Using a structural genetic model to derive map distances and correlation

Chris Salahub
University of Waterloo

March 14, 2022

1 Introduction

Genetic research today routinely considers the entirety of a genome to identify the regions most strongly related to measured physical traits. The goal is to associate measured genome sequences, the *genotype*, with physical characteristics, the *phenotype*. Computational and methodological advances in the pursuit of these *quantitative trait loci* (QTLs) have distinguished *genomics* as its own field. Central to genomics is the *genome-wide association study* (GWAS), where many *markers*, sequences of nucleotides at known positions on the genome, are measured. Extracting useful results from markers and measured traits is a complicated task which has motivated decades of statistical and biological research. Interested individuals are left to sort through this research in order to understand genomics.

$$\mathbf{G} \xrightarrow{\ select\ } \mathbf{S} \xrightarrow{\ annotate\ } \mathbf{T} \xrightarrow{\ encode\ } \mathbf{X} \xrightarrow{\ summarize\ } \mathbf{z}$$

Figure 1: A structural model of a GWAS.

To aid in this, Figure 1 draws on the literature to present a simple structural model of the process taken in GWAS to convert raw marker measurements into a form which can be used to identify QTLs. It identifies four key steps (*selection*, *annotation*, *encoding*, and *summarization*) between five increasingly abstract representations of the genome, where the genome is defined by Doerge et al. (1997) as all heritable material potentially passed from parent to offspring. By highlighting the abstractions and steps in plain language, this simplified model provides a convenient map to guide the understanding of GWAS. This is not a replacement for surveys such as Uffelmann et al. (2021); Tam et al. (2019), instead it provides a guiding structural framework with exceptional explanatory power to facilitate understanding of other papers in the field.

The model starts with **G**, the whole genome of an individual organism. Genetic information is stored in DNA, a long molecule consisting of a sequence of four *nucleotide bases*: guanine, cytosine, adenine, and thymine. A *diplodic* individual inherits one version or *variant* of a complete DNA sequence from each parent, and so has two copies in all *somatic* (i.e. non-reproductive) cells. Though it can be represented as one long sequence, DNA is actually structured into *chromosomes*, separate strands of DNA which contain only a part of the sequence. As most genetic research concerns diplodic species, this will be implicitly assumed.

It is usually not feasible or desirable to design a study around the measurement of all of **G**, and so the *select* step chooses regions to measure. These regions are represented by **S**. Often **S** consists of a series of *single nucleotide polymorphisms* (SNPs), single nucleotide substitutions in a known sequence at a known position. In human studies this is supported by SNP databases such as NCBI (2021) which document hundreds of millions of common SNPs in the human genome. Only a small proportion of these are estimated to occur frequently enough in the population to be useful in a GWAS, perhaps 15 million according to Koboldt et al. (2013). Modern SNP arrays can simultaneously identify roughly one million of these per array, see LaFramboise (2009); Tam et al. (2019), and most GWAS will measure an array's worth of SNPs. *Linkage disequilibrium*, effectively the correlation between regions of the genome, facilitates inference to regions outside of those selected in **S**. While third generation genome sequencing technologies allow for entire genomes to be sequenced, as noted in Heather and Chain (2016); Hasin et al. (2017); Uffelmann et al. (2021), persistent high costs of next generation technologies and more than a decade of SNP microarray development leave microarrays as the dominant measurement method.

After selecting SNPs to obtain **S**, researchers must *annotate* the raw data. The raw signal produced by an SNP microarray is fluorescence, with different degrees of fluorescence corresponding to a different genotypes. Converting the fluorescent areas of an array to a genotype is a challenging problem and has developed in tandem with the arrays themselves. Early models used non-parametric clustering techniques on the signal from several microarray pores, but more complex hidden Markov and Bayesian models have also been developed. LaFramboise (2009) details some of these. Whatever method is used, the selected regions are assigned genotypes in **T** denoted with capital or lowercase letters at each SNP, as in Siegmund and Yakir (2007); Visscher and Goddard (2019).

Finally, relationships between **T** and an observed trait or within **T** itself are quantified by converting each annotated SNP to a number. To do this, GWAS first *encode* each SNP variant with a numeric value and then *summarize* the pairs at each location into a number. Typically no distinction is made between these steps: Lander and Botstein (1989); Cheverud (2001); Siegmund and Yakir (2007) detail the *dominance* and *additive* summaries by moving directly from a genotype to a numeric value. It is useful for clarity and full generality to separate the two distinct steps involved in this process, however.

This paper presents the details of this structural model. By using mathematical notation for each of the abstractions, a framework with extraordinary explanatory power is devised. Section 2 provides a detailed explanation of the model with all the necessary notation. The model is then used in a novel derivation of the Haldane *map distance*, a common measure used to locate SNPs, in Section 3. The utility of the model if further demonstrated in Section 4, where the model is used to derive the correlation between markers under classic breeding population settings. This results in a convenient expression of the correlation between markers in any genetic study. Finally, Section 6 compares the results of this derivation directly to panel data in mice.

2 A structural genetic model

The structural model starts with

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

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 $N_P = 3,234,830,000$. $\mathbf{G}$ represents the whole genome of an individual, with both the maternal and paternal variants of all chromosomes placed sequentially in adjacent columns. Both of these variants are complete, double-stranded sequences of DNA, but nucleotides pair uniquely. Adenine binds exclusively with thymine and guanine exclusively binds with cytosine. Therefore $\mathbf{g}_1$ and $\mathbf{g}_2$ record the pattern only for one of the two DNA strands for each column, the complementary strand being implied by this sequence.

Rather than address the whole genome, GWAS typically deal with a selected subset of segments on chromosomes of interest. This is represented by

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

with $K \ll N_P$. This corresponds to a mapping from $\mathbf{G} \to \mathbf{S}$ where K rows of $\mathbf{G}$ are chosen or sampled to create $\mathbf{S}$. These rows are not typically selected at random, but are motivated by previous work and databases of SNPs and other known markers. Most commonly, then, the mapping $\mathbf{G} \to \mathbf{S}$ is a selection of $M < K$ disjoint sequences from $\mathbf{G}$.

In the case of SNPs, the markers are most often *biallelic*, i.e. the population is dominated by two different sequences or *alleles* at the marker. These are often denoted using two different letters, such as A and B, or analogously the uppercase and lowercase version of the same letter, such as A and a. Converting the measured markers to letters is called annotation, which maps $\mathbf{S} \to \mathbf{T}$ with

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

Denoting the ith position of $\mathbf{t}_j$ as t_{ij}, $t_{lj} = A$ and $t_{mj} = A$ do not represent identical sequences at positions l and m. Instead this indicates that the sequences annotated by the capital at each position are present at their respective positions.

These annotated variants in $\mathbf{T}$ might next be converted to a numeric form. This is a mapping $\mathbf{T} \to \mathbf{X}$ such that

$$\mathbf{X} := [\mathbf{x}_1 | \mathbf{x}_2], \ \mathbf{x}_1, \ \mathbf{x}_2 \in \mathbb{R}^M.$$

Commonly this is even more restricted with $\mathbf{x}_j \in \{0, 1\}^M$ where

$$x_{ij} = \begin{cases} 1, & \text{if } t_{ij} = A \\ 0, & \text{if } t_{ij} = a \end{cases}, \tag{1}$$

is an indicator of the presence of the allele denoted with a capital.

