Human Population Genetics and Genomics
https://doi.org/10.47248/hpgg2505040008

Original Research

The Quantitative Genetics of Human Disease: 3A Interactions—Correlation in State

Kiana Jodeiry [1,2], Andrew J. Bass [2,3,†] Michael P. Epstein [2,3], David J. Cutler [2,3,*]

1. Department of Psychology, Emory University, Atlanta, Georgia 30322, USA; Email: kiana.jodeiry@emory.edu;
2. Center of Computational and Quantitative Genetics, Emory University, Atlanta, Georgia 30322, USA; Emails: andrew.jay.bass@emory.edu (A.J.B.); mpepste@emory.edu (M.P.E.);
3. Department of Human Genetics, Emory University, Atlanta, Georgia 30322, USA

† Current address: Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, UK

* **Correspondence:** David J. Cutler; Email: djcutle@emory.edu

Cite This Article:
Jodeiry, K, Bass, AJ, Epstein, MP, Cutler, DJ. The Quantitative Genetics of Human Disease: 3A Interactions—Correlation in State. Hum Popul Genet Genom. 2025;5(4):0008. https://doi.org/10.47248/hpgg2505040008

Received: 30 Oct 2024
Accepted: 19 Dec 2025
Published: 26 Dec 2025

Abstract

The third section of an anticipated four paper series distinguishes two different forms of genetic interactions. In this, the first paper of our discussion on genetic interactions, we describe interactions arising from correlation between genotypic and/or environmental states. In the second paper, we will describe interactions arising from non-additivity between uncorrelated factors (epistasis). We illustrate the ways in which correlations in allelic and genotypic state dramatically alter our intuitions and our quantitative genetic estimates, demonstrating why they must be explicitly accounted for. Departures from Hardy-Weinberg equilibrium are understood as a form of interaction caused by correlations in allelic state within a locus. We review the effects of ancestry on correlation in state within a locus and correlation in state between loci (linkage disequilibrium), with the latter contributing to test statistic inflation in association studies. Relatedness is understood as correlation in allelic state due to recent ancestry. Here we show that population structure, *i.e.*, ancestry, is most simply understood as causing correlation in state between factors, and we demonstrate methods to estimate quantitative genetics quantities while accounting for the correlation in state of alleles induced by complex ancestry.

Keywords: quantitative genetics; human disease; genetic interactions; population structure; linkage disequilibrium

1. Introduction

At its essence, the goal of human genetics is to discover genes that causally influence disease initiation or progression. At the genome-wide scale, case-control association studies approach this goal by identifying statistical relationships (*i.e.*, correlations) between single nucleotide polymorphisms (SNPs) and disease status. The simplest possible analytic strategies for performing association studies often assume the data have the property of being identical and independently distributed (although there are many more recent strategies for considering the joint effects of multiple variants in a region on phenotype [1–4]). The simplest approach assumes each random variable (in this context, genotypes at SNPs) is mutually independent, sharing the same probability distribution with other variables, and correlations in the state of genotypes is usually treated as some sort of confounding to be "removed".

When conducting GWAS and fine-mapping association studies, researchers often seek to reduce the influence of factors frequently viewed as artifacts: departures from Hardy-Weinberg Equilibrium (HWE), linkage disequilibrium, population structure, and relatedness among the individuals under study. These factors are usually viewed as departures from model assumptions that can bias test statistics or parameter estimation. While this view is certainly true and often extremely useful, here we will view all of these elements as related manifestations of the same phenomena, correlation in state between genetic factors. Departures from Hardy-Weinberg equilibrium are understood as a form of interaction caused by correlation in allelic state within a locus. Linkage disequilibrium is understood as correlation in allelic state between different loci. Relatedness is understood as correlation in allelic state due to recent ancestry. Complex ancestry over many generations (population structure) is understood from its effects on correlation in state within and between loci. Population structure refers to the presence of subgroups of individuals in the sample under study, such as subgroups with different ancestral backgrounds, who differ systematically across loci in their allele frequencies. In this paper, we demonstrate how variation in ancestry can create correlation in allelic state between every site in the genome, violating the assumption of independence of genetic factors, and acting as the fundamental cause of GWAS test statistic inflation. From our perspective, we will view the effects of population structure on genetic association studies as a form of interaction created by correlations in genotypic state, later contrasting it with non-additive interactions between uncorrelated markers (epistasis) in the second paper of our discussion of interactions [5].

2. Materials and Methods

We begin with a reminder [6] of our Kempthorne [7] inspired definition of "effect". The effect of being in any particular state of some factor, or any collection of states, is defined to be the average phenotype of individuals who are in that state or collection of states. The effect of an A_1 allele is the average phenotype of individuals who possess the A_1 allele. The effect of the combination of the A_1 allele and environmental state e_1 is the average phenotype of individuals who have the

combination of the A_1 allele along with environmental state e_1. An interaction is said to exist whenever the effect of a combination of states is not equal to the sum of the individual state effects. It is helpful to distinguish two forms of interaction. Interactions can arise from correlation in the state of two factors, *i.e.*, the chance of having allele A_1 might be correlated with the chance of being in environmental state e_1—these are the interactions discussed in this paper. Alternatively, the state of factors might be independent, but the effect of two or more factor states is different from the sum of the individual state effects: the independent factors are "non-additive"—we discuss these interactions in the next paper of this series [5]. When the two factors are both genetic, correlation in genetic state is often called "linkage disequilibrium" when considering allelic states at two different loci and "population structure" when considering allelic states within a locus, but these terms are far from consistently applied or used.

We find it important to distinguish correlation in state from non-additivity of independent factors, because the results of these two forms of interaction can be vastly different. Non-additivity among independent factors does not change individual effect sizes, but always increases the total variance. Correlation in state is far more complex and can cause changes in individual state effects (changes in means) as well as increases or even decreases in the total variance relative to the sum of the individual state variances. We first encountered this when examining linkage disequilibrium (LD) – the correlation in genotypic state between two different SNPs – where we observed LD to change the mean effects of individual alleles and make the total genetic variance less than the sum of the individual locus variances [6].

Departures from Hardy-Weinberg equilibrium

The first paper in this series [6] attempted to derive major quantitative genetics results with as few assumptions as possible, yet throughout the entirety of that presentation, one major assumption remained: Hardy-Weinberg Equilibrium (HWE). We chose to embrace HWE because it greatly simplified an already complex presentation, and HWE is required to reach many of the "usual" representations of key results *e.g.,* the additive variance due to a locus is $2pq\beta^2$. Nevertheless, even this assumption is not necessary for any important quantity to be well-defined, and, in fact, relaxation of HWE was part of both Fisher's original derivation [8] as well as extensively treated by Kempthorne in his 1955 series of papers [7,9–11].

Here we understand departures from HWE as a form of interaction. It is an interaction caused by correlation in allelic state within a genotype. Conversely, independence in the state of the two alleles in a single genotype will be viewed as the definition of HWE. Notice something subtle. We often call ourselves population geneticists. Inherent in the name is the notion of a population, a group of individuals, and population geneticists intuitively and reflexively think of HWE as a property of a group of individuals. While all this is both useful and true, to better understand the effects of population structure, and ancestry more broadly, it will be helpful to think of and define HWE in terms of individuals. If the states of the two alleles in an individual are independent, then that individual is in HWE, otherwise that individual is not. Over any collection of individuals, if the states of the two alleles are

uncorrelated in the collection, then that collection of individuals is in HWE, otherwise the collection is not. This collection need not have any "breeding" relationships in the traditional population genetics sense. A locus in a collection of individuals is in HWE when there is no correlation in allelic state within that locus.

Using the notation from the first paper in this series [6], let v be a diploid locus, with two alleles A_{v_0} and A_{v_1}. First, pick an individual at random. The person has genotype $G_v \in \{A_{v_0}A_{v_0}, A_{v_0}A_{v_1}, A_{v_1}A_{v_1}\}$. Then, pick an allele at random from that individual. Let A_v be the picked allele and A_v^* be the other (not picked) allele. Let $\Pr[A_v = A_{v_0}] = p_v$ and $\Pr[A_v = A_{v_1}] = q_v$. Note that we have oriented allelic labels so that A_{v_1} is the minor allele ($p_v \geq q_v$). If we think of our individual as coming from some sort of population, p_v and q_v are the frequencies of the two alleles in this population.

