Notes on epidemiological genetics

April 2022

Foreword

These notes are based on lectures given by David Clayton (Florence, 2005). Back in 2009, I spent two full years working on statistical genetics and genome-wide association studies (GWAS). I read a lot of material on statistical and genetic epidemiology, Bioconductor, and related stuff. Among the statistical literature, there were a few books on genetic epidemiology and biostatistics, the latter focusing on the analysis of bioarrays essentially. The Lancet hasn't published its series on statistical analysis of GWAS yet. Finally, the most interesting material I found were handouts written by David Clayton, who I met at a conference some years later. His other publications were always inspiring, but I found his lecture notes on epidemiological genetics enlightning in many respects. Since they vanished from the interweb, I made a quick one-shot handout for my own memory.

Other interesting materials in population genetics and evolutionary models are given below:

- Notes on Population Genetics (GitHub)
- Lecture Notes on Computational and Mathematical Population Genetics
- Applied Population Genetics

The basis of genetics

Overview

Genetics is the study of traits or factors in plants, animals or humans that are heritable. It has long been believed that inheritance was a blending of parental characteristics, and Mendel developed a theory in which this mechanism involves random transmission of ”discrete units of information” called genes. This theory assumes two essential hypotheses: (1) when a parent passes one of two copies of a gene to its offspring, these are transmitted with probability 1/2, and (2) different genes are inherited independently of one another. The later point was erroneous.

A genes corresponds to a sequence of DNA, which is composed of strings of bases, and we distinguish four nucleotide bases denoted A, T, G and C. The two strands of DNA in the double helix structure are complementary (we talk about the sense and anti-sense strands): A binds with T, and G binds with C. Additionally, 3-base sequences, also known as codons, code for amino acids and sequences of amino acids form proteins. Even if a gene codes for a protein, it also has sections concerned with the expression and regulation of genes, and RNA processing.

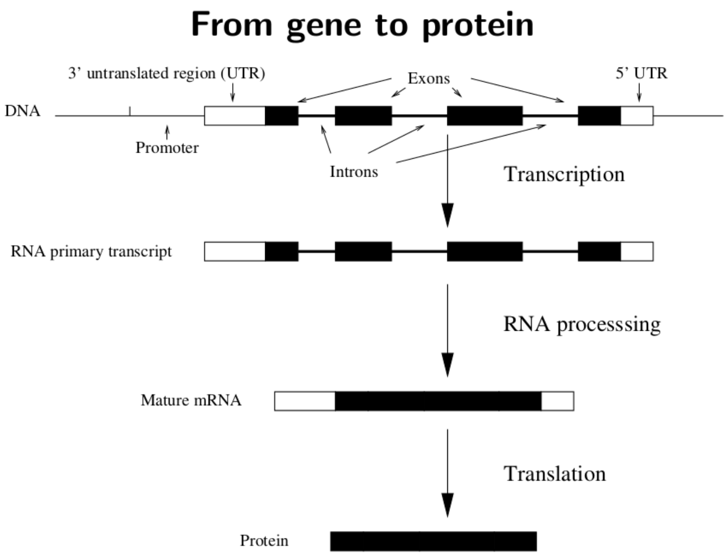


Figure 1: From gene to protein

Mutations and polymorphisms occur in genomes. The process of mutation describes the way new variants of a gene arise, while as a noun we use mutation to describe a rare variant of gene. Polymorphisms are more common variants, since most mutations will disappear but some will achieve higher frequencies due either to random genetic drift or to selective pressure. The most common forms of variants are:

- repeated sequences of 2, 3 or 4 nucleotides (microsatellites)
- single nucleotide polymorphisms (SNPs) in which one ″letter″ of the code is altered
- exonic SNPs may or may not cause an amino acid change

The human genome consists of about 3×10^9 base pairs and contains about $25,000$ genes. Much of the DNA is either in introns or in intragenic regions. Cells containing two copies of each chromosome are called diploid (most human cells). Cells that contain a single copy are called haploid. Humans have 23 pairs of chromosomes – 22 autosomal pairs and one pair of sex chromosomes. Females have two copies of the X chromosomes while males have one X and one Y chromosome.

All chromosomes have a stretch of repetitive DNA called the centromere. This plays an important role in chromosomal duplication before cell division. If the centromere is located at the extreme end of the chromosome, that chromosome is called acrocentric. If the centromere is in the middle of the chromosome, it is termed metacentric. The ends of the chromosomes that are not centrometric are called telomeres.

Are males at a disadvantage due to less gene product (for the hundreds of genes on the X chromosome)? In fact, females only have one active copy, the other being switched off during early embryonic development. This phenomenon of X-inactivation appears to be random: with very few exceptions, it cannot be predicted whether the paternal or maternal copies are inactivated in a given cell.

According to Mendelian transmission, one copy of each gene is inherited from the mother and one from the father – the two copies need not be identical. Mendel postulated that mother and father each pass one of their two copies of each gene independently and at random. Thus if, at a given locus, the father carries alleles a and b and the mother carries c and d, the offspring may be a/c, a/d, b/c or b/d – each with probability $1/4$. However, transmission of genes at two different positions, or loci, on the same chromosome may not be independent. If not, they are said to be linked.

A collection of linked loci (i.e., loci that tend to be inherited together) is called a haplotype. Immediately before the cell division that leads to gametes, parts of the homologous chromosomes may be exchanged. An individual with haplotypes $A - B$ and $a - b$ may produce gametes $A - B$ and $a - b$, or $A - b$ and $a - B$. This process is called recombination. The probability of recombination during meiosis is termed the recombination fraction, and is usually denoted by θ.

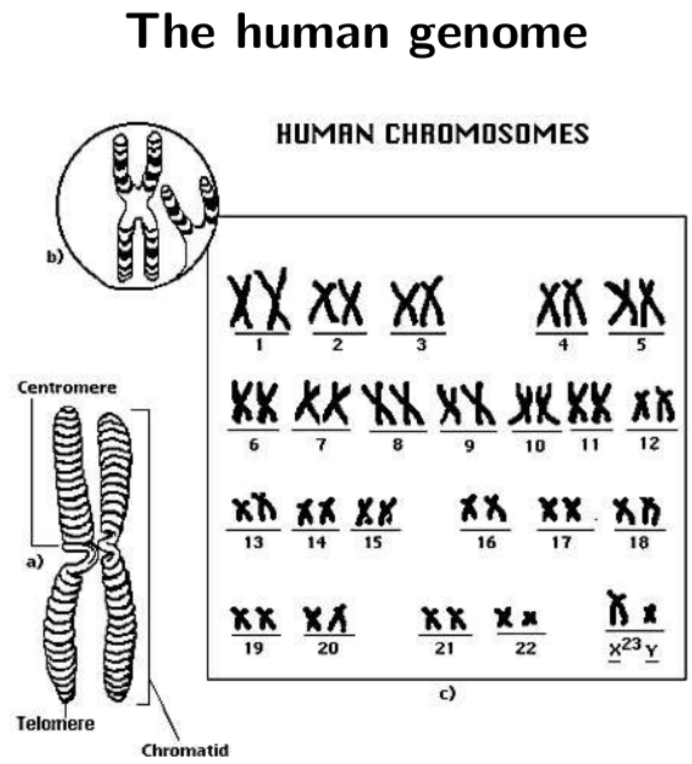
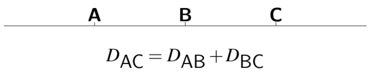


Figure 2: The human genome

What has been described historically, and above, as recombination should, more properly, be called cross-over. Although cross-over is indeeed caused by breaking and rejoining of chromosomes, they more often rejoin nearly the same way around. Often a short segment of DNA ($<$ 50 base pairs) is exchanged. This is known as gene conversion.

The greater the physical distance between the two loci, the more likely it is that there will be recombination – it is this which allows mapping of genes. A simplified model is that loci can be arranged along a line in such a way that at each meiosis, recombinations occur at a constant rate. Then genetic distances are as shown in the margin Figure.



In the simplest model, the relationship between recombination frequency and genetic distance is given by Haldane's map function:

$$D_{AB} = -\frac{1}{2}\log_e(1 - 2\theta_{AB}).$$

The unit of genetic distance is called a Morgan. At each meiosis the expected number of recombinations is, by definition, one per Morgan. On average, 1 cM corresponds to about 10^6 bases. The total length of the human genome is 33 Morgans ($\approx 3 \times 10^9$ bases). In practice, things are more complicated:

- ”interference”: the model of independence of recombinations does not fit – it predicts too many recombinations close together
- ”hot spots”: uneven relationship between physical and genetic distances
- sex differences: recombination more frequent in females

A genetic locus is polymorphic if it can exit in different forms (alleles). Genetic variation arises in a number of ways: insertions, deletions, single nucleotide polymorphisms (SNPs), tandem repeat sequences, copy number polymorphisms. Polymorphisms are created by random mutations within the DNA sequence. Many polymorphisms have no functional consequence, but can be used to build framework maps of the chromosomes. Some tandem repeat markers may have 20 or more distinct alleles, but SNP’s are (almost always) diallelic.

Because human cells are diploid, there are two alleles at each genetic locus. This pair of alleles is called the individual’s genotype at that locus. This pair of alleles is called the individual’s genotype at that locus. If the two alleles are the same, the individual is said to be homozygous at the locus. If they are different, he/she is said to be heterozygous. The heterozygosity of a marker is defined as the probability that two alleles chose at random are different. If π_i is the (relative) frequency of the i-th allele,

$$\text{Heterozygosity} = 1 - \sum_i \pi_i^2.$$

If alleles i and j have relative frequencies π_i and π_j, then, under random mating, the genotype frequencies are

$$\begin{aligned}\Pr(i/j) &= 2\pi_i\pi_j \quad (i \neq j)\\ \Pr(i/i) &= \pi_i^2.\end{aligned}$$

This is termed Hardy-Weinberg equilibrium (HWE). Even if genotype frequencies are not initially in HWE, they will return to HWE in a single generation of random mating. Deviation from HWE indicates population stratification and/or admixture – or (more likely) genotyping errors.

The phenotype is the characteristic (e.g. eye colour) that results from having a specific genotype. Often we require probability models to describe phenotypic expression of genotypes. Probabilities of phenotype conditional upon genotype are called penetrances. In many cases, the same phenotype can result from a variety of different genotypes (sometimes termes phenocopies). Equally, the same gene may have several different phenotypic manifestations (plieotrophy).

If a single copy of an allele results in the same phenotype as two copies irrespective of the second allele, the allele is said to be dominant over the second allele. Likewise, an allele which must occur in both copies of the gene to yield the phenotype is termed recessive. Alleles which correspond to mutations which destroy the coding of a protein tend to be recessive. If the phenotype for genotype i/j is intermediate between the phenotypes for i/j and j/j, the alleles i and j are codominant.

When a phenotype is controlled by two genes tere may be epistasis. This was originally defined to mean that the genotype at one locus masks the phenotypic expression of the other. The term is often used quite loosely for a complex "interaction" between two genes or, as we shall see, rather precisely in a mathematical sense.

Example: Dominance, epistasis, and blood groups: The ABO locus has three alleles. The A and B alleles are codominant, while O is recessive: A must be present for A enzyme to be produced; B must be present for B enzyme to be produced; if neither is present (O/O genotype, neither enzyme is produced. A further locus controls production of a precursor antigen, H. The A and B enzymes act on this to produce the A and B antigens: Subjects who are h/h produce no H antigen – this is the "Bombay phenotype" – such subjects are indistinguishabel from O/O subjects. So, two parents who are phenotypically type O and type B respectively can produce a type AB offspring.

Introduction to mathematical population genetics

The geneic architecture of today's populations has been shaped by a history of random mutation and recombination. The stochastic history of mutations shapes the frequency spectrum of variants – this is a source of much current controversy in genetic epidemiology. The stochastic history of recombinations shapes the patterns of linkage disequilibrium (LD) – crucially important in the design of genetic association studies.