Finally, $\mathbf{X}$ may be converted into a vector

$$\mathbf{z} \in \mathbb{R}^M$$

summarizing the individual's inherited variants. There are many common mappings $\mathbf{X} \to \mathbf{z}$. The *dominance mapping* takes $z_i = \max\{x_{i1}, x_{i2}\}$, the *homozygous mapping* uses $z_i = I_{x_{i2}}(x_{i1})$, and the *additive map* is $\mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are given according to Equation 1 and $I_y(x)$ is the indicator function

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

The additive map gives $\mathbf{z} \in \{0, 1, 2\}^M$ so z_i is equal to the count of copies of A at the ith marker across both of an individual's inherited variants.

Figure 1 displays this model, with descriptive names added to each mapping. In the first step, $\mathbf{G} \to \mathbf{S}$, we *select* segments of the entire genome to obtain the marker sequences or interest. The next step, $\mathbf{S} \to \mathbf{T}$, *annotates* the chosen markers by indicating which of the common alleles is present for that marker. These annotations are then converted to numeric values, or *encoded*, in the step $\mathbf{T} \to \mathbf{X}$. Finally, we *summarize* the matrix $\mathbf{X}$ into a vector $\mathbf{z}$ with some row-wise operation.

3 Deriving map distance

The structural model presented in Section 2 has incredible explanatory power beyond being a clear guide to GWAS data. With only a few assumptions, the model shows the Haldane map distance to be a corollary of the structure of DNA and mechanics of inheritance as understood today. This is in contrast to the typical derivation of map distance, which is based on a differential equation agnostic to the structure of genetic material, as in Kosambi (1943) and Xu (2013). The derivation of the Haldane map is outlined here. We begin with a simple sketch of sexual reproduction.

3.1 Sexual reproduction

Sexual reproduction is the recombination of the genome of two parents to create offspring genetic distinct from both. To discuss sexual reproduction, a distinction must be made between reproductive or *sex* cells, e.g. sperm, and non-reproductive or *somatic* cells. Sex cells contain only one variant of the genome, while somatic cells contain two.

Recall that there exist two variants of each chromosome within every somatic cell, a paternal variant and a maternal variant. Introduce two new matrices to represent the maternal and paternal genomes of which $\mathbf{G}$ is the offspring:

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

where $\mathbf{m}_1, \mathbf{m}_2, \mathbf{f}_1, \mathbf{f}_2 \in \mathcal{B}^{N_P}$. These represent the genomes of the parents of $\mathbf{G}$. Crudely, sexual reproduction is the construction of $\mathbf{G}$ from one random column of $\mathbf{M}$ and one random column of $\mathbf{F}$. So, $\mathbf{G}$ could be $[\mathbf{m}_1 | \mathbf{f}_2]$, for example.

The real mechanism is much more complex. During meiosis, the production of sex cells, the columns of $\mathbf{M}$ and $\mathbf{F}$ are perturbed. Rather than being inherited by $\mathbf{G}$ in the same form as in $\mathbf{M}$ and $\mathbf{F}$, regions in $\mathbf{f}_1$ may swap with regions in $\mathbf{f_2}$ and the same may occur with $\mathbf{m}_1$ and $\mathbf{m}_2$. This occurs either due to the *independent assortment of chromosomes* or due to the *crossing over* of variants.

Independent assortment is a direct consequence of the structure of the genome in somatic cells. Each chromosome is a separate molecule and so when sex cells are created, the parental variant inherited by offspring is independent of other chromosomes. This means that the paternal and maternal variants of a chromosome are equally likely to be passed on regardless of the variant another chromosome passes on.

Additionally, these variants may not be inherited identically as they appear in $\mathbf{M}$ or $\mathbf{F}$. There is a chance that the variants in a parent physically cross over each other while separating to form sex cells. Occasionally, this crossing results in a swap of the entire chromosome on either side of the cross, creating two completely new variants to pass on.

the physical process of producing sex cells cause the swapping of column entries of $\mathbf{M}$ and $\mathbf{F}$ for large, contiguous sections of rows, and the resulting swapped columns are then inherited by $\mathbf{G}$ as usual. These swaps are known as cross oversPhysically, this process is a result of the condensation of chromosomes of $\mathbf{M}$ and $\mathbf{F}$ into structures called *chromatids* which pair along the centre of a meiosing cell. The resulting proximity creates the opportunity for these variants to physically cross over each other. Occasionally, this occurs in such a way that entire sections of the genome are switched between them. Without the genetic recombination resulting from this, $\mathbf{G}$ would inherit one unmodified variant from each of $\mathbf{M}$ and $\mathbf{F}$, preventing totally new combinations of genetic material from occurring.

These variants are not

3.2   Modelling cross overs

To model crossing over, begin by making the assumption that crossing over occurs independently for each chromosome, and will affect directly only that chromosome's variants.[1] Consider a vector $\mathbf{h} \in \{1, \ldots, C\}^{N_P}$ which denotes the chromosome of each row of $\mathbf{M}$[2]. For simplicity, assume that for all $i \leq j$, $h_i \leq h_j$, that is all base pairs of a chromosome appear in adjacent rows. As crossing over occurs independently for each chromosome, a crossing over event in chromosome c, say, will affect only those rows of $\mathbf{M}$ where $\mathbf{h} = c$. For the moment, then, consider the case where $\mathbf{h}$ is a vector of ones, that is the case of a single chromosome.

Under this setting, let a cross over begin at the ith base pair. That is to say, suppose the chromosomes physically cross at the ith base pair.[3] Each variant is consquently separated into two parts: the part up to, but not including, the ith base pair, and the part from the ith base pair until the end. These two parts are then swapped between the variants, so that the first part of one variant forms a new chromosome with the second part of the other. Whenever the verb "begin" is used in the context of an index in a cross over, it will refer to this conceptualization, which corresponds to a swap of the values in the first $i - 1$ rows of $\mathbf{M}$. In order to track cross overs, introduce an indicator vector $\mathbf{v} = (v_1, \ldots, v_{N_P})^{\mathsf{T}}$ where

$$V_i = \begin{cases} 1 & \text{if a cross over beginning at base pair } i \text{ occurs,} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

and define $\boldsymbol{\pi}$ so that $\pi_i = P(V_i = 1)$. This can be done without loss of generality, as the order of crossing over events does not affect the final chromosome in this setting. Any chromosome in a sex cell for which a cross over has occurred is known as *recombinant*.

---

[1]Dependence between cross overs is known as *interference*.

[2]Or, equivalently, $\mathbf{F}$.

[3]We have an additional simplfying assumption here: that the chromosome will always be aligned such that the ith position on one variant will match with the ith on the other during a cross over.

As we rarely sequence the entire genome of an individual's somatic and sex cells, we will seldom see **M** and its recombinant forms. Instead, **S** is derived from **G** and $\mathbf{M}_S$ and $\mathbf{F}_S$ from **M** and **F**. From this, swaps of the markers of **S**, $\mathbf{M}_S$, and $\mathbf{F}_S$ between generations are used to estimate the number of sex cells containing recombinant chromosomes.[4] However, this method tells us nothing of how many cross over events occurred between any two markers. Any odd number of events leads to a swap, while any even number will be undetectable. Under this setting, the true count of indices i for which $v_i = 1$ cannot be known, and hence the π_i cannot be estimated individually.

