sum of additive and dominance variances, or equivalently the sum of the variance due to the regression, and the residual variance due to dominance:

$$\begin{aligned} \mathrm{var}(G_{i\ell}) &= (-a_\ell - \mu)^2(1 - p_\ell)^2 \\ &\quad + (d_\ell - \mu)^2 2p_\ell(1 - p_\ell) + (a_\ell - \mu)^2 p_\ell^2 \\ &= 2p_\ell(1 - p_\ell)(a_\ell + (1 - 2p_\ell)d_\ell)^2 \\ &\quad + (2p_\ell(1 - p_\ell)d_\ell)^2 = v_{g\ell} + v_{d\ell} \end{aligned} \tag{4}$$

(see Crow & Kimura, 1970 pp. 117–119). These variances sum over the multiple loci contributing to the trait: $V_g = \sum_\ell v_{g\ell}$, $V_d = \sum_\ell v_{d\ell}$. The total phenotypic variance $V_t = \mathrm{var}(y_i)$ is the sum of additive, dominance, and environmental variances: $V_t = V_g + V_d + V_e$.

1.2 | Correlations due to relatedness in a random-mating population

Beyond the fundamental demonstration that a model of Mendelian variants maintains genetic variance in a population, the stated purpose in *F18* was to explain correlations between relatives of different kinds, using this same model of Mendelian variants with causal effects. Under random mating, and allowing for dominance Fisher derived the correlations in genetic effects, using the joint distribution of genotypes at a diallelic locus in a pair of relatives. He considered first a parent and offspring, then an ancestor and descendant. He extended this to any unilateral relationship, where only one pair of homologous alleles is correlated between the two individuals. In this case, only the additive component of genetic variance enters the expression. Finally, he considered bilateral relatives, such as siblings and double-first-cousins, where the individuals are related through both parents so there is an additional dimension of genotypic dependence which involves the dominance variance. Later in the paper, *F18* extended these results to the case of multiple alleles at a locus.

Finally, in this first part of the paper, Fisher extends the results to include additive epistatic interactions between pairs of loci. As he comments. "*there is no biological reason for supposing that the [joint genotypic effects] should be exactly represented by the deviations formed by adding [the single-locus effects].*" Interestingly, he comments that although more complex interactions may exist, these are unlikely to produce any statistically detectable effect. The same view has been taken in much of the more recent literature. On the one hand, biological systems are often nonlinear, and interest in using epistasis in models for complex traits or for disease prediction continues. It is also argued that, even if epistatic contributions are individually small, the collective contribution may be large, resulting in heritability estimates that are higher in closer relatives (Falconer &

MacKay, 1996). On the other hand, in analyses of data, estimates of the contribution of epistatic variance to the total genetic variance are typically small. Mäki-Tanila and Hill (2014) show that while multilocus epistatic effects make substantial contributions to the additive variance, they do not lead to substantial contributions to the nonadditive component of genetic variance.

Often trait values are transformed, and adjusted for covariates, with the goal of improving Normality and linearity. The height and other human stature measurements considered by Pearson and Lee (1903) and analyzed by *F18* do not require transformation: Adjusted for sex within a homogeneous population, height has a close to Normal distribution. Sverdlov and Thompson (2018) show that complex traits that satisfy certain conditions in their genetic and environmental components can be well represented by a linear genetic model after appropriate transformation, despite underlying biological complexity. They determine conditions which together define a boundary between systems suitable and unsuitable for linear modeling.

The focus of *F18* was on the genetic effects on a quantitative trait, but in analyses of observed correlations between relatives, the denominator is total phenotypic variance. Hence a major impact on the observed correlation is the contribution of the environment, or in Fisher's words "*arbitrary external causes independent of heredity*". *F18* modeled the effect of the environment as a constant addition V_e to phenotypic variance, but in reality, closer relatives share not only more genetic effects but also more of their environment. For many traits, this is likely to be a major contribution to the observation that the similarity of close relatives, as compared with remote relatives, is greater than can be explained by genetic effects alone. Additionally, gene–environment interactions are another source of nonlinearity in modeling the observable value of a quantitative trait.

2 | QUANTITATIVE TRAIT VARIATION UNDER ASSORTATIVE MATING

2.1 | Variances and covariances: Overview

Fisher recognized that under his model the high correlations sometimes observed for traits such as body size and height could not be met under the assumption of random mating, and thus the major focus of his paper is then on the effects of assortative mating. He stated as "obvious," the increase in variance and the gametic phase disequilibrium (LD) that characterize such a mating structure, and derived the appropriate equations.

Fisher's paper is notoriously difficult to follow, although the annotated version of Moran and Smith (1966) clarifies many points. In *F18*, Fisher works directly with the joint genotypic arrays of relatives, which are significantly simplified by considering gene identity by descent (IBD). Given the pedigree-based probabilities (k_0, k_1, k_2) of relatives sharing 0, 1 or 2 genes IBD (Cotterman, 1940), Fisher's complicated genotypic probability arrays for the random-mating case are easily derived.

Wright (1921), in an early version of his path analysis methods, was the first to use IBD probabilities directly in analyzing quantitative trait variation. In his approach, only additive effects are considered: Dominance and epistasis are absent, and environmental effects are independent. Path analysis leads to much simpler derivations of Fisher's results for equilibrium correlations between relatives under assortative mating, but is not a generative model in the sense of Fisher's explicit gene effects.

Crow and Felsenstein (1968) use the basic IBD concept and parameters of Wright (1921) to rederive Fisher's results. However, their variance components are, as for Fisher, explicit functions of additive effects and dominance deviations for genetic variants at multiple loci. In this paper, hereafter denoted *CF68*, rather than considering genotypic arrays, they study the allelic correlations between homologous genes within individuals and between mates, and between nonhomologous genes on a single haplotype, on the two haplotypes within an individual, and on haplotypes in mates. *CF68* consider first the case where the trait is the additive result of effects at a large number of possibly linked loci, with multiple alleles, of varying effect, and varying allele frequencies. Under this model, and assuming positive assortative mating based on mate phenotypes, they derive expressions for the increase in phenotypic variance in the population, for the buildup in LD, and for the increase in homozygosity. Extending to situations with dominance, and adding in the effects of environment, *CF68* further confirm Fisher's results that assortative mating increases additive genetic variance, but not environmental variance. They note that, contrary to the claim of *F18* the dominance variance does increase due to the increase in homozygosity, but that this effect is minimal when there are causal variants at many loci.

A very clear exposition that follows the IBD approach of *CF68* is given by Nagylaki (1982), who gives also a clear introductory summary with very useful references to the relevant literature at that time. His formulation is closer to Wright's in that he considers regression

equations and conditional expectations rather than specific gene effects, and considers only additive effects without dominance. He shows that to derive the equilibrium correlations between relatives, it suffices to suppose that the regression of individual phenotype on genotype is linear and that the regression of individual phenotype on mate phenotype is linear. However, for the population phenotypic variance, it is also required that the regression of allelic effects on individual phenotype is linear. Together, these requirements effectively restrict distributions of allelic and environmental effects to the multivariate Normal case. Using conditional expectation arguments, Nagylaki provides clear derivations of Fisher's results for several types of relatives. A particular point of emphasis made by Nagylaki (1982) is that all the equilibrium correlations are independent of the genetic linkage map, as stated by *F18* and shown also by *CF68*.

