Using a structural genetic model to derive map distances and correlation

Chris Salahub
University of Waterloo

February 17, 2022

1 Introduction

In the last few decades the field of *genetics*, which investigates of the impact of individual genes, has given way to *genomics*, where the entirety of the genome is considered simultaneously. Facilitated by advances in gene microarrays, it is now routine for researchers to measure nucleotide sequences at hundreds of thousands or millions of positions on a genome LaFramboise (2009). This shift has proven trend-setting for other fields, where technological advances are ushering similar expansions in scope Hasin et al. (2017). These expansions in measurement have motivated new applications of statistics, and beg for greater participation by statisticians and mathematicians to solve the pressing problems of multiple testing they create Doerge et al. (1997); Doerge and Churchill (1996); Galwey (2009); Reshef et al. (2018).

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 (2017). Such complete and detailed accounts with the biology and statistics separated are unquestionably important, but fail to present a simple and unified picture of genomics for researchers with a statistical background.

The goal of this paper is to present a simple structural model for this purpose. Starting from the known organization of DNA, each of the steps taken to generate

A framework that provide extraordinary explanatory power.

- Elevate differences
- Emphasize the virtue of the paper: a general introduction that is accessible to any statistician using a general structural framework which makes the identification of assumptions and methodology clear

2 A structural genetic model

A natural motivation for the computation of an effective number of tests is that of genomic association studies, as these studies typically involve large numbers of tests performed on dependent variables. For the purpose of modelling and discussion, we introduce a matrix $\mathbf{G}$ to represent the entire genome of an individual. $\mathbf{G}$ contains both the maternal and paternal variants of all chromosomes placed sequentially, with the maternal variant in one column and the paternal in another. As nucleotides bind uniquely, we can consider recording the pattern only for one of the two DNA strands for each column. This gives $\mathbf{G} = [\mathbf{g}_1|\mathbf{g}_2]$ where $\mathbf{g}_1,\ \mathbf{g}_2 \in \mathcal{B}^{N_P} = \{\text{adenine, guanine, cytosine, thymine}\}^{3{,}234{,}830{,}000}$ would be the base pair sequences of an individual's paternal and maternal genome variants, respectively, if the entire genome were recorded.[1]

Of course, sequencing the entire human genome is rarely done. A genome-wide association study (GWAS) instead deals with a selected subset measuring only segments of genetic material on chromosomes of interest. Call this subset $\mathbf{S} = [\mathbf{s}_1|\mathbf{s}_2]$ where $\mathbf{s}_1,\ \mathbf{s}_2 \in \mathcal{B}^K,\ K \ll N_P$. In our model, this corresponds to a mapping from $\mathbf{G} \to \mathbf{S}$ where K rows of $\mathbf{G}$ are chosen or sampled to create $\mathbf{S}$.[2]

Note that these rows are not selected at random. GWASs typically focus on the measurement of *markers*: contiguous DNA sequences with a known location and length on the genome. Consequently, the mapping $\mathbf{G} \to \mathbf{S}$ takes the form of the selection of $M < K$ disjoint blocks of adjacent rows of $\mathbf{G}$.

These markers are commonly chosen to be *biallelic* as well, that is having two primary variants or *alleles* in the population which are traditionally denoted by a capital and lowercase letter, such as A and a. Thus we can annotate the genome, which maps $\mathbf{S} \to \mathbf{T}$ with $\mathbf{T} = [\mathbf{t}_1|\mathbf{t}_2]$ such that $\mathbf{t}_1,\ \mathbf{t}_2 \in \{A, a\}^M$. If we denote the i^{th} position of $\mathbf{t}_j$ as t_{ij}, note that $t_{lj} = A$ and $t_{mj} = A$ does not mean that they represent the same sequence in $\mathbf{S}$, but instead indicates that the respective dominant alleles are present at these positions.[3]

Next, the annotated variants in $\mathbf{T}$ might be converted to a numeric form. This numeric conversion is a mapping $\mathbf{T} \to \mathbf{X}$, with $\mathbf{X} := [\mathbf{x}_1|\mathbf{x}_2]$ where $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathbb{R}^M$. A common choice is $\mathbf{x}_j \in \{0, 1\}^M$ 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 dominant allele.