3.3 Simplifying Assumptions

Fortunately, if we only care about the recombination of two particular markers on the genome, estimating individual π_i values is unnecessary. Consider two such positions, j and k with $j < k$, and note that cross overs beginning at $j+1, j+2, \ldots, k-1, k$ all result in these positions being split between variants. Now, motivated by identifiability, assume that $\pi_j = \pi_{j+1} = \cdots = \pi_{k-1} = \pi_k = \pi_{j:k}$. Under this assumption, the number of cross overs beginning in $j+1, j+2, \ldots, k-1, k$, is given by the binomial expression

$$P(N_c = n_c) = \binom{k-j}{n_c} \pi_{j:k}^{n_c} (1 - \pi_{j:k})^{k-j-n_c},$$

where N_c is a random variable giving the count of cross overs in the region. For convenience, let $r = k - j$ and $\pi = \pi_{j:k}$, which gives

$$P(N_c = n_c) = \binom{r}{n_c} \pi^{n_c} (1 - \pi)^{r-n_c}.$$

Note here that r is a unitless count of base pairs between positions j and k. Recognizing that markers are often separated by a great number of base pairs, and so r will typically be very large, we next take the limit of this expression as $r \to \infty$:

$$\lim_{r \to \infty} P(N_c = n_c) = \lim_{r \to \infty} \binom{r}{n_c} \pi^{n_c} (1 - \pi)^{r-n_c}.$$

At this point, an arbitrary substitution can be made. Consider the substitution $\pi = \frac{\beta d(j,k)}{r} := \frac{\beta d}{r}$. By this substitution, the probability π is reparameterized by a rate parameter, β, a distance measure, $d(j,k)$, and the r base pairs separating j and k. As the units of β and d will always result in a unitless product, the choices of β and d are a matter of individual discretion. Any convenient distance d can be chosen and will invoke a corresponding β. If physical distance, for example in angstroms, were used, then β would correspond to a rate of cross overs per unit length. One could alternatively use $d(j,k) = k - j$ to get a rate per base pair. As such a substitution is arbitrary, it gives a great deal of flexibility to choose a convenient set of units for measurement or understanding. Performing the substitution,

---

[4]The proportion of sex cells produced with such a swap is called the *recombination rate* for the pair of markers.

$$\lim_{r \to \infty} P(N_c = n_c) = \lim_{r \to \infty} \frac{r(r-1)\dots(r-n_c)}{n_c!} \left(\frac{\beta d}{r}\right)^{n_c} \left(1 - \frac{\beta d}{r}\right)^{r-n_c}$$

$$= \lim_{r \to \infty} \frac{r^{n_c} + O(r^{n_c-1})}{n_c!} \left(\frac{\beta d}{r}\right)^{n_c} \left(1 - \frac{\beta d}{r}\right)^{r-n_c}$$

$$= \lim_{r \to \infty} \frac{r^{n_c} + O(r^{n_c-1})}{r^{n_c}} \left(\frac{(\beta d)^{n_c}}{n_c!}\right) \left(1 - \frac{\beta d}{r}\right)^{r-n_c} \tag{3}$$

$$= \frac{(\beta d)^{n_c}}{n_c!} \lim_{r \to \infty} \frac{r^{n_c} + O(r^{n_c-1})}{r^{n_c}} \left(1 - \frac{\beta d}{r}\right)^{r-n_c}$$

$$= \frac{(\beta d)^{n_c}}{n_c!} e^{-\beta d},$$

which is the Poisson limit theorem for the binomial distribution.

Recall that if N_c is odd, it will result in a swap of markers j and k between variants, while if N_c is even, there will be no resultant swap. Define the recombination probability $p_r(d)$, which gives the probability of observing a swap for positions j and k with distance $d(j,k) := d$ between them. Then $p_r(d)$ is given by a sum of all odd terms of the above distribution. Therefore,

$$p_r(d) = \sum_{l=0}^{\infty} \frac{(\beta d)^{2l+1}}{(2l+1)!} e^{-\beta d}$$

$$= e^{-\beta d} \sum_{l=0}^{\infty} \frac{(\beta d)^{2l+1}}{(2l+1)!} \tag{4}$$

$$= e^{-\beta d} \left(\frac{e^{\beta d} - e^{-\beta d}}{2}\right)$$

$$= \frac{1}{2} \left(1 - e^{-2\beta d}\right).$$

A final substitution converts Equation 4 to a form familiar to researchers in genomics. Setting $\beta = \frac{1}{100}$ so that each each unit increase in d corresponds to a 0.01 increase in the expected number of crossing over events gives us Haldane's formula for the *map distance* in *centiMorgans*. These distance units are defined

purely by the assumptions leading to Equation 3 and the relationship specified in Equation 4 with the choice of $\beta = \frac{1}{100}$, rather than reflecting a true physical distance.[5]

4   Genetic correlation

Recall $\mathbf{z}$ as depicted in Figure 1 and described in the beginning of Section 1. Typically, when the notion of the correlation between markers is discussed in a GWAS, what is actually under discussion is the observed correlation matrix of this vector in a particular population (e.g. Cheverud (2001)). As a consequence, this quantity is key in the determination of the effective number of tests in GWASs. Despite the prevalence of simulation to compute this matrix in the literature, it can also be determined analytically.

For clarity, let $\mathbf{z}$ indicate an instance of the random vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_M)^\mathsf{T}$. We let the random vector $\mathbf{Z}$ follow the distribution of the summarized values $\mathbf{z}$ in a particular population. This population may be real, as is the case when this modelling is used in practice, or purely hypothetical, as will be the case in the following analysis.

Return to the annotated matrix $\mathbf{T}$ and consider two markers at row indices j and k. Introduce $\mathbf{c}$, which is defined similarly to $\mathbf{h}$ earlier, but now indicates chromosomal membership for the markers in $\mathbf{T}$ rather than the base pairs in $\mathbf{G}$. As every marker is contained within a single chromosome, $\mathbf{c}$ is always unambiguously defined.

There are two cases. Either j and k are on the same chromosome, that is $h_j = h_k$, or they are not, and so $h_j \neq h_k$. If these markers are not on the same chromosome, we can use the assumptions of Section 3.2 to see immediately that there will be no correlation between Z_j and Z_k, as these markers will assort independently with their different chromosomes. If they are on the same chromosome, let $d(j, k) = d$ be the distance between them measured in centiMorgans. Denote the dominant and recessive alleles with A and a respectively for j and use B and b analogously for k. Assume that the pairwise association of these markers in the population is of interest, i.e. that we can ignore all other markers on this chromosome in our analysis. Under this setting, we may consider a radically simplified $\mathbf{T}$, with 2 rows rather than M and taking the form

$$\mathbf{T} = \begin{bmatrix} A & a \\ b & B \end{bmatrix},$$

where the letters placed above are merely demonstrative. A simplified version of $\mathbf{X}$ follows immediately from this $\mathbf{T}$. Consider

$$\mathbf{X} = \begin{bmatrix} x_{j1} & x_{j2} \\ x_{k1} & x_{k2} \end{bmatrix},$$

with all entries in $\{0, 1\}$. As was the case for $\mathbf{z}$, we can treat these lowercase entries as realizations of random variables X_{rs}, $r \in \{j, k\}$, $s \in \{1, 2\}$.[6] Using this form of $\mathbf{X}$, we can consider $Cor(Z_j, Z_k)$ for the

---

[5]One may imagine this distance measure was chosen out of pure convenience. J.B.S. Haldane devised the approximation in 1919, at which point the chemical structure of DNA was still a mystery, but the concept of recombination had been revealed through patterns of inheritance.