Let p_i denote the relative frequency of a given variant in a population of N_t chromosomes in generation t. The chromosome population in generation $(t+1)$ is generated by drawing a sample of N_{t+1} chromosomes with replacement from the N_t chromosomes in generation t. Allele frequencies "drift" randomly – but very slowly when N_t is large. To make the model realistic, N_t must be interpreted as the effective population size – much smaller than the true population size.

Figure 3 show the history of $n = 7$ chromosomes at a single site.

The rate of colaescences is proportional to number of pairs, $(n(n+1)/2$, and inversely proportional to effective population size N_t.

Most mutations are recent and have low frequency. Such consider-

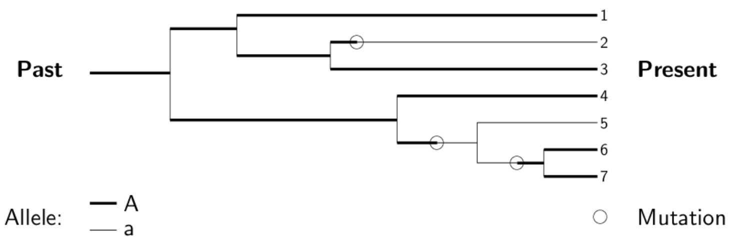


Figure 3: The coalescent model

ations have led some to believe that complex disease with a heritable component will be influenced by very many, individually rare, genetic variants. These would be difficult to detect in epidemiological studies. Simple models are complicated by effects of selection and of bottlenecks – reduction of population to a very small number followed by rapid expansion. These models can be extended to describe spectrum of multiple polymorphisms within a chromosome section, allowing for recombination between sites.

Segregation of traits in families

We shall consider a trait Y which, for now, is quantitative. The variance of the trait is the mean squared deviation of Y from the population mean, $\bar{Y}$:

$$\text{Variance}(Y) = \text{Mean}\{(Y - \bar{Y})^2\}$$

The covariance of the trait between two subjects is the mean of the products of their deviations from the population mean:

$$\text{Covariance}(Y) = \text{Mean}\{(Y_1 - \bar{Y}) \times (Y_2 - \bar{Y})\}$$

The correlation coefficient is the covariance scaled to lie between -1 and $+1$ by dividing by the trait variance.

Imagine individuals in a population sorted into groups so that, within each group, individuals are genetically identical at all loci relevant to a trait of interest. The environmental component of variance is the variance between trait values for subkects with the same genotype (i.e. the within-group variance). The genetic component of variance is the difference between this and the total variance (i.e. the between-group variance). It is also equal to the covariance between trait values in genetically identical individuals. The heritability of the trait is the ratio of genetic variance to the total variance.

Interpretation: because it is a ratio we must be careful about interpreting heritability as measuring the ”importance” of genetic influences. But substantial heritability points to an ”experiment of nature” which genetic epidmiologists can exploit.

Estimation: the definition is in terms of a ”thought experiment”. Real opportunities are scarce:

- Twin studies: compare trait correlations for monozygotic and for dizygotic twins
- Adoptee studies: compare traits correlation for true sibs and for adoptees
- More general family studies: we can infer heritability using covariance structure analysis – but we rely heavily on mathematical models.

  At the heart of this argument is an analysis of variance (Fisher, 1918). Imagine that the genetic effect is mediated by one locus (the ”trait locus”) with m alleles. If we could tabultae the trait mean by maternal and paternal allele, we would have an $m \times m$ table with $(1, \ldots, i, \ldots m)$ rows and $(1, \ldots, j, \ldots m)$ columns. Usually, we would only observe the traingular table, folded on the diagonal.

Assuming equality of effects of maternal and paternal alleles, the table is symmetric about the diagonal. Under H-W equilibrium, this is a "balanced design" – row and column assignments are independent.

Table 1: Trait means and frequencies

M	P	
	1 ... j ... m	Total
1		
...		
i	$\mu_{ij}(p_i p_j)$	$\mu_j = \sum_j p_j \mu_{ij}(p_i)$
...		
m		
Total	$\mu_j(p_j)$	$\mu = \sum_{ij} p_i p_j \mu_{ij}$

$$
\begin{aligned}
&= \text{Mean}(Y - \mu_{ij})^2 + \quad \text{"Environmental"} \\
&= \sum_i p_i(\mu_i - \mu)^2 + \sum_j p_j(\mu_j - \mu)^2 + \quad \text{"Additive"} \\
&= \sum_{ij} p_i p_j(\mu_{ij} - \mu_i - \mu_j + \mu)^2 \quad \text{"Dominance"}
\end{aligned}
$$

Analysis of variance:

Table 2: Analyse of variance table

Source	Component of variance	
Between rows (maternal alleles)	"Additive" genetic	σ^2_{Add}
Between columns (paternal alleles)		
Interaction (non-additivity)	"Dominance" genetic	σ^2_{Dom}
Subjects within cells (genotypes)	"Environmental" genetic	σ^2_{Env}

The additive component of variance is the variance "explained" by a model in which maternal and paternal alleles have simple additive effects on the mean trait value. The dominance component represents residual genetic variance not explained by a simple sum of effects (Table 2).

Note Fisher's use of a standard term of classical genetics, dominance, in a new way.

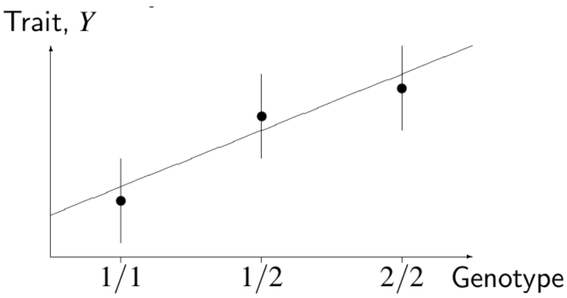


Figure 4: Example of a diallelic locus

In Figure 4, the environment variance is represented by the vertical bars. The total genetic variance is the variance between genotype means:

- Additive component is that due to the regression line
- Dominance component is that about the regression line

Two individuals who share two alleles IBD at the trait locus are genetically identical in so far as that trait is concerned. The covariance between their trait values is the total genetic variance $\sigma^2_{Gen} = \sigma^2_{Add} + \sigma^2_{Dom}$. Two individuals who share one allele IBD at the trait locus share the genetic effect of that allele. The covariance between their trait values is half the additive component of variance, $\sigma^2_{Add}/2$. Two individuals who share zero alleles IBD at the trait locus are effectively unrelated. The covariance between their trait values is zero. All this assumes there is no shared environmental influences.

But the degree of relationship between two individuals determines the probabilities of being 0, 1, or 2 IBD. In general, the covariance between trait values in two relatives is

$$z_1 \frac{\sigma^2_{Add}}{2} + z_2(\sigma^2_{Add} + \sigma^2_{Dom}) = 2\Phi\sigma^2_{Add} + z_2\sigma^2_{Dom},$$

where Φ is the kinship coefficient. The dominance compoennt is frequently small so that covariance (and hence correlation) is proportional to the kinship coefficient, Φ.

Fisher considered the model in which the effects of several loci combined additively. In this case, additive and dominant components at each locus also combine additively:

- Overall heritability is the sum of components due to each locus,
- Relationship between trait covariance (correlation) and family relationship is the same as for a single locus.

Fisher used the word epistasis in a new way, to denote deviation from additive effects. If effects are more than additive (some would term this ”synergism”), the correlation falls off faster with decreasing kinship.

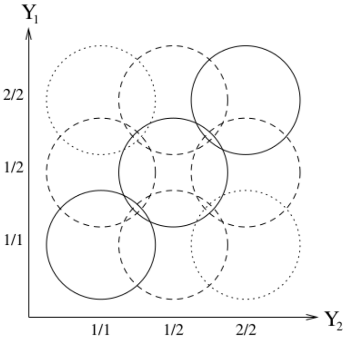
No. alleles shared IBD	2	1	0	Φ	r
	z_2	z_1	z_0		
Self, MZ twins	1	0	0	1/2	H
Parent-offspring	0	1	0	1/4	H/2
Full siblings, DZ twins	1/4	1/2	1/4	1/4	H/2
Half siblings	0	1/2	1/2	1/8	H/4
Uncle-nephew	0	1/2	1/2	1/8	H/4
Grandchild-grandparent	0	1/2	1/2	1/8	H/4
Double 1st cousins	1/16	6/16	9/16	1/8	H/4
First cousins	0	1/4	3/4	1/1	H/8

Table 3: IBD sharing, kinship, and trait correlation by relationship

Table 3 assumes no inbreeding, zero ”dominance”, and no ”epistasis”. A simple estimate of H (heritability) from twin studies is $2(r_{MZ} - r_{DZ})$.

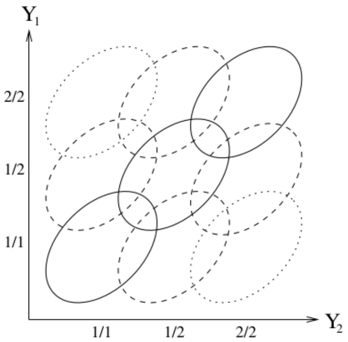
In the model for polygenic inheritance, the trait is determined by the sum of very many small effects of different genes. The distribution of the trait in two relatives, Y_1 and Y_2, is bivariate normal – an elliptical could of points. The correlation is determined by the degree of relationship (IBD probabilities) and the heritability.

Alternatively, if inheritance of the trait were due to a single major locus, the bivariate distribution for two relatives would be a mixture of circular clouds of points.



- Spacing of cloud centres depends on additive and dominance effects
- Marginal distributions of genotypes depend on allele frequency
- Tendency to fall along diagonals depends on IBD status (hence on relationship)

In the Morton-Maclean model, the trait is determined by additive effects of a single major locus plus a polygenic component. The bivariate distribution for two relatives is now a mixture of elliptical clouds.



This model can be fitted to trait values for individuals in pedigrees, using the method of maximum likelihood. It is necessary to allow for the manner in which pedigrees have been recruited into the study, or ”ascertained” – pedigrees in the study may be skewed, either deliberately or inadvertently, towards those with extreme trait values for one or more family members. Segregation analyses were often overinterpreted – the results depend on very strong model assumptions:

- additivity of effects (major gene, polygenes, and environment)
- bivariate normality of distribution of trait given genotype at the major locus

Aggregation of discrete traits, such as diseases in families have been studied by an extension of the Morton-Maclean model. Assume a latent ”liability” to disease behaves as a quantitative trait, with a mixture of major gene and polygene effects. When liability exceeds a threshold, disease occurs. This model may be fitted by maximum likelihood, although ascertainment corrections can be troublesome. As in the quantitative trait case, this approach relies upon strong modelling assumptions.

A less model-based approach is to study risk in relatives of diseased probands. These are termed recurrence risks. Recurrence risks are usually expressed relative to the general population risk, and denoted by λ_R where R denotes the relationship with the proband. We can use Fisher's (1918) results to predict the relationship between recurrence risk and relationship to affected probands, by considering a trait coded $Y = 0$ for healthy and $Y = 1$ for disease. Then,

$$\text{Population mean}(Y) = \Pr(Y = 1) = \text{Population risk, K}$$

An alternative algebric expression for the covariance is

$$\text{Covariance}(Y_1, Y_2) = \text{Mean}(Y_1 Y_2) - \text{Mean}(Y_1)\text{Mean}(Y_2)$$

and $\text{Mean}(Y_1 Y_2)$ is the probability that both members are affected. From this it follows that

$$\frac{\Pr(Y_2 = 1 \mid Y_1 = 1}{K} = \frac{\Pr(Y_2 = 1 \& Y_1 = 1}{K^2} = 1 + \frac{\text{Covariance}(Y_1, Y_2)}{K^2}$$

This is the relative recurrence risk, λ_R! As before, the covariance between Y_1, Y_2 depends on the IBD probabilities (i.e. the type of relationship), so that

$$\lambda_R - 1 = \frac{2\Phi\sigma^2_{Add} + z_2\sigma^2_{Dom}}{K^2}$$

Assuming no inbreeding and single gene, Table 4 gives the recurrence risks according to IBD sharing and relationship R.

