Submitted to Statistical Science

A structural model for genome-wide association studies

Chris Salahub,

Abstract. A structural model of the genome incorporating a modern understanding of the genome and common practice in modern genome-wide association studies is devised. This model demonstrates that the Haldane map distance is a direct consequence of the structure of the genome as it is understood today. The model also supports the derivation of an expression for correlation, which is compared to data resulting from the BSB mouse cross. The correlation test plot is introduced for this comparison and indicates a close agreement of the model and reality. Chromosome 4 displays noteworthy departures which warrant further investigation.

Key words and phrases: genetic, genomic, association, correlation, map distance, genome-wide association studies.

1. INTRODUCTION

Genetic research today routinely considers the entirety of a *genome*, that is all heritable material potentially passed to offspring as in [6], to identify regions strongly related to measured 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 must sort through this voluminous research in order to understand genomics.

To aid in this, Figure 1 draws on the literature to present a structural model of the process of converting 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. By highlighting these steps and abstractions in plain language, this model provides a clear map to guide the understanding of GWAS. While it is no replacement surveys such as [25, 24] or the literature they outline, it supplies a guiding structural framework with exceptional explanatory power to facilitate understanding of papers in the field.

Chris Salahub is a PhD Candidate in Statistics, Statistics and Acturial Science, University of Waterloo, Waterloo, Canada (e-mail: csalahub@uwaterloo.ca).

The model starts with $\mathbf{G}$, 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 is implicitly assumed.

It is usually not feasible or desirable to design a study around the measurement of all of $\mathbf{G}$, and so the *select* step chooses regions to measure. These regions are represented in $\mathbf{S}$. Often $\mathbf{S}$ 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 [19] 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 [14]. Modern SNP arrays can simultaneously identify roughly one million of these per array, see [16, 24], and most GWAS will measure only one array of SNPs. *Linkage disequilibrium*, effectively the correlation between regions of the genome, facilitates inference to regions outside of those selected in $\mathbf{S}$. While third generation genome sequencing technologies allow for entire genomes to be sequenced, as noted in [11, 10, 25], persistent high costs of next generation technologies and more than a decade of SNP array development leave arrays as the dominant measurement method.

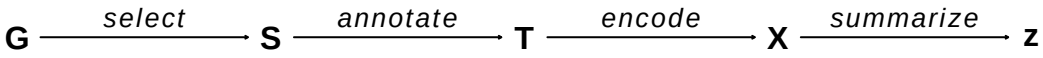


Fig 1: A structural model of genomic association studies.

After selecting SNPs to obtain $\mathbf{S}$, researchers must *annotate* the raw data. The raw signal produced by a SNP array 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 array sections, but more complex hidden Markov and Bayesian models have also been developed. [16] details some of these. Whatever method is used, the selected regions are assigned genotypes in $\mathbf{T}$. Often these will be denoted with capital or lowercase letters at each SNP, as in [23, 26].

Finally, relationships between $\mathbf{T}$ and an observed trait or within $\mathbf{T}$ 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: [17, 3, 23] 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 this structural model. By using mathematical notation for each abstraction, a framework with extraordinary explanatory power is devised. Section 2 provides an explanation of the model with all the necessary mathematical 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 is further demonstrated in Section 4, where it is used to derive the correlation between markers under classic genetic population settings. This results in an expression of the correlation between markers in any genetic study, which is simulated in Section 5 under settings from the literature. Finally, Section 6 simulates the model and compares the results to theoretical expectations and experimental 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 $N_P \approx 3,234,830,000$. $\mathbf{G}$ represents the whole genome of an individual, with all chromosomes placed sequentially in two adjacent columns corresponding to the maternally and paternally inherited variants. Thought both of these variants are complete double-stranded sequences of DNA, 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 is implied by this sequence and the unique binding of nucleotides.

Rather than address the whole genome, GWAS typically deal with a selected subset of segments 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 $K \ll N_P$. The mapping $\mathbf{G} \to \mathbf{S}$ chooses K rows of $\mathbf{G}$ to create $\mathbf{S}$. This mapping is very seldom a random one. Previous work and databases of SNPs or other known markers motivate the choice of rows. Most commonly, then, the mapping $\mathbf{G} \to \mathbf{S}$ is a non-random selection of $M < K$ disjoint sequences from $\mathbf{G}$.

In the case where $\mathbf{S}$ contains only SNPs, the markers are most often *biallelic*, i.e. the population is dominated by two different sequences or *alleles* at the marker. These can be 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, a mapping $\mathbf{S} \to \mathbf{T}$ with

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

Denoting the i^{th} 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 i^{th} 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 the genome, as in [15] and [28]. The derivation of the Haldane map is outlined here, starting with a simple sketch of sexual reproduction.

3.1 Sexual reproduction

Sexual reproduction is the recombination of the genomes of two parents to create offspring genetically distinct from both. A distinction must be made between reproductive or *sex* cells, e.g. sperm, and somatic cells. While somatic cells contain two variants of the genome, sex cells contain only one. When two sex cells combine, each provides its own variant to the offspring that results. Inheritance is mediated by the creation of sex cells, which itself involves the random selection of variants contained within somatic cells by meiosis.

To track the parental variants which may be inherited, introduce two 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 $\mathbf{m}_1, \mathbf{m}_2, \mathbf{f}_1, \mathbf{f}_2 \in \mathcal{B}^{N_P}$. 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 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 variant of one chromosome inherited by offspring is independent of other chromosomes inherited from the same parent. This means that either of the paternal and maternal variants of a chromosome is equally likely to be passed on regardless of which variant is passed on for another chromosome.

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.

3.2 Modelling cross overs

Both crossing over and the independent assortment of chromosomes occur within each parental genome independently of the other parent, and so only one of the two needs to be considered in modeling cross overs. Suppose it is $\mathbf{M}$.

We start with the assumption that genetic recombination is totally independent between chromosomes. Specifically, chromosomes not only assort independently but crossing over occurs independently on each chromosome and will affect only that chromosome's variants. This assumption can be thought of as a slightly stronger version of independent assortment. Therefore consider a vector

$$\mathbf{h} \in \{1, 2, \ldots, C\}^{N_P}$$

for $C \in \mathbb{N}$ which denotes the chromosomal membership of each row of $\mathbf{M}$. For simplicity, set $h_i \leq h_j$ for all $i \leq j$. In other words, all base pairs of a chromosome appear in adjacent rows with some specified ordering of the chromosomes. Assuming cross overs occur independently for each chromosome, a cross over in chromosome c, say, will affect only those rows of $\mathbf{M}$ where $\mathbf{h} = c$. Start with the simplest case, where $\mathbf{h}$ is a vector of ones. This is the case of a single chromosome, which can be extended to the entire genome by considering every other chromosome in the same way.

For this single chromosome, consider a cross over beginning at the i^{th} base pair. This means the two variants of the chromosome physically cross at the i^{th} base pair. Assume that the variants are always perfectly aligned so that the i^{th} position on one variant will match with the i^{th} on the other during a cross over. 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 a cross over is said to "begin at index i", it will refer to this sort of crossing: a swap of the columns for the first $i-1$ rows of $\mathbf{M}$. Introduce an indicator vector

$$\mathbf{V} = (V_1, \ldots, V_{N_P})^{\mathsf{T}}$$

where

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

and define $\boldsymbol{\pi}$ so that $\pi_i = P(V_i = 1)$. This can be done without loss of generality, as the order of cross overs in time does not affect the final chromosome. Any chromosome, offspring, or sex cell for which any cross overs have occurred is called *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, just as $\mathbf{S}$ is derived from $\mathbf{G}$, $\mathbf{M}_S$ and $\mathbf{F}_S$ are derived from $\mathbf{M}$ and $\mathbf{F}$ respectively. Swaps of the markers of $\mathbf{M}_S$ and $\mathbf{F}_S$ as inferred from $\mathbf{S}$ are then used to estimate the number of sex cells containing recombinant chromosomes. The proportion of sex cells produced with such a swap is called the *recombination rate* for the pair of markers.

