On genetic correlation

Christopher Salahub
University of Waterloo

May 1, 2022

1 Introduction

A structural model of genetics can be constructed which represents the genome of an individual by a 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

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

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).

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

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

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

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

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

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

Repeating this for every individual in a population gives n vectors $\mathbf{z}_1$, $\mathbf{z}_2$, $\ldots$, $\mathbf{z}_n$. This generates n observations of each marker which can similarly be summarized in the vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M \in \mathbb{R}^n$. These can be placed in an $n \times M$ matrix so that each individual's vector takes up a row to give

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1{}^\mathsf{T} \\ \vdots \\ \mathbf{z}_n{}^\mathsf{T} \end{bmatrix} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M],$$

which has the pairwise correlation matrix

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

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

- the $\mathbf{z}_i$ are offspring of identically annotated parents,
- cross overs and independent assortment are the only sources of recombination, and
- cross overs occur independently within chromosomes

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

$$r^2(X_j,X_k) = I_{\{c_j\}}(c_k)\,\gamma(1-2p_r) \tag{3}$$

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 parents crossed to generate the population.

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

$$Corr(Z_j, Z_k) = I_{\{c_j\}}(c_k)\,\gamma e^{-2\beta d(j,k)}. \tag{4}$$

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.

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.