On genetic correlation

Christopher Salahub
University of Waterloo

May 3, 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 the section after this cross to the end between the variants. This can actually occur several times on the same chromosome, in fact. These cross overs correspond to swaps of sections of the columns in $\mathbf{T}$.

1.1 Genetic correlation

The annotation matrix $\mathbf{T}$ serves as the basis to quantify association in genetic research, whether between markers or with observed physical traits. This quantification is the primary goal of genome-wide association studies as surveyed in 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 with $\mathbf{T}$, a more common approach is to encode and summarize $\mathbf{T}$ numerically and compute the sample correlations, given by

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

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 such encoding is the additive encoding and summary. First, A is replaced by 1 and a by 0. Row-wise addition of this indicator of A is then performed to obtain a 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 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

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

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 in question. When considering a wild population, this distribution is likely to be unknown. Certain theoretical populations, however, prescribe the distribution of Z_j. 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 from identically annotated parents,
- the parent annotations are known,
- 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 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 determined by the annotated matrices of the parents crossed to generate the population. These annotated matrices can be written

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

where all entries are 0 or 1 without a loss of generality. γ is then most succinctly defined using the difference matrix

$$\mathbf{\Delta} = \begin{bmatrix} f_{j1} - f_{j2} & m_{j1} - m_{j2} \\ f_{k1} - f_{k2} & m_{k1} - m_{k2} \end{bmatrix} := \begin{bmatrix} \delta_{jF} & \delta_{jM} \\ \delta_{kF} & \delta_{kM} \end{bmatrix}, \tag{6}$$

which records the difference between the two variants in each parent at each marker position.

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

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

$$r^2(Z_j, Z_k) = I_{c_j}(c_k)\,\gamma e^{-2\beta d(j,k)}. \tag{7}$$

2 The distribution of correlation

While the point estimates of Equations 7 and 5 are useful

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.