However, the recombination rate for a pair of markers tells us nothing of how many cross over events occurred between them. Any odd number of events leads to a swap, while any even number will be undetectable. With this restricted view, 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 the recombination of two particular markers on the genome is all we care about, estimating individual π_i values is unnecessary. Consider two such positions, j and k with $j < k$, and note that cross overs beginning at any of $j+1, j+2, \ldots, k-1, k$ all result in these positions being split between variants. For identifiability assume that $\pi_j = \pi_{j+1} = \cdots = \pi_{k-1} = \pi_k = \pi_{j:k}$. Let N_c be a random variable counting the number of cross overs beginning in the interval $\{j+1, j+2, \ldots, k-1, k\}$. Then

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

if we assume the cross overs occur independently. For convenience, let $r = k - j$ and $\pi = \pi_{j:k}$, which gives

$$P(N_c = n_c) = \binom{r}{n_c} \pi^{n_c} (1-\pi)^{r-n_c}, \tag{3}$$

where r is a unitless count of base pairs between positions j and k.

Recall that $N_P \approx 3,234,830,000$ in humans. This large number of base pairs spread over the 23 human chromosomes means that two markers will typically be separated by a great number of base pairs, and so r will be very large. Indeed, examples in [20], [22], and [8] typically have thousands or tens of thousands of base pairs between marker locations. Therefore, consider 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, a substitution can be made:

$$\pi = \frac{\beta d(j,k)}{r} := \frac{\beta d}{r},$$

with $\beta, d(j,k) \in \mathbb{R}$. This substitution reparametrizes the probability π with 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 arbitrary. Any 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 $d(j,k) = k - j$ and use a rate per base pair. This flexibility gives a great deal of freedom to choose a convenient set of units for measurement or understanding.

The substitution also leads to a substantial simplification, as

$$\begin{aligned} &\lim_{r\to\infty} \tbinom{r}{n_c} \left(\frac{\beta d}{r}\right)^{n_c} \left(1 - \frac{\beta d}{r}\right)^{r-n_c} \\ &= \lim_{r\to\infty} \frac{r^{n_c} + O(r^{n_c-1})}{n_c!} \left(\frac{\beta d}{r}\right)^{n_c} \left(1 - \frac{\beta d}{r}\right)^{r-n_c} \\ &= \frac{(\beta d)^{n_c}}{n_c!} \lim_{r\to\infty} \frac{r^{n_c} + O(r^{n_c-1})}{r^{n_c}} \left(1 - \frac{\beta d}{r}\right)^{r-n_c} \\ &= \frac{(\beta d)^{n_c}}{n_c!} e^{-\beta d} = \lim_{r\to\infty} P(N_c = n_c), \end{aligned} \tag{4}$$

the Poisson limit approximation 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 swap in the chromosome passed on. 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 from Equation 3. Taking the simplification of Equation 4 gives

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

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

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

$$= \frac{1}{2} \left(1 - e^{-2\beta d} \right) = p_r(d). \tag{5}$$

A final substitution converts Equation 5 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* or cM. By accounting for the structure of the genome and making a number of simplifying assumptions, the model from Section 2 gives a classic result of genetics without any reference to the population-level differential equation used in its original derivation. Indeed, it indicates this population level differential equation is a direct corollary of the structure of the genome. This powerful derivation can be taken a step further to compute new theoretical results.

4. GENETIC CORRELATION

[3, 18, 8] all present results based on the *correlation between markers*. Recall $\mathbf{z}$ as depicted in Figure 1 and described in the beginning of Section 1. For these papers, the *correlation between markers* refers to the observed correlation matrix of the vector $\mathbf{z}$ in a particular population. While the motivation of these authors is adjustment for multiple dependent testing, the importance of correlation in defining linkage disequilibrium makes the correlation structure of the genome a matter of general interest. Using the model of Section 2 and results of Section 3 this matrix can 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 individual markers are not split over chromosomes, $\mathbf{c}$ 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, the assumptions of Section 3.2 dictate that there will be no correlation between Z_j and Z_k, as these markers will assort independently alongside their respective chromosomes. If they are on the same chromosome, let $d(j,k) = d$ be the distance between them measured in cM as in Equation 5. Denote the alleles of j with A and a respectively 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 $X_{rs}, r \in \{j, k\}, s \in \{1, 2\}$. Consider $Cor(Z_j, Z_k)$ for the 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

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

with rows indexed by j and k where all entries are once again in $\{0, 1\}$. Assume here that $\mathbf{F}_X$ and $\mathbf{M}_X$ are known constants.[1] Additionally, introduce the difference matrix

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

This matrix will be useful in representing the correlation between Z_j and Z_k. Finally, assume that the variation in $\mathbf{Z}$ results purely from the recombination by crossing over and independent assortment.

[1] There are theoretical populations where this is true such as the F_2 intercross, where $f_{j1} = m_{j1} = f_{k1} = m_{k1} = 1$ and $f_{j2} = m_{j2} = f_{k2} = m_{k2} = 0$.

Begin with the expectation of $\mathbf{Z}$. Assuming no preferential inheritance of either variant, 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

$$Var(Z_j) = E[(X_{j1} + X_{j2})^2] - E[Z_j]^2$$

$$= \frac{1}{4}\Big[(f_{j1} + m_{j1})^2 + (f_{j2} + m_{j1})^2 + (f_{j1} + m_{j2})^2 + (f_{j2} + m_{j2})^2 - (f_{j1} + f_{j2} + m_{j1} + m_{j2})^2\Big].$$

This can be simplified to give

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

$$= \frac{1}{4}\left[\delta_{jF}^2 + \delta_{jM}^2\right] \tag{8}$$

Analogously,

$$Var(Z_k) = \frac{1}{4}\left[\delta_{kF}^2 + \delta_{kM}^2\right]. \tag{9}$$

Considering the covariance:

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

$$= Cov(X_{j1}, X_{k1}) + Cov(X_{j1}, X_{k2}) + Cov(X_{j2}, X_{k1}) + Cov(X_{j2}, X_{k2}). \tag{10}$$

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

$Cov(X_{j1}, X_{k2})$ and $Cov(X_{j2}, X_{k1})$ can be evaluated immediately. 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 $\mathbf{F}$ and $\mathbf{M}$, respectively. These covariances therefore measure the amount of *inbreeding* in a population, that is the degree to which parents tend to have the same genotype. In settings with unknown parents or when a population is being considered [4] quantify these covariances with the coefficient r. With known parents, as in our case, these diagonal values are independent of each other and therefore uncorrelated. This can be confirmed by tedious algebra. Explicitly, $Cov(X_{j1}, X_{k2}) = Cov(X_{j2}, X_{k1}) = 0$.

$Cov(X_{j1}, X_{k1})$ and $Cov(X_{j2}, X_{k2})$ measure the covariance of encodings on the same variant, and so cannot be so easily dismissed. 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 equal probabiliy of inheritance of variants 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

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

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

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

$$= \frac{1 - 2p_r(d)}{4}\delta_{jF}\delta_{kF}. \tag{11}$$

The same logic gives

$$Cov(X_{j2}, X_{k2}) = \frac{1 - 2p_r(d)}{4}\delta_{jM}\delta_{kM}. \tag{12}$$

We obtain the covariance of Z_j and Z_k by adding the above and Equation 10. Substituting Equations 11 and 12 and $Cov(X_{j1}, X_{k2}) = Cov(X_{j2}, X_{k1}) = 0$ gives

$$Cov(Z_j, Z_k) = \frac{1 - 2p_r(d)}{4}\Big[\delta_{jF}\delta_{kF} + \delta_{jM}\delta_{kM}\Big]. \tag{13}$$

Finally, Equations 8, 9, and 13 can be combined to determine the correlation:

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

$$= \frac{(1-2p_r(d))\left[\delta_{jF}\delta_{kF} + \delta_{jM}\delta_{kM}\right]}{\sqrt{\left(\delta_{jF}^2 + \delta_{jM}^2\right)\left(\delta_{kF}^2 + \delta_{kM}^2\right)}}$$

$$:= (1-2p_r(d))\gamma = Corr(Z_j, Z_k), \tag{14}$$

where

$$\gamma = \frac{\left[\delta_{jF}\delta_{kF} + \delta_{jM}\delta_{kM}\right]}{\sqrt{\left(\delta_{jF}^2 + \delta_{jM}^2\right)\left(\delta_{kF}^2 + \delta_{kM}^2\right)}}. \tag{15}$$

So, the correlation is 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 5 in place of $p_r(d)$:

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

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

$$= \gamma e^{-2\beta d}, \tag{16}$$

and so using the Haldane map distance the correlation between Z_j and Z_k decays exponentially in $d(j,k)$ with an intercept γ determined by the parents being crossed. As the entries in $\mathbf{M}_X$ and $\mathbf{F}_X$ are all 0 or 1, the differences in $\mathbf{\Delta}$ are all -1, 0, or 1. There are therefore $3^4 = 81$ potential γ values, though most of these are not unique. 17 of these are undefined, corresponding to cases where $Var(Z_j) = 0$ or $Var(Z_k) = 0$. Table 1 summarizes the frequency of different γ values for the remaining 64 combinations. Only five values are possible and they are symmetrically distributed about 0. A number of population settings for γ are of particular interest due to their use throughout history in mouse breeding experiments. These will be outlined individually.