[6]No additional notation is used to denote whether $\mathbf{X}$ is the random matrix or a realization, as this distinction is not relevant in the analysis presented here.

population resulting from an arbitrary cross of two parents. Then $\mathbf{X}$ implies a $\mathbf{Z}$ of

$$\mathbf{Z} = \begin{bmatrix} Z_j \\ Z_k \end{bmatrix} = \begin{bmatrix} X_{j1} + X_{j2} \\ X_{k1} + X_{k2} \end{bmatrix}.$$

The mechanics of sexual reproduction outlined in Section 3.1 and the genotype of the parents crossed to create $\mathbf{X}$ determine the distribution of Z_j and Z_k. Recall $\mathbf{M}$ and $\mathbf{F}$ introduced alongside sexual reproduction. Introduce simplified, annotated forms of these matrices here to represent the paternal and maternal encodings

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

where all entries are once again in $\{0,1\}$. Begin by assuming that $\mathbf{F}_X$ and $\mathbf{M}_X$ are known constants[7], and the statistical properties of $\mathbf{Z}$ result purely from the mechanics of reproduction.

Begin by considering the expectation of $\mathbf{Z}$. Due to the independent assortment of chromosomes, X_{j1} is equally likely to be either f_{j1} or f_{j2}, and so takes a uniform distribution over these two possibilities. A similar logic for all other entries in $\mathbf{X}$ applies, and so

$$E[\mathbf{Z}] = \begin{bmatrix} E[X_{j1}] + E[X_{j2}] \\ E[X_{k1}] + E[X_{k2}] \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} f_{j1} + f_{j2} + m_{j1} + m_{j2} \\ f_{k1} + f_{k2} + m_{k1} + m_{k2} \end{bmatrix},$$

from which it follows

$$Var(Z_j) = E[(X_{j1} + X_{j2})^2] - E[Z_j]^2$$
$$= \frac{1}{4} \left[(f_{j1} + m_{j1})^2 + (f_{j2} + m_{j1})^2 + (f_{j1} + m_{j2})^2 + (f_{j2} + m_{j2})^2 \right]$$
$$- \frac{1}{4}(f_{j1} + f_{j2} + m_{j1} + m_{j2})^2.$$

This can be simplified to give

$$Var(Z_j) = \frac{1}{4} \left[(f_{j1} - f_{j2})^2 + (m_{j1} - m_{j2})^2 \right]. \tag{5}$$

Analogously,

$$Var(Z_k) = \frac{1}{4} \left[(f_{k1} - f_{k2})^2 + (m_{k1} - m_{k2})^2 \right]. \tag{6}$$

Considering the covariance:

$$Cov(Z_j, Z_k) = Cov(X_{j1} + X_{j2}, X_{k1} + X_{k2})$$

$$= Cov(X_{j1}, X_{k1}) + Cov(X_{j1}, X_{k2}) + Cov(X_{j2}, X_{k1}) + Cov(X_{j2}, X_{k2}). \tag{7}$$

---

[7]There are many theoretical populatins where this is true, such as the F_2 intercross, where $f_{11} = m_{11} = f_{21} = m_{21} = 1$ and $f_{12} = m_{12} = f_{22} = m_{22} = 0$.

So the covariance is re-expressed as a sum of four terms, each of which can then be considered in turn.

This can be further simplified by considering $Cov(X_{j1}, X_{k2})$ and $Cov(X_{j2}, X_{k1})$. Both of these terms measure the covariance between values on the diagonals of $\mathbf{X}$, that is the covariance between the maternally and paternally donated variants of the genome inherited from $\mathbf{F}_X$ and $\mathbf{M}_X$, respectively. These covariances therefore measure the degree of *inbreeding* in a population, the degree to which parents tend to have the same genotype. The restricted setting of known parents derived here makes these diagonal values independent of each other and therefore uncorrelated, which can be confirmed by tedious algebra. Explicitly, $Cov(X_{j1}, X_{k2}) = Cov(X_{j2}, X_{k1}) = 0$. In unrestricted settings Crow and Kimura (1970) quantify this with the coefficient r under *assortative mating*, the tendency of individuals to choose mates with a similar phenotype.

The second pair of terms, $Cov(X_{j1}, X_{k1})$ and $Cov(X_{j2}, X_{k2})$, measure the covariance of encodings on the same variant, and so cannot be so easily reduced. Instead, consider $Cov(X_{j1}, X_{k1})$ and expand:

$$Cov(X_{j1}, X_{k1}) = E[X_{j1}X_{k1}] - E[X_{j1}]E[X_{k1}].$$

The independent inheritance of variants from a parent gives $E[X_{j1}] = \frac{1}{2}(f_{j1} + f_{j2})$ and $E[X_{k1}] = \frac{1}{2}(f_{k1} + f_{k2})$. Next consider $E[X_{j1}X_{k1}]$.

There are four possible values of $X_{j1}X_{k1}$, corresponding to inheritance of either of the two parental variants with or without recombination. If no recombination occurs, an event with probability $1 - p_r(d)$, either $f_{j1}f_{k1}$ or $f_{j2}f_{k2}$ is inherited with equal probability. If a cross over between j and k leads to recombination, then either $f_{j1}f_{k2}$ or $f_{j2}f_{k1}$ is passed on with equal probability. Accounting for these four possibilities gives

$$E[X_{j1}X_{k1}] = (1 - p_r(d))\left(\frac{1}{2}f_{j1}f_{k1} + \frac{1}{2}f_{j2}f_{k2}\right) + p_r(d)\left(\frac{1}{2}f_{j1}f_{k2} + \frac{1}{2}f_{j2}f_{k1}\right).$$

Combining this with the expectations of X_{j1} and X_{k1} gives

$$Cov(X_{j1}, X_{k1}) = E[X_{j1}X_{k1}] - E[X_{j1}]E[X_{k1}]$$

$$= (1 - p_r(d))\left(\frac{1}{2}f_{j1}f_{k1} + \frac{1}{2}f_{j2}f_{k2}\right) + p_r(d)\left(\frac{1}{2}f_{j2}f_{k1} + \frac{1}{2}f_{j1}f_{k2}\right)$$
$$- \frac{1}{4}(f_{j1} + f_{j2})(f_{k1} + f_{k2}) \tag{8}$$

$$= \frac{1}{4}(1 - 2p_r(d))(f_{j1}f_{k1} + f_{j2}f_{k2} - f_{j2}f_{k1} - f_{j1}f_{k2})$$

$$= \frac{1}{4}(1 - 2p_r(d))(f_{j1} - f_{j2})(f_{k1} - f_{k2}).$$

The same logic can be applied to $Cov(X_{j2}, X_{k2})$ to obtain

$$Cov(X_{j2}, X_{k2}) = \frac{1}{4}(1 - 2p_r(d))(m_{j1} - m_{j2})(m_{k1} - m_{k2}). \tag{9}$$