Finally, $\mathbf{X}$ might be converted into a vector $\mathbf{z} \in \mathbb{R}^M$, summarizing each individual's pair of variants into a single vector to use for linear modelling or other analysis. How exactly we perform the mapping $\mathbf{X} \to \mathbf{z}$ can differ greatly[4], but a very common choice is the *additive map* $\mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2$ with $\mathbf{x}_1$ and $\mathbf{x}_2$ given according to Equation 1. This choice means $\mathbf{z} \in \{0, 1, 2\}^M$, and z_i is equal to the count of copies of A at the i^{th} marker across both of an individual's variants.

This entire toy model is displayed in Figure 1, 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.

[1] In cells, this is not how genetic information is typically organized. Rather than occurring as one long strand, DNA exists within cells in many contiguous, but separate, sections called *chromosomes*.

[2] Choosing these rows is one of the key challenges of GWASs, as the true region of interest, or quantitative trait locus, is generally not known in advance.

[3] Dominant here is meant in the genetic sense, rather than to suggest a relative frequency.

[4] Consider the *dominance mapping* of $z_i = \max\{x_{i1}, x_{i2}\}$ or the *homozygous mapping* of $z_i = I_{x_{i1}=x_{i2}}$.

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. 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.

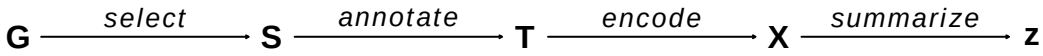


Figure 1: A diagram of the model described.

Recall the motivation of the exercise, that of identifying the number of effective tests based on the relationships between them. It should be noted, then, that the numerization and summarization steps are not strictly necessary. 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.

3 Deriving map distances

3.1 Sexual reproduction

An important consequence of the organization of the genome into chromosomes is the *independent assortment* TODO: Cite this? Where is the principle of independent assortment outlined? of chromosomes during sexual reproduction. That is, the variant of one chromosome inherited by progeny does not impact the inheritance of another. Within each chromosome, however, the process of *crossing over* presents an additional mechanism of variability.

For clarity, consider briefly the process of meiosis. Recall that there exist two variants of each chromosome within every somatic[5] cell, a paternally provided variant and a maternally provided variant. We introduce two new matrices to represent the maternal and paternal genomes of which $\mathbf{G}$ is the offspring: $\mathbf{M} = [\mathbf{m}_1|\mathbf{m}_2]$ 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 mother and father of $\mathbf{G}$ respectively. So, $\mathbf{G}$ could be $[\mathbf{m}_1|\mathbf{f}_2]$, for example. The particular method of inheritance in sexual reproduction corresponds to the construction of $\mathbf{G}$ from one random column of $\mathbf{M}$ and one random column of $\mathbf{F}$. Hence, $\mathbf{G}$ has one maternally provided variant and one paternally provided variant.

The mechanism is slightly more complex. During the production of sex cells used for sexual reproduction by meiosis, the columns of $\mathbf{M}$ and $\mathbf{F}$ may be perturbed. Rather than being inherited by $\mathbf{G}$ in the same form as in $\mathbf{M}$ and $\mathbf{F}$, the physical process of producing sex cells can cause the swapping of column entries

[5]Somatic cells can be thought of as "normal" cells. Somatic cells are the cells which constitute an organism's body, in contrast to *sex cells* or *gametes* which are used for sexual reproduction.

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 overs, and are the mechanism behind *genetic recombination*.[6]

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.[7] Consider a vector $\mathbf{h} \in \{1, \ldots, C\}^{N_P}$ which denotes the chromosome of each row of $\mathbf{M}$[8]. 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 i^{th} base pair. That is to say, suppose the chromosomes physically cross at the i^{th} base pair.[9] Each variant is consquently separated into two parts: the part up to, but not including, the i^{th} base pair, and the part from the i^{th} 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*.

As we rarely sequence the entire genome of an individual's somatic and sex cells, we will seldom see $\mathbf{M}$ and its recombinant forms. Instead, $\mathbf{S}$ is derived from $\mathbf{G}$ and $\mathbf{M}_S$ and $\mathbf{F}_S$ from $\mathbf{M}$ and $\mathbf{F}$. From this, swaps of the markers of $\mathbf{S}$, $\mathbf{M}_S$, and $\mathbf{F}_S$ between generations are used to estimate the number of sex cells containing recombinant chromosomes.[10] 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.

[6] Physically, 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.