Table 4: IBD sharing, kinship, and relative recurrence risk

No. alleles shared IBD	2	1	0	Φ	$\lambda_R - 1$
	z_2	z_1	z_0		
Self, MZ twins	1	0	0	1/2	σ^2_{Gen}/K^2
Parent-offspring	0	1	0	1/4	$\sigma^2_{Add}/2K^2$
Full siblings, DZ twins	1/4	1/2	1/4	1/4	$\sigma^2_{Add}/2K^2 + \ldots$
Half siblings	0	1/2	1/2	1/8	$\sigma^2_{Add}/4K^2$
Uncle-nephew	0	1/2	1/2	1/8	$\sigma^2_{Add}/4K^2$
Grandchild-grandparent	0	1/2	1/2	1/8	$\sigma^2_{Add}/4K^2$
Double 1st cousins	1/16	6/16	9/16	1/8	$\sigma^2_{Add}/4K^2 + \ldots$
First cousins	0	1/4	3/4	1/1	$\sigma^2_{Add}/8K^2$

Additive/dominance variance components are functions of penetrance. Assume disease is caused by a single diallelic locus, allele frequencies p and $1 - p$, and genotype relative risks are given in Table 5.

Table 5: Genotype relative risk and recurrence risk

Genotype	Probability	Relative risk	
1/1	$(1-p)^2$	1	(Reference)
1/2	$2p(1-p)$	1+δ	
2/2	p^2	1+2δ	

Then, $\sigma^2_{Add}/K^2 = 2p(1-p)\delta^2$ and $\sigma^2_{Dom} = 0$. Also, $\lambda_{PO} = \lambda_S = \lambda_{DZ} = 1 + p(1-p)\delta^2$,a nd $\lambda_{MZ} = 1 + 2p(1-p)\delta^2$. Under these assumptions, $\frac{\lambda_{MZ}-1}{\lambda_{DZ}-1} = 2$.

So far we have assumed that familial aggregation of disease is due to a single locus. If several genes are involved, the patterns of recurrence risk depend on how they act together. Two simple models have been studied to explore this:

- Genetic heterogeneity: a model for "parallel" action of genes – variation in any one of several genes can lead to disease susceptibility

- Additive model: with this model, the probability of disease is approximately predicted by a sum of effects of each locus

$$\lambda_R - 1 \approx \frac{2\Phi\sigma^2_{Add} + z_2\sigma^2_{Dom}}{K^2}$$

Here σ^2_{Add} and σ^2_{Dom} represent sum of additive and dominance contributions from each locus. Each locus makes an additive contribution to $\lambda_R - 1$. Thus, under the additive model, the relative magnitudes $\lambda_R - 1$ for different relatives is maintained. In particular, $(\lambda_{MZ} - 1)/(\lambda_{DZ} - 1) = 2$.

An alternative model is one in which genes act synergistically, as in classical epistasis. This yields a mutliplicative model, in which the overall penetrance is given by a product of effects from different genes. Then it is easily shown that recurrence risks also obey a multiplicative model: $\lambda_R = \prod_j \lambda_R^{(j)}$, where $\lambda_R^{(j)}$ represents the contribution of the \$j\$-th locus. This gives a more rapid decline of risk with distance of relationship. For example, a disease caused by two genes acting multiplicatively, each with zero dominance variance and $\lambda_{MZ} = 9$ (so that $\lambda_{MZ} - 1 = 8$), is illustrated in Table 6.

Table 6: Epistasis and synergism

Relative	Gene1	Gene 2	Overall
MZ twin	9	9	81
1st degree	5	5	25
2nd degree	3	3	9
3rd degree	2	2	4

A real example: recurrence risks for multiple sclerosis

Estimated lifetime risks in families of MS cases registered in the E.Anglian region are given in Table [[]].

Here, MZ twin risk $\approx 1/3$ and general population risk $\approx 1/800$.

Relative	N	Cases	Risk
Sibling	1350	43	.038
Parent	1239	25	.020
Offspring	1057	6	.018
Uncle/Aunt	2584	21	.009
Niece/Nephew	1776	10	.016
Cousin	3404	23	.009

Table 7: Recurrence risks for multiple sclerosis

Probability and identity by descent in Families

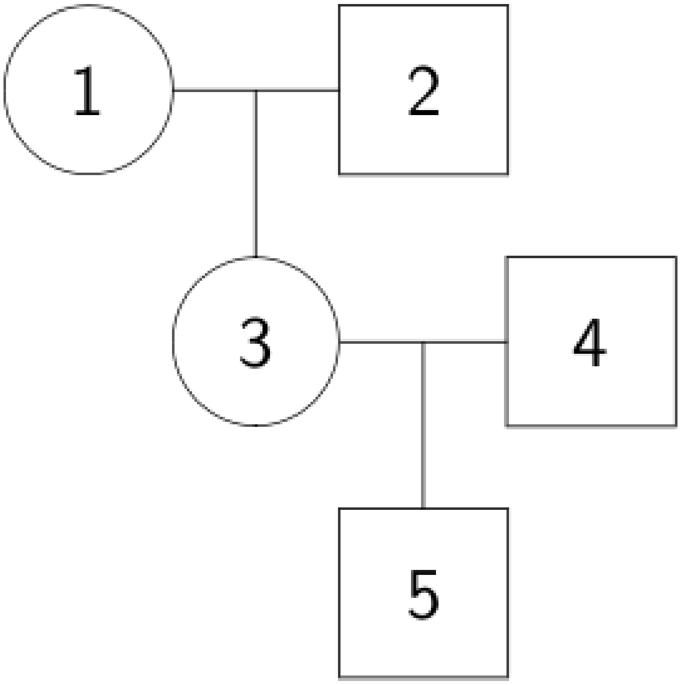


Figure 5: Probability calculations on pedigrees

Pedigree members 1, 2 and 4 are founders, while 3 and 5 are descendants. The probability of all five genotypes is

$$\text{Pr}(\text{Founder genotypes}) \times \text{Pr}(\text{Transmission to descendents})$$

Probabilities of genotypes are defined as relative frequencies in a very large (infinite) population. Programs usually assume random mating and Hardy-Weinberg equilibrium – the four alleles carried by parents are as if sampled independently from a single population.

Two genes which are copies of a common ancestral gene are said to be identical by descent (IBD). For example, in Figure 6, subjects 3 and 4 share one gene IBD (the paternal allele, *a*). But, in these families, they share respectively 0 and 2 genes IBD.

We only know IBD status if all genotypes are observed and the two parents have, between them, four different alleles. In this next case subkects 3 and 4 share a gene identically by state (IBS), but none IBD (their *a* alleles are from different parents).

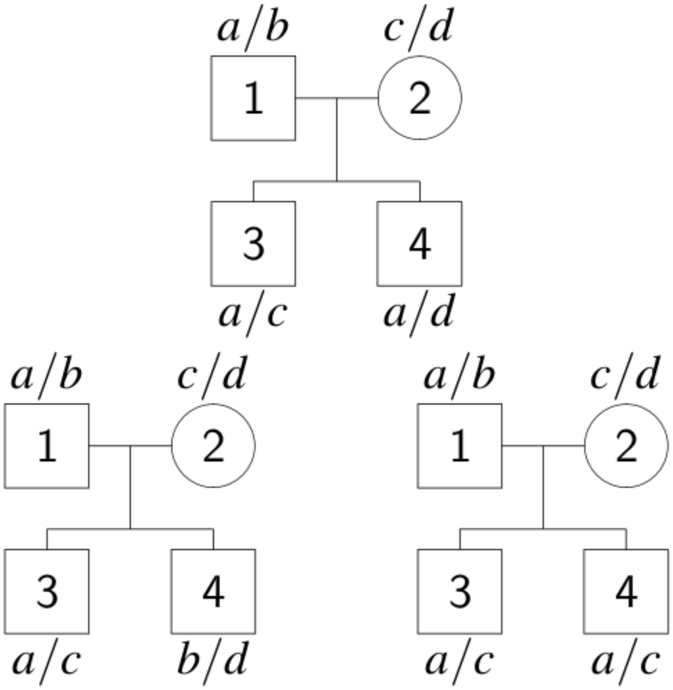
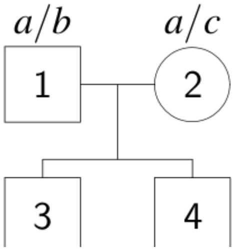
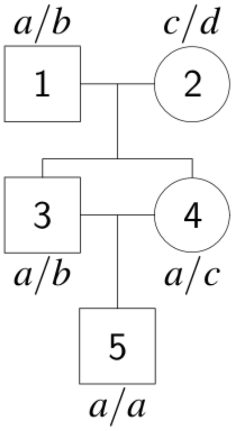


Figure 6: Identity by descent

Can an individual share a gene IBD with himself? Only if there is inbreeding.

Given only pedigree structure we can calculate probabilities of IBD states using only Mendel(s first law: a parent will transmit either gamete with probability $\frac{1}{2}$. Inbreeding coefficient is defined as the probability that the two alleles within a single inidivudal are IBD. In our example, this is $\frac{1}{4}$ (if the founders are not themselves inbred) – for each of the four grandparental genes there is a probability of $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$ of transmission of both copies to the grandchild, so the total probability that the grandchild's two genes are IBD is [1]





$$4 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

When we are unable to assign IBD sharing we can assess the probability that two individuals share 0, 1 or 2 genes IBD. These probabilities are often denoted by (z_0, z_1, z_2) – but there are several types of IBD probabilities depending on the information available. Prior IBD probabilities are the probabilities of IBD sharing conditional only upon the relationship between the two subjects.

Sibling pair (figure in margin) – what are prior values of (z_0, z_1, z_2)? Say that the first sib inherited a and c alleles:

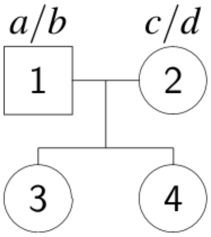
- 2-IBD: probability that second sib also inherits a and c is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
- 1-IBD: probability that second sib inherits b and d is also $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

[1] Exercise: What is the inbreeding coefficient for a child of a first cousin marriage (you can assume that founders are not inbred – a usual assumption for probability calculations in human populations)?

- 0-IBD: probability that second sib is a/d or b/c is $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

Since the labelling of alleles is arbitrary, this argument holds regardless of which alleles the first sib inherits. Thus, for two siblings:

$$z_0 = \frac{1}{4}, \quad z_1 = \frac{1}{2}, \quad z_2 = \frac{1}{4}$$



Each cousin must inherit one allele from each pair of grandparents. [2] In either case, there are four equally probable alternatives. The probability that they inherit the same grandparental copy from both sides is $1/4 \times 1/4 = 1/16$. This is the probability that they are 2-IBD. The probability that they inherit different alleles from the two sides is $3/4 \times 3/4 = 9/16$. This is the probability that they are 0-IBD. Thus the probability of 1-IBD is $1 - 1/16 - 9/19 = 3/8$.

Consider one gene at a given locus picked at random from each of two relatives. The kinship coefficient (denoted by Φ) is defined as the probability that these two genes are IBD. Given no inbreeding:

- if they are 2-IBD, probability $= \frac{1}{2}$,
- if they are 1-IBD, probability $= \frac{1}{4}$,
- if they are 0-IBD, probability $= 0$,

so that $\Phi = \frac{1}{2}z_2 + \frac{1}{2}z_1$, half the average proportion of genes shared IBD. But see Table 3.

[2] Exercise: (double first cousins) What are the probabilities of sharing 2, 0 alleles IBD? Hence, what is the probability of sharing 1 allele IBD?

Parent-of-origin effects

An important aspect of transmission/disiquilibrium studies is the ability to differentiate the effects of alleles according to their parent-of-origin. When origin is known, we will write genotype as m/p, where m is the allele inherited from the mother and p the allele inherited from the father. In the diallelic case, we are interested to determine if the risk associated with genotype 1/2 is the same as that for 2/1. It is relatively simple to restrict the TDT calculations to transmissions from mothers, or from fathers; there is only one type of family which presents problems.