We obtain the covariance of Z_j and Z_k by adding the above and Equation 7. Substituting Equations 8 and 9 and $Cov(X_{j1}, X_{k2}) = Cov(X_{j2}, X_{k1}) = 0$ gives

$$Cov(Z_j, Z_k) = \frac{1}{4}(1 - 2p_r(d))\left[(f_{j1} - f_{j2})(f_{k1} - f_{k2}) + (m_{j1} - m_{j2})(m_{k1} - m_{k2})\right]. \tag{10}$$

Finally, Equations 5, 6, and 10 can be combined to determine the correlation:

$$Corr(Z_j, Z_k) = \frac{Cov(Z_j, Z_k)}{\sqrt{Var(Z_j)Var(Z_k)}}$$

$$= \frac{\frac{1}{4}(1 - 2p_r(d))\left[(f_{j1} - f_{j2})(f_{k1} - f_{k2}) + (m_{j1} - m_{j2})(m_{k1} - m_{k2})\right]}{\frac{1}{4}\sqrt{\left[(f_{j1} - f_{j2})^2 + (m_{j1} - m_{j2})^2\right]\left[(f_{k1} - f_{k2})^2 + (m_{k1} - m_{k2})^2\right]}} \tag{11}$$

$$= (1 - 2p_r(d))\frac{(f_{j1} - f_{j2})(f_{k1} - f_{k2}) + (m_{j1} - m_{j2})(m_{k1} - m_{k2})}{\sqrt{\left[(f_{j1} - f_{j2})^2 + (m_{j1} - m_{j2})^2\right]\left[(f_{k1} - f_{k2})^2 + (m_{k1} - m_{k2})^2\right]}}$$

$$:= (1 - 2p_r(d))\gamma.$$

So, the correlation is seen to be a product of $(1 - 2p_r(d))$, which depends on the markers in question, and a factor γ, which depends on the parents being crossed. An even simpler expression is obtained by substituting the Haldane recombination probability from Equation 4 in place of $p_r(d)$:

$$Corr(Z_j, Z_k) = (1 - 2p_r(d))\gamma$$

$$= \left(1 - 2\left[\frac{1}{2}\left(1 - e^{-2\beta d}\right)\right]\right)\gamma \tag{12}$$

$$= \gamma e^{-2\beta d},$$

and so the using the Haldane map distance the correlation between Z_j and Z_k decays exponentially in $d(1, 2)$ with a constant γ determined by the parents being crossed.

As $f_{rs}, m_{rs} \in \{0, 1\}$, the differences used to determine γ are all either -1, 0, or 1. There are therefore $3^4 = 81$ potential γ values, those most of these are not unique.

Recalling that j and k were restricted to be markers on the same chromosome, this pairwise result can be immediately generalized to the correlation matrix $\mathbf{Z}$ for the markers measured on an entire genome. For markers on the same chromosome, non-trivial correlations will be proportional to $1 - 2p_r(d)$ where $p_r(d)$ indicates the probability of recombination as a function of the distance between markers. Based on the independent assortment of different chromosomes, the correlations will be zero for any pair j and k not on the same chromosome.

In other words, if $h_j = h_k$, Equation 11 dictates the correlation between Z_j and Z_k. On the other hand, if $h_j \neq h_k$ the correlation between Z_j and Z_k will be zero. This implies a block diagonal structure

corresponding to the chromosomes with correlations dictated by the probability of recombination within each chromosome. Most generally

$$Corr(Z_j, Z_k) = I_{\{h_j\}}(h_k)\,\gamma(1 - 2p_r(d)), \tag{13}$$

and under the Haldane model Equation 12 gives

$$Corr(Z_j, Z_k) = I_{\{h_j\}}(h_k)\,\gamma e^{-2\beta d(j,k)}. \tag{14}$$

There are two particular population structures for which γ is of interest, due to their popularity.[8]

First, consider the F_2 *intercross* design. This design considers the population resulting from the cross of $\mathbf{M}_X$ and $\mathbf{F}_X$ with

$$\mathbf{F}_X = \mathbf{M}_X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

and so corresponds to the setting where all the differences in γ are 1. Therefore, $\gamma_{\text{inter}} = 1$ and so the correlation for an intercross population is given by $Corr(Z_j, Z_K) = I_{h_j=h_k}e^{-2\beta d(j,k)}$.

For the F_2 *backcross*, a slightly different setting is used. Here we have a cross between $\mathbf{M}_X$ and $\mathbf{F}_X$ defined as

$$\mathbf{F}_X = \begin{bmatrix} f & f \\ f & f \end{bmatrix}, \text{ and } \mathbf{M}_X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

In this setting, both differences defined on $\mathbf{F}_X$ are 0 while both of those defined on $\mathbf{M}_X$ are 1. This immediately gives $\gamma_{\text{back}} = 1$, and so the correlation for the backcross population is given by the exact same expression as that of the intercross population, $Corr(Z_j, Z_K) = I_{h_j=h_k}e^{-2\beta d(j,k)}$.[9]

Other interesting cases involve

$$\mathbf{F}_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ or } \mathbf{M}_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

as these can result in $\gamma < 0$, and so a negative correlation. For example, if we have

$$\mathbf{F}_X = \mathbf{M}_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

then $\gamma = -1$, while taking

$$\mathbf{F}_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } \mathbf{M}_X = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix},$$

---

[8]The populations of these crosses are entirely hypothetical in humans, but through successive inbreeding over many generations such homogeneity is commonly created in mouse models.

[9]As an aside, note that names *intercross* and *backcross* are references to breeding practices. The $\mathbf{F}$ and $\mathbf{M}$ of the intercross can be thought of as the result of a cross between an individual which is entirely dominant and an individual which is entirely recessive. The in the intercross, two children resulting from this dominant/recessive cross are crossed, while in the backcross, a child is crossed with a parent. Hopefully, one appreciates why such odd patterns of breeding are only pursued in mice.

gives $\gamma = -\frac{1}{\sqrt{2}}$.

Many other settings lead to no measured correlation. Take

$$\mathbf{F}_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } \mathbf{M}_X = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix},$$

or

$$\mathbf{F}_X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } \mathbf{M}_X = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix},$$

5   Simulating the model

The model outlined in Figure 1 also suggests a straightforward set of structures which can be used for simulation. Motivated by the focus on correlation in determining M_{eff}, these simulations focused on simulating the correlation of $\mathbf{z}$ for both the F_2 intercross and F_2 backcross populations.