Associate with the random process of picking an individual and picking an allele a Bernoulli random variable $\mathrm{I}_{A_v} \in \{0,1\}$ which is an indicator that the picked allele was A_{v_1}. Thus, $\mathrm{I}_{A_v} = 1$ if $A_v = A_{v_1}$. The two alleles within a given genotype are perfectly symmetrical, and therefore the chance that A_v^* had been the picked allele is exactly the same as A_v, so $\Pr[A_v^* = A_{v_0}] = p_v$ and $\Pr[A_v^* = A_{v_1}] = q_v$. Associate with this random process of the state of allele A_v^* the Bernoulli random variable $\mathrm{I}_{A_v^*} \in \{0,1\}$. The first two central moments of these variables are

$$\begin{aligned}
\mathrm{E}[\mathrm{I}_{A_v}] &= q_v. && (1)\\
\mathrm{E}[\mathrm{I}_{A_v^*}] &= q_v. && (2)\\
\mathrm{Var}[\mathrm{I}_{A_v}] &= p_v q_v. && (3)\\
\mathrm{Var}[\mathrm{I}_{A_v^*}] &= p_v q_v. && (4)\\
\mathrm{Cov}[\mathrm{I}_{A_v}, \mathrm{I}_{A_v^*}] &= \mathrm{E}[\mathrm{I}_{A_v}\mathrm{I}_{A_v^*}] - \mathrm{E}[\mathrm{I}_{A_v}]\mathrm{E}[\mathrm{I}_{A_v^*}] && (5)\\
&= \Pr[G_v = A_{v_1}A_{v_1}] - q_v^2, && (6)
\end{aligned}$$

because $\mathrm{E}[\mathrm{I}_{A_v}\mathrm{I}_{A_v^*}]$ equals 1 if the individual's genotype was $A_{v_1}A_{v_1}$ and zero otherwise. We will refer to C_v as the covariance in allelic state at locus v for our given individual. From this we derive the probability of all three genotypic states in that individual ($f_{v_{ij}}$).

$$\begin{aligned}
C_v &= \mathrm{Cov}[\mathrm{I}_{A_v}, \mathrm{I}_{A_v^*}]. && (7)\\
q_v &= \Pr[G_v = A_{v_1}A_{v_1}] + \frac{\Pr[G_v = A_{v_0}A_{v_1}]}{2}. && (8)\\
f_{v_{11}} &= \Pr[G_v = A_{v_1}A_{v_1}] && (9)\\
&= q_v^2 + C_v. && (10)\\
f_{v_{01}} &= \Pr[G_v = A_{v_0}A_{v_1}] && (11)\\
&= 2p_v q_v - 2C_v. && (12)\\
f_{v_{00}} &= \Pr[G_v = A_{v_0}A_{v_0}] && (13)\\
&= 1 - (\Pr[G_v = A_{v_1}A_{v_1}] + \Pr[G_v = A_{v_0}A_{v_1}]) && (14)\\
&= p_v^2 + C_v. && (15)
\end{aligned}$$

If our individual comes from a collection of individuals who all have the same chance of having an A_{v_0} allele (the same frequency of A_{v_0}) and the same covariance in allelic state, then the above gives the frequency of the three genotypes in this collection, a population if you will, without any reference to breeding structure or heredity. If the allelic states are independent of each other, then $\mathrm{Cov}[\mathrm{I}_{A_v}, \mathrm{I}_{A_v^*}] = 0$, and the familiar Hardy-Weinberg proportions result in this "population". Departure from HWE is covariance in allelic state, and leads to a population with genotype frequencies as above. Thus, the formulae above can be viewed as the most "generalized" form of Hardy-Weinberg [12]. Notice there can be too many or two few heterozygotes, relative to $2p_v q_v$, depending on the sign of C_v.

Within this population of individuals defined by their shared allele frequency and covariance in allelic state, we can define our usual measures of genotypic and allelic effects at a locus in accordance with the first paper in this series [6]. Here, the genetic effect $\gamma_{v_{ij}}$ is the average phenotype of individuals with the $A_{v_i}A_{v_j}$ genotype, *i.e.*, the conditional expectation of phenotype given genotype. α_{v_0} and α_{v_1} are the allelic effects (also called additive effects) of the A_{v_0} and A_{v_1} alleles, respectively, and are similarly defined as the average phenotype of individuals who possess the allele. β_v is the difference in allelic effects between the two alleles and is naturally interpreted as the consequence of substituting an A_{v_1} allele for an A_{v_0} allele on phenotype. β_v is often termed the effect as locus and is commonly estimated in a linear regression or related framework. $\delta_{v_{ij}}$ is the "dominance deviation" of genotype $A_{v_i}A_{v_j}$ and reflects the deviation from additivity due to dominance at this locus.

$$\begin{aligned}
\gamma_{v_{ij}} &= \mathrm{E}[P|G_v = A_{v_i}A_{v_j}]. && (16)\\
\alpha_{v_0} &= \mathrm{E}[P|A_v = A_{v_0}] && (17)\\
&= p_v\gamma_{v_{00}} + q_v\gamma_{v_{01}} + \frac{C_v(\gamma_{v_{00}} - \gamma_{v_{01}})}{p_v}. && (18)\\
\alpha_{v_1} &= \mathrm{E}[P|A_v = A_{v_1}] && (19)\\
&= p_v\gamma_{v_{01}} + q_v\gamma_{v_{11}} + \frac{C_v(\gamma_{v_{11}} - \gamma_{v_{01}})}{q_v}. && (20)\\
\beta_v &= \alpha_{v_1} - \alpha_{v_0}. && (21)\\
\delta_{v_{ij}} &= \gamma_{v_{ij}} - (\alpha_{v_i} + \alpha_{v_j}). && (22)\\
V_{g_v} &= f_{v_{00}}\gamma_{v_{00}}^2 + f_{v_{01}}\gamma_{v_{01}}^2 + f_{v_{11}}\gamma_{v_{11}}^2. && (23)\\
V_{a_v} &= 4f_{v_{00}}\alpha_{v_0}^2 + f_{v_{01}}(\alpha_{v_0} + \alpha_{v_1})^2 + 4f_{v_{11}}\alpha_{v_1}^2. && (24)\\
V_{d_v} &= f_{v_{00}}\delta_{v_{00}}^2 + f_{v_{01}}\delta_{v_{01}}^2 + f_{v_{11}}\delta_{v_{11}}^2. && (25)\\
V_{I_{ad}} &= V_{g_v} - (V_{a_v} + V_{d_v}). && (26)
\end{aligned}$$

V_{g_v} is the genetic contribution to phenotypic variance due to this locus, V_{a_v} is the additive variance, and V_{d_v} is the dominance variance. $V_{I_{ad}}$ is the "total interaction" between the additive and dominance components within this locus, calculated as the difference between the total genetic variance and the sum of the additive and dominance variances, and is not a variance. We are fast approaching the very uncomfortable realization that all of these definitions are fundamentally a function of C_v, the covariance in allelic state.

To a population geneticist, the term dominance describes the phenotype of the heterozygote. Locus v is said to be additive if the average phenotype of a heterozygote is exactly halfway between the average phenotypes of the two homozygotes, $\gamma_{v_{01}} = (\gamma_{v_{00}} + \gamma_{v_{11}})/2$. In this population genetics sense, additivity means $\gamma_{v_{01}} - \gamma_{v_{00}} = \gamma_{v_{11}} - \gamma_{v_{01}}$. Call this difference $\tilde{\beta}_v$. For a locus without correlation in allelic state ($C_v = 0$), the allelic effects simplify to $\alpha_{v_0} = p_v\gamma_{v_{00}} + q_v\gamma_{v_{01}}$ and $\alpha_{v_1} = p_v\gamma_{v_{01}} + q_v\gamma_{v_{11}}$. Note that for an additive locus without correlation in allelic state, the difference in allelic effects $\beta_v = \alpha_{v_1} - \alpha_{v_0}$ is equal to the difference between the average heterozygote phenotype and that of either of the two homozygotes, $\tilde{\beta}_v$. For an additive locus, when $C_v \neq 0$, this difference $\tilde{\beta}_v$ no longer is equal to the difference in allelic effects β_v, because the allelic effects are a function of the correlation in allelic state.

