On genetic correlation

Christopher Salahub
University of Waterloo

May 6, 2022

1 Introduction

A structural model of genetics can be constructed which represents the genome of a diplodic individual by a two-column matrix

$$\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2], \ \mathbf{g}_1, \mathbf{g}_2 \in \mathcal{B}^{N_P}$$

where $\mathcal{B} = \{\text{adenine, guanine, cytosine, thymine}\}$ is the set of nucleotide bases and N_P is the length of the genome. In humans $N_P \approx 3,234,830,000$. Rather than measuring the whole genome, select M disjoint sequences of interest, called markers, with total length K and record these in

$$\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2], \ \mathbf{s}_1, \ \mathbf{s}_2 \in \mathcal{B}^K.$$

In most cases these disjoint segments are chosen from known single nucleotide polymorphisms, or SNPs, which account for the majority of variation in the coding of the human genome. Typically, SNPs are biallelic, and so take only one of two versions in the population. $\mathbf{S}$ can therefore be summarized into the M SNPs it represents by annotating which allele is present at each location. This can be done using upper- and lowercase letters, for example, to give

$$\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2], \ \mathbf{t}_1, \ \mathbf{t}_2 \in \{A, a\}^M. \tag{1}$$

These letters do not represent the same sequence when used at different locations, but rather only indicate which of the two alleles is present at a particular SNP. This annotated matrix serves as the basis of most genetic research, with conventions in notation and modelling going back to Mendel (1866) and Fisher (1919).

Genetic research is focused on the heritability of different traits. More directly, this means associating measurable physical traits such as height, eye colour, response to a drug, or the presence of a disease with the entries of $\mathbf{T}$. This must be done in spite of potentially confounding relationships present between different entries in $\mathbf{T}$ due to the process of inheritance itself. To account for inheritance, take the annotated matrices of the parents

$$\mathbf{F}_T = [\mathbf{f}_1, \mathbf{f}_2], \ \text{and} \ \mathbf{M}_T = [\mathbf{m}_1, \mathbf{m}_2], \tag{2}$$

thereby extending this structural model back a generation. We can now meaningfully talk about inheritance itself. Most crudely, inheritance involves the combination of independently donated variants from each of $\mathbf{F}_T$ and $\mathbf{M}_T$. Two additional processes may perturb the variants: independent assortment and cross overs.

Independent assortment is a well-known phenomenon in genetics, see Siegmund and Yakir (2007). While Equations 1 and 2 present variants as long columns of sequential base pairs, inside of cells these variants are actually organized into separate contiguous sections called chromosomes. Chromosomes within a parent are donated independently of each other. So while offspring may receive the variant from the first column on one chromosome, they can receive the variant from the second column on another. Let $\mathbf{c}$ be a vector of length M denoting the chromosomal membership of each marker. For marker indices j and k, independent assortment means that the variant donated at position j is independent of that at k if $c_j \neq c_k$ within a parent.

Cross overs add additional variation by perturbing sections for which $\mathbf{c}$ is constant. Within chromosomes, it is possible for the variants to physically cross at a base pair and swap between variants the sections after this cross to the end of the chromosome. This can actually occur several times on the same chromosome. In $\mathbf{T}$ cross overs result in swaps of sections of the columns where $\mathbf{c}$ is constant. Cross overs recombine the genome to create completely new variants.

2 Genetic correlation

To quantify association some measure of association must be applied to $\mathbf{T}$, $\mathbf{F}_T$, and $\mathbf{M}_T$. This quantification is a primary goal of genome-wide association studies, see Uffelmann et al. (2021); Tam et al. (2019); Wang et al. (2005). While many of the measures in Goodman and Kruskal (1979) could be used directly, a more common approach is to encode and summarize these annotated matrices numerically and compute **observed sample** correlations, given by

$$\widehat{r}^2(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}} \tag{3}$$

for $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$. If $\mathbf{x}$ and $\mathbf{y}$ are treated as realizations of the random variables X and Y respectively, this is the sample estimate of the **theoretical** correlation

$$r^2(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \tag{4}$$

These can then be used to understand the structure of the genome and its relation to physical traits, as in Cinar and Viechtbauer (2021); Li and Ji (2005); Nyholt (2004); Cheverud et al. (2001).