2.2 | Equilibrium variance and covariances under assortative mating

We here present some of the main results of *F18* following the re-derivation by *CF68*, where details may be found. In a random-mating population, the total trait variance may be written as the sum of the genic (additive), dominance, and environmental variances:

$$V_t = V_g + V_d + V_e$$

assuming independence of environmental and genetic factors. As argued by *F18*, assortative mating will increase V_g, but, with a large number of contributing causal variants, the effect on V_d is negligible. Eventually, after many generations of assortative mating, the population reaches an equilibrium. Denoting this population at equilibrium under assortative mating by *EAM*, the corresponding variances are σ_g^2, σ_d^2, and σ_e^2, and the total phenotypic variance in the EAM is

$$\sigma_t^2 = \sigma_g^2 + \sigma_d^2 + \sigma_e^2. \tag{5}$$

Here $\sigma_d^2 \approx V_d$ and $\sigma_e^2 = V_e$, but we use the notation σ^2 to distinguish more clearly the equilibrium under assortative mating from the initial random-mating population.

For the EAM population, the narrow-sense (additive/genic) heritability is $h^2 = \sigma_g^2/\sigma_t^2$, and the broad-sense (genetic/genotypic) heritability is $H^2 = (\sigma_g^2 + \sigma_d^2)/\sigma_t^2$. Focusing on direct assortative mating for phenotype, suppose the correlation between mates is ρ_m. Then the correlation in genotypic trait contributions between mates is $\rho_m H^2$ and between their genic values is $A = \rho_m h^2$. Note that the A of *F18* is the $\hat{A}$ of *CF68*, and that, in contrast to *F18* and *CF68*, we work in terms of heritability in the *EAM* population. As assortative mating proceeds, there is a buildup in gametic phase LD, and an increase in trait variance. At equilibrium, $\sigma_g^2 \approx V_g/(1 - \rho_m h^2) = V_g/(1 - A)$ (Wright, 1921). Assuming $\sigma_d^2 = V_d$ and $\sigma_e^2 = V_e$, the total phenotypic variance increases:

$$\begin{aligned}\sigma_t^2 &= \sigma_g^2 + \sigma_d^2 + \sigma_e^2 \approx \sigma_g^2 + (V_t - V_g)\\ &= V_g(A/(1 - A)) + V_t = V_t + A\sigma_g^2\end{aligned}$$

CF68 rederive Fisher's formulae for correlations between pairs in a variety of pedigree relationships, both unilateral and bilateral. Because of the direct phenotypic correlation ρ_m between mates, an individual who deviates by a unit amount from the mean will have a mate who deviates (on average) by an amount ρ_m. The mean parental phenotypic deviation is $\frac{1}{2}(1 + \rho_m)$. The mean deviation of genic values in the offspring is thus $\frac{1}{2}h^2(1 + \rho_m)$ and this is then also the phenotypic correlation ρ_p between parent and offspring:

$$\rho_p = \frac{1}{2}h^2(1 + \rho_m). \tag{6}$$

The mean deviation of the genic value of the offspring's mate is then $\frac{1}{2}h^2(1 + \rho_m)A$, so the mean for this couple is $\frac{1}{2}h^2(1 + \rho_m)\frac{1}{2}(1 + A)$, and this is also the mean deviation for their offspring. That is, the correlation between grandparent and grandchild is $(\frac{1}{2})^2h^2(1 + \rho_m)(1 + A)$. Likewise, each additional generation, with a correlation A between the genic values of mates, given an inflation $(1 + A)$ over the random-mating Mendelian $\frac{1}{2}$. The correlation between an individual and nth generation descendant is $(\frac{1}{2}h^2(1 + \rho_m))(\frac{1}{2}(1 + A))^{n-1}$. As *F18* notes, the relative effect of assortative mating increases for more distant relatives.

Considering only the additive effects, each relevant mating in a collateral relationship gives also an inflation $(1 + A)$ over the random-mating formula. For example, for an aunt–niece pair there are two relevant matings, and the correlation is $\frac{1}{4}h^2(1 + A)^2$. However, unlike in the random-mating case, the dominance contribution $D = \sigma_d^2/\sigma_t^2$ also enters into the formulae, because assortative mating reduces the variance within a sibship in a way that affects more distant collateral relatives. Accepting *F18* that the dominance variance is little affected by assortative mating, the correlation ρ_s between sibs is

$$\rho_s = D/4 + h^2(1 + A)/2 = (1/4)(H^2 + h^2(1 + 2A)) \tag{7}$$

because $H^2 = D + h^2$. (The first form is that given by *CF68* while *F18* uses the latter.) This transmits to the aunt–niece pair as $\frac{1}{4}h^2(1 + A)^2 + \mathrm{DA}/4$, and to (unilateral) first cousins as $\frac{1}{8}h^2(1 + A)^3 + \mathrm{DA}^2/16$. Bilateral relatives such as double-first-cousins provide not only terms in D, but also additional terms in A even if $D = 0$. Nagylaki (1982) also considers other examples such as step-parents and step-sibs, who are have no common ancestors, but whose phenotypes are correlated through the mating correlations of parents.

2.3 | The inheritance of height

In this paper, we use height to trace history of application of quantitative genetic theory. Height is one of the most studied human quantitative traits, and has been the canonical example of a quantitative genetic trait from Pearson and Lee (1903) to Lello et al. (2018). It is easily measured, and is probably a trait for which a model of a very large number of contributing loci each of very small effect is most applicable. Within a homogeneous population, and adjusted for age and sex, it has a Normal distribution. However, it is also subject to substantial environmental effects, apparent population differences, and assortative mating within populations.

F18 used the data of Pearson and Lee (1903) to exemplify his theory. Based on these data, he gives values $\rho_m = 0.2804$ and $\rho_p = 0.5066$, for the phenotypic correlation ρ_m between mates and ρ_p between parent and offspring. From Equation (6),

$$\rho_p = (1/2)(1 + \rho_m)h^2,$$

where $h^2 = \sigma_g^2/\sigma_t^2$ is the (narrow-sense) heritability in the equilibrium population. This provides an estimated $h^2 = 0.7913$. The correlation in additive genetic values of mates, $A = \rho_m h^2$, is estimated as 0.2219.

The observed correlation between siblings reported by Fisher (1918) is $\rho_s = 0.5433$. So from Equation (7) the proportion of total variance due to genetic factors is estimated as

$$H^2 = 4 \times 0.5433 - 0.7913(1 + 2 \times 0.2219) = 1.03.$$

Because $H^2 \approx 1$, this indicates that the proportion of total variance due to genetic effects is essentially 100%. Fisher concluded that height is minimally affected by environmental effects shared and not shared by relatives. However, as noted by *CF68* this assumes that environmental correlations for sibs are no greater than for parent and offspring. This is unlikely to be true, and shared environment may play a greater role, and dominance a lesser one, than Fisher concluded.

Following *CF68* we have also an analysis of the variance in height for the EAM population, based on these data. In terms of the values in a random-mating population. We have $V_g = \sigma_g^2(1 - A)$ and $V_t = \sigma_t^2 - A\sigma_g^2$. Thus the heritability in the random-mating population would be

$$\frac{V_g}{V_t} = \frac{\sigma_g^2(1 - A)}{(\sigma_t^2 - A\sigma_g^2)} = \frac{h^2(1 - A)}{(1 - Ah^2)}$$

Assortative mating has increased additive genetic variance by a factor $1/(1 - A) = 1.285$. The heritability in the random-mating population would be 0.747, rather than the 0.791 observed in the EAM population. Then also $V_t = \sigma_t^2(1 - Ah^2) = 0.824\sigma_t^2$. The total variance is increased by a factor $1/0.824 = 1.21$.