$$\begin{aligned} \beta_v = \alpha_{v_1} - \alpha_{v_0} &= p_v(\gamma_{v_{01}} - \gamma_{v_{00}}) + q_v(\gamma_{v_{11}} - \gamma_{v_{01}}) \\ &\quad + \frac{C_v}{pq}[q_v(\gamma_{v_{01}} - \gamma_{v_{00}}) + p_v(\gamma_{v_{11}} - \gamma_{v_{01}})] \end{aligned} \tag{27}$$

$$= p_v\tilde{\beta}_v + q_v\tilde{\beta}_v + \frac{C_v(q_v\tilde{\beta}_v + p_v\tilde{\beta}_v)}{p_vq_v} \tag{28}$$

$$= \tilde{\beta}_v\left(1 + \frac{C_v}{p_vq_v}\right). \tag{29}$$

In our Kempthorne-inspired interpretation [6], effects are defined as the expectation of phenotype given the context in which they occur, in this case a population with a particular covariance in allelic state. In individuals from a different population with a different covariance in allelic state, the effect of an allele will be different. In this interpretation, β_v is the "true" difference in allelic effects in a collection of individuals with covariance in allelic state C_v. However, we can envision a different, idealized collection of individuals, identical to the first in all ways except that this idealized collection is in HWE, and therefore $C_v = 0$ in this idealized group of people. From a Falconer perspective, the effect of an allele is a physically determined entity, independent of its context. Thus, in a Falconer inspired presentation of these results, we would be tempted to call $\tilde{\beta}_v$ the "real" difference in allelic effects, and β_v the "estimated" value because the collection is not in HWE. Similarly, call $\tilde{V}_{g_v}$, $\tilde{V}_{a_v} = 2pq\tilde{\beta}^2$, $\tilde{V}_{d_v} = 0$, and $\tilde{V}_{I_{ad}} = 0$ the variance components of this locus in the idealized population. **Figure 1** plots these components in the real population with $C_v \neq 0$ scaled to their values in the ideal population for all non-zero components in the ideal population without covariance in allelic state, and unscaled otherwise. Thus, for this locus, we have plotted V_{g_v}, V_{a_v}, and $2pq\beta^2$ in the real population with non-zero covariance in allelic state divided by their values in the ideal population, and V_{d_v} and $V_{I_{ad}}$ in the real population as their values in the ideal population are zero. Plotted data is for a locus with $q_v = 0.5$ and $\tilde{\beta} = 1$.

Correlation in state changes everything. A locus that would be completely additive in a population in HWE now has both additive and dominance variance. The total genetic variance of the locus is no longer the sum of the additive and dominance variances. In other words, covariance in allelic state induces dominance variance as well as total interaction variance, the latter of which can be negative. In a population

with covariance in allelic state, the additive variance, defined as V_{a_v}, is no longer the same as $2pq\beta^2$. Positive covariance in allelic state increases estimates of V_{g_v}, V_{a_v}, and $2pq\beta^2$, and negative covariance in allelic state decreases these estimates. We define the term heritability to be the additive variance divided by the total phenotypic variance, but heritability is no longer an easy-to-understand predictor of the resemblance between relatives, as the correlation between relatives will be a function of C_v in each relative, which could differ between them. Correlation in allelic state can break down all of our hard earned intuitions and have unpredictable effects on estimates of within-locus variance components, as such, departures from HWE must be modeled explicitly.

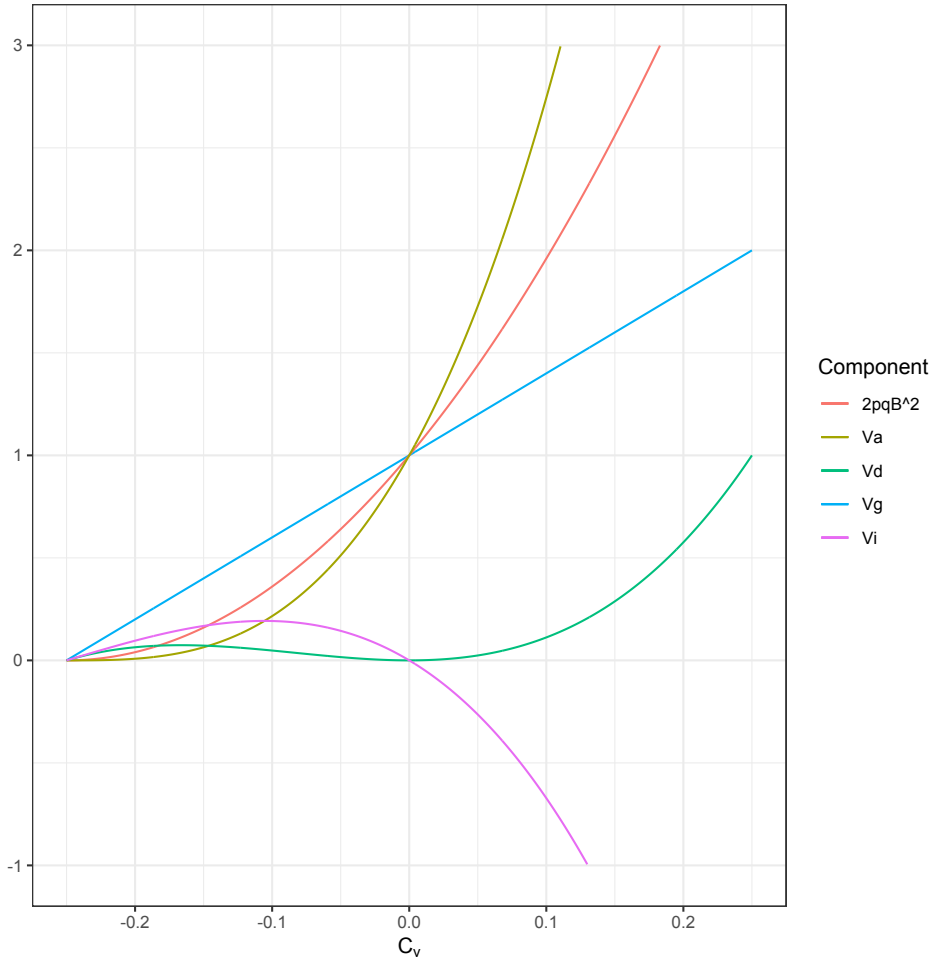

Figure 1 Within locus variance components. Variance components are represented as the ratio of their values in the real population with non-zero covariance in allelic state to their values in an idealized population in HWE, except V_d and $V_{I_{ad}}$ which are their values in the real population as they would be 0 in the idealized population.

3. Results and Discussion

We will now demonstrate that ancestry can cause correlations in allelic state within a locus that are identified as departures from HWE, as well as correlations in allelic state between loci that are identified as apparent linkage disequilibrium. We additionally show that within-locus correlations in allelic state in a collection of individuals in HWE are a measure of relatedness.

3.1 Ancestry affects the correlation in state within a locus

Imagine a collection of individuals who differ in their ancestry. We can think of this collection as a population, but we do not mean to assert anything in particular about breeding within this collection. Pick an individual i at random from this collection. Individual i has two alleles at our diploid locus v. Call those alleles A_{v_i} and $A^*_{v_i}$. Let q_{v_i} be the frequency of the A_{v_1} allele among individual i's ancestors. Let C_{v_i} be the covariance in allelic state for our chosen individual i. Consider C_v to be the covariance in allelic state in the entire collection of individuals from whom we picked individual i. To find this, we randomize over the choice of individual i and use the law of total covariance in Equation 31. In this collection of individuals,

$$\begin{aligned} C_v &= \mathrm{Cov}[\mathrm{I}_{A_v}, \mathrm{I}_{A^*_v}] && (30)\\ &= \mathrm{E}[\mathrm{Cov}[\mathrm{I}_{A_v}, \mathrm{I}_{A^*_v}|i]] + \mathrm{Cov}[\mathrm{E}[\mathrm{I}_{A_v}|i], \mathrm{E}[\mathrm{I}_{A^*_v}|i]] && (31)\\ &= \mathrm{E}[C_{v_i}] + \mathrm{Cov}[q_{v_i}, q_{v_i}] && (32)\\ &= \mathrm{E}[C_{v_i}] + \mathrm{Var}[q_{v_i}]. && (33) \end{aligned}$$

Thus, in this collection of individuals, the covariance in allelic state is the average covariance in allelic state within individuals, plus the variance in allele frequency across the ancestries of the individuals in the collection. Notice that we now think of q_{v_i} as the frequency of the A_{v_1} allele in the ancestors of i, a personalized allele frequency for individual i, if you will [13], and we can sensibly discuss $\mathrm{Var}[q_{v_i}]$ as the variation in this personalized frequency across individuals. Similarly we think of C_{v_i} as a property of i's immediate ancestry. Are the parents of i related to one another? Is there some form of assortative mating occurring between i's parents? Any sort of correlation between genotype and the identity of i's parents (inbreeding / assortative mating) will lead to $\mathrm{E}[C_{v_i}]$ being non-zero. Modeling these factors' effect on phenotype might conveniently be done here. If we assume that within individuals there is no covariance in allelic state, *i.e.*, that individuals are not inbred or the result of assortative mating, in a formal sense assuming that the two alleles are independent within an individual, then we arrive at the simplified

$$C_v = \mathrm{Var}[q_{v_i}]. \quad (34)$$

Since variances are non-negative, we see that absent assortative mating (or some other factor that would create negative covariance in allelic state within individuals), this collection of individuals who vary in their ancestry will have a positive C_v, and a deficit of heterozygotes relative to HWE. This is essentially a restatement of the Wahlund effect [14], albeit without explicit reference to discrete populations. A somewhat more modern and familiar quantification of this concept is as Wright's F_{ST} [15],