One possibility is the additive encoding and summary. First, for all of $\mathbf{T}, \mathbf{F}_T$, and $\mathbf{M}_T$ A is replaced by 1 and a by 0. Row-wise addition of this indicator of A can then be performed on $\mathbf{T}$ to obtain the vector

$$[z_1, z_2, \ldots, z_M]^\mathsf{T} \in \{0, 1, 2\}^M.$$

Repeating this for every individual in a population gives n such vectors. Equivalently, we obtain n observations of each of the M markers. Denote the j^{th} marker measurement on the i^{th} individual as z_{ij}, then measurements over a population can be placed in the $n \times M$ matrix

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \ldots & z_{1M} \\ z_{21} & z_{22} & \ldots & z_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \ldots & z_{nM} \end{bmatrix} := [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M],$$

where

$$\mathbf{z}_j = [z_{1j}, z_{2j}, \ldots, z_{Mj}]^{\mathsf{T}}.$$

Each $\mathbf{z}_j$ is a vector of n realizations of the random variable Z_j distributed according to the distribution of the j^{th} marker in the population. When considering a wild population, this distribution is likely to be unknown. If $\mathbf{M}_T$ and $\mathbf{F}_T$ are known constants for the entire population, however, the distribution of Z_j is dictated by cross overs and independent assortment alone. In either case $\mathbf{Z}$ has an **observed** pairwise correlation matrix

$$\widehat{\mathbf{R}} = \begin{bmatrix} Var(\mathbf{z}_1) & \widehat{r}^2(\mathbf{z}_1, \mathbf{z}_2) & \widehat{r}^2(\mathbf{z}_1, \mathbf{z}_3) & \ldots & \widehat{r}^2(\mathbf{z}_1, \mathbf{z}_{M-1}) & \widehat{r}^2(\mathbf{z}_1, \mathbf{z}_M) \\ \widehat{r}^2(\mathbf{z}_2, \mathbf{z}_1) & Var(\mathbf{z}_2) & \widehat{r}^2(\mathbf{z}_2, \mathbf{z}_3) & \ldots & \widehat{r}^2(\mathbf{z}_2, \mathbf{z}_{M-1}) & \widehat{r}^2(\mathbf{z}_2, \mathbf{z}_M) \\ \widehat{r}^2(\mathbf{z}_3, \mathbf{z}_1) & \widehat{r}^2(\mathbf{z}_3, \mathbf{z}_2) & Var(\mathbf{z}_3) & \ldots & \widehat{r}^2(\mathbf{z}_3, \mathbf{z}_{M-1}) & \widehat{r}^2(\mathbf{z}_3, \mathbf{z}_M) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \widehat{r}^2(\mathbf{z}_M, \mathbf{z}_1) & \widehat{r}^2(\mathbf{z}_M, \mathbf{z}_2) & \widehat{r}^2(\mathbf{z}_M, \mathbf{z}_3) & \ldots & \widehat{r}^2(\mathbf{z}_M, \mathbf{z}_{M-1}) & Var(\mathbf{z}_M) \end{bmatrix}.$$

For an arbitrary entry in this matrix, say $\widehat{r}^2(\mathbf{z}_j, \mathbf{z}_k)$, let c_j and c_k indicate the respective chromosomes of markers j and k. and suppose that these markers have a probability of recombination of p_r. If we assume that