2.4 | Assortment due to population structure

Also established in the early population genetics literature is that genetic differences across subpopulations cause an increased genotypic variance, increased homozygosity, and LD relative to a random-mating homogeneous population (Wahlund, 1928). *CF68* draw analogies between phenotypic assortative mating and inbreeding, whether due to population subdivision or consanguineous marriages. Gimelfarb (1981) shows that not only does assortative mating affect correlations between relatives, but that it also itself affects the frequencies of certain types of consanguineous marriages. Relative to inbreeding, assortative mating for a trait with multiple contributing variants causes a smaller increase in homozygosity but a larger increase in phenotypic variance. Inbreeding per se does not cause a systematic change in haplotype frequencies, unlike assortative mating which creates associations favoring haplotypes contributing to extreme trait values. Of course, while inbreeding affects the genotypic structure at all segregating loci, at equilibrium under assortative mating only causal loci are affected. However, for a trait such as height, this difference is moot.

In general, there are many potential causes of apparent assortative mating as evidenced by a positive correlation between the trait values of mates. The environment may produce effects on traits that are subject to assortment, while shared environment throughout marriage can also result in correlations between mates. Assortment may not be directly on the trait phenotype of interest, but on an associated trait, and *F18* comments that if the assortment is not directly on the

phenotype, the association will appear "*somewhat masked by environmental effects in the observed [mate] correlation*." Population subdivision itself creates apparent assortative mating, in that individuals within subdivisions are more genetically similar than individuals in different subdivisions. In this case, the causes of correlations between mates are more directly related to allelic or genotypic values than to phenotype.

In fact, *F18* develops three "theories" of assortative mating, corresponding to assortment on the basis of phenotype, on the basis of genotypic values, and on the basis of the genic (additive genetic) values $\mu + \beta_\ell x_{i\ell}$ of Equation (2). Under the three cases, the correlations in phenotypic, genotypic, and genic values of mates, required to produce an observable mate phenotypic correlation ρ_m are, respectively, ρ_m, ρ_m/H^2, and ρ_m/h^2. Here again $h^2 = \sigma_g^2/\sigma_t^2$ is the narrow-sense heritability at equilibrium in the assortative mating population, and $H^2 = (\sigma_g^2 + \sigma_d^2)/\sigma_t^2$. The correlations between relatives differ in the three cases. As derived by *F18*, the parent–offspring correlations are respectively:

$$h^2(1 + \rho_m)/2,\ h^2(1 + \rho_m/H^2)/2 \quad \text{and} \quad (h^2 + \rho_m)/2$$

F18 comments that the third case leads to "*results in some respects more intelligible and in accordance with existing knowledge*." In this case, using again the same data as before, $\rho_m = 0.2804$, $\rho_p = 0.5066$, *F18* calculates $h^2 = (2\rho_p - \rho_m) = 0.7328$, $A = \rho_m/h^2 = 0.3826$, and $(1 + A)/2 = 0.691$. The sib correlation is again given by Equation (7), but in terms of ρ_p and ρ_m this now takes the form

$$\begin{aligned} \rho_s &= (1/4)(H^2 + ((2\rho_p - \rho_m) + 2\rho_m) \\ H^2 &= 4\rho_s - 2\rho_o - \rho_m \end{aligned}$$

Substituting $\rho_s = 0.5433$ and the other two correlations as before, $H^2 = 0.8796$. The value of broad-sense heritability is reduced from the earlier value of 100% but is still high, leaving only a small portion for the environment.

Even in the classic data of Pearson and Lee (1903), the effects of population subdivision and differential environments were probably greater than Fisher realized. Cultural and socioeconomic differences, resulting in differences in childhood nutrition, have a significant impact on traits such as height (Cole, 2003). In current large-scale studies of heterogeneous populations, the differences are likely no less. Unless population structure is corrected for, genetic and environmental differentiation will contribute to apparent high correlations between closer relatives as compared with their more distant kin.

3 | FROM GENETIC MARKERS TO GENOME-WIDE SNPs

3.1 | Variance component models for heritability

From *F18* onwards, correlations between relatives have informed heritability studies. However, the goals and methods for the analysis of quantitative genetic traits diverged between human genetic studies and livestock breeding and agriculture. In animal and plant breeding, the primary goal of heritability studies was to assess selection potential and to design optimal selection programs. The variance component and path analysis approaches initiated by Wright (1921) were widely used from Kempthorne (1957) to Falconer and MacKay (1996).

In human genetics also, correlations between relatives were used to estimate the components of genetic variation and the relative contributions of genes and environment. However, in natural populations, environmental effects are harder to control or measure. Twin studies offered a way forward (Hopper, Foley, White, & Pollaers, 2013), and have been widely used especially in behavioral genetics (Plomin, DeFries, Knopik, & Neiderhiser, 2014). In the absence of other relatives, or of twins reared apart, a necessary assumption is that the effect of the shared environment is the same for dizygous (DZ) and monozygous (MZ) twin pairs. With more types of relatives available, more components are identifiable, but, for example, if the only bilateral relatives in the sample are sibs, dominance variance cannot be distinguished from the shared sib environment. Hopper (1993) gives a thorough discussion of both of the flexibility and limitations of these models as applied in medical genetic studies, while Wang, Guo, He, and Zhang (2011) have given a more recent analysis of both identifiability of effects and the properties of likelihood-based statistics that are used to test for effects.

A pedigree provides a variety of relationship types, and Lange, Westlake, and Spence (1976) first developed methods for likelihood-based segregation analysis of data on small pedigrees using variance component models. Across individuals, the trait data are multivariate Normal, so that these models can be applied to data on large and complex pedigrees, and it is also possible to fit models of shared environment and allow covariances to depend on measured covariate factors (Hopper, 1993). However, here we consider only additive genetic effects. In this case, following the notation of Section 2.2, the total phenotypic variance is σ_t^2, the additive genetic (heritable) variance is σ_g^2, and $h^2 = \sigma_g^2/\sigma_t^2$. The environmental contribution to variance σ_e^2 is often assumed uncorrelated between individuals. Then, disregarding

inbreeding and shared environment, under the simple additive model, and on a known pedigree, the variance of the vector of trait values on a set of N relatives is

$$2\Phi\sigma_g^2 + \mathbf{I}\sigma_e^2 = \sigma_t^2(2\Phi h^2 + (1 - h^2)\mathbf{I}), \quad (8)$$

where 2Φ is the matrix of pedigree-based expected pairwise genome-wide IBD proportions.

For Normal data with the covariance structure (8), the expected log-likelihood is most conveniently analyzed through the eigenvalues λ_i, $1 = 1, ...N$ of 2Φ. This was used by Thompson and Shaw (1990) to develop an EM algorithm for the estimation of h^2. More recently, Blangero et al. (2012) have used the same approach to give the noncentrality of the χ_1^2 test statistic for testing for nonzero h^2 as $-\sum_1^N \log(1 + h^2(\lambda_i - 1))$. Raffa and Thompson (2016) have given more general results enabling confidence interval estimates for h^2 by inverting the χ_1^2 test statistic for testing a hypothesized null value h_0^2. They approximate the noncentrality parameter for the χ_1^2 test statistic in terms of the variance of the log-eigenvalues of 2Φ. If the geometric mean ν of the eigenvalues is close to 1, their approximation can be simplified to

$$\frac{1}{2}\sum_{i=1}^{N}(\log(\lambda_i) - \log(\nu))^2\,(h^2 - h_0^2)^2. \quad (9)$$

Hence accuracy of estimation of h^2 is very sensitive to multiple small eigenvalues of 2Φ, for which the absolute values of $\log(\lambda_i)$ are large.

