

Linkage and Genetics

Daniel F. Simola

10 May 2005

1 Tutorial on population genetics

Key concepts:

1. Hardy-Weinberg equilibrium
2. Selection
3. Random genetic drift, and the balance between drift and selection (neutrality)

1.1 Hardy-Weinberg

“The Hardy-Weinberg equilibrium means that sexual reproduction does not cause a constant reduction in genetic variation in each generation; on the contrary the amount of variation remains constant generation after generation, in the absence of other disturbing forces. The equilibrium is the direct consequence of the segregation of alleles at meiosis in heterozygotes,” - Richard Lewontin.

Assuming random mating, no selection (allele frequencies constant), and no mutation, the genotype frequencies of the alleles are given by the binomial expansion of the allele frequencies: $(x + (1 - x))^2 = x^2 + 2x(1 - x) + (1 - x)^2 = 1$.

A population having genotypic frequencies X (A_1A_1), $2Y$ (A_1A_2), and Z (A_2A_2) achieves, after one generation of random mating, stable genotypic frequencies x^2 , $2x(1 - x)$, and $(1 - x)^2$, where $x = X + Y$ and $1 - x = Y + Z$. If the initial frequencies $X, 2Y, Z$ are already of the form $x^2, 2x(1 - x), (1 - x)^2$, then these are stable for all generations.

Thus if you are given genotype frequencies ($X, 2Y, Z$), you can calculate allele frequencies ($x, 1 - x$), but the reverse is not true unless you assume something like Hardy-Weinberg.

Most populations are close to Hardy-Weinberg equilibrium, because fitness doesn't change on a small time-scale.

1.2 Natural Selection

Hardy-Weinberg assumes selective neutrality; ie, both alleles have equal fitness. However this is quite an assumption. There are three typical schemes for incorporating fitness: homozygote advantage, heterozygote advantage, and homozygote disadvantage.

Using the general fitness scheme, the frequency of A_1 and A_2 after one generation is now described by the following:

$$x' = w_{11}x^2 + w_{12}2x(1 - x)$$

Table 1: Fitness schemes

A_1A_1	A_1A_2	A_2A_2
w_{11}	w_{12}	w_{22}
$1 + s$	$1 + sh$	1
$1 - s_1$	1	$1 - s_2$

$$(1 - x)' = w_{12}1/2x(1 - x) + w_{22}(1 - x)^2$$

Note that selection against a recessive disease allele results in only a slow decrease in that allele's frequency, because the genotype is not generated as often.

1.3 Mutation-selection balance

Suppose there is now mutation, where A_1 mutates to A_2 at rate u and A_2 mutates to A_1 at rate v . Then, if there is no selection, $x' = (1 - u)x + v(1 - x)$. It is assumed that u, v are $O(1/n)$. A stable equilibrium is reached when $x' = x = \frac{v}{u+v}$.

Suppose both mutation and selection occur. Intuitively mutation against a selectively favorable allele will prevent saturation of that allele in the population, since there is a constant force applied against the allele. Hence the idea of balance between mutation and selection. Using different fitness schemes, you can get different expressions describing this balance.

1.4 Random genetic drift and bottlenecks

The smaller the population size, the more pronounced become the effects of drift. This occurs because not all possible genotypes can be realized each generation, so a "sampling" of genotypes takes place. This sampling may be biased, and partly explains the prevalence of high frequency genetic diseases such as cystic fibrosis and Tay-Sachs. Notably, a population bottleneck can significantly distort the allele frequencies of a population, over just a few generations, due to the compounding of random sampling at each generation.

2 Genetic linkage in humans

There are two key concepts in linkage analysis:

1. How do you use linkage to map genetic loci using meiotic recombination?
2. What are LOD scores, what don't they tell you?

Linkage is the relatively close physical association between a set of genetic markers which lie on the same chromosome. A more general term for close association, which also applies when markers do not share the same chromosome, is an epistatic interaction. In general linkage simply indicates an association, and says nothing about the expected identity and likelihood of the markers in an association.