- the population are all offspring of the known matrices $\mathbf{M}_T$ and $\mathbf{F}_T$;
- cross overs and independent assortment are the only sources of recombination;
- cross overs occur with perfect alignment across variants; and
- cross overs occur independently within chromosomes

then it can be shown that the **expected** theoretical correlation is given by

$$r^2(Z_j, Z_k) = I_{c_j}(c_k)\,\gamma(1 - 2p_r) \tag{5}$$

where

$$I_y(x) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise} \end{cases}$$

is the indicator function and $\gamma \in \{-1, -1/\sqrt{2}, 0, 1/\sqrt{2}, 1\}$ is a constant depending on the entries of $\mathbf{M}_T$ and $\mathbf{F}_T$. Ignoring the other markers which do not contribute to this pairwise correlation, the encodings of the annotated matrices $\mathbf{M}_T$ and $\mathbf{F}_T$ can be written

$$\mathbf{F}_X = \begin{bmatrix} f_{j1} & f_{j2} \\ f_{k1} & f_{k2} \end{bmatrix} \text{ and } \mathbf{M}_X = \begin{bmatrix} m_{j1} & m_{j2} \\ m_{k1} & m_{k2} \end{bmatrix}, \tag{6}$$

where all entries are 0 or 1. γ is most succinctly defined using the difference matrix

$$\mathbf{\Delta} = \begin{bmatrix} f_{j1} - f_{j2} & m_{j1} - m_{j2} \\ f_{k1} - f_{k2} & m_{k1} - m_{k2} \end{bmatrix} := \begin{bmatrix} \delta_{jF} & \delta_{jM} \\ \delta_{kF} & \delta_{kM} \end{bmatrix} \tag{7}$$

recording the difference between the two variants in each parent at each marker position. Using these differences

$$\gamma = \frac{\left[\delta_{jF}\delta_{kF} + \delta_{jM}\delta_{kM}\right]}{\sqrt{\left(\delta_{jF}^2 + \delta_{jM}^2\right)\left(\delta_{kF}^2 + \delta_{kM}^2\right)}}.$$

If we additionally assume that cross overs occur uniformly over the interval j to k and that this interval is sufficiently large, the map distance of Haldane (1919) arises automatically from this model. Supposing the interval j to k has an arbitrary length $d(j,k)$ measured in reference to a uniform recombination rate $\beta \in \mathbb{R}$, the probability of recombination is given by

$$p_r(d(j,k)) = \frac{1}{2}\left(1 - e^{-2\beta d(j,k)}\right)$$

and so

$$r^2(Z_j, Z_k) = I_{c_j}\left(c_k\right)\gamma e^{-2\beta d(j,k)}. \tag{8}$$

3 The distribution of correlation

While the point estimates of Equations 8 and 5 are useful to understand the mean behaviour of the correlations for large populations, they do not communicate the theoretical distributions. This makes an assesment of fit for any particular $\widehat{\mathbf{R}}$ difficult, requiring repeated simulations to obtain an empirical null distribution. As this results from treating $\mathbf{M}_X$ and $\mathbf{F}_X$ as constant in Equation 9, the random matrices

$$\mathcal{F}_X = \begin{bmatrix} F_{j1} & F_{j2} \\ F_{k1} & F_{k2} \end{bmatrix} \text{ and } \mathcal{M}_X = \begin{bmatrix} M_{j1} & M_{j2} \\ M_{k1} & M_{k2} \end{bmatrix}, \tag{9}$$

resolve this problem. Each of the entries in these matrices can be treated as a Bernoulli random variable with some probability of being 1. The matrices $\mathbf{M}_X$ and $\mathbf{F}_X$ can be repurposed to parametrize these variables by giving the expectations of each, thereby generalizing the constant case.