When offspring is 1/1 we know that both parents have transmitted allele 1. Similarly for allele 2, when the offspring is 2/2. But, when the offspring is 1/2 we know that one parent has transmitted allele 1 and one has transmitted allele 2 – but we don't know which. Triads in which mother, father, and child have identical genotype can be included in the simple TDT, but not in an analysis by parent-of-origin.

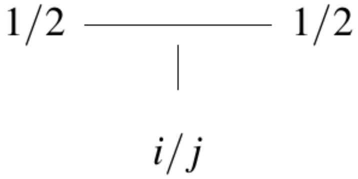
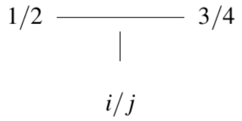


Figure 7: Testing for differential parental transmission



$i/j =$ 1/3 each with probability 0.25
1/4
2/3
2/4

Remember the factorization of transmission probabilities (Figure 7). Under the multiplicative model, RR for genotype i/j, $\theta_{i/j} = \Phi_i \Phi_j$.

$$\Pr(\text{Child is i/j}) = \frac{\Phi_i \Phi_j}{\Phi_1\Phi_3 + \Phi_1\Phi_4 + \Phi_2\Phi_3 + \Phi_2\Phi_4} = \frac{\Phi_i}{\Phi_1 + \Phi_2} \times \frac{\Phi_j}{\Phi_3 + \Phi_4}$$

This argument extends naturally to the case where maternal and paternal alleles carry different Φ's. Under the multiplicative model, maternal and paternal transmissions are independent.

This in turn suggests a simple contingency table analysis. For a diallelic marker we have a 2 × 2 table with allele transmitted (1, 2) in rows and heterozygous (mothers, fathers) in columns. There are two problems:

- Independence would only hold under the strict multiplicative model, and
- Omission of intercross triads with heterozygous offspring creates dependence between transmissions

One proposal is simply to omit all transmissions from intercross families – each triad then only contributes one informative transmission. This is the transmission asymmetry test (TAT).

An alternative analysis is ti use the full conditional likelihood – equivalent to case vs three ”pseudo-controls”. ”Intercross” triads require special treatment:

- omit sets based around heterozygous affected offspring (cases),
- but we must also omit heterozygous pseudo-controls from other sets. . .
- a 1/1 case has a single 2/2 control and vice versa – these sets are uninformative about parent-of-origin effects

For a diallelic locus, we can show that the information for a parent-of-origin effect comes from two 2 × 2 tables (Table 8).

Mother		1/1	1/2		2/2	1/2
Father		1/2	1/1		1/2	2/2
Child	1/1	a_1	b_1	1/2	a_2	b_2
	1/2	c_1	d_1	2/2	c_2	d_2

Table 8: Information on parent-of-origin in a CPG analysis

The two odds ratios, $(a_1 \times d_1)/(b_1 \times c_1)$ and $(a_2 \times d_2)/(b_2 \times c_2)$, both estimate the relative risk $\theta = \pi_{2|1}/\pi_{1|2}$. CPG analysis is equivalent to Mantel-Haenszel pooled test for association across the two tables. The TAT analyses the single collapsed table – only valid when parental genotype does not affect risk (except via child's genotype).

Another approach, suggested by Weinberg (AJHG, 61:229-235), is also based on counts of triads of parents and an affected offspring. Consider the frequency of mating types. For a diallelic locus, we build a 2 × 2 table with mother and father's alleles in rows and columns. In the population, diagonally opposite frequencies should be eqaul. After selection by affected offspring, the ratio of such frequencies should reflect corresponding offspring risk ratios. This assumption can be termed parental symmetry, or parental exchangeability.

Mother	Father	Child
1/1	1/1	1/1
1/2	1/2	1/1, 1/2, 2/2
2/2	2/2	2/2
1/2	1/1	1/1
1/1	1/2	1/1
1/2	1/1	1/2
1/1	1/2	1/2
2/2	1/1	1/2
1/1	2/2	1/2
2/2	1/2	1/2
1/2	2/2	1/2
2/2	1/2	2/2
1/2	2/2	2/2

Table 9: The 15 possible case-parent triads

In Table 9, first group of triads are not (directly) informative about parent-of-origin effects. Remaining 5 pairs have asymmetries between parental genotypes which might relate to risk and, therefore, to frequency of the triads. Weinberg suggest conditioning upon mating type and affected offspring genotype, regarding the ”response” variable as being which parent is which.

Table 10: The parental asymmetry test (TAT)

Mother	Father	Child	Risk	Odds
1/1	1/1	1/1	$\pi_{1/1}$	1
1/1	1/2	1/1	$\pi_{1/1}$	
1/2	1/1	1/2	$\pi_{2/1}$	θ
1/1	1/2	1/2	$\pi_{1/2}$	
2/2	1/1	1/2	$\pi_{2/1}$	θ
1/1	2/2	1/2	$\pi_{1/2}$	
2/2	1/2	1/2	$\pi_{2/1}$	θ
1/2	2/2	1/2	$\pi_{1/2}$	
2/2	1/2	2/2	$\pi_{2/2}$	1
1/2	2/2	2/2	$\pi_{2/2}$	

In Table 10, in middle three groups, count a as total number in which child genotype was 2/1, and b as the total number in which it was 1/2. ML estimate of θ is a/b. Chi-squared test (1-df) is $(a-b)^2/(a+b)$. However, this procedure assumes that maternal genotype has no direct effect on the risk of disease in the child (because the case vs. pseudo-control method conditions upon parental genotype, that method makes no such assumption).

Table 11 shows the effect of maternal genotype. Note that the thrid pair of triads are uninformative in the CPG analysis – the assumption of parental symmetry allows additional data to be used. Weinberg proposed that we can estimate θ, ψ_1, ψ_2 by logistic regression, and calculate a likelihood ratio test for $\theta = 1$.

Table 11: Allowing for a (mutliplicative) effect of maternal genotype

Mother	Father	Child	Risk	Odds
1/1	1/1	1/1	$\pi_{1/1,1/2}$	ψ_1
1/1	1/2	1/1	$\pi_{1/1,1/1}$	
1/2	1/1	1/2	$\pi_{2/1,1/2}$	$\theta\psi_1$
1/1	1/2	1/2	$\pi_{1/2,1/1}$	
2/2	1/1	1/2	$\pi_{2/1,2/2}$	$\theta\psi_2$
1/1	2/2	1/2	$\pi_{1/2,1/1}$	
2/2	1/2	1/2	$\pi_{2/1,2/2}$	$\theta\psi_2/\psi_1$
1/2	2/2	1/2	$\pi_{1/2,1/2}$	
2/2	1/2	2/2	$\pi_{2/2,2/2}$	ψ_2/ψ_1
1/2	2/2	2/2	$\pi_{2/2,1/2}$	

Using logistic regression, ”indicator variables” σ, m_1 and m_2 are used to account for maternal origin and genotype effects (Table 12). There is no intercept, 5 data points and 3 parameters, so that there are 2 df for ”fit”, i.e. deviation from multiplicative effects of origin and maternal genotype.

Mother	Father	Child	Risk	Response	σ	m_1	m_2
1/1	1/1	1/1	$\pi_{1/1,1/2}$	1	0	+1	0
1/1	1/2	1/1	$\pi_{1/1,1/1}$	0			
1/2	1/1	1/2	$\pi_{2/1,1/2}$	1	+1	+1	0
1/1	1/2	1/2	$\pi_{1/2,1/1}$	0			
2/2	1/1	1/2	$\pi_{2/1,2/2}$	1	+1	0	+1
1/1	2/2	1/2	$\pi_{1/2,1/1}$	0			
2/2	1/2	1/2	$\pi_{2/1,2/2}$	1	+1	-1	+1
1/2	2/2	1/2	$\pi_{1/2,1/2}$	0			
2/2	1/2	2/2	$\pi_{2/2,2/2}$	1	0	-1	+1
1/2	2/2	2/2	$\pi_{2/2,1/2}$	0			

Table 12: Logistic regression

Interaction between maternal and child genotypes can masquerade as a parent-of-origin effect (Table 13). In this example, allele 2 carries RR of θ – but only if mother is $1/1$. This makes the same predictions as the model underlying the PAT – except for the fourth comparison. For $\theta > 1$, this scenario looks very much like excess paternal transmission. In general, it is rather difficult to distinguish "imprinting" effects from interaction between mother and child genotypes.

Mother	Father	Child	Risk	Odds
1/2	1/1	1/1	$\pi_{1/1}$	1
1/1	1/2	1/1	$\pi_{1/1}$	
1/2	1/1	1/2	$\pi_{2/1}$	$1/\theta$
1/1	1/2	1/2	$\pi_{1/2}$	
2/2	1/1	1/2	$\pi_{2/1}$	$1/\theta$
1/1	2/2	1/2	$\pi_{1/2}$	
2/2	1/2	1/2	$\pi_{2/1}$	1
1/2	2/2	1/2	$\pi_{1/2}$	
2/2	1/2	2/2	$\pi_{2/2}$	1
1/2	2/2	2/2	$\pi_{2/2}$	

Table 13: Interaction between maternal and child genotypes

Weinberg's method must be expected to be more efficient – particularly if we can assume no effect of maternal genotype. Additional assumption of a priori parental exchangeability allows use of $1/1 + 2/2$ matings. The genralization of these methods to loci with more than 2 alleles is relatively straightforward. Generalization to > 1 affected sibs per family is not straightforward – we get bias as well as wrong standard errors.

We can extend the case/pseudo-control analysis to incorporate the additional information available by assuming exchangeability of parental genotypes. We condition upon the two parental genotypes but not on their order. The likelihood is then equivalent to comparing the case with 7 pseudo-controls (Figure 8). Again, generalization to > 1 affected offspring is not straightforward.

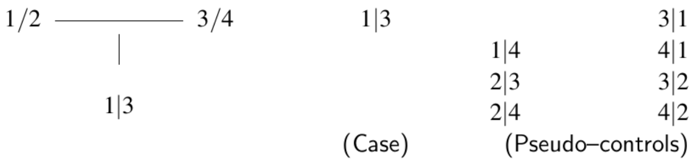
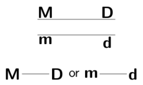


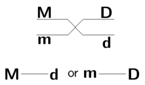
Figure 8: Conditioning on exchangeable parental genotype

Family-based studies and the TDT

wip

Linkage studies





If two loci, M and D, are linked, the probability that they be passed down together as a haplotype depends upon the probability of recombination during meiosis. For unlinked loci (e.g., on different chromosomes), the recombination fraction $\theta = 0.5$ while, for completely linked loci $\theta = 0$. If D is an unknown disease gene and M is an observed marker locus, and if M and D are linked, the marker will segreagate in the same way as disease.

In general, different marker alleles will segregate with disease in different families.

In order to be useful for mapping disease genes, marker loci must have two characteristics:

1. they should be highly polymorphic, so that their segregation in families can be tracked accurately, and

2. their locations should be known accurately

The most useful marekrs for this purpose are microsatellite markers – repeated nucleotide sequences. The most frequently occuring microsatellites are dinucleotide repeats, but tri- and tetranucleotide repeats are preferred since they can be typed more accurately.

Genetic disatnces between markers are determined by typing them in a standard set of families, e.g., the 3-generation CEPH pedigrees. These estimates are subject to error. With the completion of the human genome project, physical location of markers are now known. However, there is an uneven relationship between physical distance (bases) and recombination fractions and genetic distance (cM). At least we can now be fairly confident of the order of markers along a chromosome.

We observe pedigrees in which more than one member has disease. For quantitative traits we measure markers and trait values in families and see whether marker and trait segregate similarly. Again we may choose to ascertain families according to trait values if we wish – e.g., strongly concordant or discordant sibs; this will often result in increased power to detect linkage. Two appraoches to the analysis of linkage studies have grown up, usually termed ”parametric” and ”non-parametric”.