γ	Frequency
-1	8
$-\frac{1}{\sqrt{2}}$	16
0	16
$\frac{1}{\sqrt{2}}$	16
1	8

TABLE 1
Frequency of γ values across the 64 combinations for which correlation is defined

The first of these is the the F_2 *intercross* design. *Cross* here is short for sexual reproduction, not crossing over. This design considers the population resulting from the cross of $\mathbf{M}_X$ and $\mathbf{F}_X$ with

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

This corresponds to a setting where all the differences in γ are 1 and so $\gamma_{\text{inter}} = 1$.

The next is the F_2 *backcross*. Here we have a cross between $\mathbf{M}_X$ and $\mathbf{F}_X$ defined as

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

where $f \in \{0, 1\}$. In this setting, both differences defined on $\mathbf{F}_X$ are 0 while both of those defined on $\mathbf{M}_X$ are 1. This gives $\gamma_{\text{back}} = 1$, the same as that of the intercross population.

Other interesting cases without historical basis 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

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

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

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

for example. Note that these negative values are somewhat arbitrary. The encoding of 1 or 0 for particular alleles at a marker is not prescribed, but is rather an analytical choice. Therefore in any of these cases the encoding could be switched to give a positive γ of the same value freely.

Finally, these results can be extended to the whole genome. Recalling that j and k were restricted to be markers on the same chromosome, this pairwise result can be generalized to the correlation matrix of $\mathbf{Z}$ for markers measured on an entire genome. For markers on the same chromosome correlations will be proportional to

$1 - 2p_r(d)$, where $p_r(d)$ is 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 $c_j = c_k$, Equation 14 dictates the correlation between Z_j and Z_k. On the other hand, if $c_j \neq c_k$ 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

$$Corr(Z_j, Z_k) = I_{\{c_j\}}(c_k)\,\gamma(1 - 2p_r(d)), \tag{17}$$

and under the assumptions leading to the Haldane model Equation 16 gives

$$Corr(Z_j, Z_k) = I_{\{c_j\}}(c_k)\,\gamma e^{-2\beta d(j,k)}. \tag{18}$$

5. SIMULATING THE MODEL

The correlation results of Equation 18 are simulated by combining the model in Section 2 with the map distance derivation of Equation 5. A structure which mirrors $\mathbf{T}$ is first created. It consists of two columns of annotated biallelic markers which may be on separate chromosomes. Intra-chromosome distances for those on the same chromosome are specified together with a function to generate recombination probabilities given these distances. By default, these distances are cMs and probabilities are given by Equation 5.

A populaton can be generated from a pair of these matrices. For each individual offspring in the populuation, a few steps occur. First, cross overs are simulated using independently drawn Bernoulli random variables with probabilities given by the distances between markers. If a cross over occurs all intra-chromosomal rows of the parental matrix are swapped above the cross over index. Next the chromosomes of each offspring are selected from each parent independent of the selection of other chromosomes or the other parent. Simulating in this way creates dynamics consistent with the model in Section 2. Each individual genome generated can then be encoded and summarized before the population-wide correlation matrix is computed.

Previous literature motivates the particular simulation settings used here. [3] investigates the correlation between markers by simulating a single chromosome with equidistant markers. All combinations of chromosome lengths of 50, 75, and 100 cM with markers equidistant at 50, 25, 12.5, and 6.25 cM were simulated for populations of 500 F_2 intercross offspring. [17] instead simulates twelve chromosomes of length 100 cM with markers every 20 cM along each for a population of 250 F_2 backcross offspring.

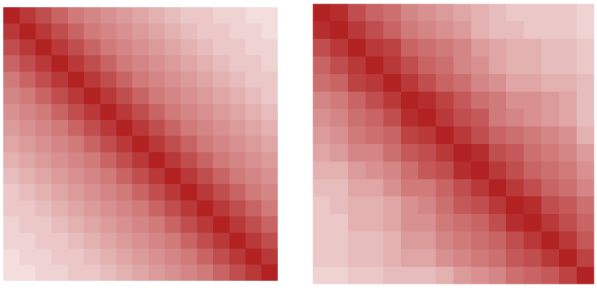
Departing from a reference to distances in cM or base pairs, [18] set their simulation scenarios using the genetic r^2 measure as defined in [12], which is exactly Pearson's product moment correlation for the two by two contingency table case. This difference is meaningful, as [23] note that r^2 is not constant over generations. After k generations it is given by

$$r_k^2 = [1 - p_r]^{2k} r_0^2$$

for two markers with $r^2 = r_0^2$ initially and a probability of recombination of p_r. Unlike cM or base pairs, which are constant over generations, r^2 eventually goes to zero.

Nonetheless, [18] simulate 10 independent regions within each of which 5 markers are placed such that adjacent markers have an r^2 of 0.8 between them. This design is analogous to that of [17], despite the difference in description.

The simulations of [3] and [17] were recreated using the implementation detailed above. Specifically, these were the 100 cM chromosome with 6.25 cM separated markers of [3] and the twelve 100 cM chromosomes with 20 cM separated markers of [17]. The resulting simulated correlation matrices and theoretical correlation matrices are visualized side by side using heatmaps in Figures 2 and 3. In each heat map, the position of a square corresponds to the position of the corresponding correlation in the correlation matrix, and darker squares have a larger magnitude than lighter squares. Blue squares indicate negative correlation while red squares indicate a positive correlation.

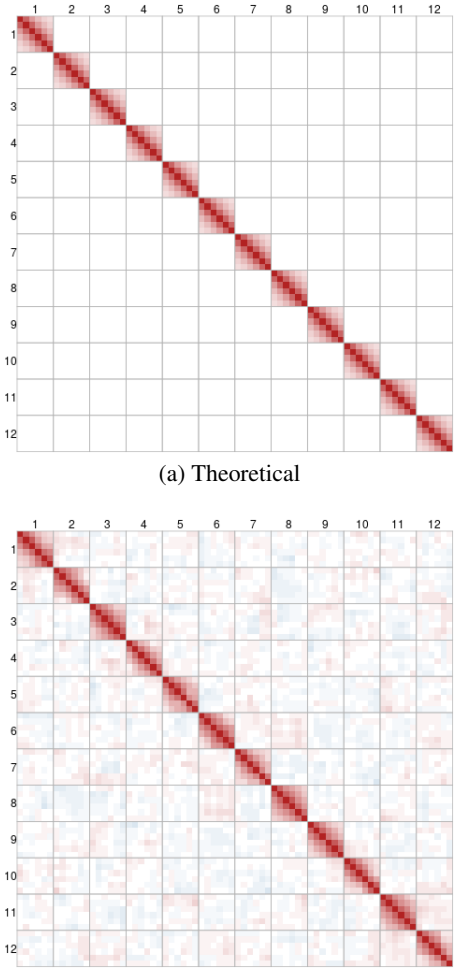


(a) Theoretical (b) Simulated

Fig 2: The correlation matrices of a population of 500 F_2 intercross offspring measured on a 100 cM chromosome with markers each 6.25 cM apart.

Figure 2(a) displays a pattern of constant off-diagonal lines of decreasing value, as expected from Equation 5. Roughly the same pattern is seen in Figure 2(b), though it is much noisier. Rather than having clear constant lines along each off-diagonal, Figure 2(b) has regions of similar values which occur across several off-diagonal lines. This leads to the appearance of large squares of more strongly related values, a pattern absent from Figure 2(a).