Additional random variables can be introduced to account for cross overs and the equal probability of either variant being donated. Let L_F and L_M be Bernoulli random variables which are 1 if the left column is donated by the corresponding random matrix. Similarly take C_F and C_M as Bernoulli random variables which are 1 if recombination due to cross overs occurs in the corresponding random matrix. Therefore

$$L_F, L_M \sim Bern\left(\frac{1}{2}\right) \tag{10}$$

with $L_F \perp\!\!\!\perp L_M$, and

$$C_F, C_M \sim Bern\left(p_r(d(j,k))\right) \tag{11}$$

with $C_F \perp\!\!\!\perp C_M$.

Denoting the relevant marker encodings of the offspring of $\mathcal{F}_X$ and $\mathcal{M}_X$ by

$$\mathcal{X} = \begin{bmatrix} X_{j1} & X_{j2} \\ X_{k1} & X_{k2} \end{bmatrix}. \tag{12}$$

We then get the stochastic representations

$$\begin{aligned}
X_{j1} &= (1 - C_F)\Big[L_F F_{j1} + (1 - L_F)F_{j2}\Big] + C_F\Big[(1 - L_F)F_{j1} + L_F F_{j2}\Big] \\
X_{j2} &= (1 - C_M)\Big[L_M M_{j1} + (1 - L_M)M_{j2}\Big] + C_M\Big[(1 - L_M)M_{j1} + L_M M_{j2}\Big] \\
X_{k1} &= (1 - C_F)\Big[L_F F_{k1} + (1 - L_F)F_{k2}\Big] + C_F\Big[(1 - L_F)F_{k1} + L_F F_{k2}\Big] \\
X_{k2} &= (1 - C_M)\Big[L_M M_{k1} + (1 - L_M)M_{k2}\Big] + C_M\Big[(1 - L_M)M_{k1} + L_M M_{k2}\Big]
\end{aligned}$$

which correspond to the values in the cases of left and right donation both with and without recombination. These terms can be simplified by collecting terms based on the entries of $\mathcal{F}_X$ and $\mathcal{M}_X$. Using X_{j1}, for example:

$$\begin{aligned}
X_{j1} &= \Big(L_F + C_F - 2C_F L_F\Big)F_{j1} + \Big(1 - L_F - C_F + 2C_F L_F\Big)F_{j2} \\
&= \Big(L_F^2 + C_F^2 - 2C_F L_F\Big)F_{j1} + \Big(1 - L_F^2 - C_F^2 + 2C_F L_F\Big)F_{j2} \\
&= (L_F - C_F)^2 F_{j1} + \Big[1 - (L_F - C_F)^2\Big]F_{j2}
\end{aligned}$$

as $L_F, C_F \in \{0, 1\}$. This can be simplified further by considering $(L_F - C_F)^2 := D_F$. As $L_F - C_F \in \{-1, 0, 1\}$ is a random difference, $D_F \in \{0, 1\}$ is a Bernoulli random variable. More specifically, considering the probability of $L_F - C_F \in \{-1, 1\}$ gives

$$D_F \sim Bern\left(\frac{p_r(d(j,k))}{2} + \frac{1}{2}(1 - p_r(d(j,k)))\right) = Bern\left(\frac{1}{2}\right).$$

This is unsurprising. Cross overs affect both variants symmetrically, and so there is no reason to expect one to be favoured over the other in inheritance. Extending this logic to the other entries of $\mathcal{X}$ and defining D_M similarly to D_F gives

$$\begin{aligned}
X_{j1} &= D_F F_{j1} + (1 - D_F)F_{j2} \\
X_{j2} &= D_M M_{j1} + (1 - D_M)M_{j2} \\
X_{k1} &= D_F F_{k1} + (1 - D_F)F_{k2} \\
X_{k2} &= D_M M_{k1} + (1 - D_M)M_{k2}
\end{aligned}$$

Noting that the random variables in $\mathcal{F}$ and $\mathcal{M}$ are all within $\{0, 1\}$, each of these is a mixture of two Bernoulli random variables.