Parametric linkage analysis: Assume a diallelic disease gene with known allele frequencies and penetrances. For multipoint analyses (using multiple markers), also assume the inter-marker genetic distances to be known. Only the location of the disease gene, and hence the recombination fractionw ith marker(s), is regarded as unknown. We investigate the support for different (genetic) locations of the disease gene by calculating the likelihood for different values of the recombination fraction(s) – the probability of data hiven θ. It is conventional to express this in terms of the log (base 10) of the ratio of the likelihood to the likelihood at $\theta = 0.5$ – the maximized LOD score (MLS).

In contrast with segregation studies, in the analysis it turns out not to matter how we ascertained pedigrees – the part of the likelihood whihc depends on ascertainment doesn't depend on θ so the MLS is unaffected. However, the MLS can be strongly affected by what we assume for penetrance, causal allele frequency, etc., and we can't generally estimate these from linkage studies due to ascertainment of families. This approach is most suitable for simple "Mendelian" traits, e.g., diseasescaused by a single gene with high penetrance, but has been extended to deal with genetic heterogeneity, i.e., where some pedigrees are linked to one gene and other pedigrees to another, as in the case of *BrCa1* and *BrCa2* breast cancer families.

Non-parametric or model-free approaches: If marker and trait loci are linked, affected relative pairs will have more IBD sharing at the marker locus than we would expect given their relaionship. There is no explicit model for inheritance, although decisions must be taken about how to score IBD sharing and how much emphasis to give to different types of relative pairs. The most common type of study is the affected sib pair study – largely because these are the most readily available. Parents may be collected and genotyped. Their availability improves the accuracy of IBD estimates when markers are not fully polymorphic, and protects against genotyping errors.

The simplest scoring system is to compare observed and expected numbers of genes shared IBD between affected pairs in the pedigree. An unselected sibling pair is expected to share $0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$ IBD. When we observe an affected sib pair to be 0-, 1- or 2-IBD we score it -1, 0, or $+1$ respectively. More usually we don't know IBD status precisely. Observed and expected IBD sharing is calculated by

$$0 \times z_0 + 1 \times z_1 + 2 \times z_2$$

where z's are "prior" and "posterior" IBD sharing probabilities respectively.

An example is shown in Figure 9.

Given these data, $z_1 = z_2 = 0.5$ so that, *a posteriori*, the expected number of genes shared IBD is $0.5 + 0.5 \times 2 = 1.5$. Thus, this sibship contributes $+0.5$ to the total score. The non-parametric linkage (NPL) score is a t or z statistic calculated by dividing the total score by its standard deviation.

In order to calculate NPL scores, it is only necessary to calculate posterior IBD probabilities under the null hypothesis ($z_0 = z_2 = 0.25$,

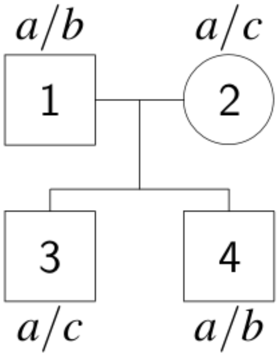


Figure 9: Example of IBD sharing

$z_1 = 0.5$). We can also calculate the likelihood under this assumption. We can also maximize the likelihood with respect to the prior IBD probabilities, and hence calculate an MLS. Often the likelihood is maximized over values of z_0, z_1, z_2 consistent with plausible genetic models – the ”possible triangle restriction”. MLS scores calculated in this way are not strictly comparable with those from parametric linkage analyses.

Parametric and non-parametric methods face the same problems of computing the probability of various inheritance patterns within pedigrees. Single point analyses consider the problem of a single marker plus the unobserved trait locus while multipoint analyses consider all the markers simultaneously. Computation can be laborious owing to having to consider possible recombination at each meiosis and by missing data; some pedigree members are unobserved for trait and/or markers. Even if markers are observed, their phase is not directly observed but must be inferred.

Here is an illustration for the case of unknown phase. We observe only that the father is A/a and B/b at two loci, as shown in Figure 10.

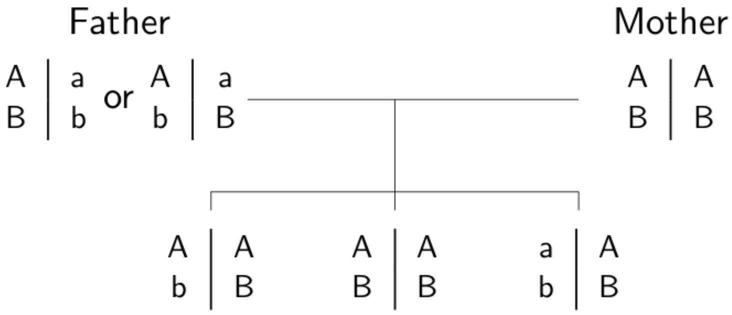


Figure 10: The case where phase is not known

Depending on the paternal phase, there is either one or two recombinations. If the two phases are equally probable the likelihood contribution is:

$$\theta(1-\theta)^2 + \theta^2(1-\theta) = \theta(1-\theta).$$

There are two widely used algorithms:

1. Elston-Stewart: deals with few markers (≤ 5) but large pedigrees
2. Lander-Green: deals with many markers but with pedigrees of limited depth (≤ 4 generations)

The former method is used in the classical linkage analysis programs which implement parametric linkage analysis: LINKAGE, FASTLINK, VITESSE. The latter method is used in programs which implement non-parametric methods: MAPMAKER/SIBS, GENEHUNTER, ALLEGRO, MERLIN. All programs use a standard ”preped” data structure.

The preped file layout is as follows:

1. Pedigree identifier
2. Member identifier within pedigree
3. Identifier of father of this person (or zero for a founder)
4. Identifier of mother of this person (or zero for a founder)
5. Sex (1 = Male, 2 = Female)
6. Disease status (1 = Unaffected, 2 = Affected)
7. . . . pairs of values containing the two alleles at each marker

By convention, missing data items are coded as zero.

There are strengths and weaknesses. Linkage analysis is unaffected by allelic heterogenity ($>$ 1 causal variant in a disease susceptibility gene) – there is no assumption that the same marker allele segregates with disease in different families. ”Mendelian” traits:

- Only one or two loci are involved, but causal variants are highly penetrant
- Large pedigrees with multiple cases of disease are observed
- Parametric linkage analysis is efficient, and the pattern of segregation is clear so that plausible models can be chosen

For ”complex” traits:

- Disease susceptibility varies due to variation in several (perhaps many) loci

- Common variants with small effects and/or rare variants with larger (though still modest) effects may be involved
- Typically we do not observe large multiply-affected pedigrees – recurrence risks fall off quickly with decreasing kinship
- Parametric linkage methods can be misleading if we search for the model which gives the biggest linkage peaks.

The probabilities that two relatives are affected given their IBD state is given in Table 14, where λ_{MZ} and $\lambda_{\text{P/O}}$ refer to part of the RRR attributable to this locus.

IBD state	$\Pr(Y_1 = 1\text{and}Y_2 = 1 \mid \text{IBD-state})$
2	$K^2 + \sigma^2_{\text{Add}} + \sigma^2_{\text{Dom}} = K^2\lambda_{\text{MZ}}$
1	$K^2 + \sigma^2_{\text{Add}}/2 = K^2\lambda_{\text{P/O}}$
0	K^2

Table 14: Power of affected relative pair studies

By applying Bayes rule, the probabilities of IBD states 0, 1, 2 for two relatives of type R given that they are both affected are $z_0(1/\lambda_{\text{R}})$, $z_1(\lambda_{\text{P/O}}/\lambda_{\text{R}})$ and $z_0(\lambda_{\text{MZ}}/\lambda_{\text{R}})$, where (z_0, z_1, z_2) are the prior IBD probabilities for relatives of this type.

For sib pairs, we have

$$\frac{1}{4\lambda_{\text{Sib}}}, \frac{\lambda_{\text{P/O}}}{2\lambda_{\text{Sib}}}, \frac{\lambda_{\text{MZ}}}{4\lambda_{\text{Sib}}}$$

Example:

Genotype	Frequency	Relative Risk	
1/1	25%	1.0	(reference)
1/2	50%	1.5	
2/2	25%	2.0	

Table 15: Example of affected relative pair studies

Here, $\lambda_{\text{P/O}} = \lambda_{\text{Sib}} = 1.0625$, and $\lambda_{\text{MZ}} = 1.125$ so that the probabilities that an affected sib pair share 0-, 1- or 2-IBD at the disease locus are 0.235, 0.5, and 0.265. Even if marker were totally informative, and tightly linked to the disease locus, huge sample sizes are necessary to detect difference from $(0.25, 0.5, 0.25)$ sharing.

Increased IBD sharing at a disease-susceptability locus falls off with distance from the locus. The rate at which this happens determines how closely spaced a set of markers must be to cover the whole genome.

- The more distant the relative pairs, the more rapidly increased IBD sharing falls off with distance – there are more intervening meioses and, therefore, more opportunities for recombination.
- For sib pair studies, a 400 marker set, corresponding approximately to 10cM spacing, is adequate.

A corollary to this is that, the closer the relationship between the affected relative pairs, the less certain we can be about the true position of a disease locus. With sib pairs we might be able to implicate a 10 to 20 cM region – corresponding to about 10 to 20 million nucleotide bases.

A whole-genome screen involves a high degree of multiple testing. It has been calculated that a LOD score peak of 3 corresponds to a whole-genome p-value of 0.05 – one false positive in 20 genome screens. The equivalent NPL score is ˜ 3.7, corresponding to an uncorrected p-value of 1×10^{-4} (one-sided). Complex diseases may have modest total λ_{Sib}, shared between several/many loci; if we need to achieve such standards of proof, the power of linkage studies may be modest, even with very large numbers of sib pairs.

The MRC BRIGHT study: A collaborative study between Aberdeen, Cambridge, Glasgow, Leicester, London (Barts), Oxford, and CNG Paris. First phase was an affected sib pair linkage study. ”Cases” defined as falling within the top 5% of the BP distribution with onset before age of 60. Second phase was a TDT trio collection, now extended to add a population-based control group from each centre. Results were first published in The Lancet, June 2003.

Initial power calculations (Mark Lathrop & Joe Terwilliger) were as follows: Assumed 30% heritability for blood pressure variation, shared between 5 loci (although the proposal is for a clinical endpoint, not a QTL), and 80% power to detect linkage at genome-wide 5% level using 300 equally spaced markers (average heterozygosity 80%). Initial target was 1500 affected sibling pairs – without parents.

Table 16: The linkage study families

No. families	Full-sibs	Half-sibs	Affected sib pairs
1361	1	0	1361
6	2	0	12
150	3	0	450
1	4	0	4
21	6	0	126
2	10	0	20
1	15	0	15
1	21	0	21
43	0	1	43
1	0	3	3
11	1	2	33
1	6	4	10

On chromosome 5p we have an MLS of 2.21, while on chromosomes 6q and 9q the MLS is 3.00 and 2.37.

Given the estimates of IBD sharing probabilities at each location, we can estimate the corresponding values of λ_{Sib} (assuming no dominance variance). Confidence intervals can be obtained by bootstraping (Table 17).