3.2 | Mapping quantitative trait loci

In human genetics, once a trait was shown to have a genetic component, a goal was often to use linkage studies to map the quantitative trait loci (QTL) q at which there are variants of major effect. One approach to the mapping of these QTL is to relate the estimated probabilities of IBD at specific marker loci j to correlations or differences in trait values of relatives. Haseman and Elston (1972) regressed the squared difference between sib phenotypes on the proportion of marker alleles shared IBD at a marker locus. Under an additive genetic model, the expectation of the slope of the regression line is

$$-2(1 - 2\theta_{qj})^2\sigma_q^2, \quad (10)$$

where θ_{qj} is the recombination fraction between marker locus j and trait locus q, and σ_q^2 is the additive genetic variance attributable to variants at the trait locus q. An important feature of the Haseman-Elston approach is that, by modeling a statistic quadratic in the phenotypic values, the variance σ_q^2 becomes embedded in the regression coefficient (10), a mean measure, rather than in the covariance matrix (for example, Equation (8)). There have been many newer versions of Haseman-Elston regression over the years, to increase power or to deal with ascertainment (Sinha & Gray-McGuire, 2007; Wang & Elston, 2004), but we do not pursue these here.

With the advent of genome-wide genetic marker maps in the 1980s and 1990s, methods of pedigree-based mapping advanced, but only with the methods of Almasy and Blangero (May 1998) did the explicit QTL effects of *F18* and the variance model of Wright (1921) become fully combined in the human gene mapping literature. Their models allow for not only additive and dominance effects, but also interactions, and other variance components such as those due to shared environment. However, in practice, analysis is usually restricted to local and genome-wide additive effects. Equation (8) for the covariance matrix for the vector of trait observations now includes also a locus-specific term. On a set of relatives, the covariance matrix for their trait values becomes

$$\sum_{q=1}^{Q}\widehat{\Pi_q}\sigma_q^2 + 2\Phi\sigma_g^2 + \mathbf{I}\sigma_e^2, \quad (11)$$

where $\widehat{\Pi_q}$ is the matrix of estimated pairwise proportions of IBD at locus q with additive genetic variance contribution σ_q^2, and the remaining terms of the covariance are as in Equation (8). Since trait values are directly modeled, in (11) σ_q^2 enters into the covariance matrix, rather than into the regression coefficient of (10). As genetic marker data became increasingly available, the estimates $\widehat{\Pi_q}$ were also readily obtained, for example by use of programs such as *Merlin* (Abecasis, Cherny, Cookson, & Cardon, 2002).

3.3 | Finding Fisher's causal variants with genome-wide association studies

As it became possible to identify increasing millions of single-nucleotide polymorphisms (SNPs) in human genomes, it became possible, in theory, to identify the multiple causal variants postulated by *F18*, or at least variants that are in strong association (LD) with causal variants. This is the goal of genome-wide association studies (GWAS). The basic GWAS approach takes no account of either relationships between sample members, nor among loci across the genome. Each SNP is tested for association with a phenotype of interest.

Considering only the effect on phenotype y_i of variants at a single locus ℓ, and in the absence of dominance, the model of *F18* in Equations (1) and (2), becomes

$$y_i = G_{i\ell} + e_i = \mu + \beta_\ell x_{i\ell} + e_i, \tag{12}$$

where $G_{i\ell}$ is the genetic effect at locus ℓ in individual i, and e_i is an independent environmental effect, $\text{var}(e_i) = \sigma_e^2$. The variance of x_{il} is $2p_l(1-p_l)$, so under the fitted linear model the regression slope estimate is Normal with mean β_ℓ and variance $\sigma_e^2/(2Np_\ell(1-p_\ell))$. In testing the null hypothesis of zero effect at locus ℓ the noncentrality of the χ_1^2 statistic is

$$\begin{aligned} 2Np_\ell(1-p_\ell)\beta_\ell^2/\sigma_e^2 &= N\sigma_\ell^2/\sigma_e^2 \approx N\sigma_\ell^2/(\sigma_t^2-\sigma_\ell^2) \\ &= Nh_\ell^2/(1-h_\ell^2). \end{aligned} \tag{13}$$

because $\sigma_\ell^2 = 2p_\ell(1-p_\ell)\beta_\ell^2$ (Equation (4)). Because we disregard dominance, the total phenotypic variance of Equation (5) is $\sigma_t^2 \approx \sigma_\ell^2 + \sigma_e^2$, and heritability $h_\ell^2 = \sigma_\ell^2/\sigma_t^2$. In a GWAS study for a trait such as height, the genetic variance attributable to any one QTL will be small, so then $(1-h_\ell^2) \approx 1$, and (13) becomes Nh_ℓ^2.

If the marker j tested is not itself the QTL ℓ but is in association with it, the power to detect the QTL is reduced, as in Equation (10). In the context of LD and population samples, the linkage term $(1-2\theta_{qj})^2$ of (10) is replaced by the LD measure of allelic association $r_{\ell j}^2$ between the QTL ℓ and marker locus j. The expected regression slope in regressing phenotype on marker j is $r_{\ell j}\beta_\ell$. The proportion of phenotypic variance explained by the SNP marker j is $r_{\ell j}^2 h_\ell^2$, so the noncentrality of the χ_1^2 test statistic becomes $Nh_\ell^2 r_{\ell j}^2$ (Chen, 2014). As in the comparison of Equations (10) and (11), there is a key difference between the regression of Haseman and Elston (1972) and the GWAS regression (12). The first models squared phenotypic differences so that σ_q^2 enters into the expectation of the regression coefficient (10), while (12) models phenotype so that σ_t^2 enters the standard deviation of the regression coefficient and the noncentrality (13) of the χ_1^2 distribution. Chen (2014) provides a much fuller exposition and comparison.

Early GWAS suffered from small sample sizes and inadequate power. Larger ones, using data from multiple or inhomogeneous populations, suffer from genetic heterogeneity. In both case, the testing of large numbers of SNPs causes problems of multiple testing. Determining an appropriate level of genome-wide significance and methods of genomic control to account for population heterogeneity were much-discussed issues (Devlin & Roeder, 1999). There is, of course, a huge literature in this area, but that is not the topic here. While many significant and replicable associations were detected by GWAS, the proportion of trait heritability accounted for by these genes was often very low (Manolio et al., 2009).

Low GWAS-based estimates of heritability are particularly apparent for human height. Although the analysis of *F18* may overestimate heritability, other studies of close relatives have also provided estimates of the order of 80%. Even allowing for inflation by effects of shared environment or epigenetic factors that also contribute to correlations in close relatives, this is far higher than detected GWAS effects can account for. Lango et al. (2012) undertook a meta-analysis of 46 earlier studies, with a combined total of data on over 183,000 individuals. Their approach selected SNPs representing 180 loci each showing a robust significant signal of association with human height. In their analysis, these SNPs explain only 10% of the phenotypic variation in height, while they estimate that unidentified common variants of similar effect size would increase this to 16%. However, this would still be only 20% of the presumed heritable variation. The larger more recent meta-analysis of Yengo et al. (2018) of 700,000 individuals finds almost 3,300 "near-independent" SNP variants with statistically significant effects on height. However, these GWAS SNPs still explain less than 25% of the phenotypic variance in an independent sample of similar ethnicity.

4 | FROM GENETIC MAPS TO GENOMES

4.1 | Realized relatedness

As denser marker data became increasingly available, it became possible to estimate proportions of genome shared IBD both at specific hypothesized causal loci, and also genome-wide. While pedigree relationships provide location-specific probabilities and genome-wide expectations, meiosis is highly variable and human genomes are short (Thompson, 2013), so that realized proportions of genome shared IBD by relatives will differ from pedigree expectations. Genetic marker data can be used to estimate IBD either in conjunction with the pedigree relationship, or in the absence of any known relationships. In an analysis of the heritability of height from sib-pair data, Visscher et al. (2006) proposed replacing pedigree-based kinship coefficients with an estimate of the realized proportions of genome shared IBD, potentially gaining information from the variation in IBD across sib pairs.