The summary vector has entries

$$\begin{aligned}
Z_1 &= X_{j1} + X_{j2} \\
Z_2 &= X_{k1} + X_{k2}
\end{aligned}$$

and so understanding the distribution of the entries of $\mathcal{X}$ is key to characterizing the distribution of this summary. First note that $D_F \perp\!\!\!\perp D_M$ are assumed to be independent of $\mathcal{F}$ and $\mathcal{M}$. While it is possible the variant present affects the probability of recombination, such complexity would have a minor effect at most. Therefore

$$\begin{aligned}
E[X_{j1}] &= \frac{1}{2}\Big(E[F_{j1}] + E[F_{j2}]\Big) \\
E[X_{j2}] &= \frac{1}{2}\Big(E[M_{j1}] + E[M_{j2}]\Big) \\
E[X_{k1}] &= \frac{1}{2}\Big(E[F_{k1}] + E[F_{k2}]\Big) \\
E[X_{k2}] &= \frac{1}{2}\Big(E[M_{k1}] + E[M_{k2}]\Big)
\end{aligned}$$

where the expectations of each entry of $\mathcal{M}$ and $\mathcal{F}$ are population parameters. Using the conditional variance decomposition

$$\begin{aligned}
Var(X_{j1}) &= E\Big[Var(X_{j1}|D_F)\Big] + Var\Big(E[X_{j1}|D_F]\Big) \\
&= \frac{1}{2}\Big[Var(F_{j1}) + Var(F_{j2})\Big] + \frac{1}{2}\Big(E[F_{j1}]^2 + E[F_{j2}]^2 - \big(E[F_{j1}] + E[F_{j2}]\big)^2\Big) \\
&= Var(F_{j1}) + Var(F_{j2}) - E[F_{j1}]E[F_{j2}]
\end{aligned}$$

From Equations **??** and **??**, we get

$$E[Z_1] = \frac{1}{2}\Big(E[F_{j1}] + E[F_{j2}] + E[M_{j1}] + E[M_{j2}]\Big) \tag{13}$$

$$E[Z_2] = \frac{1}{2}\Big(E[F_{k1}] + E[F_{k2}] + E[M_{k1}] + E[M_{k2}]\Big). \tag{14}$$

While we can decompose the variance conditionally as

$$Var(Z_1) = E\Big[Var(Z_1|D_F, D_M)\Big] + Var\Big(E[Z_1|D_F, D_M]\Big).$$

$$E\Big[Var(Z_1|D_F, D_M)\Big] = \frac{1}{4}Var(F_{j1} + M_{j1}) + \frac{1}{4}Var(F_{j1} + M_{j2}) + \frac{1}{4}Var(F_{j2} + M_{j1}) + \frac{1}{4}Var(F_{j2} + M_{j2})$$

References

James M Cheverud, Ty T Vaughn, L Susan Pletscher, Andrea C Peripato, Emily S Adams, Christopher F Erikson, and Kelly J King-Ellison. Genetic architecture of adiposity in the cross of LG/J and SM/J inbred mice. *Mammalian Genome*, 12(1):3–12, 2001.

Ozan Cinar and Wolfgang Viechtbauer. *poolr: Methods for Pooling P-Values from (Dependent) Tests*, 2021. URL https://CRAN.R-project.org/package=poolr. R package version 1.0-0.

Ronald A Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

Leo A Goodman and William H Kruskal. *Measures of association for cross classifications*. Springer, 1979.

JBS Haldane. The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics*, 8(29):299–309, 1919.

J Li and L Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221–227, 2005.

Gregor Mendel. Versuche uber pflanzen-hybriden. *Verhandlungen des naturforschenden Vereins in Brunn*, 4:3–47, 1866.

Dale R Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765–769, 2004.

David Siegmund and Benjamin Yakir. *The statistics of gene mapping*. Springer Science & Business Media, 2007.

Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.

Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21, 2021.

William Wang, Bryan J Barratt, David G Clayton, and John A Todd. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6(2):109–118, 2005.