For both populations, 100,000 individuals were simulated under three separate settings. In all settings, the simulated genomes consisted of 20 markers with differing structure. In the first setting, call it setting (a), the simplest possible case was examined: that of markers spaced evenly along a single chromosome. A distance of distance of 15 cM between adjacent markers was chosen for this setting. Setting (b) also investigated markers on a single chromosome, but these markers were no longer equidistant. Instead, markers one through six were separated by adjacent distances of 2 cM. 6 through 11 by adjacent distances of 5 cM, and similarly 11 through 16 were separated by 10 cM and 16 through 20 by 20 cM. Finally, in setting (c) a genome of two chromosomes was investigated, with the markers on the first placed 5 cM apart and those on the second 15 cM apart. The resulting correlation matrices for these settings are displayed in Figure 2 for the F_2 backcross population and Figure 3 for the F_2 intercross population.
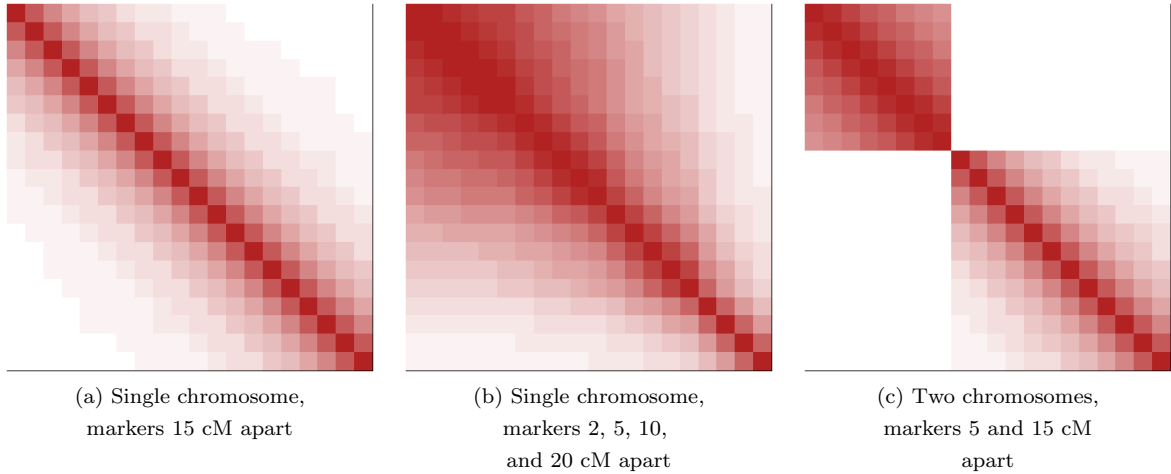

(a) Single chromosome,
markers 15 cM apart

(b) Single chromosome,
markers 2, 5, 10,
and 20 cM apart

(c) Two chromosomes,
markers 5 and 15 cM
apart

Figure 2: F_2 backcross correlation matrices for the three settings simulated.

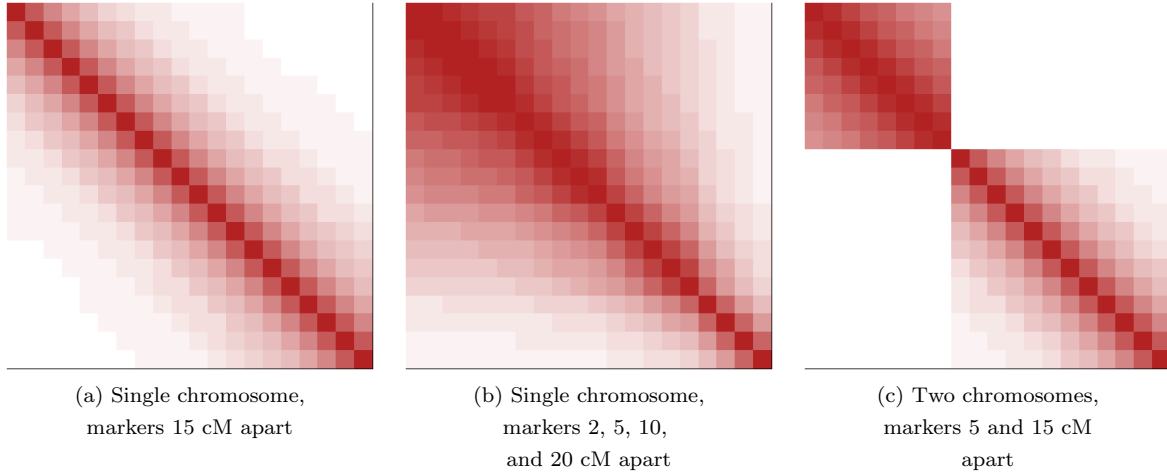|(a) Single chromosome, markers 15 cM apart|(b) Single chromosome, markers 2, 5, 10, and 20 cM apart|(c) Two chromosomes, markers 5 and 15 cM apart|

Figure 3: F_2 intercross correlation matrices for the three settings simulated.

The patterns present in Figures 2 and 3 are exactly as expected from Section 4. There is an almost perfect similarity between the backcross and intercross design, with only small local variations present between them, in particular for setting (a). All settings show a clear maximum along the main diagonal and a regular decay for off-diagonal elements. Indeed, setting (b) demonstrates how this decay is faster for more distantly spaced markers than nearer ones with its distinct fan pattern. Setting (c) additionally demonstrates the lack of correlation between different chromosomes with its striking block structure for both the backcross and the intercross.

Such simulations are a confirmation of the earlier analysis performed, but were much more time consuming than using the analytical results. This suggests that the current standard method of simulating large populations is a rather inefficient way of achieving the same result as the analytical prescription of Section 4. Conveniently, the prescription ignores the particular setting: both the backcross and intercross behave identically when viewed through the lens of correlation.

5.1 Implementation

TODO: Lose this section?

In order to make the above procedures repeatable, easily read, and flexible, the code for these simulations was implemented in R using the S3 class system. The core construct was a class meant to reflect $\mathbf{X}$.

Objects of the `genome` class are lists with two elements: `alleles` and `dists`. The `alleles` element is a list of two column matrices, where each matrix represents the value of $\mathbf{X}$ for a particular chromosome. `dists` is a list of vectors of the same length as `alleles`. Each vector in `dists` gives the distances between the alleles in the corresponding element of `alleles`. A function called `abiogenesis` allows for the convenient creation of a genome through the specification of distances and allele values.

14

The function `sex` is then used to cross any pair of `genome` objects with the use of a `meiosis` helper function. `sex` accepts an arbitrary distance function which is passed to `meiosis` to convert the `dists` elements of the `genomes` into probabilities of crossing over. By default, the Haldane map distance conversion of Equation 4 is used. Random Bernoulli trials for each of these probabilities then determines the locations of cross over events, and sections of the columns of `alleles` are swapped accordingly.

The correlation of repeated swaps is then computed via `popCorrelation`. As the choice of scoring is at the discretion of the analyst, it accepts not only a list of `genome` objects representing the result of repeated crosses, but also a scoring function which accepts a single `genome` and returns a $\mathbf{z}$ value. By default, the additive scoring of $\mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2$ is used.

The score function generates the $\mathbf{z}$ values for each `genome` provided to `popCorrelation`, and correlations between the elements of these $\mathbf{z}$ are computed. Finally, the `image` wrapper `corrImg` displays a correlation matrix rearranged so that the main diagonal is consistent with the typical arrangement of correlation matrices. The code for this implementation can be found at https://github.com/Salahub/genetic-model.

6   Comparing the model to reality

Though simulation confirms that population correlations generated under the model of sexual reproduction described in 3.1 match the predictions of Equation 14, this does not mean it reflects genetic mechanics with fidelity. Evaluating the extent to which this is the case requires data.

Luckily, Cheverud (2001) cites earlier work by Cheverud et al. (2001) in which the two pure mouse strains were used to generate an F_2 intercross population. Cheverud (2001) reproduces a correlation matrix of $\mathbf{z}$ under the additive mapping for this population, providing the opportunity to compare Equation 14 to observed correlations generated under the same setting. Figure 4 displays the results of this comparison.