Figure 3 displays the [17] setting with the addition of guide lines to aid in reading the plot. These guide lines extend from labels on the top and left sides of the plot to



(a) Theoretical

(b) Simulated

Fig 3: The correlation matrices of a population of 250 F_2 backcross offspring measured on twelve 100 cM chromosomes with markers 20 cM apart on each.

indicate the chromosome of each marker. These lines create square borders along the diagonal which distinguish the intra-chromosome correlations within their borders from the inter-chromosome correlations outside of them. As suggested by Equation 5 and shown in Figure 3(a), Figure 3(b) has a stark block diagonal structure which agrees with these diagonal squares. The simulation therefore agrees very well with theory in this aspect. Within the chromosomes, there is also good agreement between Figure 3(a) and Figure 3(b). Both have decreasing correlations along the off-diagonal lines, with Figure 3(b) displaying similar departures from Figure 3(a) as Figure 2(b) does from Figure 2(a).

A more interesting noise pattern is seen between chromosomes outside the blocks in Figure 3(b). Unlike the strictly positive correlations seen in Figure 2, both negative and positive correlations are observed. Though many chromosomes show consistent patterns between their markers, with all correlations either positive or negative as between chromosomes 9 and 6 or 11 and 12, many have more complicated relationships. Between chromosomes 7 and 3, for example, both negative and positive correlations are observed between markers which are larger than the smallest intra-chromosomal correlations within 2.

6. COMPARING THE MODEL TO REALITY

Though simulation confirms that population correlations generated under the model of Section 2 match the predictions of Equation 18, the true test of any model must involve empirical measurements. This requires data other than in [17] and [3], who only perform simulations.

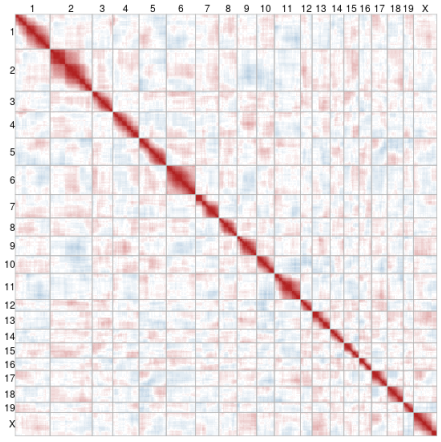
The Mouse Genome Database (MGD) of [2] provides annotation data for more than a dozen mouse populations resulting from crosses of known breeds or *strains*. The database also provides the references needed to determine cM distances between markers measured in these experiments. All of these resources are publicly provided at the Mouse Genome Informatics website: www.informatics.jax.org.

Considering only those experiments with complete observations leaves several data sets. Two of these investigate an identical population setting: the *BSB mouse cross* of [7]. BSB mice are those resulting from the F_2 backcross of the C57BL/6J and *Mus Spretus* mouse strains, detailed respectively in [13] and [5]. The first of these crosses is the *JAX BSB* cross of [21] and the second is the *UCLA BSB* cross of [27]. Both the JAX and UCLA BSB cross data were downloaded from the the JAX database.

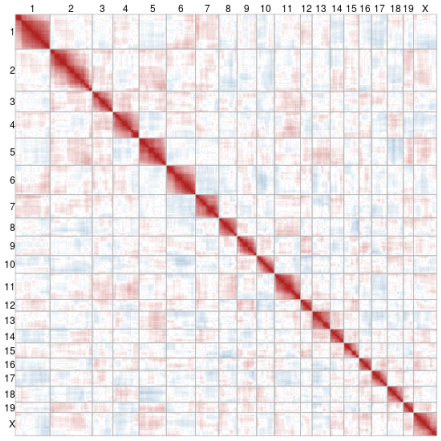
The data sets require further cleaning before being used, however. First, any markers which do not have a known position along a chromosome in cMs are removed from both data sets. Following that, any individual mice with incomplete data are excluded. For the JAX BSB data this leaves 94 mice annotated at 1496 markers while the UCLA BSB data has 66 mice annotated at 111 markers. The correlation matrices for these data sets are displayed in Figures 4(a) and 5(a) respectively.

To determine the expected distribution of these correlations, the cM positions of measured markers were used to simulate 10,000 crosses under each of the JAX BSB and UCLA BSB settings using the methods of Section 5. Figures 4(b) and 5(b) display example correlation matrices from one such simulated population. For each setting, the quantile of the each experimental pairwise correlation was then computed using the 10,000 simulated crosses. Figures 4(c) and 5(c) display those quantiles which are less than 250 in blue and those which are greater than 9,750 in red for their respective settings. These correspond to unadjusted two-sided 95% confidence rejection regions for each correlation.

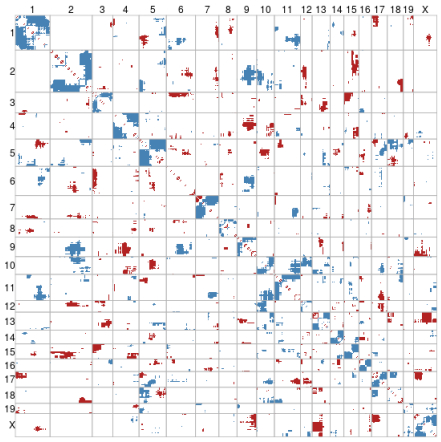
Qualitatively, the simulated examples show good agreement to experimental results. In both Figures 4 and 5 the patterns of correlation between chromosomes are similar



(a) Experiment

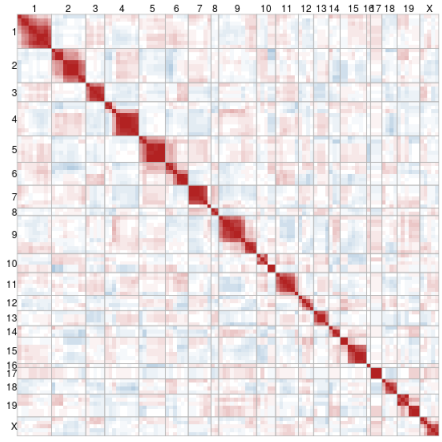


(b) Simulated example

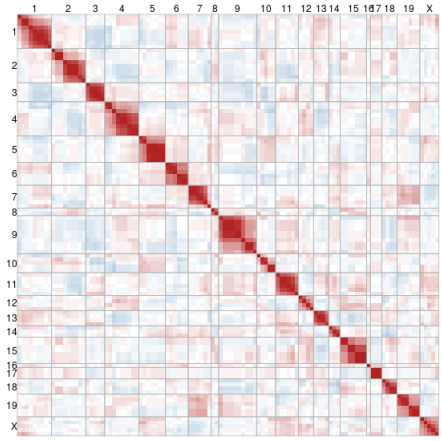


(c) Observed quantiles

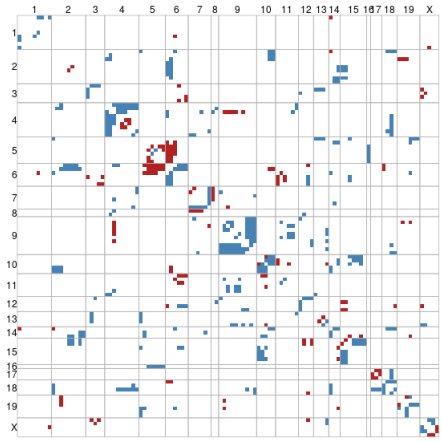
Fig 4: Observed and simulated correlations for the MGD JAX BSB cross from [21]. (c) displays quantiles determined from 10,000 simulated crosses, those quantiles less than 250 are shaded blue and those greater than 9,750 are shaded red.



(a) Experiment



(b) Simulated example



(c) Observed quantiles

Fig 5: Observed and simulated correlations for the MGD UCLA BSB cross from [27]. (c) displays quantiles determined from 10,000 simulated crosses, those quantiles less than 250 are shaded blue and those greater than 9,750 are shaded red.

between experiment and simulation. Figures 4(c) and 5(c) additionally suggest that these patterns are little more than noise. The regions of unusually strong correlations shaded in red do not appear to follow any clear pattern, nor do the patterns of unusually weak correlations shaded in blue.