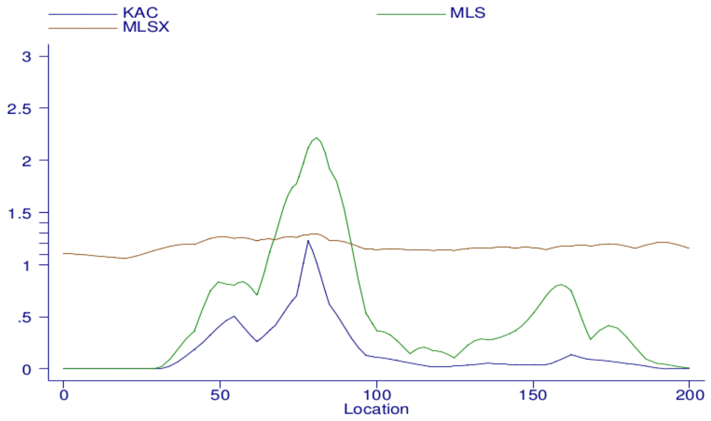


Figure 11: Chromosome 5

Table 17: Confidence intervals by bootstrap

Chrm.	MLS	$\widehat{\lambda}_{Sib}$	SE	95% CI
5p	2.21	1.056	0.024	1.020–1.114
6q	3.00	1.083	0.037	1.038–1.195
9q	2.37	1.163	0.057	1.057–1.286

We should expect these estimates to be biased upwards (selection of most significant results).

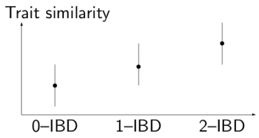
As for disease traits, there are parametric and non-parametric approaches to the analysis.

Parametric approach: (variance components)

- Trait normally distributed conditional upon causal genotype
- We can fit a variance components model to pedigree data and hence calculate MLS statistics for linkage
- This approach is highly dependent on validity of the normality assumption

Non-parametric analysis: Look for association between IBD state and trait similarity in relative pairs.

In the Haseman-Elston method of analysis, we collect pairs of relatives (usually siblings). Given a very highly polymorphic marker we could classify each pair as 0-, 1-, or 2-IBD. Then, we relate trait similarity measured either by (minus) squared difference in trait value, $-(Y_1 - Y_2)^2$, or by product of deviations, $(Y_1 - \bar{Y})(Y_2 - \bar{Y})$, to IBD sharing.



- More generally we cannot assign IBD status with certainty
- Use estimated IBD sharing score: $1 \times \Pr(1 - \text{IBD} = +2 \times \Pr(2 - \text{IBD})$., where the probabilities are posterior probabilities of IBD state given marker data.

Power is increased by sampling extremes of trait similarity, e.g. sibs who have aither very similar or very different trait values (concordant and discordant relative pairs). The analysis requires some modification – better to take IBD state as the dependent variable and trait similarity as independent variable. Note that variance component methods are invalidated by sampling on trait similarity.

Linkage and association

Linkage studies are not powerful against the smalle effects we are comint to expect in disease genetics. Increasingly attention has been directed to association studies which look for association between phenotype and genotype at the population level:

- direct approach: test candidate causal polymorphisms, e.g. an SNP which leads to an amino acid substitution
- indirect approach: test marker loci which may be associated with the causal variant at the population level

Then, in contrast with linkage, the same marker allele(s) track the causal variant across the whole population. Association between loci at the population level is called allelic association.

Haplotype relative frequencies between two loci A and B are defined as in Table 18.

Table 18: Haplotype relative frequencies

A	B		
	1	2	·
1	p_{11}	p_{12}	$p_{1\cdot}$
2	p_{11}	p_{12}	$p_{2\cdot}$
·	$p_{\cdot 1}$	$p_{\cdot 2}$	1

There is allelic association between loci A and B if $P_{ij} \neq P_{i\cdot} timesP_{\cdot j}$. Reasons for allelic association may be different:

- linkage disequilibrium: allelic association due to close proximity of loci
- stratification: the population has two distinct subpopulations, with different allele frequencies at both loci (Figure in margin)
- admixture: due to interbreeding of two founder populations. Proportion of genes derived from the different populations varies between individuals

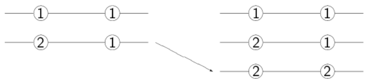
$$\begin{array}{cc|c} .49 & .21 & .7 \\ .21 & .09 & .3 \\ \hline .7 & .3 & 1 \end{array} + \begin{array}{cc|c} .09 & .21 & .3 \\ .21 & .49 & .7 \\ \hline .3 & .7 & 1 \end{array} \rightarrow \begin{array}{cc|c} .29 & .21 & .5 \\ .21 & .29 & .5 \\ \hline .5 & .5 & 1 \end{array}$$

There is also a phenomenon known as erosion of linkage disequilibrium by recombination. An individual could receive $i - j$ haplotype directly from a parent or he/she could receive a recombinant haplotype formed by crossover of parental $i - x$ and $y - j$ haplotypes. It follows that the expected change in haplotype frequency between one generation, P_{ij}, and the next, $P^{\star}_{ij}$, is

$$P^{\star}_{ij} = (1 - \theta)P_{ij} + \theta P_{i\cdot}P_{\cdot j}.$$

Assuming allele frequencies to stay the same, the disequilibrium coefficients $P^{\star}_{ij} - P_{i\cdot}P_{\cdot j} = (1 - \theta)(P_{ij} - P_{i\cdot}P_{\cdot j})$ decay deometrically. If θ is small, this is approximately exponential.

Let's start from the beginning. There once was no variation at locus B and there were only two $A - B$ haplotypes. Then there was a mutation at B on one chromosome (Figure in margin). The $2 - 2$ haplotype will either die out or "drift" to an appreciable frequency in the population, with or without selective pressure.



Before the first recombination, the haplotype frequencies will be as follows (Table 19):

Table 19: Haplotype relative frequencies before first recombination

A	B		
	1	2	·
1	$p_{1\cdot}$	0	$p_{1\cdot}$
2	$p_{\cdot 1} - p_{1\cdot}$	$p_{\cdot 2}$	$p_{2\cdot}$
·	$p_{\cdot 1}$	$p_{\cdot 2}$	1

The initial disequilibrium coefficient reads:

$$\begin{aligned} P_{11} - P_{1\cdot}.P_{\cdot 1} &= P_{1\cdot} - P_{1\cdot}.P_{\cdot 1} \\ &= P_{1\cdot}(1 - P_{\cdot 1} \\ &= P_{1\cdot}.P_{\cdot 2} \end{aligned}$$

Disequilibrium then decays due to recombination. Lewison's D' measure expresses the disequilibrium coefficient relative to this initial value:

$$D' = \frac{P_{11} - P_{1\cdot}.P_{\cdot 1}}{P_{1\cdot}.P_{\cdot 2}}.$$

In order to estimate D' we construct a 2×2 contingency table, and plug the observed relative frequencies into the formula for D'. But we must arrange the table correctly – with the initially absent haplotype in the correct celll. In practice we must guess which one this is from the current data – a procedure which leads to (upward) bias. It is then easily calculated from the observed and expected frequencies of the conventional chi-squared test: In case two of the cells have $O < E$, choose the one with the smallest O – we assume that this represents the most recent haplotype. Then $D' = 1 - O/E$.

If d is the genetic distance between the loci and the recombination fraction is small, $(1 - \theta) \approx e^{-d}$ and, if the (younger) mutation is t generations old, the expected value of D' is given by the Malecot model:

$$\mathbb{E}(D') = e^{-td}.$$

Allowing for further mutation and for the upward bias in the estimate,

$$\mathbb{E}(D') = Ae^{-td} + B.$$

However, this is only an expectation – the real values are created by a random process; observed values of D' vary considerably around the model.

In practice we do not observe haplotypes; we observe the genotype at the two loci. Our data is a 3×3 contingency table with 1/1, 1/2 and 2/2 in rows and columns, instead of 1 and 2 as before, so that 8/9 genotype combinations can be resolved into pairs of haplotypes. The ninth could be either $\begin{matrix} 1-1 \\ 2-2 \end{matrix}$ or $\begin{matrix} 1-2 \\ 2-1 \end{matrix}$. The EM algorithm is used to iteratively carry out the following calculations:

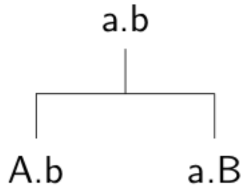
1. Resolve phase-ambiguous genotypes by splitting them between the two assignments in ratio given by their relative probability using current estimates of haplotype frequencies

2. Count assignments to obtain new haplotype frequencies

A different measure of LD is provided by the coefficient r^2. Recall that D' is derived from population genetic considerations; there is no intention that $D' = 1$ should imply that the two loci carry the same information. The most important index with this property is

$$r^2 = \frac{(P_{11} - P_{1\cdot}P_{\cdot 1})^2}{P_{1\cdot}P_{2\cdot}P_{\cdot 1}P_{\cdot 2}}.$$

Note that r^2 can be small even when D' is 1. A wild-type $a.b$ haplotype is modified by single mutations at each locus, introducing new alleles A and B. There is no subsequent recombination. The haplotype phylogeny and the haplotype frequencies are given in the following Table and in the margin Figure.



	b	B
a	90	10
A	10	0

Table 20: Haplotype relative frequencies before first recombination

From the Table above, we have $r = -0.111$ and $r^2 = 0.012$. The low r^2 arose because the mutations occured on different branches of the chromosome ancestry.

Cross-overs occur randomly at each meiosis with rate 1 per Morgan. Consider the segment around a fixed locus $(\cdot)$ on chromosomes inherited IBD by two siblings, as shown in Figure 12. (x indicates cross-overs.)

In the "sum process", cross-overs occur at the rate of 2 per Morgan. Some simple probability theory shows that the distribution of this distance is $\chi^2_4/4$. For two subjects in a large closed population, in which all pairs of subjects are IBD at each locus if we go back N generations, the distribution of the shared segment length is $\chi^2_4/(4N)$ (Mean = $1/N$, SD $\approx 0.7\times$ Mean).

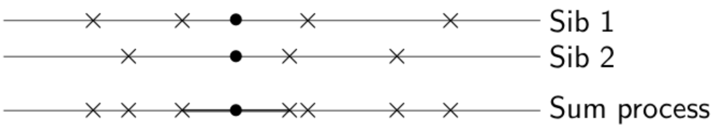


Figure 12: Another view of lD: Haplotype sharing

It seems that LD is even more variable than this. There are at least two reasons:

1. the ”coalescence time” back to common ancestor is, itself, very variable,
2. sperm typing experiments have revealed that recombination does not occur with equal probability at all points in the genome – there are ”hot” and ”cold” spots.

Recently it has been suggested that the genome falls into ”blocks”, with little haplotype diversity within blocks: Mean block size seems to be about 14 kb in Caucasians and 8 kb in Africans, but, again, this is very variable; there are blocks up to 200 kb in size. However, block boundaries may be indistinct.

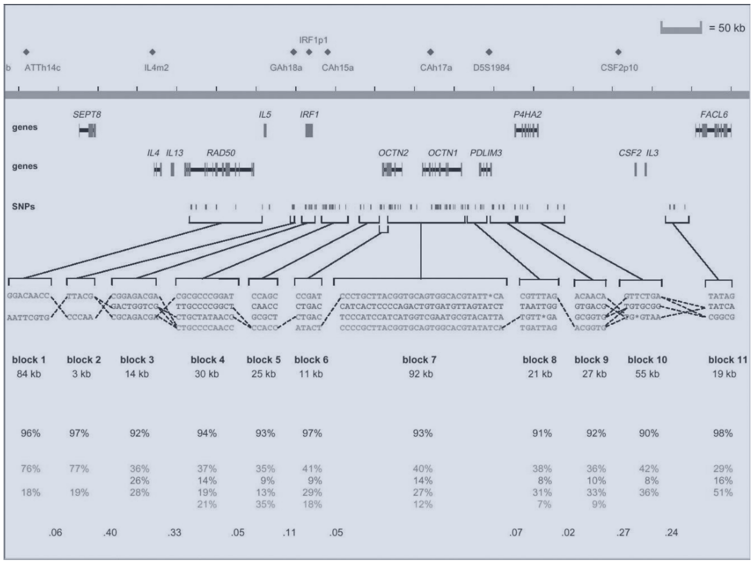


Figure 13: Hot spots (From Daly et al., Nature Genetics 2001)

A consequence of the lack of haplotype diversity in regions of strong LD is that there is considerable redundancy – most polymorphisms (and haplotypes) in such a region can be predicted from a smaller set. Johnson et a. (Nature Genetics, 2001) coined the term ”haplotype tagging” SNPs, or ”htSNPs” for such a set. Choice of tags can be integrated within an SNP discovery program:

1. Validate SNPs derived from database, and identify new SNPs by

sequencing as much of the region as feasible in a small number of subjects (32?, 48?)