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

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

[9] We have an additional simplfying assumption here: that the chromosome will always be aligned such that the i^{th} position on one variant will match with the i^{th} on the other during a cross over.

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

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,

$$\begin{aligned}
\lim_{r\to\infty} P(N_c = n_c) &= \lim_{r\to\infty} \frac{r(r-1)\dots(r-n_c)}{n_c!}\left(\frac{\beta d}{r}\right)^{n_c}\left(1-\frac{\beta d}{r}\right)^{r-n_c} \\
&= \lim_{r\to\infty} \frac{r^{n_c}+O(r^{n_c-1})}{n_c!}\left(\frac{\beta d}{r}\right)^{n_c}\left(1-\frac{\beta d}{r}\right)^{r-n_c} \\
&= \lim_{r\to\infty} \frac{r^{n_c}+O(r^{n_c-1})}{r^{n_c}}\left(\frac{(\beta d)^{n_c}}{n_c!}\right)\left(1-\frac{\beta d}{r}\right)^{r-n_c} \\
&= \frac{(\beta d)^{n_c}}{n_c!}\lim_{r\to\infty} \frac{r^{n_c}+O(r^{n_c-1})}{r^{n_c}}\left(1-\frac{\beta d}{r}\right)^{r-n_c} \\
&= \frac{(\beta d)^{n_c}}{n_c!}e^{-\beta d},
\end{aligned} \tag{3}$$

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,

$$\begin{aligned}
p_r(d) &= \sum_{l=0}^{\infty} \frac{(\beta d)^{2l+1}}{(2l+1)!}e^{-\beta d} \\
&= e^{-\beta d}\sum_{l=0}^{\infty} \frac{(\beta d)^{2l+1}}{(2l+1)!} \\
&= e^{-\beta d}\left(\frac{e^{\beta d}-e^{-\beta d}}{2}\right) \\
&= \frac{1}{2}\left(1-e^{-2\beta d}\right).
\end{aligned} \tag{4}$$

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.[11]

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 $c_j = c_k$, or they are not, and so $c_j \neq c_k$. 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 and it must be understood that their case is arbitrary. 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_{ij}, $i, j \in \{1, 2\}$.[12] Using this form of $\mathbf{X}$, we can consider $Cor(Z_j, Z_k)$ for the population

[11] 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.

[12] 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.

resulting from an arbitrary cross of two parents. For ease of notation, and without a loss of generality, suppose $j = 1$ and $k = 2$. Then $\mathbf{X}$ implies a $\mathbf{Z}$ of

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} X_{11} + X_{12} \\ X_{21} + X_{22} \end{bmatrix}.$$

The distribution of these X_{ij} is entirely determined by the mechanics of sexual reproduction outlined in Section 3.1 and the genotype of the parents crossed to create $\mathbf{X}$. 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

$$\mathbf{F}_X = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix}, \text{ and } \mathbf{M}_X = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix},$$

where all entries are once again in $\{0, 1\}$. Note that these entries, unlike those in $\mathbf{X}$ and $\mathbf{Z}$, will never be treated as random. It is assumed that $\mathbf{F}_X$ and $\mathbf{M}_X$ are known constants for any particular setting[13], and the statistical properties of $\mathbf{Z}$ result purely from the mechanics of reproduction.

Begin by considering the expectation of $\mathbf{Z}$. Due to independent assortment of chromosomes, X_{11} is equally likely to be either f_{11} or f_{12}, and so takes a uniform distribution over these two possibilities. A similar logic for all other X_{ij} applies, and so

$$\begin{aligned} E[\mathbf{Z}] &= \begin{bmatrix} E[X_{11}] + E[X_{12}] \\ E[X_{21}] + E[X_{22}] \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} f_{11} + f_{12} + m_{11} + m_{12} \\ f_{21} + f_{22} + m_{21} + m_{22} \end{bmatrix}, \end{aligned}$$

from which it follows

$$\begin{aligned} Var(Z_1) &= E[(X_{11} + X_{12})^2] - E[Z_1]^2 \\ &= \frac{1}{4}\left[(f_{11} + m_{11})^2 + (f_{12} + m_{11})^2 + (f_{11} + m_{12})^2 + (f_{12} + m_{12})^2\right] \\ &\quad - \frac{1}{4}(f_{11} + f_{12} + m_{11} + m_{12})^2. \end{aligned}$$