In the animal-breeding area, versions of the genomic relatedness matrix or (GRM) have been widely used for breeding value prediction replacing the pedigree-based matrix of Equations (8) and (11) (Hayes, Visscher, &

Goddard, 2009; van Raden, 2008). The GRM measures the pairwise genotypic similarity between individuals. Given genotype data at M SNPs, and population reference allele frequencies p_ℓ, ($\ell = 1,...,M$), let $x_{i\ell}$ be the allelic dosage (0, 1 or 2 copies) of the reference SNP allele for individual i at locus ℓ. Then the expectation of $x_{i\ell}$ is $2p_\ell$ and, in the absence of inbreeding and under random mating, the variance is $2p_\ell(1-2p_\ell)$. The general GRM form is

$$\Omega_{ik} = \sum_{\ell=1}^{M} w_\ell \frac{(x_{i\ell} - 2p_\ell)(x_{k\ell} - 2p_\ell)}{2p_\ell(1-p_\ell)} = \sum_{\ell=1}^{M} w_\ell S_{i\ell} S_{k\ell}, \tag{14}$$

where w_ℓ are a set of nonnegative weights summing to 1, and $S_{i\ell} = (x_{i\ell} - 2p_\ell)/\sqrt{2p_\ell(1-2p_\ell)}$ is the standardized genotype of individual i at locus ℓ.

In an infinite idealized population, the expected value of $S_{i\ell}S_{k\ell}$ is $2\Phi_{ik}$, so any weights w_ℓ in (14) provide an unbiased estimate of relatedness relative to the current population. While the most usual form has $w_\ell = 1/M$ for all ℓ, other forms have been used for robustness against extreme p_ℓ-values (van Raden, 2008), for greater statistical efficiency, and/or to accommodate LD (Speed, Hemani, Johnson, & Balding, 2012; Wang, Sverdlov, & Thompson, 2017). Methods to estimate location-specific IBD probabilities between individuals from population data have also been established (Brown, Glazner, Zheng, & Thompson, 2012). If local IBD is estimated across the genome, then clearly a genome-wide estimate also follows. Wang et al. (2017) compare the results of several approaches to the estimation of genome-wide relatedness in the absence of pedigree information.

4.2 | Using all SNPs: Total additive genetic variance

Trait-associated SNPs identified by GWAS (Section 3.3) usually explain only a small fraction of heritable variation. Variants that are either rare or of small effect will not be identified. In their development of the GCTA approach, Yang, Lee, Goddard, and Visscher (2011) use large population samples of individuals each genotyped at a large number of SNPs, genome-wide. To accommodate all SNPs, a random effects model is used, with each SNP's additive effect a_ℓ having mean 0 and variance σ^2. The model is the linear mixed model for the phenotype vector $\mathbf{y}$:

$$\begin{aligned} \mathbf{y} &= \mathbf{Z}\gamma + \mathbf{S}\mathbf{a} + \mathbf{e} \\ \mathrm{var}(\mathbf{y}) &= \sigma^2 \mathbf{S}\mathbf{S}' + \sigma_e^2 \mathbf{I}, \end{aligned} \tag{15}$$

where $\mathbf{S}$ is the $N \times M$ matrix of $(S_{i\ell})$ in (14) and γ is a set of fixed effects coded in $\mathbf{Z}$, such as age, sex, or components of population structure.

Then from Equation (14) with $w_\ell = 1/M$ for all ℓ, $\Omega = \mathbf{S}\mathbf{S}'/M$ and the total variance explained by all SNPs is $\sigma_g^2 = M\sigma^2$, so that (15) becomes

$$\mathrm{var}(\mathbf{y}) = \sigma_g^2 \Omega + \sigma_e^2 \mathbf{I}.$$

The model is then directly analogous to the classical pedigree-based quantitative trait analysis (eEquation (8)), and can be fit using REML methods (Falconer & MacKay, 1996). The only difference is that the pedigree-based matrix 2Φ is replaced by the marker-based estimate Ω. The form of Ω as an average over SNPs ℓ, allows the SNPs to be subdivided. Thus the total genetic variance σ_g^2 may be partitioned into the contributions from different regions of the genome, simply by partitioning Ω. For example, the heritable variance may be partitioned by chromosome (Yang et al., 2011) or by functionally defined categories of SNPs (Gusev et al., 2014).

In the GCTA approach, Ω is regarded as the realized relatedness in a population sample of remotely related individuals. The model in principle admits the use of close relatives, but these would dominate the phenotypic correlations, and would be more subject to effects of shared environment which are not included in the model. Whether the pedigree relatedness (Equation (11)) or the GRM (Equation (15)) is used to model the covariances, any additive effects shared by close relatives will contribute to the estimate of heritability. In practice, therefore, the method is applied to large population samples of N individuals not known to be related. While GCTA aims to exploit the structure of remote relatedness within a population, it must also guard against heterogeneity among subpopulations. Typically, therefore, several leading eigenvectors of a Principal Components Analysis (PCA) of the genotypic variation are included in the fixed effects γ (Yang et al., 2011).

Using a variance matrix based on 294,831 SNPs in 3,925 remotely related individuals, Yang et al. (2010) explain 45% of the phenotypic variance in height, so some heritability remains "missing." Yang et al. (2010) suggest that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, including multiple rare alleles of small effect. However, there are many other possible causes of low levels of correlation between remote relatives as compared with the parent–offspring, fraternal, avuncular, and cousin relationships of *F18* and pedigree studies. In addition to shared environment and other familial effects, potential genetic causes include epistasis between causal variants

(Zuk, Hechter, Sunyaev, & Lander, 2012). Increasing the number of SNPs included in the GRM (14) may not capture all relevant heritable variation. Even if based on whole-genome sequence, it may not capture possible functional genetic similarities of segments of genome shared by descent: In remote relatives, such segments are few, but often quite large (Thompson, 2013). Epigenetic factors and chromosomal structural variants may also have effects that decay faster with decreasing relatedness than is predicted for Mendelian factors.

As numbers of individuals and numbers of SNPs included in analyses are increased, the issues with PCA adjustments for population structure in the model (15) become greater. Conomos, Reiner, Weir, and Thornton (2016) develop a more flexible approach to consider samples from individuals with diverse and more complex ancestry, including admixture. Their adjustment for population and relatedness structure partitions the genotypic correlations between individuals into separate components representing more distant and more recent common ancestry. Heckerman et al. (2016) show that including additional random effects in the model (15) using spatial location as a surrogate for unmeasured environmental factors reduces estimates of additive genetic heritability. However, there are also issues in adjusting for population heterogeneity, since this heterogeneity may be itself a reflection of population-level relatedness between individuals and indistinguishable from effects of gene–environment interaction.

Increasing the numbers of individuals increases the number of variants in the sample, including rare variants, and hence captures more of the genetic variance. However, it also brings greater heterogeneity of both genes and environment, and also computational challenges in the dimension of the matrix Ω. Whether pedigree-based (8) or genotype-based (15), estimates of heritability are sensitive to the structure of the covariance matrix. Precision of estimation is dependent on the eigenvalues of this matrix and especially on the small eigenvalues (Equation (9)). As the number of individuals becomes large, the eigenvalues of the GRM will typically have a highly skewed distribution, with many small values Kumar, Feldman, Rehkopf, and Tuljapurkar (2016). Uncertainties in the estimation of these eigenvalues leads to uncertainty in the resulting estimates of h^2.