$$F_{ST} = \frac{\mathrm{Var}[q_{v_i}]}{\mathrm{E}[q_{v_i}](1 - \mathrm{E}[q_{v_i}])}, \quad (35)$$

where $\mathrm{E}[q_{v_i}]$ is the average frequency of the A_{v_1} allele in the ancestors of this collection of individuals. In the Wrightian conception, we might identify $E[C_{v_i}]$ as proportional to F_{IS}. There is a long history in human genetics of quantifying

the variation in allele frequency across ancestry in terms of F_{ST} [16–18], and it is well established that in humans, when ancestry varies across different continents, $F_{ST} \approx 0.1 - 0.2$. When ancestry varies widely within a single continent, $F_{ST} \approx 0.05 - 0.1$, and it tends to be less across smaller geographic regions. In the population genetics context, it is widely understood that variation in allele frequency due to ancestry leads to departures from HWE. In the quantitative genetics context, it is less widely appreciated that variation in allele frequency due to ancestry (**Figure 1**) can play havoc with everything, including creating dominance variance and making the additive variance larger than the total genetic variance, because negative covariance is created between the additive and dominance contributions. This is the first, albeit minor, reason all quantities in human genetics should be estimated with some procedure (linear mixed model / inclusion of PCs as covariates, *etc.*) that attempts to condition on the ancestry of the studied individuals. Variation in allele frequency *within* a single locus does not, though, create "false positive" association. From Equation 29 above, we have $\beta = \tilde{\beta}(1 + F_{ST})$. If we interpret effects in the Kempthorne sense, we would say the effect size differs between a population in HWE versus one out of it. In the Falconer sense, where we might think of $\tilde{\beta}$ as the "true" β, we would say that departures from HWE inflate our estimate of $\tilde{\beta}$. Of course, if $\tilde{\beta}$ is zero, this has no effect on association test statistics. However, variation in ancestry has effects on estimates of β that go far beyond those of departures from HWE.

3.2 Ancestry affects the correlation in state between loci

All effect sizes in human genetics should be estimated via a method that accounts for individuals' ancestry - because variation in ancestry creates correlation in state between every site in the genome. This correlation in state *between* loci caused by variation in ancestry among study participants causes substantial false positive association in genetic association studies if uncorrected. To understand why this is, consider two diploid loci, v and w. Let I_{A_v} and I_{A_w} be the indicators that a randomly picked allele is A_{v_1} and A_{w_1}, respectively. Recall [6] that the standard measures of linkage disequilibrium D and r^2 can be written as

$$D = \mathrm{Cov}[\mathrm{I}_{A_v}, \mathrm{I}_{A_w}], \tag{36}$$

$$r^2 = \frac{\mathrm{Cov}^2[\mathrm{I}_{A_v}, \mathrm{I}_{A_w}]}{\mathrm{Var}[\mathrm{I}_{A_v}]\mathrm{Var}[\mathrm{I}_{A_w}]}. \tag{37}$$

Now imagine a collection of individuals whose ancestry differs from one another. Each individual i in this collection has some ancestry where q_{v_i} and q_{w_i} are the frequency of the A_{v_1} and A_{w_1} alleles among i's ancestors. Let D_i be the standard measure of linkage disequilibrium between the loci found in individual i, a personalized linkage disequilibrium for i.

$$\begin{aligned}
\mathrm{E}[I_{A_v}] &= \mathrm{E}[\mathrm{E}[I_{A_v}|i]] && (38)\\
&= \mathrm{E}[q_{v_i}]. && (39)\\
\mathrm{E}[I_{A_w}] &= \mathrm{E}[q_{w_i}]. && (40)\\
\mathrm{Var}[I_{A_v}] &= \mathrm{Var}[\mathrm{E}[I_{A_v}|i]] + \mathrm{E}[\mathrm{Var}[I_{A_v}|i]] && (41)\\
&= \mathrm{Var}[q_{v_i}] + \mathrm{E}[q_{v_i}(1-q_{v_i})] && (42)\\
&= \mathrm{Var}[q_{v_i}] + \mathrm{E}[q_{v_i}] - \mathrm{E}[q_{v_i}^2] && (43)\\
&= \mathrm{Var}[q_{v_i}] + \mathrm{E}[q_{v_i}] - (\mathrm{Var}[q_{v_i}] + \mathrm{E}[q_{v_i}]^2) && (44)\\
&= \mathrm{E}[q_{v_i}](1-\mathrm{E}[q_{v_i}]). && (45)\\
\mathrm{Var}[I_{A_w}] &= \mathrm{E}[q_{w_i}](1-\mathrm{E}[q_{w_i}]). && (46)\\
D &= \mathrm{Cov}[I_{A_v}, I_{A_w}] && (47)\\
&= \mathrm{E}[\mathrm{Cov}[I_{A_v}, I_{A_w}|i]] + \mathrm{Cov}[\mathrm{E}[I_{A_v}|i], \mathrm{E}[I_{A_w}|i]] && (48)\\
&= \mathrm{E}[D_i] + \mathrm{Cov}[q_{v_i}, q_{w_i}]. && (49)\\
r^2 &= \frac{(\mathrm{E}[D_i] + \mathrm{Cov}[q_{v_i}, q_{w_i}])^2}{\mathrm{E}[q_{v_i}](1-\mathrm{E}[q_{v_i}])\mathrm{E}[q_{w_i}](1-\mathrm{E}[q_{w_i}])}. && (50)
\end{aligned}$$

Note that Equations 38 and 39 follow from the law of total expectation, Equation 44 uses the definition of variance ($\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$), and Equation 48 uses the law of total covariance. Thus, we arrive at the result that D in the collection of individuals is the average D within individuals, plus the covariance in allele frequencies over ancestries [19]. If loci v and w are not physically linked (loci on different chromosomes, say), then we expect that $E[D_i] = 0$, and this simplifies to

$$\begin{aligned}
D &= \mathrm{Cov}[q_{v_i}, q_{w_i}]. && (51)\\
r^2 &= \frac{\mathrm{Cov}^2[q_{v_i}, q_{w_i}]}{\mathrm{E}[q_{v_i}](1-\mathrm{E}[q_{v_i}])\mathrm{E}[q_{w_i}](1-\mathrm{E}[q_{w_i}])}. && (52)
\end{aligned}$$

We conclude that when a group of individuals varies in ancestry, there will appear to be linkage disequilibrium (covariance in allelic state between loci) at any pair of loci that vary in frequency over ancestry, even if those loci are physically unlinked. From the fact that all correlations are bound between -1 and 1, we know that

$$\begin{aligned}
-\sqrt{\mathrm{Var}[q_{v_1}]\mathrm{Var}[q_{w_1}]} \leq D \leq & \sqrt{\mathrm{Var}[q_{v_1}]\mathrm{Var}[q_{w_1}]}. && (53)\\
r^2 &= \frac{D^2}{\mathrm{E}[q_{v_i}](1-\mathrm{E}[q_{v_i}])\mathrm{E}[q_{w_i}](1-\mathrm{E}[q_{w_i}])} && (54)\\
&\leq \frac{\mathrm{Var}[q_{v_1}]\mathrm{Var}[q_{w_1}]}{\mathrm{E}[q_{v_i}](1-\mathrm{E}[q_{v_i}])\mathrm{E}[q_{w_i}](1-\mathrm{E}[q_{w_i}])} && (55)\\
&= F_{ST_v}F_{ST_w}, && (56)
\end{aligned}$$

where F_{ST_v} and F_{ST_w} are Wright's measures of F_{ST} at locus v and w respectively. Thus, r^2 between unlinked loci is bounded by the product of F_{ST} at the individual loci. If in this collection of individuals, there are only two different ancestries with any appreciable difference in allele frequency (the precise situation treated in [19]), then equality necessarily holds and $D = Cov[q_{v_i}, q_{w_i}] = \pm\sqrt{Var[q_{v_i}]Var[q_{w_i}]}$ regardless of

the proportions of those ancestries (e.g., the respective percentages of European and African ancestry, say). To see this, call q_{v_1}, q_{w_1} and q_{v_2}, q_{w_2} the allele frequencies in the two different ancestries, and let c_1 and c_2, with $c_1 = 1 - c_2$, be the fraction of individuals in this collection whose ancestries are from each type. Let $\delta_{q_v} = q_{v_1} - q_{v_2}$ and $\delta_{q_w} = q_{w_1} - q_{w_2}$.