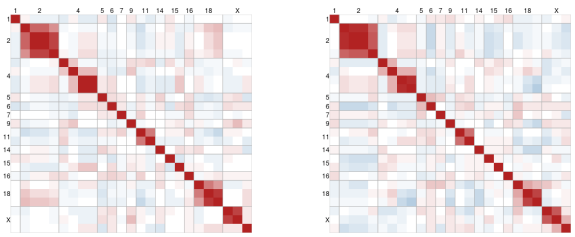
The similarity continues within chromosomes. Figures 4(c) and 5(c) are generally not shaded within chromosomes. In particular, very little of the region close to the diagonal is shaded. The most noteworthy pattern in either sub-plot occurs in the corners of the diagonal squares indicating chromosomes in Figure 4(c). Many of these corners are shaded blue, suggesting these distant intra-chromosome correlations are less than might be expected in the JAX BSB cross in many cases. The pattern of shading is suggestive of block structures within chromosomes where contiguous sections are fit well by the model but may have more complex dynamics between them.

A likely explanation is the non-independence, or *interference*, of cross overs. [1] evaluated the pattern of cross overs in the [21] experiment, the basis of the JAX BSB data. Their results suggest that cross overs are not fully independent. Most mouse chromosomes are much less than 100 cM in length, yet cross overs rarely occur within 20 cM of each other and fewer cross overs than expected tend to occur on the same chromosome. This interference will have little impact on the correlation between markers with short distances between them, as more than one cross over event is unlikely to occur in a short interval. Markers separated by longer distances are impacted by this observed interference to a greater extent, as the observed number of double cross overs will be less than expected. This increases chance that distant markers will be separated in meiosis by a single crossover, leading to a weaker correlation than predicted.

That said, it is important not to over-interpret this pattern. It is not repeated in Figure 5 and the shading of quantiles has not been adjusted to account for the many multiple tests performed in each plot. In order to get a greater sense of this experimental departure from theory, the common markers measured between the UCLA BSB and JAX BSB data were identified and the correlation matrices computed for these common locations. Having two experimental replicates should reduce the impact of one unusual measurement. These correlations are displayed in Figure 6. Most chromosomes have only one marker measured in common between these experiments, but chromosomes 2, 4, and 18 have several.

A more detailed exploration is given by further simulation. The common markers on chromosomes 2 and 4 were used to generate 10,000 simulated crosses of 80 mice, the average of the JAX and UCLA BSB cross populations. Correlation matrices were computed for each population. Additionally, 10,000 populations of each of the JAX and UCLA crosses were simulated independently, and the average of these correlation matrices computed for each pair. As the JAX and UCLA crosses were carried out far from each other years apart, these experiments are assumed to be independent.

For any pair of markers this gives two resulting distributions. One showing the distribution of correlation



(a) JAX BSB data (b) UCLA BSB data

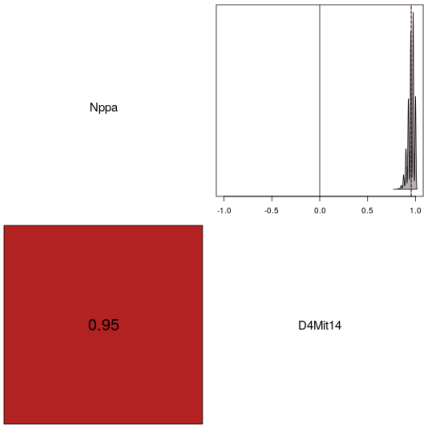
Fig 6: Pairwise correlations for the common marker positions of [21] and [27].

over repeated crosses and one showing the distribution of the correlations averaged across two populations. Both of these are compared to the JAX and UCLA data using the two-by-two layouts of Figure 7. Each subplot of Figure 7 consists of four cells. Along the diagonal, each cell provides one MGD marker symbol to display the pair of markers being compared, in this case Nppa and D4Mit14. Above the diagonal a distribution is displayed, in all cases the kernel density estimate (KDE) of the distribution of interest. Below the diagonal a numeric summary with some informative shading is displayed.

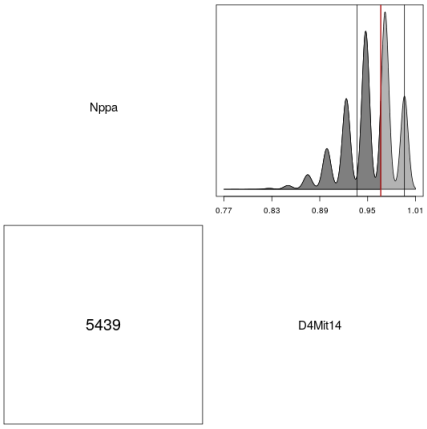
Figure 7(a) displays the KDE on a plot reflecting the range of possible correlations in the upper cell. The solid black line in this cell is drawn at the theoretical value predicted by Equation 5 and the dashed red line at the mean over all simulations. The lower cell reports the mean value of these correlations shaded with the divergent palette of Figures 2 to 6.

Figure 7(b) displays the KDE in more detail by placing it in a plot with limits set by the estimate in the upper cell. Two solid lines are added at the correlations computed from the JAX and UCLA BSB data and a thicker red line is added at their mean, below which the density is shaded. The lower cell of Figure 7(b) communicates the quantile of this red line, that is how many simulated correlations fall in the shaded region. As in Figures 5 and 4, this cell is shaded blue if the quantile is less than 250 and red if it is greater than 9,750.

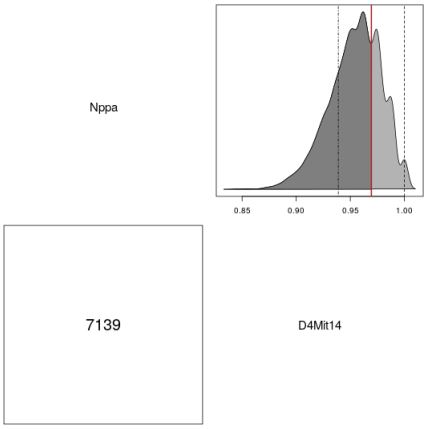
Figure 7(c) is identical to 7(b), but with a more appropriate KDE. To model the distribution of the mean of the JAX and UCLA correlations for a marker, 10,000 independent BSB crosses of both 66 and 94 mice were simulated. For each independent pair, the correlation is computed on both populations and averaged. This gives 10,000 realizations of the average correlation of the JAX and UCLA data under the model and so a more appropriate quantile value. Additional information is encoded by line type. The JAX correlation is marked with a dashed line while the UCLA data is marked with a dot-dashed line.



(a) Distribution of correlations



(b) JAX and BSB correlations against simulated populations of 80



(c) JAX and BSB correlations against paired populations of 66 and 94

Fig 7: The pairwise units used to construct the correlation distribution and correlation test plots.

Figure 7 facilitates a critical assessment of the model's fit for Nppa and D4Mit4. Figure 7(a) reiterates the close agreement of simulation and theory: the simulated mean and the theoretical value are both at 0.95. Across all simulations, the KDE shows universally strong, positive correlations above 0.5. The experimental values do not seem out of the ordinary range of these simulated correlations in Figure 7(b), nor does their mean. This suggests the model works well for this pair. Figure 7(c) reinforces this conclusion. The quantile of the mean is near the centre to the distribution of analogous means of paired, independent populations.

The subplots of Figure 7 are expanded into larger displays by adding more markers along the diagonal, treating the original subplots as single pairwise units repeated over the larger array. Expanding to include all common markers on chromosomes 2 and 4 between the JAX and UCLA BSB data gives an eight by eight matrix. Along the diagonal, all eight marker symbols are displayed. For any cell, the corresponding pair of markers is found by tracing along its row and column until the diagonal is reached. Interpretation then follows as in the two-by-two case. In an array, the pattern across all markers can be seen at a glance, especially in the lower cells, which can then guide the inspection of specific pairs in more detail. Such an expansion of Figure 7(a) gives the *correlation distribution plot* of Figure 8, while for Figure 7(c) the *correlation test plot* of Figure 9 results.

In Figure 8, the distributions of simulated correlations are generally symmetric and unimodal with centres at the theoretical correlation. Indeed, the means of simulated correlations agree with theory to two decimals for all pairwise correlations. The shape and spread of the distribution of correlations seems highly dependent on the proximity of a pair of markers. Markers which are close together in cM and have a high correlation display very little variation across the simulations relative to markers which are further apart on the same chromosome or are on different chromosomes. Additionally, the KDEs of the close markers have separate peaks and roughness indicative of clusters of correlation values. Figure 9 shows this as well, but it has been smoothed somewhat by averaging.

Of the twenty eight lower cells, six are shaded. The first of these, between markers D2Mit22 and a, is perhaps misleading. The observed quantiles are computed by counting the values less than or equal to that observed, but this pair has an observed mean correlation of 1. It is therefore necessarily larger than or equal to all other mean correlations, despite having an identical correlation as 291 simulated means. This shading should therefore be ignored, as the value is not so unusual.