4.3 | Using all SNPs: LD-score regression

The method of LD-score regression (Bulik-Sullivan, Loh, Finucane, Ripke, & Yang, 2015) takes a different approach to estimate the total additive genetic variance σ_g^2 or $h^2 = \sigma_g^2/\sigma_t^2$, using genotype data for large numbers of SNPs in a large number of individuals. The method does not require individual phenotypes, but instead uses the values of the GWAS χ_1^2 test statistics testing for an effect of each SNP variant (Equation (13)). The method also does not require inversion or even computation of the realized relatedness matrix Ω (Equation (14)). As population sample size N and numbers of SNPs M become ever larger this is a significant computational advantage. The model for the vector of phenotypic observations $\mathbf{y}$ on the observed individuals is the original model of *F18* (Equations (1) and (2)):

$$\mathbf{y} = \mathbf{S}\beta + \mathbf{e}, \tag{16}$$

where again $\mathbf{S}$ is the $N \times M$ matrix of $(S_{i\ell})$, the standardized genotype of individual i at SNP marker ℓ (Equation (14)). The phenotypic values are also standardized: The elements y_i have mean 0 and variance 1. The terms $\mathbf{S}$, β, and $\mathbf{e}$ in (16) are all considered random, with mean 0. The variance of $\mathbf{e}$ is $(1 - h^2)\mathbf{I}$ (Equation (8)). The total heritable variance explained by M SNPs is h^2 so the variance of β is $(h^2/M)\mathbf{I}$.

In LD-score regression, the view of $\mathbf{S}$ is orthogonal to that of GCTA, in that $S_{i\ell}$ are considered independent over individuals i, but the dependence of $S_{i\ell}$ among SNPs ℓ is at the core of the approach. The LD correlation measure $r_{\ell j}$ is the expected value of $S_{i\ell}S_{ij}$, and the LD-score of variant ℓ is defined as $L_\ell = \sum_{j=1}^{M} r_{\ell j}^2$. The regression slope estimate is $b_\ell = (1/N)S_\ell' y$, where S_ℓ is the $N \times 1$ vector of elements $S_{i\ell}$. With the centered variables, the expected value of b_ℓ is 0, and the expected value of the GWAS test statistic for SNP ℓ is $N\mathrm{var}(b_\ell)$. Given $\mathbf{S}$, the expected value of b_ℓ is again 0 so that the variance of b_ℓ becomes the expected conditional variance give $\mathbf{S}$. Then as shown in Bulik-Sullivan et al. (2015) (Supporting Information Note),

$$\begin{aligned} N^2\mathrm{var}(b_\ell|S) &= \mathrm{var}(S'_\ell y|\mathbf{S}) = S'_\ell \mathrm{var}(y|\mathbf{S})S_\ell \\ &= S'_\ell((h^2/M)SS' + (1 - h^2))S_\ell \\ &= (h^2/M)(S'S_\ell)'(S'S_\ell) + N(1 - h^2). \end{aligned}$$

The $M \times 1$ vector $(S'S_\ell)$ has components $N\hat{r}_{\ell j} = \sum_{i=1}^{N} S_{ij}S_{i\ell}$, where $\hat{r}_{\ell j}$ is the genotypic LD correlation between loci ℓ and j. Then

$$N\mathrm{var}(b_\ell|S) = (Nh^2/M)\sum_{j=1}^{M} \hat{r}_{\ell j}^2 + (1 - h^2).$$

Note that just as Equation (16) is a multivariate form of Equations (2) and (12), this variance derivation is a multivariate form of Equation (13).

The sample squared correlation $\hat{r}_{\ell j}^2$ will overestimate the population $r_{\ell j}^2$: $\mathrm{E}(\hat{r}_{\ell j}^2) \approx r_{\ell j}^2 + (1 - r_{\ell j}^2)/N$. Summing

over SNPs j, we obtain the expected GWAS test statistic for effects at locus ℓ:

$$N\ \mathrm{E}(\mathrm{var}(b_\ell|S)) = (Nh^2/M)(L_\ell + M/N) + (1 - h^2) = (Nh^2/M)L_\ell + 1 \quad (17)$$

In practice the summation over SNPs j in computing L_ℓ is restricted to M SNPs within (say) 1 cM of SNP ℓ. It is assumed that beyond this range any sample LD is the result of sampling variation, population heterogeneity, or factors such as selection.

As for the original GWAS analyses, population heterogeneity inflates the expected value of the χ^2_1 test statistic. Heterogeneity adds a term which depends on both the phenotypic and the allelic frequency differences among subpopulations, but the term (Nh^2L_j/M) in (17) remains unchanged (Bulik-Sullivan et al., 2015). Thus regressing GWAS summary test statistics on LD scores provides an estimate of h^2 that is robust to population structure.

As for GCTA, partitioning of heritability by functional annotation is possible using a stratified version of LD-score regression (Finucane et al., 2015). In this case, the partitioning of SNPs is within the LD-score L_ℓ. The term $(h^2/M)\sum_j r^2_{\ell j}$ is replaced by $\sum_C \tau_C \sum_{j\in C} r^2_{\ell j}$, where τ_C is the per-SNP heritability in category C. Rather than a regression on L_ℓ to estimate the single h^2, there is then a multiple regression of the the GWAS χ^2_1 test statistics on the $\sum_{j\in C} r^2_{\ell j}$ for categories C.

The method of LD-score regression produces higher estimates of heritability than earlier analyses using only GWAS SNPs that meet genome-wide significance, but still less than those of the variance component methods. Evans et al. (2018) provides an overview of many current methods for estimation of h^2, and comparisons of performance under different types of genotypic data, different causal variant frequencies, and different population structure models. These are of interest, but the focus here is the continuing difference of perspective between Fisher's models of effects (including LD effects) associated with individual genetic variants, as compared with Wright's variance component approach in which these are subsumed.

4.4 | IBD, state, or function?

Whether phrased in terms of genic effects of multiple variants as in *F18* or via the path coefficients of Wright (1921), the idea of an underlying population in which alleles descend to relatives, thereby creating genotypic and phenotypic correlations is fundamental to almost all methods of genetic analysis of quantitative traits. For defined relationships, or within a defined pedigree, this model is explicit in the assumed pedigree or relationship structure.

With recent population-based methods, in the absence of a defined pedigree, the interpretation of models of phenotypic variation are less clear. While the GRM $\mathbf{\Omega}$ may be considered an estimate of the relatedness by descent relative to the current population, it is in fact simply a measure of genotypic or allelic correlation. Powell, Visscher, and Goddard (2010) propose that, rather than considering IBD relative to some past time-point, IBD should be *defined* via these correlations in allelic type between gametes or between individuals. Nonetheless, these correlations are still viewed as deriving from the within-population structure of descent from common ancestors to descendants. More recent discussions of the relationship between IBD as defined by the ancestral coalescent of genome and the SNP-based measured of relatedness based on extant population data are given by Thompson (2013) and by Speed and Balding (2015).

In the case of GWAS and related methods, there is no explicit dependence on a model of gene descent in the population. However, although the individuals are considered independent, the LD associations that are the basis of the methods again arise from descent. In the case of local LD, variants arising on a given genetic background remain in association over many generations, due to the generation-to-generation descent of genome in very large blocks. GWAS methods are based on this local LD between causal variants and SNP genotypes. Differentiation among populations results in long-range LD, which must be accounted for. The method of LD-score regression uses only local SNPs in computing an LD score, because, as for the GWAS statistics that it uses, it depends on the LD maintained by the shared descent of variants over many generations.