$$\begin{aligned}
\mathrm{Cov}[q_{v_i}, q_{w_i}] &= \mathrm{E}[q_{v_i} q_{w_i}] - \mathrm{E}[q_{v_i}]\mathrm{E}[q_{w_i}] && (57)\\
&= c_1 q_{v_1} q_{w_1} + c_2 q_{v_2} q_{w_2} - (c_1 q_{v_1} + c_2 q_{v_2})(c_1 q_{w_1} + c_2 q_{w_2}) && (58)\\
&= c_1 c_2 (q_{v_1} - q_{v_2})(q_{w_1} - q_{w_2}) && (59)\\
&= c_1(1 - c_1)\delta_{q_v}\delta_{q_w}. && (60)\\
\mathrm{Var}[q_{v_i}] &= \mathrm{E}[q_{v_i}^2] - \mathrm{E}[q_{v_i}]^2 && (61)\\
&= c_1 q_{v_1}^2 + c_2 q_{v_2}^2 - (c_1 q_{v_1} + c_2 q_{v_2})^2 && (62)\\
&= c_1(1 - c_1)\delta_{q_v}^2. && (63)\\
\mathrm{Var}[q_{w_i}] &= c_1(1 - c_1)\delta_{q_w}^2. && (64)\\
D = \mathrm{Cov}[q_{v_i}, q_{w_i}] &= \pm\sqrt{\mathrm{Var}[q_{v_i}]\mathrm{Var}[q_{w_i}]}. && (65)
\end{aligned}$$

Recall that we assumed v and w are independent loci (unlinked). Thus, in any collection of individuals, when there are only two ancestries significantly contributing variance in allele frequency, D between every pair of sites in the genome will be approximately plus-or-minus the square root of the product of the variance in allele frequency at each site. These variances are defined relative to the ancestry of the individuals in the collection under consideration. Differing individuals with differing ancestry necessarily have differing variance in their allele frequency.

Human population structure can be a major cause of "false positive" association between genetic variation and phenotype whenever there is allele frequency variation in the ancestors of the individuals examined. This has been known for decades, motivating the development of a large range of complementary methods to correct for these effects (***e.g.,*** TDT, genomic control, structure, principal component analysis, linear mixed models) [20–24]. To understand more quantitatively exactly what is occurring, imagine attempting to estimate the contribution of locus v to phenotype, and describing that effect as β_v. Let $\tilde{\beta}_v$ be the effect of this locus if there were no variation in ancestry among the studied individuals. Thus, $\tilde{\beta}_v$ is the effect of this locus in an idealized collection of individuals without variation in ancestry. β_v is the effect in the real collection under study (the true effect is $\tilde{\beta}_v$, versus the estimated effect, β_v, in the Falconer sense). Finally assume that $\tilde{\beta}_v = 0$. So, in the absence of variation in ancestry, this locus has no effect on phenotype. In the first paper in this series [6], we showed that for any pair of additive loci v and w, in the absence of any interactions other than correlation in allelic state

$$\beta_v = \tilde{\beta}_v + \frac{D_{v,w}}{p_v q_v}\tilde{\beta}_w, \qquad (66)$$

where $D_{v,w}$ is the standard linkage disequilibrium measure between loci v and w. Now imagine the collection of all loci w_k contributing an additive effect, $\tilde{\beta}_{w_k} \neq 0$, in the idealized population, but in this case we are imagining that $\tilde{\beta}_v = 0$. The total additive variance in the idealized population is

$$\tilde{V}_A = \sum_k 2p_{w_k}q_{w_k}\tilde{\beta}^2_{w_k}. \tag{67}$$

Now consider the combined effects of all loci w_k on β_v in a population with variation in ancestry.

$$\begin{aligned}
\beta_v &= \tilde{\beta}_v + \sum_k \frac{D_{v,w_k}}{p_v q_v}\tilde{\beta}_{w_k} && (68)\\
&= \sum_k \frac{D_{v,w_k}}{p_v q_v}\tilde{\beta}_{w_k}. && (69)\\
\beta^2_v &= \left(\sum_k \frac{\mathrm{Cov}[q_v, q_{w_k}]}{p_v q_v}\tilde{\beta}_{w_k}\right)^2 && (70)\\
&\approx \frac{1}{p^2_v q^2_v}\left(\sum_k \pm\tilde{\beta}_{w_k}\sqrt{\mathrm{Var}[q_v]\mathrm{Var}[q_{w_k}]}\right)^2, && (71)
\end{aligned}$$

when only a small number of ancestries contribute significantly to variation in allele frequency. This value is necessarily bigger than zero, if there is any variation in q_v over ancestries. Recall that a natural method to test for association between locus v and phenotype would be to asses a null model where $2Np_v q_v \beta^2_v$ is approximately χ^2_1 distributed, where N is the number of individuals studied [25].

We thus arrive at the full intuition for the need to account for ancestry in human genetic studies. A naive χ^2 test for association between locus v and phenotype will not have an expectation centered at 1, as the test assumes, but instead centered around a larger value that is a complex function of the covariance in allele frequencies and V_A. This inflation will occur at every site in the genome whose allele frequency varies over ancestry, *i.e.*, potentially every site tested. While nearly all human geneticists understand that tests for association must account for ancestry, the reason for this is often only incompletely understood. Many discussions of this revolve around the idea that the mean phenotype (or disease prevalence for dichotomous traits) may differ between ancestries, and this difference in means causes correlation between ancestry and phenotype. Other discussions might suggest a "mechanism" for a correlation between phenotype and ancestry involving differing environmental states across ancestries, perhaps differing diets, rates of smoking, or other social determinants of health. While correlations between genes and mean phenotype or between genetic and environmental state can cause inflation of test statistics, they are in no sense required for the inflation to be pronounced. If there is additive variance for the trait, and if any of the alleles contributing to that variance differ in frequency among the ancestors of the study participants, then there will be test statistic inflation, and that inflation will increase with increasing sample size, N, and with increasing variance in allele frequency among the participants, F_{ST}. When effects are defined as the expectation of phenotype given the context in which they occur (*i.e.*, in a Kempthorne-inspired

presentation), this is not "inflation" *per se*, but real association between marker and phenotype, though the cause of the association is not the direct effects of the tested marker. Instead, the detected association is caused by the correlation in state between the tested marker and other variants scattered throughout the genome, *i.e.*, the association is positive indication of the interaction – the correlation in state – between this marker and other unlinked sites. Given that the goal of association studies often is to identify loci that putatively have causal influence on phenotype, this phenomenon of association caused by interactions due to correlation in allelic state is nearly universally referred to as an issue of test statistic inflation, not one of "true" interaction, and we shall do the same.

Patterns common in disease genetics and human demographic history accentuate this inflation. The key aspect to consider is the sign of $\mathrm{Cov}[q_v, q_{w_k}]\beta_{w_k}$. Since the signs of both $\mathrm{Cov}[q_v, q_{w_k}]$ and β_{w_k} could be positive or negative, one might naively assume (hope) that the terms in the sum above will alternate signs sufficiently frequently that the mean over all SNPs might approach 0. This is unlikely to be true for disease traits. Recall we have oriented β so that it is the effect of substituting the rarer allele for the more common allele. If $\beta > 0$ then the rare allele makes the trait bigger, or in the case of a disease trait, increases the risk of disease relative to the common allele. For non-disease phenotypes there may be little reason to doubt that the sign of β is "random". However, for disease traits there is good reason to believe that on average β will be positive [13], because selection is likely to keep alleles rare if they increase the risk of a disease. This is especially likely if the allele has a strong effect on disease. Human demographic history suggests that $\mathrm{Cov}[q_v, q_{w_k}]$ will also be positive most of the time. In general, a positive covariance between two variables implies that larger values at both variables occur more often than expected by chance. In this context it means the ancestry with the higher minor allele frequency at one locus tends to have the higher minor allele frequency at the other. In the second paper in this series [13], we saw that human demographic history tends to cause this exact pattern of correlation (conditional on alleles being present in two ancestries, one ancestry has systematically higher minor allele frequencies than the other). In that paper, the correlation in minor allele frequencies across unlinked sites appears to be the likely explanation for the most perplexing observations when applying polygenic risk scores across populations with differing ancestry. Here, the correlation in minor allele frequency induced by human demographic history is seen as the primary driver of test statistic inflation.

In this context, we can best understand how and why various procedures designed to account for this inflation work. Perhaps the most intuitively simple, but complex algorithmically, is the STRUCTURE / STRAT approach developed by Jonathan Pritchard and colleagues [22,26]. The idea is simple, variation in ancestry creates deviations from HWE within loci as well as correlation in allele frequency between loci, therefore, in principle, one can infer ancestral groups from the data [22] and then conditional on the inferred ancestry, perform association studies within ancestry groups in HWE [26]. In the structured association approach developed by Pritchard and colleagues, individuals are assigned to a subpopulation (possibly accounting for admixture with fractional cluster membership) using a model-based clustering program (STRUCTURE) [22] and association statistics are computed

stratifying by subpopulation (STRAT) [26]. The approach is natural, but extremely computationally challenging for large sample sizes, and a bit sensitive to a correct *a priori* determination of the number of "meaningful" ancestral groups contributing to the collection.