All of the remaining shaded cells, with lower correlations than expected, involve chromosome 4 alone. Moreoever, the black lines showing the independent experimental observations for these pairwise correlations are much less variable than other cells. They are almost identical between experiments. It was noted earlier in the discussing of Figure 4 that the model used to compute these correlations does not account for cross over interference.

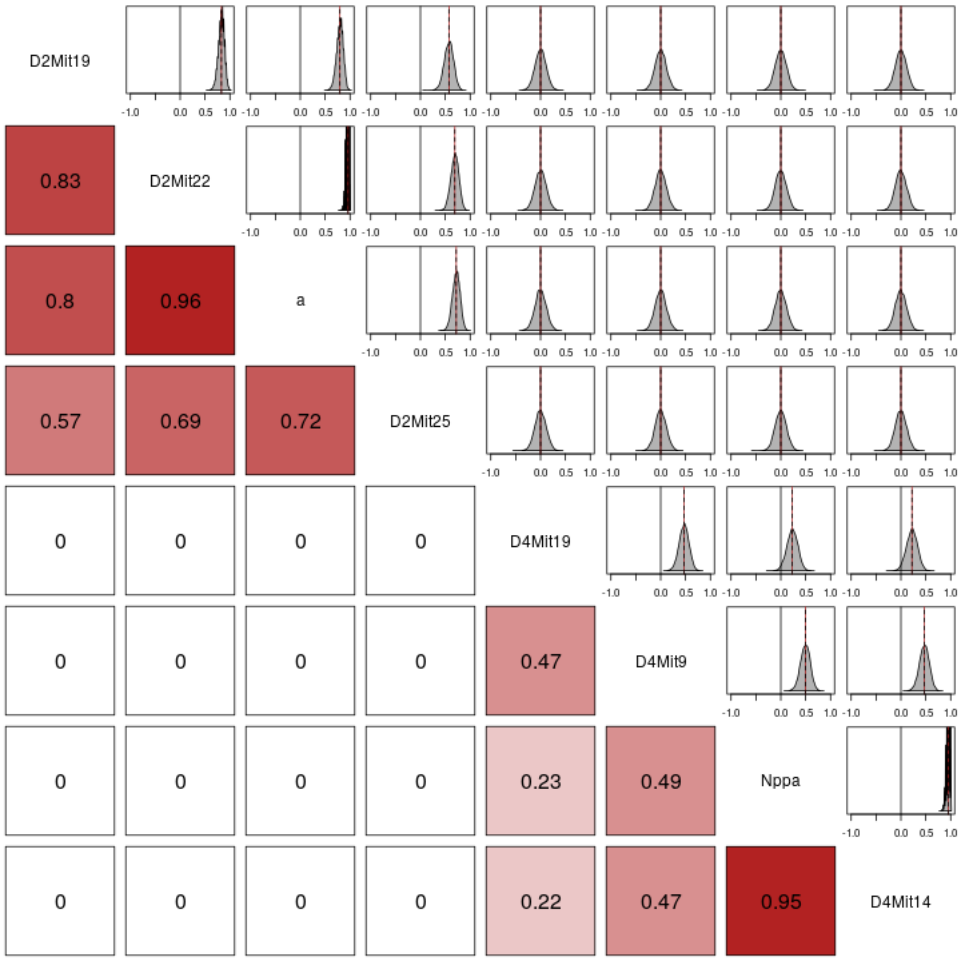


Fig 8: The *correlation distribution plot* for 10,000 simulated BSB crosses of the common markers on chromosomes 2 and 4 in the JAX and UCLA data.

The consistent departure of markers on chromosome 4 from the expectations of the model therefore suggests that chromosome 4 may experience stronger cross over intereference. This hypothesis is bolstered when chromosome 18 is added in Figure 10.

It is harder to read the cells with chromosome 18 added, but only one cell within chromosome 18 is shaded. The shaded pair, D18Mit14 and D18Mit33, does not show the close agreement over experiments as do the significant pairs of chromosome 4. Chromosome 4 is therefore noteworthy for the poorer fit of the model to its correlations and the consistency of these correlations over independent experiments.

This demonstrates the value of the correlation test plot in Figure 9 as a diagnostic. Repeated simulation under a model with no interference is used to create the distributions seen in the upper cells. Any departure from these distributions, especially a consistent departure over multiple experiments, shows where this model is inadequate. This suggests that these regions may have more cross over interference or generally more complex dynamics than others where the model fits well, and may be a priority for future measurement or research into more complicated inheritance patterns.

6.1 The impact of cross overs on correlation

Another complexity is the roughness noted in Figure 8 and repeated to a lesser extent in Figure 9. In particular the pairs D2Mit22/a and Nppa/D4Mit14 seem to cluster around the same seven values across the population in Figure 8. This suggests the correlations for highly associated markers take several discrete values in a population. This is confirmed in Figure 11 with a KDE and bar plot of correlations between Nppa and D4Mit14 placed side-by-side.

The bar plot in Figure 11(b) places lines with lengths proportional to the number of observed correlations with a given correlation at the corresponding point on the horizontal axis. Seven clear clusters are visible, and a handful of lines suggest a potential eighth. The first of these from the right is at 1, and is a single line rather than a collection of lines. The peaks of each of the following groups

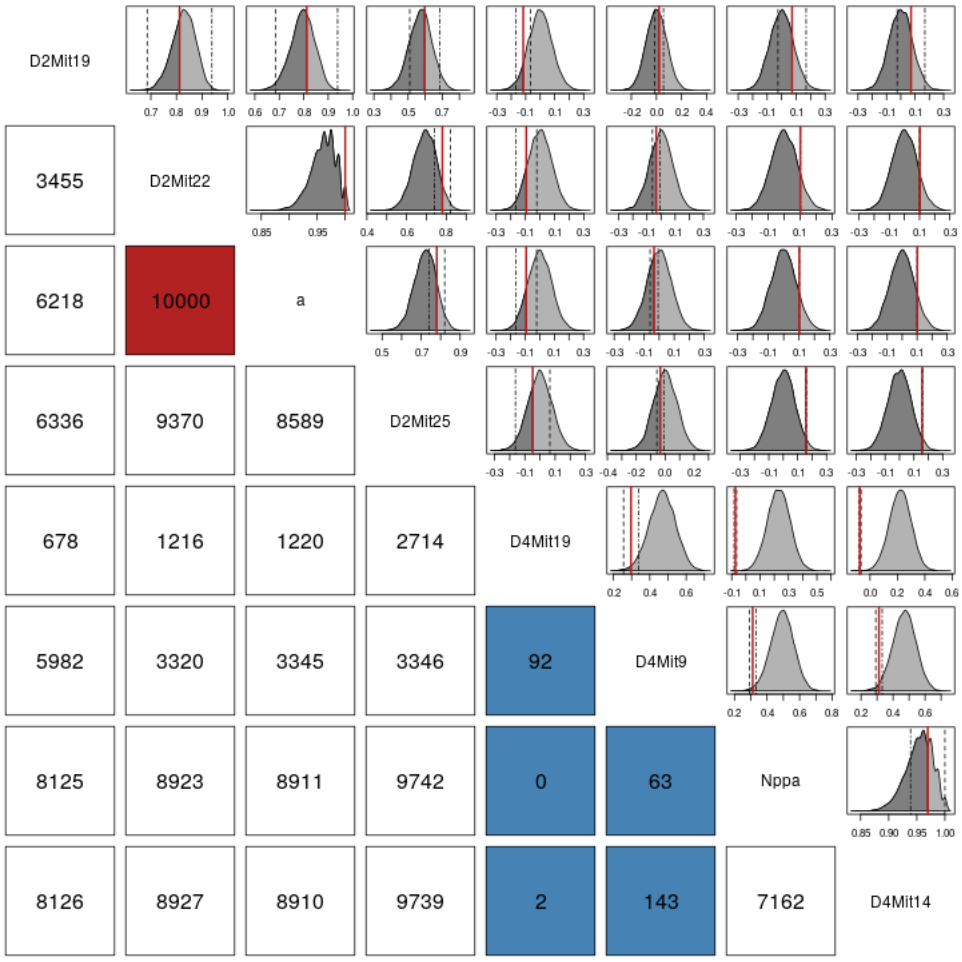


Fig 9: The *correlation test plot* for the JAX and UCLA BSB crosses. The upper cells show the distribution of 10,000 simulated averaged correlations between the JAX and UCLA BSB crosses. The experimental results are marked by black lines and their mean marked by a red line. The dashed line presents the The bottom cells give the quantile of the corresponding mean over the 10,000 simulated crosses.

of lines correspond to the peaks in Figure 11(a). This intriguing pattern can be explained by the features of the backcross being simulated and the equation for sample correlation.