Only approaches that measure identity by state (IBS), without modeling the source of this identity, avoid this population-dependence. Prediction takes this view, in that it is not only irrelevant what is the source of genetic and phenotypic variation, but also whether the predicting variants are causal. Relatives may show covariance for many reasons, both genetic and nongenetic. There may be multiple multi-allelic associated or linked causal loci, which may exhibit dominance and epistasis. Additionally, there may be environmental effects, and correlated environments, gene–environment interaction, population structure, and epigenetic effects. Nonetheless, a simple additive model, fit to such a trait, can reflect the relationship between genic and phenotypic values and provide effective phenotypic predictions. While the predictive weights attributed to specific markers may have little biological relevance, the model may have high

out-of-sample predictive accuracy. This is the approach taken in the genomic analyses of livestock and developed by de los Campos, Gianola, and Allison (2010).

Predictive accuracy is assessed by the predictive R^2, the proportion of phenotypic variance explained in the regression of observed values on the prediction. Heritability provides an upper bound on predictive R^2. As sample size tends to infinity, the bound is theoretically attainable. Where, as for height, there are large numbers of genes of small effect, then the accuracy of predictions from population samples is low. Samples containing a significant proportion of close relatives provide much higher whole-genome predictive accuracy (Makowsky et al., 2011), just as they provide much higher estimates of heritability.

Recent developments in prediction methods from population samples show improved predictive accuracy (Lello et al., 2018). On a large sample of 453,000 training-set individuals, from an initial set of 645,000 SNPs, their final height predictor contains about 20,000 active SNPs, and achieves a predictive R^2 of about 40%. Because this is close to the common-SNP heritability estimate of Yang et al. (2011), they suggest that their methods bring prediction for height close to the asymptotic bound, closing the gap between heritability estimates and predictive accuracy in large population samples. However, prediction, like GCTA and LD-score regression, is population-dependent. A predictor assigns no biological interpretation to the effects of the active SNPs, and out-of-sample accuracy provides no basis for across-population performance.

Sverdlov and Thompson (2013) develop an alternative approach that seeks to avoid this population dependence. Rather than asking: What is the correlation between the phenotypes in these individuals given they are uncle and nephew? They ask: What is the correlation between the phenotypes in these individuals given they both carry this collection of genetic variants. In this approach of identity by function, it is the sharing of functional variants that underlies phenotypic similarity. Unlike the variance component models, the model is generative in having a joint distribution of genotypes and effects, given a mutation process for the creation of variants and a trait model of the phenotypic contributions of variant effects. In this sense, it reflects directly back to *F18*, but the randomness is no longer in the process of meiosis, but in the effects of particular variants.

5 | SUMMARY: FROM THE PAST TO THE FUTURE

One hundred years ago, Fisher (1918) postulated that quantitative phenotypic variation could be explained by the additive effects of a very large number of Mendelian factors, each of very small effect. He obtained formulae for the heritable components of this variation that affect observable phenotypic correlations between related individuals. At that time, few genes were known, genetic linkage was barely established, and there were no genetic maps. Although Fisher's analysis was general in considering dominance and epistasis, it was limited by considering only equilibrium in an idealized infinite population in which genetic linkage does not affect results.

An alternate methodology was developed by Wright (1921), using what would later become his theory of path coefficients. Fisher's postulated distinct causal variants are subsumed into a single heritable variance component, the additive genetic variance. Regression modeling using Wright's approach was the foundation of much development of quantitative genetic modeling in plant and animal breeding, where the goals were the prediction of breeding values, estimation of selection potential, and design of selection programs.

Human statistical genetics took a different route, using data on related individuals to map Mendelian genes of large effect. As genetic markers and genetic maps were developed, so also was QTL mapping in human pedigrees. Although causal genes were determined for many disease traits, and some major QTL were mapped, it was only with the advent of more widely available SNP data around 2000 that the potential to find Fisher's postulated causal variants of small effect emerged. GWAS undertook this challenge, but early GWAS sample sizes and analysis methods were insufficient for the task. As numbers of individuals N, and number of markers M because larger, and as methods to account for population structure were developed, GWAS had much greater success. However, the proportion of total heritable variation that could be explained by SNP variants of significant effect remained low.

The goal of Fisher (1918) was to show that his Mendelian model could explain observed phenotypic correlations for biometric traits in humans. To explain high sib correlations in height, he developed three theories of assortative mating in parents: by phenotype, by genotype, and by genic value, the latter two being equivalent under a purely additive model. While there is some assortative mating for height, the apparent high sib correlations are likely due to shared environmental effects, as aspect not considered by Fisher. Subsequent modeling of quantitative traits in related individuals (whether of humans, plants or animals) have placed much greater emphasis on environmental effects.

Fisher's third theory, that parental correlations are at the genic level are more directly relevant today. Population substructure inflates phenotypic variance and results in genotypic correlations in remote relatives. It is these

genotypic correlations between individuals not known to be related that are exploited in the GCTA approach (Yang et al., 2011), although the broader population-level structure is accounted for by, for example, PCA adjustments. In contrast to the GCTA approach in which the focus is on correlations between individuals, the initial development of GWAS methodology took a different view. Correlations between individuals are not modeled: individuals are considered independent. Instead, the LD correlations between SNP variants resulting from population-level remote ancestry is the focus, and population-level structure is a concern addressed, for example, by methods of genomic control. This view carries over to methods of LD-score regression, which uses GWAS summary statistics, and in which population-level structure does not impact the regression slope.

Rather than detecting individual variants, both GCTA and LD-score regression share the goal of explaining the total additive genetic heritability of a quantitative trait. To do so, they model the combined effects of all SNPs or even ultimately all variations in the genome sequence. While the contribution of individual variants will seldom be detectable, there are multiple variants within a functional gene; genes having significant causal effects may be detected. However, genomes are finite. New variants arise with every meiosis, most never to achieve polymorphic frequencies. A variant present in only a single member of the sample occurs on a single genetic background with reference to other local variants and is not shared by any other individual. Populations are also finite. The models of population genetics relate to probabilities in a hypothetically infinite population, or to probabilities over the process of evolution in a population. Evolution happened once only, all populations are related, and as sample size approaches population size, even the meaning of many population parameters becomes unclear. Above all, neither populations nor genomes are in equilibrium.

For 100 years, the equilibrium, infinite-population model of Fisher (1918), with quantitative traits resulting from an infinite number of variants each of infinitesimal effect, has been a powerful theoretical model in the analysis of the total heritable variation in quantitative traits in many species. However, Fisher's ultimate goal was the *discovery of biological knowledge by quantitative methods* (Fisher, 1948). With not only the availability of whole-genome sequence, but also increasing data on structural variants, epigenetic factors, and variation in DNA transcription, there are new opportunities for quantitative analysis of heritable quantitative variation. Even though quantitative traits can often be represented by a linear genetic model, at least on a transformed phenotypic scale (Sverdlov & Thompson, 2018), the underlying biological system is often highly nonlinear. In addition to environmental and social effects, trait values of closer relatives may be more highly correlated for many biological reasons, The larger segments of DNA shared IBD by closer relatives may lead to the much greater similarity of function than the same amount of sequence identity at a population level. Resolving these issues remain exciting challenges for the future.

ACKNOWLEDGMENTS

I am grateful to Saonli Basu for many extended discussions of current methods and approaches. I am also grateful for helpful discussions with Joe Felsenstein, James Lee, Serge Sverdlov, and Ellen Wijsman. I appreciate the useful detailed comments of two referees and thank them for their perspectives and thorough reading. The work on this paper was supported in part by NIH grant R37 GM046255.