2.1 Recombination

An offspring's set of alleles in question is either **parental** (non-recombinant) or **nonparental** (recombinant), depending on the occurrence of meiotic recombination. However multiple recombination events are also possible. Then an odd number of cross over events yields a recombinant, or nonparental genotype in the offspring.

Frequency of recombination (θ) is a random variable used as an estimator of distance (d) between two genetic loci. If two loci are so far apart that an arbitrary number of recombination events may occur (loci on same or different chromosomes), then the $\theta = 1/2$.

$$\theta = \frac{\text{number of recombinants}}{\text{total number of offspring}} \tag{1}$$

The relation between θ and d is shown in Haldane's mapping function: $\theta = 1/2(1 - e^{-2d})$. This assumes the probability of a crossover is a constant. In other words, this assumes no interference.

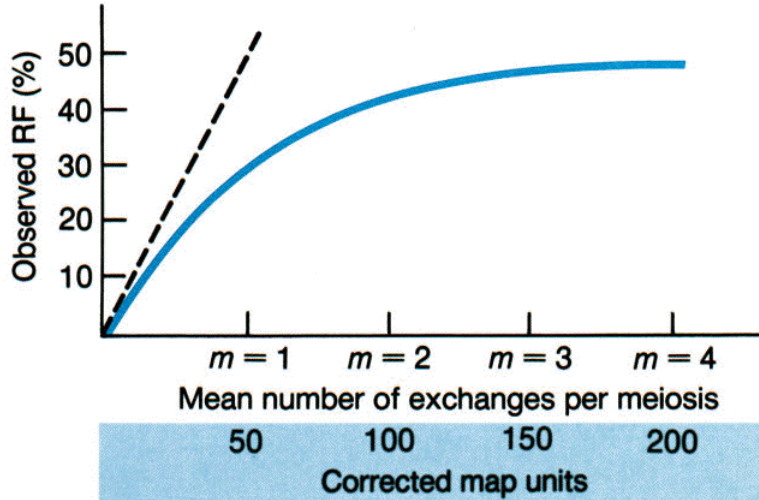


Figure 1: Haldane's mapping function. 1cM \sim 1% recombination $\sim 10^6$ base pairs in humans.

2.2 LOD scores

A LOD score is the log of the odds of observing some association between two alleles. We use the log of the likelihood ratio test statistic to test the hypothesis of linkage for two loci. We want to estimate θ , the recombination fraction between the two alleles, and the maximum likelihood function is the best estimate we can get.

In testing for linkage, you test for a significant deviation from $\theta = 1/2 =$ no linkage. In addition, the odds ratio of observed to expected θ tells you how significant is your observation.

A random variable representing the recombination between two alleles is a binomial random variable with parameter θ , the probability of recombination. Thus the likelihood of k recombinant offspring from n is the following:

$$\binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

The likelihood that the two alleles are unlinked is the following (just set $\theta = 1/2$):

$$\binom{n}{k} (1/2)^k (1/2)^{n-k}$$

The odds ratio is then

$$\frac{\binom{n}{k} \theta^k (1 - \theta)^{n-k}}{\binom{n}{k} (1/2)^k (1/2)^{n-k}} = \frac{\theta^k (1 - \theta)^{n-k}}{(1/2)^n}$$

When you observe the number of recombinants in a population, just substitute this number for k . The only remaining unknown is θ , for which you can now solve, yielding the maximum likelihood estimate for θ . (Solve by trying many values of $\theta \in [0, 1/2]$ and creating a table.) Recall above Equation 1, which is optimal for simple cases, but not as accurate when you don't know the number of recombinants exactly, or there is incomplete penetrance or variable age of onset. The LOD score is the log of this ratio.

The convention is that if the LOD score is greater than 2, linkage is suggested. If the LOD is greater than 3, linkage is established. Thus this score not only tests for deviation from 1/2 but also provides the best estimate of θ . Since this is an estimate, you typically take a 90% interval around the best estimate, by considering all values of θ where the LOD score is within 1 unit of maximum LOD. Negative LOD scores suggest two alleles are unlinked, and this is considered significant if the LOD score is less than -2.

2.3 Mapping

You can use this approach to create a genetic map of several loci, by estimating the recombination fraction for all pairs of loci, yielding relative positional information.