This can be simplified to give

$$Var(Z_1) = \frac{1}{4}\left[(f_{11} - f_{12})^2 + (m_{11} - m_{12})^2\right]. \tag{5}$$

Analogously,

[13] One such setting is the F_2 intercross, where $f_{11} = m_{11} = f_{21} = m_{21} = 1$ and $f_{12} = m_{12} = f_{22} = m_{22} = 0$.

$$Var(Z_2) = \frac{1}{4}\left[(f_{21} - f_{22})^2 + (m_{21} - m_{22})^2\right]. \tag{6}$$

Finally, considering the covariance:

$$\begin{aligned} Cov(Z_1, Z_2) &= Cov(X_{11} + X_{12}, X_{21} + X_{22}) \\ &= Cov(X_{11}, X_{21}) + Cov(X_{11}, X_{22}) + Cov(X_{12}, X_{21}) + Cov(X_{21}, X_{22}). \end{aligned} \tag{7}$$

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

Further simplifying this sum, these terms can be placed into two pairs. The first pair, $Cov(X_{11}, X_{22})$ and $Cov(X_{12}, X_{21})$, measure the covariance of values on the diagonals of $\mathbf{X}$. Consequently, they measure the covariance between marker encodings on different chromosomes, one inherited from $\mathbf{F}_X$ and the other from $\mathbf{M}_X$. As chromosomes assort and are are inherited independently from each parent, the diagonal X_{ij} values are independent of each other and therefore have zero covariance.[14] Explicitly, $Cov(X_{11}, X_{22}) = Cov(X_{12}, X_{21}) = 0$.

The second pair of terms, $Cov(X_{11}, X_{21})$ and $Cov(X_{12}, X_{22})$, measure the covariance of encodings on the same chromosome, and so cannot be so easily reduced. Instead, consider $Cov(X_{11}, X_{21})$ and rewrite

$$Cov(X_{11}, X_{21}) = E[X_{11}X_{21}] - E[X_{11}]E[X_{21}].$$

Recognize that $E[X_{11}] = \frac{1}{2}(f_{11} + f_{12})$ and $E[X_{21}] = \frac{1}{2}(f_{21} + f_{22})$ by the logic of independent assortment. Next, to evaluate $E[X_{11}X_{21}]$, consider the four possible values of this product.

A cross over may occur in $\mathbf{F}_X$ with probability $p_r(d)$, or it may not occur. Independently, either of the two chromosomes, recombinant or not, may be passed on with equal probability. So, the offspring with $\mathbf{X}$ can inherit either the first or the second variant, and either of these variants may be recombinant or not. Mathematically:

$$E[X_{11}X_{21}] = (1 - p_r(d))\left(\frac{1}{2}f_{11}f_{21} + \frac{1}{2}f_{12}f_{22}\right) + p_r(d)\left(\frac{1}{2}f_{11}f_{22} + \frac{1}{2}f_{12}f_{21}\right).$$

Combining this with the expectations of X_{11} and X_{21}, gives

[14] Tedious algebra confirms this logic, though it is not included here.

$$
\begin{aligned}
Cov(X_{11}, X_{21}) &= E[X_{11}X_{21}] - E[X_{11}]E[X_{21}] \\
&= (1 - p_r(d)) \left(\frac{1}{2} f_{11} f_{21} + \frac{1}{2} f_{12} f_{22} \right) + p_r(d) \left(\frac{1}{2} f_{12} f_{21} + \frac{1}{2} f_{11} f_{22} \right) \\
&\quad - \frac{1}{4} (f_{11} + f_{12})(f_{21} + f_{22}) \\
&= \frac{1}{4} \left(1 - 2p_r(d)\right) \left(f_{11} f_{21} + f_{12} f_{22} - f_{12} f_{21} - f_{11} f_{22}\right) \\
&= \frac{1}{4} \left(1 - 2p_r(d)\right) (f_{11} - f_{12})(f_{21} - f_{22}).
\end{aligned} \tag{8}
$$

The same logic can be applied to $Cov(X_{12}, X_{22})$ to obtain

$$
Cov(X_{12}, X_{22}) = \frac{1}{4} \left(1 - 2p_r(d)\right) (m_{11} - m_{12})(m_{21} - m_{22}). \tag{9}
$$