Figure 4: Experimental and theoretical correlation matrices for an F_2 intercross population.

The theoretical and experimental matrices look rather similar structurally, with similar trends and local patterns. Taking the difference suggests that theoretical correlation tends to overestimate the actual

correlation between z_j and z_k. In a few cases this underestimation is rather severe. TODO: Is it severe? Think about this shading more carefully: scaled, CI, p-val, or the like
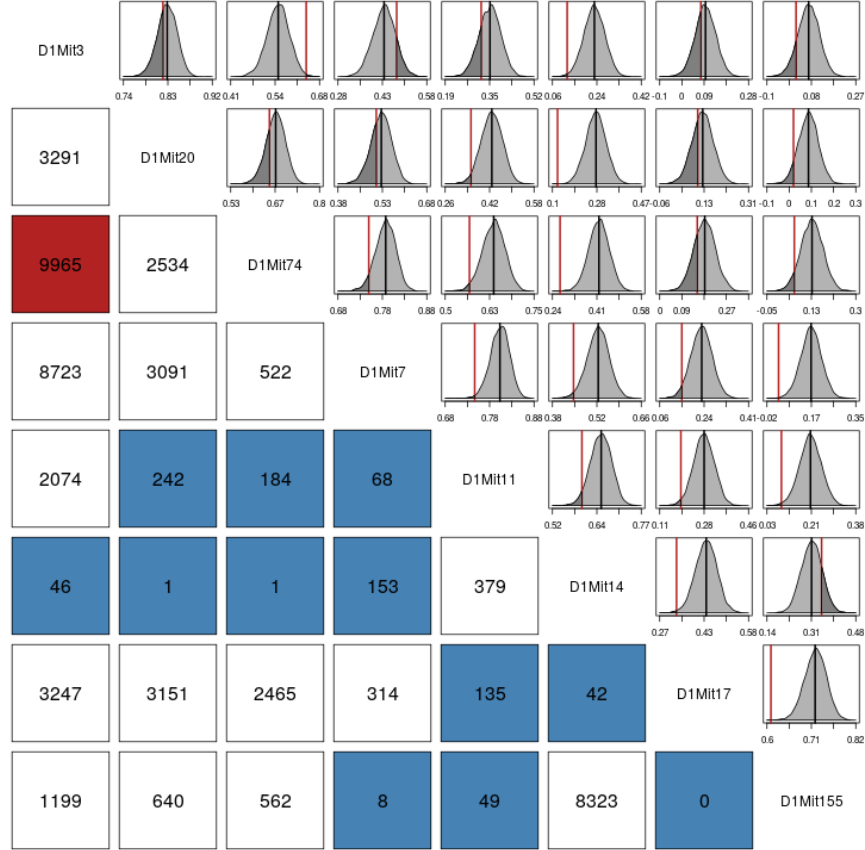

Figure 5: The correlation test plot for data from Cheverud et al. (2001) compared to simulations under theory.

That said, it performs reasonably well, especially given that Haldane (1919) proposed the distance measure at its core more than a century ago.

7   Other views

While the development of the model in Section 2 focused on heredity from a given father and mother, a radically simplified model can be considered for a population of unknown heritage. Rather than considering the values in $\mathbf{X}$ for an individual in this population in terms of the parental variants, the possible values can, instead, be considered directly when determining the pairwise relationship between two markers.
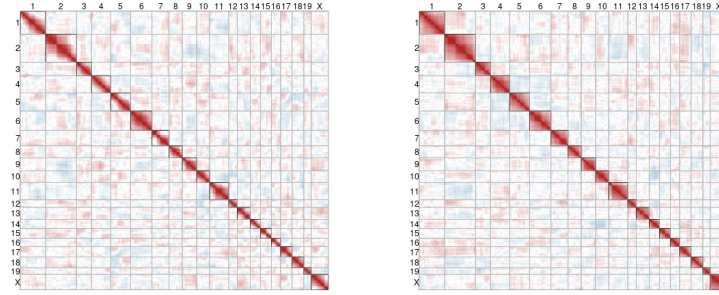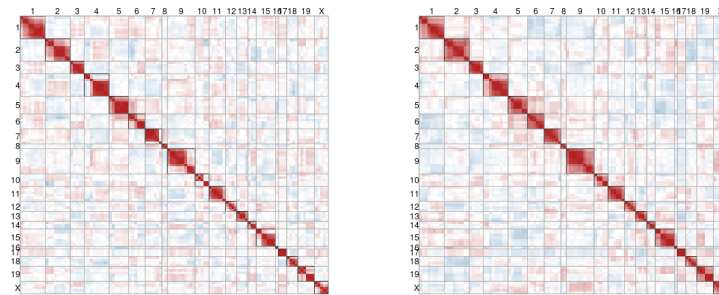
Observed   Simulated

Figure 6: Observed and simulated correlations for the JAX BSB panel from the Mouse Genome Database.

Observed   Simulated

Figure 7: Observed and simulated correlations for the UCLA BSB panel from the Mouse Genome Database.

Recall the simplified two marker versions of $\mathbf{X}$ from Section 4. Consider a column of $\mathbf{X}$, say $\mathbf{x}_1$, and note that the two entries of $\mathbf{x}_1$ only take values in $\{0, 1\}$, giving four possible combinations. For every possible combination, there is a corresponding probability of an individual in a population having such a combination on a variant. Suppose these are defined as in Table 1.

		x_{11}		Marginal
		0	1	
x_{21}	0	p_{ab}	p_{Ab}	$1 - p_B$
	1	p_{aB}	p_{AB}	p_B
	Marginal	$1 - p_A$	p_A	

Table 1: Probabilities of combinations of $\mathbf{x}_1$ in a population of individuals.

This conceptualization, which does not directly consider the parentage of a population, is dominant in the literature.

7.1 Settings in the Literature

Despite this difference, the settings introduced in Section 5 occur commonly in the literature. Among studies focusing on real data, including Galwey (2009), Nyholt (2004), and Salyakina et al. (2005), the structure of selected markers is motivated by previous work. As a consequence, these studies typically only view one chromosome, indeed one small section of a chromosome, associated with a phenotype. Additionally, the markers within this segment are typically not evenly spaced, a situation similar to setting (b) from Section 5. Typically, these distances are not reported in centimorgans, but instead in base pairs.

In contrast, simulation studies seem to be characterized by equidistant measurements on one or several chromosomes. Cheverud (2001) performs simulation on a single chromosome with equidistant markers, and records the results for several distances. This is of the same form as setting (a) from Section 5. Lander and Botstein (1989) examines a case of 12 chromosomes with markers spaced every 20 cM along each, which is a specific case of setting (c).

Departing from a reference to distances in centimorgans or base pairs, Li and Ji (2005) set their simulation scenarios using the genetic r^2 measure, defined by Hill and Robertson (1968) as

$$r^2 = \frac{(p_{AB}p_{ab} - p_{Ab}p_{aB})^2}{p_A(1 - p_A)p_B(1 - p_B)} = \frac{(p_{AB} - p_A p_B)^2}{p_A(1 - p_A)p_B(1 - p_B)}, \tag{15}$$

which is exactly Pearson's product moment correlation for the two by two contingency table case. In adopting this measure to define their simulation settings, Li and Ji (2005) use different language than other studies. Their simulation is described as an investigation of ten independent regions within which five markers are placed such that adjacent markers have an r^2 of 0.8 between them. Despite this difference in language, this design is clearly analogous to setting (c) from Section 5.