2. Estimate haplotypic structure and employ search algorithms to determine the most predictive subset of polymorphisms. Generalizations of the R^2 measure are used to assess this.

Indirect association studies: choice of markers and mutli-marker analyses

The success, and cost-effectiveness of indirect association studies depends on good choice of markers. Until recently, investigators have speculatively typed a few known polymorphisms in genes of interest. Recently it has been realised that success depends on detailed knowledge of the polymorphisms occuring in a gene and of the LD structure. We resequence candidate genes in small panels of 32–96 subjects. These data are then used to select markers for large-scale studies. Hopefully the need for this laborious step will be reduced as a result of the International HapMap Project.

An international collaboration whose aim is to map LD in four hhuman populations (`http://www.hapmap.org`). The main aim is to provide the data needed to make the best choice of SNPs for indirect association studies.

Subjects studied:

- 30 parent/child trios from Nigeria
- 30 trios from USA (European ancestry)
- 45 unrelated individuals from each of China and Japan

In first stage, 600,000 SNPs were typed. In next phase 3m SNPs will be typed (complete Autumn, 2005), rising to 4.5m.

How will ve analyse the data? What determines the power to detect a causal variant? How can markers be chosen to maximize power, given constraints on resources? We start with a general model for indirect association (Figure 14).

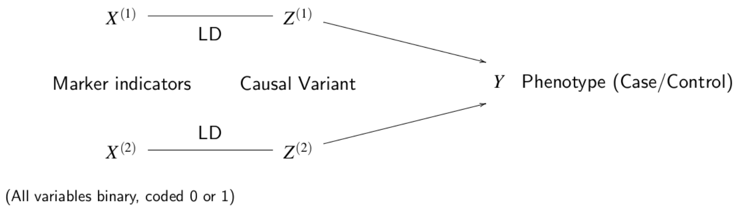


Figure 14: A model for indirect association (autosome)

$X^{(1)}, X^{(2)}$ code presence or absence of various features of the marker haplotype. Their sum, $X^{(+)}$, codes the number of times each feature occurs in the marker genotype (0, 1, or 2). We observe the indirect association between X and Y. Power depends on strength of both relationships.

The Cochran-Armitage test is basically a t-test; in large samples t^2 is approximately distributed as χ^2 on 1 df:

$$t = \left(\bar{X}^{(+)}_{\text{Cases}} - \bar{X}^{(+)}_{\text{Controls}}\right) / \sqrt{\text{Var}\left(\bar{X}^{(+)}_{\text{Cases}} - \bar{X}^{(+)}_{\text{Controls}}\right)}, \quad t^2 \sim \chi^2_1$$

A more general formula, applicable to quantitative traits, Y:

$$U = \sum_{i=1}^{N} X_i^{(+)}(Y_i - \bar{Y}), \quad U^2/\text{Var}(U) \sim \chi^2_1$$

Alternatively, we can obtain essentially the same results using regression analysis.

When we have several markers, they form a marker haplotype. X is then potentially multivariate, coding presence or absence of several features of the marker haplotype. $X^{(+)}$ then codes the number of times each feature occurs in the marker genotype (0, 1 or 2). Several authors have proposed the use of Hotelling's T^2 test – the natural geenralization

- For case-control studies, compare the vector of difference in means between the different elements of $X^{(+)}$ with its variance-covariance matrix
- In large samples, T^2 is distributed approximately as χ^2 with df = number of features coded in X

Assumptions and derivation (Figure 14):

- Generalized codominant causal model, $g(\mathbb{E}\{Y\}) = \mu + \gamma Z^{(+)}$
- Linear regression for prediction of causal variant, $\mathbb{E}\{Z\} = \kappa + \beta_1 X_1 + \beta_2 X_2 + \ldots$
- Test is derived as a score test for $\gamma = 0$, maximizing its value over the unknown regression coefficients, β (a Lagrange multiplier test)

Power is determined by the degrees of freedom and the non-centrality parameter, η, of the χ^2 test:

$$\eta = N R^2_{Z/X} R^2_{Y/Z} \quad \text{Quantitative traits}$$

$R^2_{Z/X}$ is the coefficient of determination – the % of variance of Z "explained" by X; $R^2_{Y/Z}$ is the heritability due to this causal variant. In case-control study: $\eta \approx \frac{2N_0N_1}{(N_0+N_1)} \cdot \frac{(p'-p)^2}{\bar{p}(1-\bar{p})} \cdot R^2_{Z/X}$. Finally, $p, p', \bar{p}$ are frequencies of the causal variant in controls, cases, and the whole study respectively. N_1, N_0 are numbers of cases and controls. Increasing the complexity of X increases η but also increases the degrees of freedom in the test. What is X?

In the case of single locus, testing markers one at a time by comparing allele frequencies in cases and controls. For haplotypes, we may perhaps group rarer haplotypes. In the multivariate locus, comparing the profile of allele frequency differences for several markers. Multilocus and haplotype approaches can also involve multiple testing of groups of two or three markers at a time. Optimality depends on a complex balance between multiplicity of tests, degrees of freedom and $R^2_{Z/X}$ – the ability to predict the causal variant.

To calculate genotype score $X^{(+)}$, haplotype coding requires phase resolution (Table 21): Average $X^{(+)}$ over possible phase resolutions, weighting by posterior probabilities. Locus coding does not – each marker genotype simply coded 0, 1 or 2.

Marker	Haplotype			Locus	
haplotype	X_1	X_2	X_3	X_1	X_2
1.1	0	0	0	0	0
2.1	1	0	0	1	0
1.2	0	1	0	0	1
2.2	0	0	1	1	1

Table 21: Haplotype versus multivariate locus indicators

In Figure 15, $P = 6.5 \times 10^{-8}$ (3527 cases + 3930 controls), $P = 7.3 \times 10^{-3}$ (725 families), and $P = 1.3 \times 10^{-10}$ (combined).

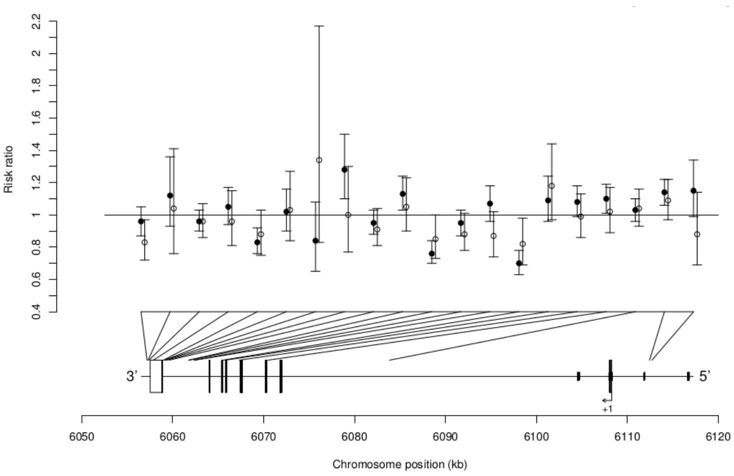


Figure 15: Locus coding: 20 tags in IL2RA (CD25)

Is simple locus scoring optimal? Tags and causal variant are close to a perfect phylogeny. This needs to be nearly true to be able to detect small effects. Test is readily extended by adding (first order) interaction terms. What about recessive variants in the codominant model? We can extend the test by adding an indicator for estimated heterozygosity at the causal locus. The non-centrality parameter/degrees-of-freedom trade-off may be better viewed locally in the phylogeny – do a series of 2- or 3-df tests, with multiple testing correction.

If \$A\$–F are tags, a causal variant, Z, would simply add a new clique – it is likely to be tagged by a relatively small subset of markers (Figure 16).

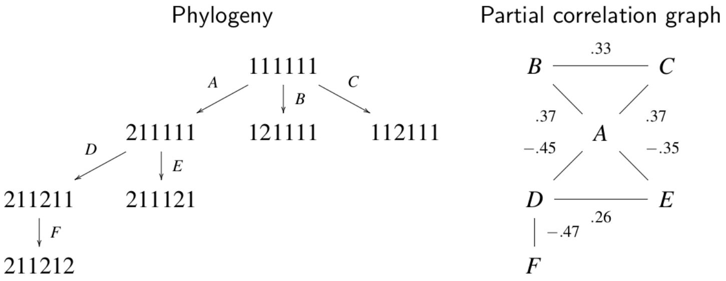


Figure 16: Perfect phylogenies and clique junction trees

As an example, consider the CTL4A central region (Figure 17), where tag SNPs are shown in blue.

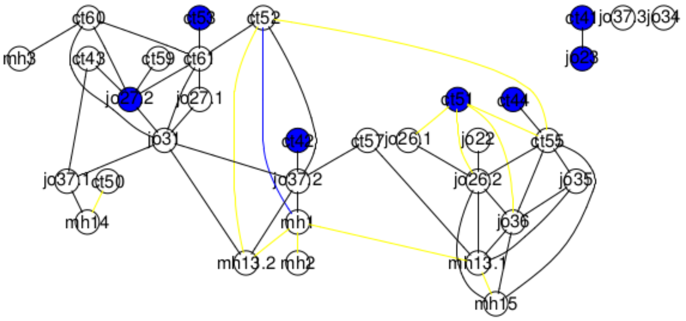


Figure 17: Example of the CTLA4 region

When choosing tags, the aim is to minimize redundancy while capturing as much information as possible. There are three main methods:

1. Single marker tagging: each known SNP has a single tag with $r^2 \geq 0.8$

2. Multi-marker tagging: multiple regression of each SNP on tag locus scores has $R \geq 0.8$

3. Haplotype tagging: multiple regression of each SNP on haplotype indicators has $R^2 \geq 0.8$

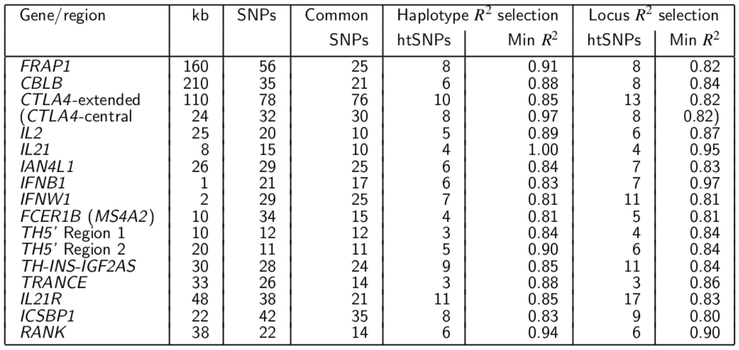
Note that multiple regression R^2 is biased upward whent he sample size is small – multi-marker tagging won't work quite as well in a future study as we would (naively) think.

Single marker tag selection using cluster analysis is computationally the most tractable method yet proposed. We carry out a cluster analysis of all SNPs in order to assign them to ”bins” such that all SNPs in a bin have hogh r^2 with each other. We select one SNP from each ”bin” as a tag. This method has been implemented efficiently enough to allow its use on a whole-genome scale – it is currebtly in use in the design of whole-genome SNP ”chips”. It is implemented in Haploview – the software for interacting with the data on the HapMap website.