We obtain the covariance of Z_1 and Z_2 by combining the above with Equation 7. Substituting Equations 8 and 9 and the zeros corresponding to covariances between chromosomes gives

$$
Cov(Z_1, Z_2) = \frac{1}{4} (1 - 2p_r(d)) \left[(f_{11} - f_{12})(f_{21} - f_{22}) + (m_{11} - m_{12})(m_{21} - m_{22}) \right]. \tag{10}
$$

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

$$
\begin{aligned}
Corr(Z_1, Z_2) &= \frac{Cov(Z_1, Z_2)}{\sqrt{Var(Z_1) Var(Z_2)}} \\
&= \frac{\frac{1}{4}(1 - 2p_r(d)) \left[(f_{11} - f_{12})(f_{21} - f_{22}) + (m_{11} - m_{12})(m_{21} - m_{22}) \right]}{\frac{1}{4} \sqrt{\left[(f_{11} - f_{12})^2 + (m_{11} - m_{12})^2 \right] \left[(f_{21} - f_{22})^2 + (m_{21} - m_{22})^2 \right]}} \\
&= (1 - 2p_r(d)) \frac{(f_{11} - f_{12})(f_{21} - f_{22}) + (m_{11} - m_{12})(m_{21} - m_{22})}{\sqrt{\left[(f_{11} - f_{12})^2 + (m_{11} - m_{12})^2 \right] \left[(f_{21} - f_{22})^2 + (m_{21} - m_{22})^2 \right]}} \\
&:= (1 - 2p_r(d))\gamma.
\end{aligned} \tag{11}
$$

So, the correlation is seen to be a product between $(1-2p_r(d))$, which depends on the markers in question, and a factor γ, which depends on the parents being crossed. Substituting Equation 4 into Equation 11:

$$
\begin{aligned}
Corr(Z_1, Z_2) &= (1 - 2p_r(d))\gamma \\
&= \left(1 - 2\left[\frac{1}{2}\left(1 - e^{-2\beta d}\right)\right]\right)\gamma \\
&= \gamma e^{-2\beta d},
\end{aligned} \tag{12}
$$

an so finally it is clear the correlation between Z_1 and Z_2 is dictated by an exponential decay in $d(1,2)$ under the Haldane model.

However, the constraint that $f_{ij}, m_{ij} \in \{0, 1\}$ means that γ can be rewritten slightly. Recognizing that the pairwise differences in Equation 11 are equivalent to indicators of inequality between terms. That is, $f_{11} - f_{12} = I_{f_{11} \neq f_{12}}$ and similarly for the other terms. γ can be expressed as

$$
\gamma = \frac{I_{f_{11} \neq f_{12}} I_{f_{21} \neq f_{22}} + I_{m_{11} \neq m_{12}} I_{m_{21} \neq m_{22}}}{\sqrt{\left(I_{f_{11} \neq f_{12}} + I_{m_{11} \neq m_{12}}\right)\left(I_{f_{21} \neq f_{22}} + I_{m_{21} \neq m_{22}}\right)}},
$$

which, while not much easier to write, is conceptually clearer. The binary nature of each f_{ij} and m_{ij} leaves only 16 distinct crosses, and so 16 values of γ. These are enumerated in in Table 1.

Table 1: An enumeration of the possible values of the indicators and γ.

$I_{f_{11} \neq f_{12}}$	$I_{f_{21} \neq f_{22}}$	$I_{m_{11} \neq m_{12}}$	$I_{m_{21} \neq m_{22}}$	γ
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	$1/\sqrt{2}$
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	$1/\sqrt{2}$
1	1	0	0	1
1	1	0	1	$1/\sqrt{2}$
1	1	1	0	$1/\sqrt{2}$
1	1	1	1	1

Surprisingly, many of the possible crosses admit a γ value of zero, suggesting that for a slight majority of possible population designs, markers will be uncorrelated. A second critical observation is that, for

the non-zero correlation populations, the structure is dictated by $1 - 2p_r(d)$, which may be modified by a factor of $\frac{1}{\sqrt{2}}$. This implies that the only relevant process for modelling correlation in this model is the probability of cross overs between given markers.