The use of r^2 by Li and Ji (2005) to specify their population parameters is somewhat curious, however. r^2 presents difficulties for measuring genetic distance over generations in any model including recombination. Consider p_{AB}, p_A, and p_B and their relationship over generations. Without selection in survival or mating, p_A and p_B will remain constant, while p_{AB} will change.

18

Suppose we have an offspring with

$$\mathbf{G} = \begin{bmatrix} 1 & g_{12} \\ 1 & g_{22} \end{bmatrix}.$$

There are two possibilities for $\mathbf{M}$ which could result in this particular $\mathbf{G}$ when the independence of variant heritability is considered:

$$\mathbf{M} = \begin{bmatrix} 1 & g_{12} \\ 1 & g_{22} \end{bmatrix}, \text{ or } \mathbf{M} = \begin{bmatrix} g_{11} & 1 \\ 1 & g_{22} \end{bmatrix}.$$

In the first of these two $\mathbf{M}$ configurations, $\mathbf{G}$ results if no recombination occurs, while in the second $\mathbf{G}$ results if recombination occurs. The first configuration occurs with probability p_{AB} and is passed on with probability $1 - p_r(d)$, as recombination disturbs the necessary variant. The second configuration occurs with probability $p_A p_B$ and is passed on with probability $p_r(d)$.

Denoting $p_{AB,k}$ as the value of p_{AB} at generation k, we then write

$$p_{AB,k} = [1 - p_r(d)]p_{AB,k-1} + p_r(d)p_A p_B,$$

from which it can easily be derived that

$$r_k^2 = [1 - p_r(d)]^2 \, r_{k-1}^2 = [1 - p_r(d)]^{2k} \, r_0^2$$

and so

$$\lim_{k \to \infty} [1 - p_r(d)]^{2k} \, r_0^2 = 0$$

whenever $p_r(d) > 0$.

In this manner, strongly associated markers A and B become less associated over time in the absence of selection pressures in sexual reproduction and survival, as noted in Siegmund and Yakir (2007). This makes the use of r^2 under any model with recombination problematic, as its use requires the specification of an initial condition and generation of interest. Li and Ji (2005) simulate data using given p_A, p_B, and r^2 values to determine p_{AB} and generate a population which matches p_A, p_B, and p_{AB}. This method therefore ignores the population characteristics of the parents and offspring, instead providing only a snapshot of the population characteristics at a particular time.

It is therefore far more natural to use a distance in these studies. Whether reported in centiMorgans, base pairs, or some other measure, these measures remain constant over generations, and govern the dynamics of recombination.

8 Conclusion

Despite this, many of the introductions to the field rely on the models of early pioneers of genetics. The works of Mendel, Pearson, Fisher, Haldane, and others in genetics were groundbreaking, but also occurred well before a modern understanding of DNA or the mechanics of inheritance Visscher and Goddard

(2019). As a result, these models do not provide a modern context. Modern textbooks and papers consequently introduce the structure of DNA and the models describing inheritance in separate sections, if the structure of DNA is addressed at all Crow and Kimura (1970); Siegmund and Yakir (2007); Xu (2013); Liu (1998). Such complete and detailed accounts with the biology and statistics separated are unquestionably important, but fail to present an accessible and unified picture of genomics for researchers with a statistical background. TODO: Move to conclusion: "this model provides a map to help understand larger works"

LaFramboise (2009) notes that microarrays remove the middle steps entirely: by taking luminance directly it is possible to inspect and relate $\mathbf{z}$ to physical characteristics without subjective intermediate steps.

This framework may not be limited to genomics. Hasin et al. (2017) note the expansion of genome-wide methods to protein and RNA sequencing. In both of these cases, the framework above applies, but has less explanatory power. One might imagine investigating proteins using this framework to see which are relatively under or over-expressed

Each of these italicized steps could be performed in a number of ways, with consequences on the final quantification of the genome's relevant features. Selection typically involves the choice of at most one million SNPs from an SNP database such as NCBI (2021). Though these databases have hundreds of millions of identified SNPs, Koboldt et al. (2013) suggests that perhaps only 15 million are common enough to be useful, and that no further useful SNPs are likely to be found. While modern microarray technology dictates the limit on the number of simultaneously selected SNPs, see LaFramboise (2009); Tam et al. (2019), studies such as Assimes et al. (2016) often choose far fewer. These rely on SNPs used in previous literature and ready-made general microarrays as selection criteria. Lander and Botstein (1989) proposes SNPs selected uniformly across the genome for agnostic studies.

There are new technologies which allow entire genomes to be sequenced, see Heather and Chain (2016); Hasin et al. (2017); Uffelmann et al. (2021), but even when these reach a cost and speed accessible to most researchers there is little doubt known markers with previous literature will still be the focus of many studies. These technologies are more likely to expand the selection pool than to make selection irrelevant.

Categorical measures of association, such as the χ^2 test, could readily be applied to $\mathbf{T}$, where each possible row combination is treated as a different category. Such measurement would likely be more computationally inefficient, but would entirely circumvent the last two steps of Figure 1.

References

Themistocles L Assimes, I-T Lee, Jyh-Ming Juang, Xiuqing Guo, Tzung-Dau Wang, Eric T Kim, Wen-Jane Lee, Devin Absher, Yen-Feng Chiu, Chih-Cheng Hsu, et al. Genetics of coronary artery disease in taiwan: a cardiometabochip study by the taichi consortium. *PLoS One*, 11(3):e0138014, 2016.

James M Cheverud. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87(1):52–58, 2001.

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.

James F Crow and Motoo Kimura. *An introduction to population genetics theory*. Harper & Row, 1970.

RW Doerge, ZB Zeng, and BS Weir. Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science*, pages 195–219, 1997.

Nicholas W Galwey. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology*, 33(7):559–568, 2009.

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.

Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18 (1):1–15, 2017.

James M Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8, 2016.

WG Hill and Alan Robertson. Linkage disequilibrium in finite populations. *Theoretical and applied genetics*, 38(6):226–231, 1968.

Daniel C Koboldt, Karyn Meltz Steinberg, David E Larson, Richard K Wilson, and Elaine R Mardis. The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38, 2013.

DD Kosambi. The estimation of map distance from recombination values. *Annals of Eugenics*, 12(1): 172–175, 1943.

Thomas LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, 37(13):4181–4193, 2009.

Eric S Lander and David Botstein. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199, 1989.

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

Ben Hui Liu. *Statistical genomics: linkage, mapping, and QTL analysis*. CRC press, 1998.

NCBI. NCBI dbSNP Build 155, 2021. URL https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi.

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.

Daria Salyakina, Shaun R Seaman, Brian L Browning, Frank Dudbridge, and Bertram Müller-Myhsok. Evaluation of Nyholts procedure for multiple testing correction. *Human heredity*, 60(1):19–25, 2005.

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.

Peter M Visscher and Michael E Goddard. From RA Fishers 1918 paper to GWAS a century later. *Genetics*, 211(4):1125–1130, 2019.

Shizhong Xu. *Principles of statistical genomics*, volume 571. Springer, 2013.