At the other end of the computational spectrum is genomic control, pioneered by Devlin and Roeder [20]. Here the approach is incredibly simple, calculate the mean (median) χ^2 test statistic, usually called λ, from a genome-wide study and divide each site's χ^2 by λ. The intuition comes from the fact that the number of SNPs contributing to additive variance is likely very large. If so, it is not unreasonable to believe that some sort of strong law of large numbers holds, and the distribution across SNPs of β_v may therefore approximate a normal distribution, and the association test statistic will approach a non-central χ^2 with $\lambda \sim \mathrm{E}[\beta_v^2]$. That genomic control robustly controls for type-1 error in genetic studies in virtually all cases [27,28] is strong evidence in support of this intuition. While undoubtedly a robust tool, genomic control is also thought to be something of a blunt instrument, destroying "signal" in the process of achieving well calibrated type-1 error rates. By ignoring the magnitude of $\mathrm{Cov}[q_v, q_{w_k}]$ and correcting for mean behavior alone, this technique penalizes all sites equally, even when an individual site v has little or no allelic covariance with any other site in the genome, perhaps because v has no substantial variation in allele frequency across ancestries.

Since the fundamental problem being solved is caused by covariance in allele frequency and additive variance for the trait, the most natural solution to this problem is to include both in a linear mixed model [24]. Here, the tested model takes the form

$$\vec{P} = \vec{S_v}\tilde{\beta}_v + \mathbf{R}V_A + \epsilon, \tag{72}$$

where $\vec{P}$ is the vector of phenotypes, $\vec{S_v}$ is a vector of minor allele counts at locus v, $\mathbf{R}$ is a matrix describing the covariance in allelic state between each pair of individuals (much more on this in 3.3), and V_A is the total additive variance. If both P and S_v have been normalized to have mean 0, to the extent that $\mathbf{R}$ fully captures the covariance in allelic state between each pair of individuals and no other interactions exist, this should produce an estimate of $\tilde{\beta}_v$, the effect of substituting the rare allele for the common allele in a collection of individuals without ancestral variation in allele frequency - the correct β in the Falconer sense, not inflated by population structure.

Linear mixed models can be quite computationally burdensome when compared to a linear model, particularly with thousands of individuals. Note that to perform certain statistical analyses within a linear mixed model framework, such as calculating individual genetic effects, the inverse of the genetic relatedness matrix $\mathbf{R}$ is needed. Calculating the inverse of a large relatedness matrix can be computationally demanding, especially with large sample sizes. To dramatically reduce this burden, many newer linear mixed model approaches (*e.g.,* [29,30]) for large datasets implement algorithms that approximate the degree of relatedness among individuals in the sample without requiring a pre-computed genetic relationship matrix *i.e.*, without explicitly calculating the covariance in allelic state between each pair of individuals. For example, a linear mixed model can be approximated with a

linear model by replacing the $\mathbf{R}$ matrix with a subset of its eigenvectors. In this way, we can view a linear model with inclusion of principle components (PCs) derived from $\mathbf{R}$ [21] as a natural approximation to the more computationally burdensome linear mixed model. These approximate approaches to linear mixed models may have comparable computational efficiency to that of a linear model.

Of course, to a geneticist, the most natural way to remove the effects of variation in allele frequency over ancestries is to ensure that the test for association being performed is not affected by correlation in allelic state at unlinked markers. This can this be achieved by focusing only on alleles transmitted from parents known to be heterozygotes [23]. The transmission-disequilibrium test (TDT) counts the number of transmissions from parents known to be heterozygotes to children with some given phenotype, generally a disease state. Because Mendelian segregation is strongly regulated during meiosis to have nearly 50-50 transmission rates, regardless of the identity of any particular allele at a given site, the observed transmission rate from heterozygous parents is a function of the penetrance of the allele for the conditioned phenotype, and this transmission rate is fundamentally independent of any correlations in state at other unlinked sites [31,32]. While this approach is virtually guaranteed to solve issues associated with ancestry, it is often far harder to collect sample sets that include both parents and offspring, and undetected genotyping error can have particularly challenging effects. Computation of the TDT statistic on trios in which one parent is missing marker genotype data increases the type-1 error rate of the statistic [33] as does genotyping error, which can cause apparent over-transmission of common alleles [34,35].

3.3 Phenotypic covariance, relatedness, and the covariance in allelic state

The first paper in this series [6] derived the covariance between individuals in a generalized form. Here we focus on a very specialized case where there is potentially correlation in genotypic state, but no other non-additive interactions of any kind - no dominance, additivity between all loci, and no correlation in state of non-genetic factors. Imagine a collection of individuals that might be a single population in HWE or could be collection of individuals with differing and complex ancestry. We are interested in two randomly picked individuals from this collection; call them individuals 1 and 2 and their phenotypes P_1 and P_2. At some locus v contributing to this phenotype, let S_{v_1} and S_{v_2} be the counts of minor alleles, A_{v_1}, in individual 1 and 2, respectively. Let $\vec{S_1}$ and $\vec{S_2}$ be the vector of these counts over all loci contributing to the phenotype. Let $\alpha_{v_{0_1}}$ and $\alpha_{v_{0_2}}$ be the average phenotype of an individual with the common allele at locus v in individuals with ancestral minor allele frequencies q_{v_1} and q_{v_2} respectively. While the allelic effect α will be a function of allele frequency, assume the difference in allelic effects $\tilde{\beta} = \alpha_1 - \alpha_0$ [13] is the same in all individuals regardless of allele frequency.

$$\mathrm{E}[P_1|\vec{S_1}] = \sum_v (\tilde{\beta}_v S_{v_1} + 2\alpha_{v_{0_1}}). \quad (73)$$

$$\mathrm{E}[P_2|\vec{S_2}] = \sum_v (\tilde{\beta}_v S_{v_2} + 2\alpha_{v_{0_2}}). \quad (74)$$

$$\mathrm{E}[P_1P_2|\vec{S_1},\vec{S_2}] = \left(\sum_v (\tilde{\beta}_v S_{v_1} + 2\alpha_{v_0})\right)\left(\sum_v (\tilde{\beta}_v S_{v_2} + 2\alpha_{v_0})\right). \quad (75)$$

$$\mathrm{Cov}[P_1,P_2|\vec{S_1},\vec{S_2}] = \mathrm{E}[P_1P_2|\vec{S_1},\vec{S_2}] - \mathrm{E}[P_1|\vec{S_1},\vec{S_2}]\mathrm{E}[P_2|\vec{S_1},\vec{S_2}] \quad (76)$$

$$= 0. \quad (77)$$

$$\mathrm{Cov}[P_1,P_2] = \mathrm{Cov}[\mathrm{E}[P_1|\vec{S_1},\vec{S_2}],\mathrm{E}[P_2|\vec{S_1},\vec{S_2}]] \quad (78)$$

$$+\mathrm{E}[\mathrm{Cov}[P_1,P_2|\vec{S_1},\vec{S_2}]]$$

$$= \mathrm{Cov}\left[\sum_v (\tilde{\beta}_v S_{v_1} + 2\alpha_{v_{0_1}}), \sum_v (\tilde{\beta}_v S_{v_2} + 2\alpha_{v_{0_2}})\right] \quad (79)$$

$$= \sum_v \tilde{\beta}_v^2 \mathrm{Cov}[S_{v_1},S_{v_2}] + \sum_{v\neq w} \tilde{\beta}_v\tilde{\beta}_w \mathrm{Cov}[S_{v_1},S_{w_2}]. \quad (80)$$

Line 77 uses the lack of dominance or any other interactions extensively. Thus, for an additive trait, the covariance between individuals is the sum over all contributing loci of $\tilde{\beta}^2$ multiplied by the covariance in minor allele counts, plus the sum of the product of two different β's, and their covariance in allelic state (LD) for pairs of sites across the genome.