Recalling that the primary indices 1 and 2 were a notational convenience to replace the arbitrary indices j and k on the same chromosome, this pairwise result can be immediately generalized to the correlation matrix $\mathbf{Z}$ for an entire GWAS. For markers on the same chromosome, non-trivial correlations will behave like $1 - 2p_r(d)$, where d indicates the pairwise distance between markers. Based on the independence 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 12 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 superimposed block diagonal structure, corresponding to the chromosomes, on an exponential decay. In a mathematically general form

$$Corr(Z_j, Z_k) = I_{\{h_j\}}(h_k)\,\gamma e^{-2\beta d(j,k)}. \tag{13}$$

There are two particular population structures for which γ is of interest, due to their popularity.[15]

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 indicators 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 indicators 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, $Corr(Z_j, Z_K) = I_{h_j = h_k} e^{-2\beta d(j,k)}$.[16]

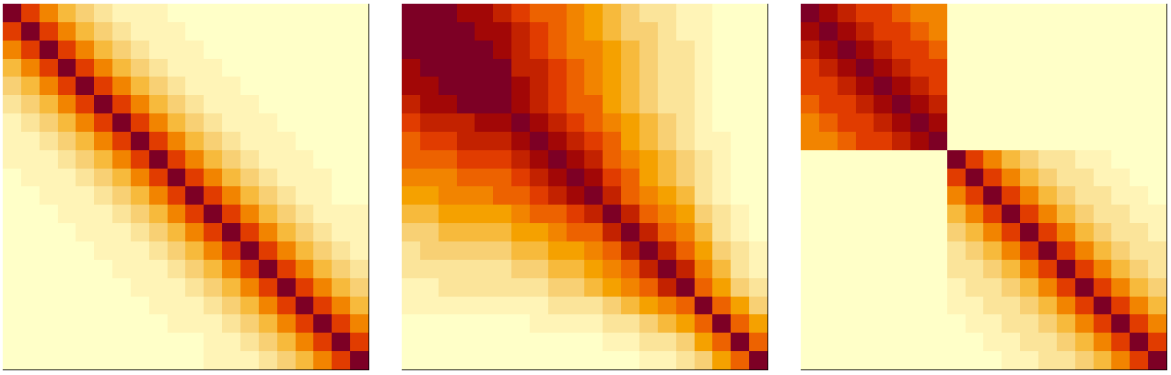
5 Simulating the model

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.

[15] 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.

[16] As an aside, note that names *intercross* and *backcross* are references to breeding practices. The **F** and **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.

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

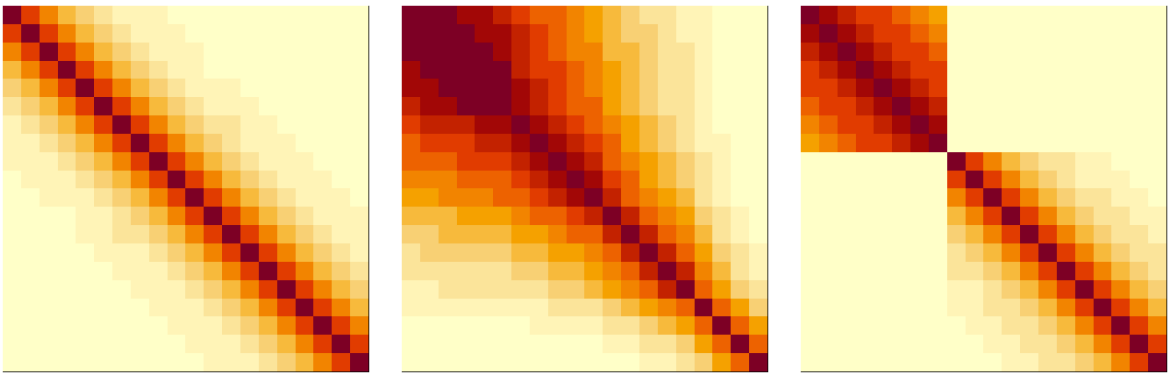
(c) Two chromosomes, markers 5 and 15 cM apart

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

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



(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.

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.

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 `genome`s 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 13, 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 13 to observed correlations generated under the same setting. Figure 4 displays the results of this comparison.