Genetic maps are used when there is little molecular information available for loci (becoming rare). You can map the genetic map to a physical map by saturating a chromosomal region with markers and testing for linkage. Linkage mapping is often used in combination with restriction fragment length polymorphisms and SNPs to identify disease genes.

3 Genetics of complex diseases

Key concepts:

- What is so different from simple Mendelian traits?
- The role of twins and penetrance
- Significance of genetic heterogeneity
- Linked markers and genome scans in clarifying the genetics

A disease is considered complex if it has a genetic component (disease is inherited) but does not follow the rules of Mendelian inheritance (you don't see segregation ratios of 1/4 or 1/2); 3-7% recurrence is typical. There are three features that make it difficult to see the effect of normal genes in complex diseases:

1. incomplete penetrance
2. multiple loci (epistatic effects, interactions)
3. genetic heterogeneity

. These features can be summed up in saying that in the case of complex diseases, given a genotype we can't predict the phenotype.

3.1 Familial aggregation

Existence of a genetic contribution implies familial aggregation, which is defined as a greater frequency of occurrence of a disease within a subset of a population than expected by chance, where chance is defined as the frequency of occurrence in the population at large. The problem with families is that they share both genetic and *environmental* factors. Two solutions are adoption studies (only environmental contribution similar) and twin studies (same genetics for monozygotic, sib-type for dizygotic, similar and different environments).

If monozygotic twins are significantly more similar than dizygotic twins for some complex phenotype, then the phenotype is considered to have a strong genetic component. The degree of similarity between twins is called the concordance rate. Twins are concordant if they both exhibit a phenotype, and discordant otherwise.

$$\text{concordance rate} = \frac{\text{number of concordant}}{\text{total}} \quad (2)$$

If this rate is less than 1, the trait has incomplete penetrance, since the genetics are identical for both twins, but only some exhibit the phenotype.

Assumptions:

- No sampling bias
- No "real world" bias: environments are similar, leading to similar behavior. Get around this by using MZ twins separated for an extended period of time.
- The disease is the result of the *same genes* and *same environmental influences* for all patients. ie, the disease has no "varieties", and is genetically and environmentally homogeneous.

3.2 Searching for complex diseases

Methods:

Linkage mapping techniques: requires no explicit knowledge of the disease genes. This is difficult because of genetic heterogeneity and incomplete penetrance, not to mention that the disease may be the result of many loci, thereby spreading the total likelihood of linkage across each locus.

Prior knowledge: helps identify candidate genes, which can be tested directly.

List of complex diseases:

1. Types I and II diabetes
2. Psychiatric diseases (schizophrenia, bipolar, autism)
3. Cardiovascular disease and stroke
4. Congenital malformations

4 SNPs and linkage disequilibrium

Genetics is closely related to variation . The phenotypic variations we observe are the result of combinations of thousands of genes, each of which has its own set of natural allelic and single nucleotide polymorphic (SNP) variations. One reason to study genetic variation is to help identify genes implicated in complex diseases, where each SNP acts as a marker which can be used to test for two kinds of genetic association: linkage and linkage disequilibrium. Linkage is defined above. **Linkage disequilibrium (LD)** is defined as a non-independent association (correlation) of genetic markers with arbitrary (non-Mendelian) inheritance patterns.

This definition can be formalized slightly by introducing the concept of a **haplotype**, which is a particular combination of markers/alleles at a given locus. We can identify haplotypes along a chromosome and categorize how many unique combinations exist at a locus in a sample of a population, and compute the probability of each haplotype appearing at that locus. Then another definition of linkage disequilibrium is the statistically unexpected appearance of a particular haplotype combination at a locus. The expected probability of this combination is derived from Mendelian inheritance patterns (1/4, 1/2).

LD is caused by the residual effect of a single **founder mutation**. Suppose at a locus there is an allele d and a flanking marker associated with that allele (linked). Then suppose a chance event mutates the allele into a disease allele D ; this event has repercussions on the other markers and alleles it is linked to. Over time recombination operates on the locus and the disease region might change somewhat, typically shortening in length. We want to find a marker which is preferentially associated with this disease allele; typically this means the marker is located in the alleles conserved region.