ORCID

Elizabeth A. Thompson http://orcid.org/0000-0003-0198-7129

REFERENCES

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, *30*, 97–101.

Almasy, L., & Blangero, J. (May 1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, *62*, 1198–1211.

Blangero, J., Diego, V. P., Dyer, T., Almeida, M., Peralta, J, Kent Jr, J. W., & Göring, H. (2012). A kernel of truth: Statistical advances in polygenic variance component models for complex human pedigrees. *Advances in Genetics*, *81*, 1–31.

Brown, M. D., Glazner, C. G., Zheng, C., & Thompson, E. A. (2012). Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics*, *190*, 1447–1460.

Bulik-Sullivan, B. K., Loh, P., Finucane, H. K., Ripke, S., & Yang, J., Schizophrenia working group of the psychiatric genomics consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47*, 291–295.

Chen, G. (2014). Estimating heritability of complex traits from genome-wide association studies using IBS-based HasemanElston regression. *Frontiers in Genetics*, *5*

Cole, T. J. (2003). The secular trend in human physical growth: A biological view. *Economics and Human Biology*, *1*, 161–168.

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free estimation or recent genetic relatedess. *American Journal of Human Genetics*, *98*, 127–148.

Cotterman, C. W. (1940). P. A. Ballonoff (Ed.), *A calculus for statistico-genetics* (1974). New York: Acaddemic Press.

Crow, J., & Kimura, M. (1970). *An introduction to population genetics theory*. New York: Harper and Row.

Crow, J. F., & Felsenstein, J. (1968). The effect of assortative mating on the genetic composition of a population. *Eugenics Quarterly*, *15*, 85–97.

de los Campos, G., Gianola, D., & Allison, D. B. (2010). Predicting genetic predisposition in humans: The promise of whole-genome markers. *Nature Reviews Genetics*, *11*, 880–886.

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*, 997–1004.

Evans, L. M., Tahmasbi, R., Vrieze, S. I., Abecasis, G., Das, S., Bjelland, D., & Keller, M. C. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, *50*, 737–745.

Falconer, D. S., & MacKay, T. M. C. (1996). *Introduction to quantitative genetics* (4th ed.). Harlow, Essex, UK: Addison Wesley Longman.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Yakir, R., Loh, P., & Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, *47*, 1228–1235.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*, 399–433.

Fisher, R. A. (1948). Biometry. *Biometrics*, *4*, 216–219.

Gimelfarb, A. (1981). Analysis of "non-traditional relationships under assortative mating. *Journal of Mathematical Biology*, *13*, 227–240.

Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjalmsson, B. J., & Xu, H., Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics*, *95*, 535–552.

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, *2*, 3–19.

Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetical Research*, *91*, 47–60.

Heckerman, D., Gurdasani, D., Kadie, C., Pomilla, C., Carstensen, T., Martin, H., & Sandhu, M. S. (2016). Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences (USA)*, *113*, 7377–1382.

Hopper, J. L. (1993). Variance components for statistical genetics: Applications in medical research to characteristics related to human diseases and health. *Statistical Methods in Medical Research*, *2*, 199–223.

Hopper, J. L., Foley, D. L., White, P. A., & Pollaers, V. (2013). Australian twin registry: 30 years of progress. *Twin Res Hum Genet.*, *16*, 34–42.

Kempthorne, O. (1957). *An introduction to genetic statistics*. New York: Wiley.

Kumar, S. K., Feldman, M. W., Rehkopf, G. H., & Tuljapurkar, S. (2016). Limitations of GCTA as a solution to the missing heritability problem. *Proceedings of the National Academy of Sciences (USA)*, *113*, E61–E70.

Lange, K., Westlake, J., & Spence, M. A. (1976). Extensions to pedigree analysis. III. variance components by the scoring method. *Annals of Human Genetics*, *39*, 485–491.

Lango, A. H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., & Hirschhorn, J. N. (2012). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*, 832–838.

Lello, L., Avery, S., Tellier, L., Vazquez, A. I., de los Campos, G., & Hsu, S. (2018). Accurate genomic prediction of human height. *Genetics*, *210*, 477–497.

Mäki-Tanila, A., & Hill, W. G. (2014). Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, *198*, 355–367.

Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vasquez, A. L., Duarte, C. W., Allison, D. B., & de los Campos, G. (2011). Beyond missing heritability; Prediction of complex traits. *PLoS Genetics*, *7*, e1002051.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747–753.

Moran, P. A. P., & Smith, C. A. B. (1966). *Commentary on R. A.Fisheras paper on "The correlation between relatives on the supposition of Mendelian inheritance."*. Cambridge, UK: Eugenics Laboratory Memoirs XLI, Cambridge Unversity Press.

Nagylaki, T. (1982). Assortative mating for a quantitative character. *Journal of Mathematical Biology*, *16*, 57–84.

Pearson, K. S., & Lee, A. (1903). On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika*, *2*, 357–462.

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2014). *Behavioral genetics* (6th ed.). New York, NY: worth Publishers.

Powell, J. E., Visscher, P. M., & Goddard, M. E. (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, *11*, 800–805.

Raffa, J. D., & Thompson, E. A. (2016). Power and effective study size in heritability studies. *Statistics in Biosciences*, *8*, 264–283.

Sinha, R., & Gray-McGuire, C. (2007). Haseman Elston regression in ascertained samples: Importance of dependent variable and mean correction factor selection. *Human Heredity*, *65*, 66–76.

Speed, D., & Balding, D. J. (2015). Relatedness in the post-genomic era: Is it still useful? *Nature Reviews Genetics*, *16*, 33–44.

Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, *91*, 1011–1021.

Sverdlov, S., & Thompson, E. A. (2013). Correlation between relatives given complete genotypes: From identity by descent to identity by function. *Theoretical Population Biology*, *88*, 57–67.

Sverdlov, S., & Thompson, E. A. (2018). The epistasis boundary: Linear vs nonlinear genotype-phenotype relationships., https://doi.org/10.1101/503466

Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, *194*, 301–326.

Thompson, E. A., & Shaw, R. G. (1990). Pedigree analysis for quantitative traits: Variance components without matrix inversion. *Biometrics*, *46*, 399–414.

vanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morely, K. I., Zhu, G., Cornes, B. K., & Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, *2*, e41.

Wahlund, S. (1928). Zuzammensetzung von Populationen und Korrelationserscheinungen vom stanpunkt der vererbungslehre aus betrachtet. *Hereditas*, *11*, 65–106.

Wang, B., Sverdlov, S., & Thompson, E. A. (2017). Efficient estimation of realized kinship from SNP genotypes. *Genetics*, *205*, 1063–1078.

Wang, T., & Elston, R. C. (2004). A modified revisited Haseman-Elston method to further improve power. *Human Heredity*, *57*, 109–116.

Wang, X., Guo, X., He, M., & Zhang, H. (2011). Statistical inference in mixed models and analysis of twin and family data. *Biometrics*, *67*, 987–995.

Wright, S. (1921). Systems of mating. III. Assortative mating based on somatic resemblance. *Genetics*, *6*, 144–161.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*, 565–569.

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*, 1–7.

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., & Weedon, M. N., The GIANT Consortium (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~ 700000 individuals of European ancestry. *Human Molecular Genetics*, *27*, 3642–3649.

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences (USA)*, *109*, 1193–1198.

How to cite this article: Thompson EA. Correlations between relatives: From Mendelian theory to complete genome sequence. *Genet. Epidemiol.* 2019;1–15.
https://doi.org/10.1002/gepi.22206