3.3.1 Relatedness for single population in Hardy-Weinberg equilibrium

If individuals 1 and 2 were drawn from a single population where all loci are unlinked and in HWE, then the second sum in 80 is 0. In a single population, if individuals 1 and 2 share fraction r of the their genome identical-by-descent (IBD) because they share recent common ancestors, and if r_0, r_1, and r_2 are the fraction of the genome where exactly $0, 1$ and 2 alleles are shared IBD, then we arrive at the the familiar

$$\mathrm{E}[S_{v_1}] = \mathrm{E}[S_{v_2}] = 2q_v. \quad (81)$$

$$\mathrm{Cov}[S_{v_1},S_{v_2}] = \sum_{i=0,1,2} E[S_{v_1}S_{v_2}|r_i] - 4q_v^2 \quad (82)$$

$$= r_0(4q_v^2) + r_1(q_v(4q_v^2 + 4p_vq_v + p_v^2) + p_vq_v^2) \quad (83)$$

$$+r_2(4q_v^2 + 2p_vq_v) - 4q_v^2$$

$$= r_1(p_v^2q_v + p_vq_v^2) + 2r_2p_vq_v \quad (84)$$

$$= p_vq_v(r_1 + 2r_2) \quad (85)$$

$$= 2p_vq_vr. \quad (86)$$

$$r = \frac{\mathrm{Cov}[S_{v_1},S_{v_2}]}{2p_vq_v}. \quad (87)$$

The covariance in genotypic state at a locus between two individuals is the relatedness between the individuals multiplied by $2pq$, for a collection of individuals in HWE. In this fashion we now see that r called relatedness is also by definition r, the correlation in genotypic state, for a collection of individuals in HWE. Quite literally, relatedness is the correlation in genotypic state. Because of this, for any

pair of individuals, we can imagine using the observed correlation in genotypic state across a collection of loci to estimate the relatedness of those two individuals. For instance, in a collection of N distinct SNPs v in HWE, we could estimate the relatedness r_{ij} between a randomly picked pair of individuals i and j as

$$\hat{r}_{ij} = \frac{1}{N}\sum_{v}\frac{\mathrm{Cov}[S_{v_i}, S_{v_j}]}{2p_v q_v} \tag{88}$$

$$= \frac{1}{N}\sum_{v}\frac{(S_{v_i}-2q_v)(S_{v_j}-2q_v)}{2p_v q_v}. \tag{89}$$

Thus, if we had a random collection of individuals drawn from a single population in HWE and had a collection of SNP genotypes in those individuals, we could estimate the relatedness between each pair of individuals directly from the genotypes. A matrix $\mathbf{R}$ whose elements r_{ij} are calculated as above is an estimate of the relatedness of those individuals over the SNPs used in the estimate. Of course, if the collection of individuals or SNPs used in the estimate is biased, it is possible the estimate of relatedness might also be biased in some fashion. Nevertheless, this method of estimation should be fundamentally robust with a sufficiently large number of SNPs in any collection of individuals in HWE.

3.3.2 Phenotypic covariance in the presence of population structure

In a collection of individuals with complex ancestries, we again think of an individual's ancestry as determining their personal allele frequency, q_{v_i}, at locus v. We further imagine individuals who do not share an ancestry are not closely related to one another so that $r_{ij} = 0$. Thus, when individual i and j are related at locus v, $q_{v_i} = q_{v_j}$. Let $I_{q_{v_{ij}}}$ be the indicator that individuals i and j share an ancestry at locus v so that they have equal personalized allele frequencies at locus v, $q_{v_i} = q_{v_j}$. $\mathrm{E}[I_{q_{v_{ij}}}]$ is the probability these two individuals have the same ancestry at this site. Letting $\bar{q}_v$ and $\mathrm{Var}[q_v]$ be the overall mean and variance in allele frequency across all individuals in the collection,

$$\mathrm{Cov}[S_{v_i}, S_{v_j}] = \mathrm{Cov}[\mathrm{E}[S_{v_i}|I_{q_{v_{ij}}}], \mathrm{E}[S_{v_j}|I_{q_{v_{ij}}}]] + \mathrm{E}[\mathrm{Cov}[S_{v_i}, S_{v_j}|I_{q_{v_{ij}}}]]. \tag{90}$$

$$\mathrm{Cov}[\mathrm{E}[S_{v_i}|I_{q_{v_{ij}}}], \mathrm{E}[S_{v_j}|I_{q_{v_{ij}}}]] = \mathrm{Cov}[2q_{v_i}, 2q_{v_j}] \tag{91}$$

$$= 4\mathrm{Cov}[q_{v_i}, q_{v_j}] \tag{92}$$

$$= 4(\mathrm{Cov}[\mathrm{E}[q_{v_i}|I_{q_{v_{ij}}}], \mathrm{E}[q_{v_j}|I_{q_{v_{ij}}}]] + \mathrm{E}[\mathrm{Cov}[q_{v_i}, q_{v_j}|I_{q_{v_{ij}}}]]) \tag{93}$$

$$= 4\mathrm{E}[I_{q_{v_{ij}}}]\mathrm{Var}[q_v]. \tag{94}$$

$$\mathrm{E}[\mathrm{Cov}[S_{v_i}, S_{v_j}|I_{q_{v_{ij}}}]] = \mathrm{E}[I_{q_{v_{ij}}}(2p_{v_i}q_{v_i}r_{ij}) + (1-I_{q_{ij}})\mathrm{Cov}[S_{v_i}, S_{v_j}|I_{q_{v_{ij}}}=0]] \tag{95}$$

$$= \mathrm{E}[I_{q_{v_{ij}}}]2\bar{q}_v\bar{p}_v r_{ij}. \tag{96}$$

$$\mathrm{Cov}[S_{v_i}, S_{v_j}] = \mathrm{E}[I_{q_{v_{ij}}}](r_{ij}(2\bar{q}_v\bar{p}_v) + 4\mathrm{Var}[q_v]) \tag{97}$$

$$= \mathrm{E}[I_{q_{v_{ij}}}]2\bar{q}_v\bar{p}_v(r_{ij} + 2F_{ST_v}). \tag{98}$$

Returning attention to the second sum in 80, begin by writing S_{v_i} as the sum of two different indicators ($I_{A_{v_i}}$ + $I_{A^*_{v_i}}$) that individual i has the A_{v_1} allele

$$\begin{aligned} S_{v_i} &= I_{A_{v_i}} + I_{A^*_{v_i}}. && (99)\\ S_{w_j} &= I_{A_{w_j}} + I_{A^*_{w_j}}. && (100)\\ \mathrm{Cov}[S_{v_i}, S_{w_j}] &= \mathrm{Cov}[I_{A_{v_i}} + I_{A^*_{v_i}}, I_{A_{w_j}} + I_{A^*_{w_j}}] && (101)\\ &= 4\mathrm{Cov}[I_{A_{v_i}}, I_{A_{w_j}}] && (102)\\ &= 4(\mathrm{E}[\mathrm{Cov}[I_{A_{v_i}}, I_{A_{w_j}}|I_{q_{v_{ij}}}, I_{q_{w_{ij}}}] && (103)\\ &\quad +\mathrm{Cov}[\mathrm{E}[I_{A_{v_i}}|I_{q_{v_{ij}}}, I_{q_{w_{ij}}}], \mathrm{E}[I_{A_{w_j}}|I_{q_{v_{ij}}}, I_{q_{w_{ij}}}]]. \end{aligned}$$

To finish we need to add the insight that when two individuals are picked at random, if they share no ancestry at a given locus, the allele draws are independent and there is no covariance in state between different loci, so that

$$\begin{aligned} \mathrm{E}[\mathrm{Cov}[I_{A_{v_i}}, I_{A_{w_j}}|I_{q_{v_{ij}}}, I_{q_{w_{ij}}}] &= \mathrm{E}[I_{q_{v_{ij}}} I_{q_{w_{ij}}} \mathrm{Cov}[q_v, q_w]] && (104)\\ &= \mathrm{E}[I_{q_{v_{ij}}} I_{q_{w_{ij}}}] \mathrm{Cov}[q_{v_i}, q_{w_i}] && (105)\\ &= \mathrm{E}[I_{q_{v_{ij}}} I_{q_{w_{ij}}}] D_{v,w}. && (106)\\ \mathrm{Cov}[\mathrm{E}[I_{A_{v_i}}|I_{q_{v_{ij}}}, I_{q_{w_{ij}}}], \mathrm{E}[I_{A_{w_j}}|I_{q_{v_{ij}}}, I_{q_{w_{ij}}}]] &= 0. && (107) \end{aligned}$$

Putting this all together with the assumption that local ancestry does not differ significantly across the genome, in other words that for all choices of v and w, $\mathrm{E}[I_{q_{v_{ij}}}] \approx \mathrm{E}[I_{q_{w_{ij}}}] = M_{ij}$, then

$$\begin{aligned} \mathrm{Cov}[P_i, P_j] &= \sum_v \mathrm{E}[I_{q_{v_{ij}}}] 2\bar{p}_v\bar{q}_v\tilde{\beta}_v^2(r_{ij} + 2F_{ST_v}) && (108)\\ &\quad + \sum_{v\neq w} 4\tilde{\beta}_v\tilde{\beta}_w \mathrm{E}[I_{q_{v_{ij}}} I_{q_{w_{ij}}}] D_{v,w}\\ &= \sum_v M_{ij}\tilde{V}_{a_v}(r_{ij} + 2F_{ST_v}) + \sum_{v\neq w} 4\tilde{\beta}_v\tilde{\beta}_w M_{ij}^2 D_{v,w}. && (109) \end{aligned}$$