Realistically there may be several markers associated with D , and in addition there may be more than one type of disease allele. In this latter case a *simple* problem is when each mutation has its own set of associated markers, and a *complex* problem is when each mutation is associated with each of a limited set of markers; in this situation there is a very weak statistical signal to identify association. Thus it is desirable to study isolated populations which may have fewer founder mutations.

4.1 SNPs

- Expected number of SNPs in human genome: 10 million
- Expected similarity of any two genomes: 99.9%
- On average, a SNP is found every 1000 bases
- This 1% of genome accounts for **all** heritable genomic variation in humans
- Average block size: 15kb (75% of genome in blocks)

- Average number of haplotypes per block: 4

Because recombination reduces association, it was thought that markers tightly linked a long time ago would display no LD, however this turns out to be false.

SNPs are individual markers, but they typically segregate in discrete physical blocks of sequence, known as **haplotype blocks**, which are 10-100kb long. A block in one person is identical to that in another, except for the SNP sites. Thus SNPs within a block are highly correlated and show significant LD (low diversity, low recombination rate). Statistical tests can compute numerical values of correlation.

The haplotype map is a genetic map whose distance metric is LD. Cf genetic map, whose distance metric is linkage. The goal of the HapMap is to identify and locate all blocks, and provide a database for medical scans.

5 Genomic Variation

5.1 single/simple nucleotide polymorphism (SNP)

- predominant molecular form of variation
- derived from and generalization of mutation
- transitions, transversions, indels, dinucleotide changes
 - transition - p-p or py-py (A->G or C->T, vice versa)
 - transversion - p-py or vice versa (A/G -> C/T, vice versa)
- may be (non)coding, (non)replacement (changes amino acid), (non)synonymous (changes codon)
- detected as candidate SNPs in pool of < 10 samples
- may affect gene function via transscriptonal/translational regulation, splicing, or RNA stability
- wild type locus = set of haplotypes
 - haplotype is distinct combination of single nucleotide types at a single locus

5.2 SNP and Population Genetics

- distributions of variation = population genetics
 - major goal of genome science
- relationship between SNPs and phenotype = quantitative genetics
- SNPs great for population genetics
 - offers highest resolution of studying variation
 - low homoplasy
 - can study noncoding sequences
- neutral theory - balance among mutation, drift, selection, migration
 - use to test for significant patterns of sequence diversity
- nucleotide diversity - average fraction of differing nucleotides between a pair of alleles
 - human - 1 SNP/1000 bases (low)
 - fly, maize - 1/100 (high)

5.2.1 identification & verification

- compare sequences
 - $\Pr(\text{detecting rare SNP of 2 alleles}) = 1 - (1-p)^{2N}$, $N = \text{---samples---}$
- assess EST quality
- verify SNP by resequencing fragment or sequencing on affy chip (by hybridization)
- variety of up-and-coming genotyping methods available

5.2.2 linkage and mapping

- linkage *vs* association
 - association: general relationship between marker and phenotype
 - linkage: relationship based solely on physical location
 - can have assoc. without linkage (FP) and linkage without assoc. (FN)
- disequilibrium - dependency (nonrandom association) among alleles
 - hardy-weinberg - among alleles at a single locus
 - linkage disequilibrium (LD) - among alleles at multiple loci
 - contingent on history - mutation/recombination, drift, selection, migration
- mapping
 - recombination mapping (linkage mapping) - basis for positional cloning
 1. localization of target to chromosomal region
 2. candidate sequencing and SNP marker identification
 3. use recombination frequency of marker to disease segregation
 - Quantitative Trait Loci (QTL) mapping
 - * many genes linked to disease/phenotype
 - * basically same as recombination mapping but test multiple loci simultaneously for statistical significance (LOD score)
 - * extends to LD mapping
 - nonparametric tests typically observed *vs* expected (chi-squared)
 - parametric tests involve maximum likelihood scoring
 - troublesome factors for disease mapping
 - * heritability, number of genes involved, penetrance & expressivity, genetic heterogeneity, population differences, admixture, environment