Recall that these correlations result from the simulation of the BSB cross, a specific backcross. In general, the backcross of the marker summaries Z_j and Z_k under the additive map can be modelled by the sexual reproduction of the annotated matrices

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

Without loss of generality, take $f = 1$. In the absence of cross overs there are only two possible values for $\mathbf{z}$: $(1,1)^\mathsf{T}$ and $(2,2)^\mathsf{T}$. Both of these values of $\mathbf{z}$ are equally likely, as the inherited variants are independently donated by each parent. Regardless of which is inherited, $Z_{ij} = Z_{ik}$ for all of the offspring indexed by i in a population of n without cross overs. The correlation, given by

(19)

$$Corr(Z_j, Z_k) = \frac{\sum_{i=1}^n z_{ij} z_{ik} - n\bar{z}_j \bar{z}_k}{\sqrt{\left(\sum_{i=1}^n z_{ij}^2 - n\bar{z}_j^2\right)\left(\sum_{i=1}^n z_{ik}^2 - n\bar{z}_k^2\right)}}$$

where $\bar{z}_l = \frac{1}{n}\sum_{i=1}^n z_{il}$, is therefore always 1.

Suppose m cross overs occur and are inherited by the last m offspring. Then there are m individuals in the population with $\mathbf{z} = (1,2)^\mathsf{T}$ or $\mathbf{z} = (2,1)^\mathsf{T}$. For simplicity, let all cross overs lead to $\mathbf{z} = (1,2)^\mathsf{T}$. Take Z_k as the *reference marker* and assume the individuals with $\mathbf{z} = (1,2)^\mathsf{T}$ would have been $\mathbf{z} = (2,2)^\mathsf{T}$ without a cross over. Let

$$Corr_{ref} := \frac{s_{jk}^* - n\bar{z}_j^* \bar{z}_k^*}{\sqrt{\left(s_{jj}^* - n(\bar{z}_j^*)^2\right)\left(s_{kk}^* - n(\bar{z}_k^*)^2\right)}}$$

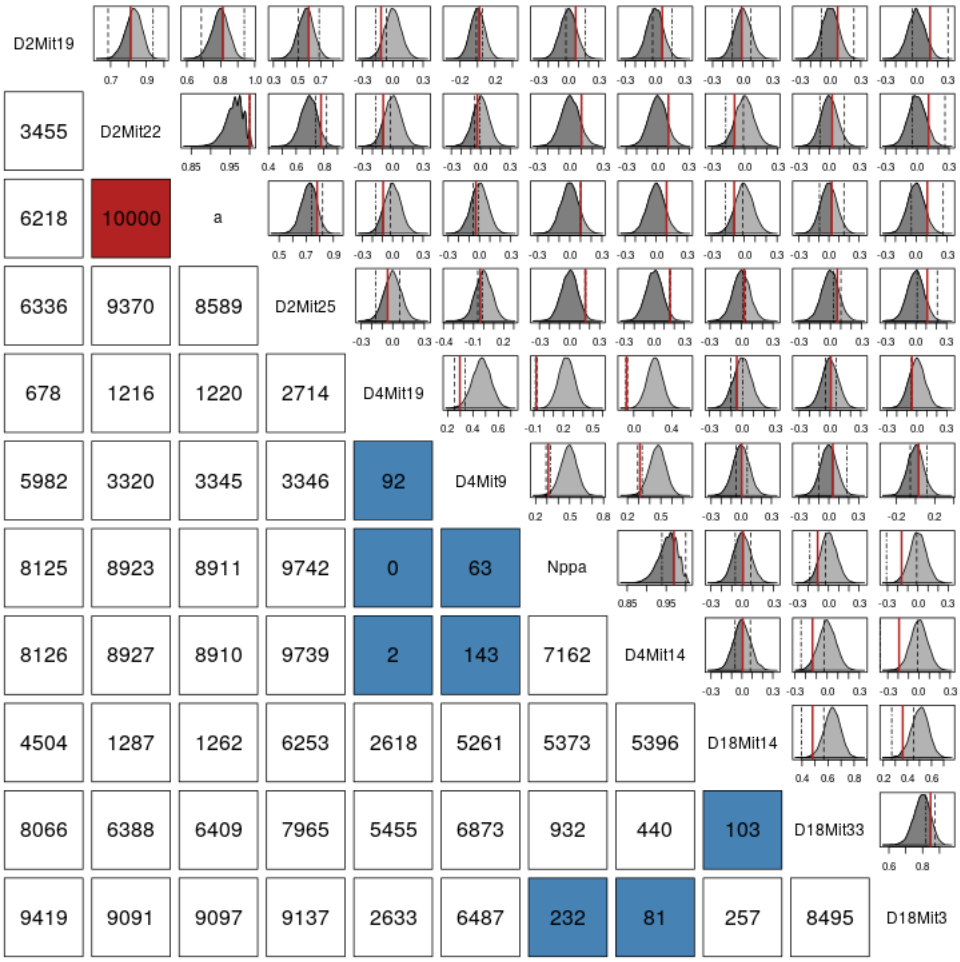


Fig 10: Adding chromosome 18's three common markers to Figure 9.

As we cannot have more cross overs under this simplification than the reference marker, we must have

$$m \leq n(\bar{z}_k^* - 1).$$

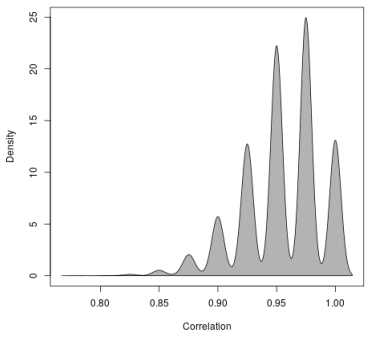
The symmetry of this problem, however, means that this does not limit the coverage of possible populations. With m cross overs as modelled here, any sum over z_{ij} in this equation will change.

The sum $\sum_{i=1}^n z_{ji}$ decreases by m since each crossed over term goes from 2 to 1. Therefore, the product $n\bar{z}_j$ will also decrease by m. Similarly, the sum $\sum_{i=1}^n z_{ij}^2$ will decrease by $3m$ as each crossed over term goes from 4 to 1. Finally, the sum $\sum_{i=1}^n z_{ji}z_{ki}$, will decrease by $2m$, as in the reference case the crossed over terms are 4, but are now 2. Together, these change Equation 6.1 to

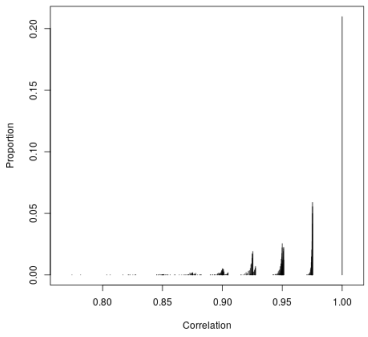
$$\frac{\left(s_{jk}^* - 2m\right) - \left(n\bar{z}_j^* - m\right)\bar{z}_k^*}{\sqrt{\left(\left[s_{jj}^* - 3m\right] - n\left[\bar{z}_j^* - \frac{m}{n}\right]^2\right)\left(s_{kk}^* - n(\bar{z}_k^*)^2\right)}}$$

$$= \frac{\left(s_{jk}^* - n\bar{z}_j^*\bar{z}_k^*\right) - m\left(2 - \bar{z}_k^*\right)}{\sqrt{\left(\left[s_{jj}^* - n(\bar{z}_j^*)^2\right] - m\left[3 - 2\bar{z}_j^* + \frac{m}{n}\right]\right)\left(s_{kk}^* - n(\bar{z}_k^*)^2\right)}}$$

$$= \frac{\left(s_{jk}^* - n\bar{z}_j^*\bar{z}_k^*\right) - m\left(2 - \bar{z}_k^*\right)}{\sqrt{\left(\left[s_{jj}^* - n(\bar{z}_j^*)^2\right] - m\left[3 - 2\bar{z}_k^* + \frac{m}{n}\right]\right)\left(s_{kk}^* - n(\bar{z}_k^*)^2\right)}}.$$