Recall that $\tilde{\beta}_v$ is the effect of this locus in an idealized collection of individuals without variation in ancestry (*i.e.*, $C_v = 0$, in HWE). As such, $\tilde{V}_{a_v} = 2\bar{p}_v\bar{q}_v\tilde{\beta}_v^2$ is the additive variance due to this locus and $\tilde{V}_A = \sum_v \tilde{V}_{a_v}$ is the total additive variance in the idealized population. If the collection of individuals is an approximately equal mixture of two major ancestry groups, so every individual shares approximately half of their ancestry with every other individual such that $M_{ij} \approx 0.5$, an almost natural looking result appears

$$\begin{aligned} \mathrm{Cov}[P_i, P_j] &\approx \sum_v \tilde{V}_{a_v}\left(\frac{r_{ij}}{2} + F_{ST_v}\right) + \sum_{v\neq w} \tilde{\beta}_v\tilde{\beta}_w D_{v,w} && (110)\\ &\approx \tilde{V}_A\left(\frac{r_{ij}}{2} + F_{ST}\right) + \sum_{v\neq w} \tilde{\beta}_v\tilde{\beta}_w D_{v,w}, && (111) \end{aligned}$$

where r_{ij} is the relatedness between individuals i and j and F_{ST} is something like the average F_{ST} across all loci contributing to the phenotype. Of course, as shown above $D_{v,w}$ is a function of the variation in allele frequency across the ancestries of the individuals in the collection.

Covariance in genotypic state can therefore be described in several ways. It is relatedness in a collection in HWE. It is also the standard measure of linkage disequilibrium, and variation in ancestry induces LD between sites. Finally it can be the primary "cause" of inflation of test statistics in genetic association studies. Overall, the resemblance between individuals with complex ancestry is a function of V_A, relatedness among those individuals, and the variation in allele frequency over ancestries.

4. Conclusions

In the two papers that comprise the third part of this series, we distinguish interactions arising from correlations in state between factors from those arising from non-additivity of independent factors.

Correlation in state of factors occurs commonly. Here we understand departures from HWE as a form of interaction caused by correlation in allelic state within a genotype, which can result from variation in allele frequency across ancestry. Correlation in state within loci has very complex impacts, and can cause changes in effects (changes in means) as well as increases or decreases in the total variance relative to the sum of the individual factor variances. In the presence of correlation in state, concepts such as dominance variance or interaction variance are often not well defined, and may appear to be negative, because non-independence of the state of factors can create negative covariance. In the presence of correlations in allelic state, a locus that would have been completely additive had it been in HWE has both additive and dominance variance; its total genetic variance is no longer the sum of the additive and dominance variance; the additive variance is no longer the same as $2pq\beta^2$, and heritability is now a function of the correlation in allelic state in each relative (which could differ between them). Correlations in allelic and genotypic state dramatically alter our intuitions and our quantitative genetic estimates, and thus should be accounted for explicitly.

Variation in ancestry can create correlation in state between every site in the genome. Whenever there is ancestry variation in a collection of examined individuals, any pair of loci that vary in frequency over ancestry will appear to be in linkage disequilibrium, even if those loci are physically unlinked. If there is additive variance for the trait, and if any of the alleles contributing to that variance differ in frequency among the ancestors of the study participants, then there will be test statistic inflation for any naive test correlating allele frequency with phenotype. Correlation in minor allele frequency induced by human demographic history is the primary driver of this inflation in association studies, and we reviewed several approaches to account for this inflation [20,22–24,26].

Correlation is genotypic state between two individuals is also a measure of the relatedness of those individuals. Departures from Hardy-Weinberg Equilibrium (HWE), linkage disequilibrium (LD), population structure, and relatedness are generally viewed by population geneticists as distinct concepts. However, as we have shown, departures from HWE, LD, and relatedness are all measures of the correlation in genotypic state among the individuals in a collection. While it is often

useful to treat the effects of each of these independently of one another, all of them can be understood as forms of genetic interaction caused by correlation in state—between alleles within a locus, between pairs of sites across the genome, and between individuals.

Declarations

Ethics Statement

Not applicable.

Consent for Publication

Not applicable.

Availability of Data and Material

Not applicable.

Funding

This study was supported by NIH Grants RF1 AG071170 and U01 DK134191.

Competing Interests

David J. Cutler is a member of the Editorial Board of the journal *Human Population Genetics and Genomics*. The author was not involved in the journal's review of or decisions related to this manuscript. The authors have declared that no other competing interests exist.

Author Contributions

All authors participated in the derivation, writing, and editing of this work.

Acknowledgments

This work has benefited from many helpful suggestions by Greg Gibson, and long conversations with Loic Yengo.

References

1. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015 Apr 17;11(4):e1004219. DOI
2. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 2013;9(8):e1003671. DOI
3. 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. Am J Hum Genet. 2011 Jul 15;89(1):82-93. DOI
4. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics. 2016 May 1;32(9):1423-1426. DOI
5. Jodeiry K, Bass AJ, Epstein MP, Cutler DJ. The Quantitative Genetics of Human Disease: 3b Interactions - Non-Additivity and Missing Heritability. Hum Popul Genet Genom. Forthcoming 2025.
6. Cutler D, Jodeiry K, Bass A, Epstein M. The quantitative genetics of human disease: 1. Foundations. Hum

Popul Genet Genom 2023; 3(4):0007. DOI
7. Kempthorne O. The Theoretical Values of Correlations between Relatives in Random Mating Populations. Genetics. 1955 Mar;40(2):153-167. DOI
8. Fisher RA. The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans R Soc Edinburgh. 1918;52:399–433.
9. Kempthorne O. The Correlation between Relatives in a Simple Autotetraploid Population. Genetics. 1955 Mar;40(2):168-174. DOI
10. Horner TW, Kempthorne O. The Components of Variance and the Correlations between Relatives in Symmetrical Random Mating Populations. Genetics. 1955 May;40(3):310-320. DOI
11. Kempthorne O. The Correlations between Relatives in Inbred Populations. Genetics. 1955 Sep;40(5):681-691. DOI
12. Gillespie, JH. Population Genetics: A Concise Guide. 2nd ed. Baltimore, Md: Johns Hopkins University Press, 2004.
13. Cutler D, Jodeiry K, Bass A, Epstein M. The Quantitative Genetics of Human Disease: 2 Polygenic Risk Scores. Hum Popul Genet Genom 2024;4(3):0008. DOI
14. Wahlund S. ZUSAMMENSETZUNG VON POPULATIONEN UND KORRELATIONSERSCHEINUNGEN VOM STANDPUNKT DER VERERBUNGSLEHRE AUS BETRACHTET. Heredity (Edinb). 1928;11(1):65-106. DOI
15. WRIGHT S. Genetical structure of populations. Nature. 1950 Aug 12;166(4215):247-249. DOI
16. Lewontin RC. The Apportionment of Human Diversity. In: Dobzhansky T, Hecht MK, Steere WC, editors. Evolutionary Biology. New York, NY: Springer US. 1972. pp. 381–398. DOI
17. Roychoudhury AK, Nei M. Human polymorphic genes: world distribution. New York: Oxford University Press. 1988.
18. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008 Feb 21;451(7181):998-1003. DOI
19. Nei M, Li WH. Linkage disequilibrium in subdivided populations. Genetics. 1973 Sep;75(1):213-219. DOI
20. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999 Dec;55(4):997-1004. DOI
21. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904-909. DOI
22. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000 Jun;155(2):945-959. DOI
23. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993 Mar;52(3):506-516.
24. Zhu X, Zhang S, Zhao H, Cooper RS. Association mapping, using a mixture model for complex traits. Genet Epidemiol. 2002 Aug;23(2):181-196. DOI
25. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J; Schizophrenia Working Group of the Psychiatric Genomics Consortium; et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015 Mar;47(3):291-295. DOI
26. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000 Jul;67(1):170-181. DOI
27. Hao K, Li C, Rosenow C, Wong WH. Detect and adjust for population stratification in population-based association study using genomic control markers: an application of Affymetrix Genechip Human Mapping 10K array. Eur J Hum Genet. 2004 Dec;12(12):1001-1006. DOI
28. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet. 2011 Jul;19(7):807-812. DOI
29. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015 Mar;47(3):284-290. DOI

30. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018 Sep;50(9):1335-1341. DOI
31. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet. 1995 Aug;57(2):455-464. DOI
32. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet. 1996 Nov;59(5):983-989.
33. Curtis D, Sham PC. A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet. 1995 Mar;56(3):811-812 DOI
34. Gordon D, Heath SC, Liu X, Ott J. A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am J Hum Genet. 2001 Aug;69(2):371-880. DOI
35. Mitchell AA, Cutler DJ, Chakravarti A. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. Am J Hum Genet. 2003 Mar;72(3):598-610. DOI