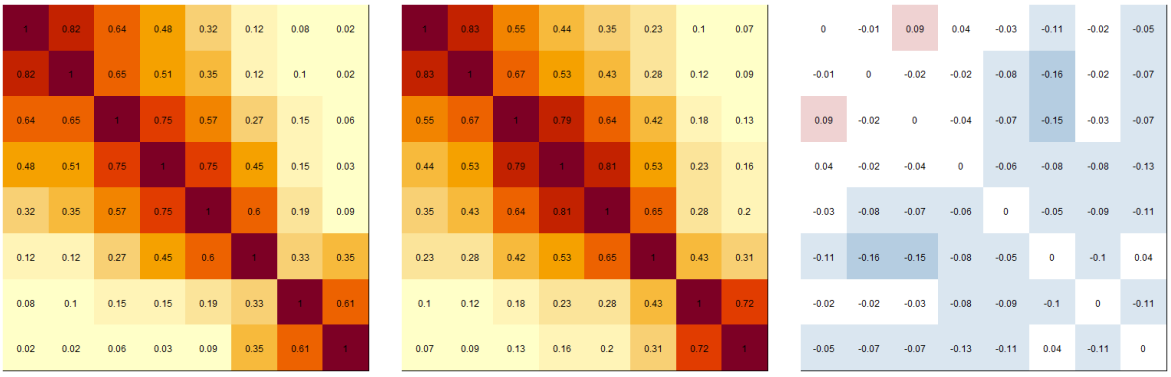


Figure 4: Experimental and theoretical correlation matrices for an F_2 intercross population and their difference.

While the theoretical and experimental matrices look rather similar structurally, with similar trends and local patterns, the difference indicates rather clearly that the theoretical correlation tends to underestimate the actual correlation between z_j and z_k. In a few cases this underestimation is rather severe.

Lingering is the question as to the cause of this discrepancy. Both model specification and random variation under the model could potentially account for this discrepancy. In order to give a greater sense of how atypical the observed difference is, we might wish to display the true difference alongside differences for data sets simulated under the model. Following the line up test described in Buja et al. (2009), Figure 5 results.

It must be noted that this is not a proper line up test following Buja et al. (2009), as the true difference was presented before this line up. Nonetheless, it gives some sense of the typical variation seen under the model. The true difference, in position nine, stands out among these, despite considerable variation in these differences across these simulated data sets. This suggests the crude model used here could be improved. That said, it performs reasonably well, especially given that Haldane (1919) proposed the distance measure at its core more than a century ago.

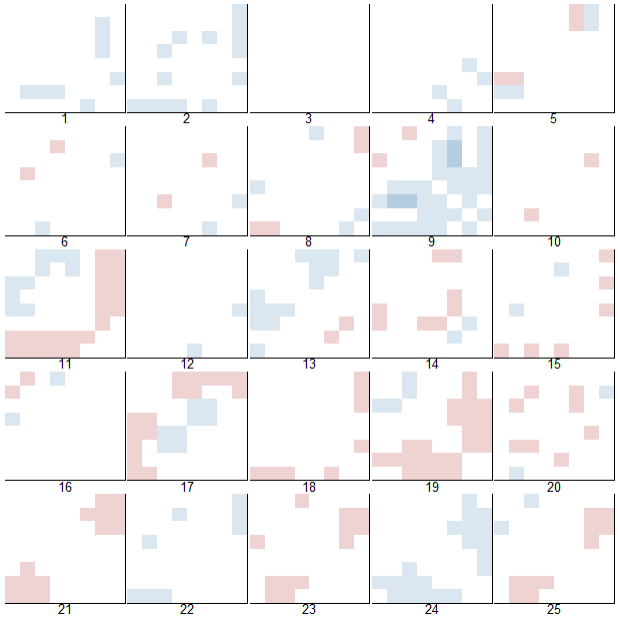


Figure 5: A line up of twenty five differences of correlation matrices, one of which is the true difference from Cheverud et al. (2001) and the others generated using Section 4.

7 Other conceptualizations

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.

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 2.

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

Table 2: 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

$$r^2 = \frac{(p_{AB}p_{ab} - p_{Ab}p_{aB})^2}{p_A(1-p_A)p_B(1-p_B)} = \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)}, \tag{14}$$

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.

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.

References

Andreas Buja, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F Swayne, and Hadley Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4361–4383, 2009.

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 and GA Churchill. Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1):285–294, 1996.

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.

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

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, 2017.

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 N Reshef, Yakir A Reshef, Pardis C Sabeti, and Michael Mitzenmacher. An empirical study of the maximal and total information coefficients and leading measures of dependence. *The Annals of Applied Statistics*, 12(1):123–155, 2018.

Daria Salyakina, Shaun R Seaman, Brian L Browning, Frank Dudbridge, and Bertram Müller-Myhsok. Evaluation of Nyholt's 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.

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

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