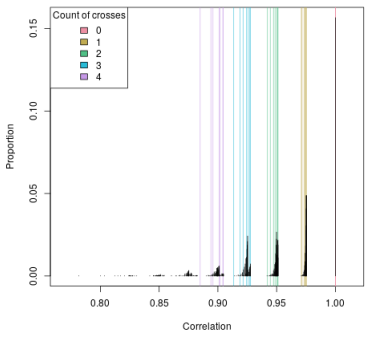


(a) KDE

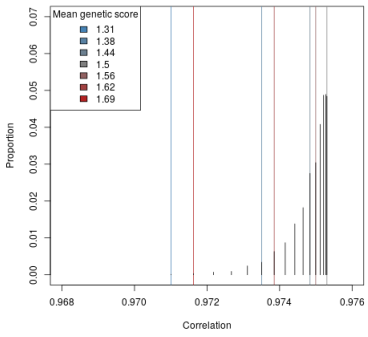


(b) Bar plot

Fig 11: The distribution of the correlation between markers Nppa and D4Mit14 over 10,000 simulated BSB crosses.



(a) Coloured by count of cross overs



(b) One cross over coloured by mean score

Fig 12: The distribution of the correlation between markers Nppa and D4Mit14 over 10,000 simulated BSB crosses.

Noting that $\bar{z}_j^* = \bar{z}_k^* := \bar{z}$ and $s_{jj}^* = s_{kk}^* = s_{jk}^* := s_{..}$ by construction, this can be further simplified to

(20)
$$Corr(Z_j, Z_k) = \frac{\left(s_{..} - n\bar{z}^2\right) - m\left(2 - \bar{z}^2\right)}{\sqrt{\left(s_{..} - n\bar{z}^2\right)^2 - m\left(3 - 2\bar{z} + \frac{m}{n}\right)\left(s_{..} - n\bar{z}^2\right)}}$$

The correlation is thereby shown to be changed primarily by the number of cross overs, but that the extent of this change is dictated by the mean additive score of the reference marker, $\bar{z}_k^*$. As $\bar{z}_k^* \in [1, 2]$, cross overs will always decrease the correlation, though the magnitude of this decrease will depend on $\bar{z}_k^*$ in a non-linear fashion.

Using the above, the correlations were computed for the cases of zero, one, two, three, and four cross overs for a selection of seven mean additive scores. Figure 12 displays these settings with coloured lines laid under the bar plot of Figure 11(b). These theoretical lines match the values observed in the simulated distribution exactly. As can be deduced from the form of Equation 20, the correlation is dominantly impacted by the number of cross overs, the mean additive score of the reference marker changes the correlation by small amounts in comparison.

The set up of Equation 20 also suggests a stochastic way to model backcross data without simulating complete crosses as done here. Let C be a random variable giving the count of cross overs in the population of size n. Then C is binomially distributed with parameters $p_r(d)$ and n. Next, let Z the be count of twos in the reference marker, which is binomially distributed with parameters $1/2$ and n for this particular backcross. The count of twos in the non-reference marker is then given by $Y = Z - C$, the difference of these two binomial counts.

7. CONCLUSION

We have presented here a structural model of genetic measurement incorporating a the current understanding of the genome and typical practice in GWAS. Using this model, the Haldane map distance was shown to be a direct corollary of the genome and mechanics of inheritance as they are currently understood. The model also supported a derivation of genetic correlation, or linkage disequilibrium, which facilitated a comparison of the model's predictions to observed data from [21] and [27]. These comparisons suggest the model fits well for most of the mark-

ers examined, and indicate chromosome 4 may have more genetic interference than others in the mouse genome.

Some novel plot matrices were created to support this investigation. The correlation distribution and correlation test plot provide an enriched version of standard correlation matrices by displaying distributional information as well as point estimates. These plots do not scale as well as the standard correlation matrices, but provide an excellent view of the pairwise relationships for a moderate number of markers. Hopefully they inspire other researchers to develop further rich custom visualizations.

All of these results demonstrate the extraordinary explanatory power and clarity of this model. By separating the steps used to create useful genetic data in Figure 1, a rich framework is created. Each step contextualizes the next, tracing a clear path from a structural representation of the genome to the numeric values used in practice. Separated in this way, different methods might be considered for each step.

Selection and annotation are currently based on SNP microarrays as outlined in [16], but as modern genome sequencing advances further as outlined in [11, 10, 25], this may change. The problem of selection will remain important even if a full sequence can be obtained. Minimizing spurious associations would still require the careful choice of regions to compare, which will certainly be annotated for convenience and clarity.

The remaining steps, which involve encoding and summarizing these annotated segments, may be skipped entirely. A plethora of categorical measures of association are outlined in literature such as [9]. Any of these might be directly applied to annotated genetic data without the need for encoding or summarization.

Finally, this work took a rather narrow view of potential encodings, summaries, and population settings. The model outlined here might immediately be applied to maps such as the dominance map on encoded markers, or to some of the known population settings outlined in Section 4 other than the backcross. Further generalization may be possible by treating $\mathbf{F}_X$ and $\mathbf{M}_X$ as random matrices with a known distribution rather than known constants, extending the results gained to a natural population rather than a specified cross.

REFERENCES

[1] Broman, K. W., Rowe, L. B., Churchill, G. A. and Paigen, K. (2002). Crossover interference in the mouse. *Genetics* **160** 1123–1131.

[2] Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., Richardson, J. E. and Database Group, t. M. G. (2019). Mouse genome database (MGD) 2019. *Nucleic acids research* **47** D801–D806.

[3] Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* **87** 52–58.

[4] Crow, J. F. and Kimura, M. (1970). *An introduction to population genetics theory*. Harper & Row.

[5] Dejager, L., Libert, C. and Montagutelli, X. (2009). Thirty years of Mus spretus: a promising future. *Trends in Genetics* **25** 234–241.

[6] Doerge, R., Zeng, Z. and Weir, B. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* 195–219.

[7] Fisler, J. S., Warden, C. H., Pace, M. J. and Lusis, A. J. (1993). BSB: a new mouse model of multigenic obesity. *Obesity research* **1** 271–280.

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

[9] Goodman, L. A. and Kruskal, W. H. (1979). *Measures of association for cross classifications*. Springer.

[10] Hasin, Y., Seldin, M. and Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology* **18** 1–15.

[11] Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* **107** 1–8.

[12] Hill, W. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and applied genetics* **38** 226–231.

[13] JAX (2022). JAX Stock #000664 Technical Report, The Jackson Laboratory.

[14] Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. and Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell* **155** 27–38.

[15] Kosambi, D. (1943). The estimation of map distance from recombination values. *Annals of Eugenics* **12** 172–175.

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

[17] Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121** 185–199.

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

[19] NCBI (2021). NCBI dbSNP Build 155.

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

[21] Rowe, L., Nadeau, J., Turner, R., Frankel, W., Letts, V., Eppig, J., Ko, M., Thurston, S. and Birkenmeier, E. (1994). Maps from two interspecific backcross DNA panels available as a community genetic mapping resource. *Mammalian Genome* **5** 253–274.

[22] Salyakina, D., Seaman, S. R., Browning, B. L., Dudbridge, F. and Müller-Myhsok, B. (2005). Evaluation of Nyholt's procedure for multiple testing correction. *Human heredity* **60** 19–25.

[23] Siegmund, D. and Yakir, B. (2007). *The statistics of gene mapping*. Springer Science & Business Media.

[24] Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20** 467–484.

[25] Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T. and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers* **1** 1–21.

[26] VISSCHER, P. M. and GODDARD, M. E. (2019). From RA Fisher's 1918 paper to GWAS a century later. *Genetics* **211** 1125–1130.

[27] WELCH, C. L., XIA, Y.-R., SHECHTER, I., FARESE, R., MEHRABIAN, M., MEHDIZADEH, S., WARDEN, C. H. and LUSIS, A. J. (1996). Genetic regulation of cholesterol homeostasis: chromosomal organization of candidate genes. *Journal of Lipid Research* **37** 1406–1421.

[28] XU, S. (2013). *Principles of statistical genomics* **571**. Springer.