For multi-marker taggong, we need to find the best subset of SNPs such that the remainder can be predcited by multiple regression with $R^2 \geq 0.8$. Regresion can be on

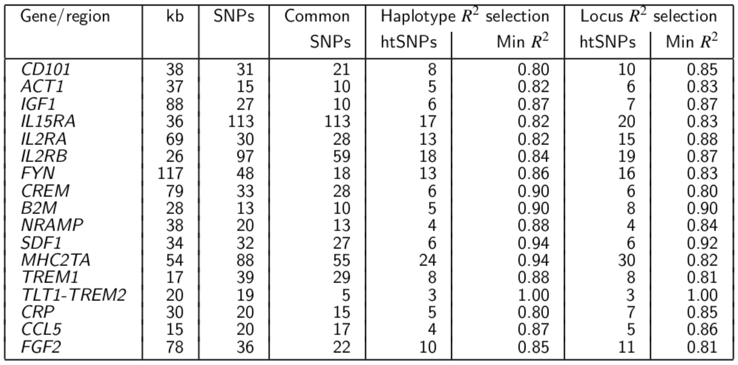
- Locus scores, providing a resonably efficient and simple analysis
- Haplotype scores, providing the minimum set of tags

Best subset searches are computationally intensive – we need to restrict the choice to a shortlist, for example using single marker tagging, perhaps with $r^2 \geq 0.5$, or simple step-up and setp-down regression strategies.



Gene/region	kb	SNPs	Common	Haplotype R^2 selection		Locus R^2 selection	
			SNPs	htSNPs	Min R^2	htSNPs	Min R^2
FRAP1	160	56	25	8	0.91	8	0.82
CBLB	210	35	21	6	0.88	8	0.84
CTLA4-extended	110	78	76	10	0.85	13	0.82
(*CTLA4*-central	24	32	30	8	0.97	8	0.82)
IL2	25	20	10	5	0.89	6	0.87
IL21	8	15	10	4	1.00	4	0.95
IAN4L1	26	29	25	6	0.84	7	0.83
IFNB1	1	21	17	6	0.83	7	0.97
IFNW1	2	29	25	7	0.81	11	0.81
FCER1B (*MS4A2*)	10	34	15	4	0.81	5	0.81
TH5′ Region 1	10	12	12	3	0.84	4	0.84
TH5′ Region 2	20	11	11	5	0.90	6	0.84
TH-INS-IGF2AS	30	28	24	9	0.85	11	0.84
TRANCE	33	26	14	3	0.88	3	0.86
IL21R	48	38	21	11	0.85	17	0.83
ICSBP1	22	42	35	8	0.83	9	0.80
RANK	38	22	14	6	0.94	6	0.90

Figure 18: Some genes/regions



Gene/region	kb	SNPs	Common	Haplotype R^2 selection		Locus R^2 selection	
			SNPs	htSNPs	Min R^2	htSNPs	Min R^2
CD101	38	31	21	8	0.80	10	0.85
ACT1	37	15	10	5	0.82	6	0.83
IGF1	88	27	10	6	0.87	7	0.87
IL15RA	36	113	113	17	0.82	20	0.83
IL2RA	69	30	28	13	0.82	15	0.88
IL2RB	26	97	59	18	0.84	19	0.87
FYN	117	48	18	13	0.86	16	0.83
CREM	79	33	28	6	0.90	6	0.80
B2M	28	13	10	5	0.90	8	0.90
NRAMP	38	20	13	4	0.88	4	0.84
SDF1	34	32	27	6	0.94	6	0.92
MHC2TA	54	88	55	24	0.94	30	0.82
TREM1	17	39	29	8	0.88	8	0.81
TLT1-TREM2	20	19	5	3	1.00	3	1.00
CRP	30	20	15	5	0.80	7	0.85
CCL5	15	20	17	4	0.87	5	0.86
FGF2	78	36	22	10	0.85	11	0.81

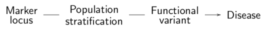
Figure 19: Some genes/regions (con’t)

Do we know enough? The strategy outlined is guaranteed to work – if we have identified all possible causal variants and typed them in our initial small sample. If we have imperfect knowledge, the choice is potentially flawed – and even the final version of HapMap falls short of resequencing.

Population-based studies of disease/gene associations

Three reasons for a genetic association:

- The locus is a functional variant, that is the association is causal
- The locus is in linkage disiquilibrium with a functional variant
- The association is due to confounding by population stratification



In respect to association studies, genetic epidemiology differs little from the classical epidemiology of behavioural and environmental risk factors. Much of this chapter revises familiar epidemiological ideas, but there are some aspects of analysis which are special to genetics. We will start by discussing studies of disease risk, and deal with quantitative traits at the end of the chapter.

Disease frequency can be measured in terms of prevalence and incidence, the latter being the preferred measure for assessing cause of disease. Both are probability measures. Prevalence is defined as the probability, $\pi(t)$, that an individual has disease at some specified point in time. Incidence is defined either in terms of the probability of developing the disease over a fixed period, or of the probability rate:

$$\text{Incidence, } \lambda(t) = \lim_{\delta \to 0} \frac{\Pr(\text{Onset of disease between } t \text{ and } t+\delta)}{\delta}$$

In epidemiology, association between disease and aetiological factors are usually expressed in terms of relative risk measures. In the simplest case, this is some measure of disease risk in exposed subjects divided by the same measure of risk in unexposed subjects. In genetic epidemiology, relative risks may be defined for genotypes, alleles, or haplotypes.

For a diallelic locus with alleles A, a, there are three genotypes: A/A, A/a, a/a. We will usually take one of these, aa say, as reference and express genotype relative risk as

$$\begin{aligned} \text{GRR}_{A/A} &= \frac{\text{Risk for A/A genotype}}{\text{Risk for a/a genotype}} = \theta_{A/A} \\ \text{GRR}_{A/a} &= \frac{\text{Risk for A/a genotype}}{\text{Risk for a/a genotype}} = \theta_{A/a} \end{aligned}$$

Allelic relative risks, Φ_A, Φ_a are defined by the multiplicative model $\theta_{i/j} = \theta_i \theta_j$ where, again one allele is taken as reference. In the diallelic case, taking a as reference so that $\Phi_a = 1$:

$$\theta_{A/A} = (\Phi_A)^2, \qquad \theta_{A/a} = \Phi_A$$

If the relative frequency of alleles i, j are f_i, f_j, the relative frequency of genotype i/j under H-W equilibrium is $2 f_i f_j$ if $i \neq j$, and $(f_i)^2$ if

$i = j$. HWE assumption implies that each subject's two chromosomes are sampled independently from the population. A sample of N independent subjects can be treated as a sample of $2N$ independent chromosomes. When there is HWE in the population and the multiplicative model holds then HWE also holds in cases of disease, but with modified allele frequencies:

$$f_i^\star = (\Phi_i f_i) / \sum_i (\Phi_i f_i)$$

Prevalence studies include cross-sectional studies and case-control studies (prevalent cases versus healthy controls). Incidence studies include prospective (cohort) studies and case-control studies (incident cases versus healthy population controls). Case-control designs are much more efficient. They have some disadvantages in the study of environmental and behavioural causes of disease, but these are not relevant to genetic associations.

Naturally, we would treat the observed genotype as the random variable, and compare its distribution between cases and controls. This becomes computationally complicated when there are many alleles or several loci. But the same answer is obtained by treating disease status (case vs. control) as a random outcome, predicted by genotype. The general method for analysis of such models is logistic regression. But simpler methods are available for simple cases.

Assuming the multiplicative model, HWE in the population, and a rare disease, we can simply count alelles in cases and controls. For a diallelic locus, see Table 22.

Table 22: Allele counting – the 2 × 2 table

Allele	Cases	Controls
A	D_A	H_A
a	D_a	H_a

We can test for association using the conventional χ^2 test (1-df). The allelic relative risk (A vs. a) is estimated by the odds ratio:

$$\frac{D_A/D_a}{H_A/H_a} = \frac{D_A/H_A}{D_a/H_a} = \frac{D_A/H_a}{D_a/H_A}$$

For rare diseases, the odds ratio estimates the allelic relative risk. In studies of prevalent cases, the relative risk refers to ratios of prevalence. In studies of incident cases, the relative risk measure is the ratio of incidence rates in the base population – the incidence rate ratio λ_A/λ_a. This does not depend on the disease being rare, but does assume

- ″incidence density sampling″ (we draw control(s) whenever a case arises)

- proportion of population exposed is stable over the duration of the study

Below are some general considerations on test statistics for $\theta = 1$.

- Likelihood ratio test: (twice) the difference between the log likelihoods at $\theta = 1$ and at $\theta = \hat{\theta}$.
- Score test: squared slope of the graph of log likelihood against $\log\theta$ at $\theta = 1$ divided by its variance, taht is U^2/V
- Wald test: squared difference between $\hat{\theta}$ (or $\log\hat{\theta}$) and the null value, divided by its variance (margin figure)

In large samples, all three appraoches lead to 1-df χ^2 tests. The conventional χ^2 tests for the 2×2 table is the score test.

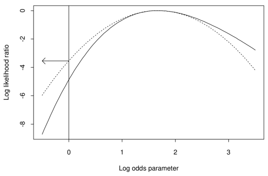
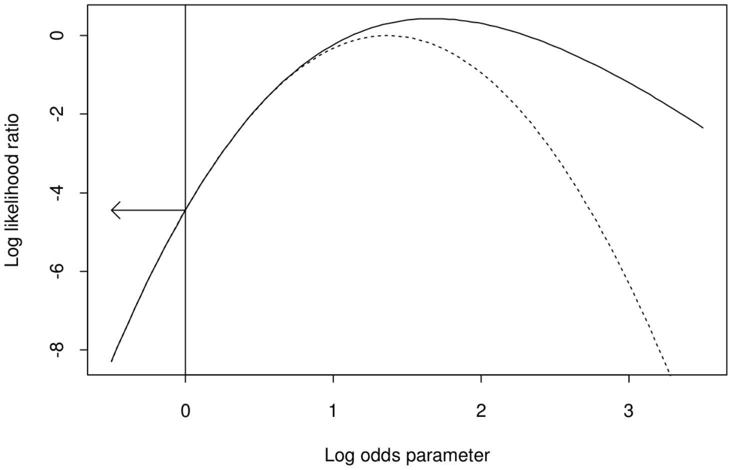


Figure 20: Score test



Allele	Cases	Controls	Total
A	D_A	H_A	N_A
a	D_a	H_a	N_a
Total	D	H	N

Table 23: Allele counting – the 2×2 table

Score: $U = D_A - D\frac{N_A}{N}$ and its variance: $V = N_A N_a DH/N^3$. Notice that U^2/V is the conventional $\sum(O - E)^2/E$ statistic.

LR and Wald tests by logistic regression: Fit a logistic regression model for case or control origin of allele against an indicator, x, taking value 1 for allele A and 0 for allele a:

Allele	x	Probability	Odds	Log Odds	
a	0	π_a	$\pi_a/(1-\pi_a)$	$\log\{\pi_a/(1-\pi_a)\}$	α
A	1	π_A	$\pi_A/(1-\pi_A)$	$\log\{\pi_A/(1-\pi_A)\}$	$\alpha + \beta$

Table 24: LR and Wald tests

The regression coefficient β is the log of the odds ratio:

$$\log(\mathrm{OR}) = \log \frac{\pi_A/(1-\pi_A)}{\pi_a/(1-\pi_a)} = \log \frac{\pi_A}{1-\pi_A} - \log \frac{\pi_a}{1-\pi_a} = \beta$$

We can estimate β and carry out LR test of $\beta = 0$ (equivalent to $\Phi = 1$).

Genotype	Cases	Controls
Leu/Leu	89	56
Leu/Pro	369	250
Pro/Pro	342	266
Total	800	572

Table 25: Example of the Pro871Leu SNP in the BrCa1 gene

The assumptions of (a) HWE in the populations (and hence in controls), and (b) the multiplicative model, means that we can treat chromosomes as independent and analyse the 2 × 2 table (Table 26). The estimated RR for the Leu allele is $\hat{\Phi}_{Leu} = (547 \times 782)/(1053 \times 362) = 1.122$ (the multiplicative model implies a RR for Leu/Leu of $(1.122)^2 = 1.259$). Likelihood ratio and score tests are 1.954 and 1.949 respectively.

Allele	Cases	Controls
Leu	547	362
Pro	1053	782
Total	1600	1144

Table 26: Example of the Pro871Leu SNP in the BrCa1 gene (con't)

Regarding subject counting, we distinguish two more 2 × 2 analyses. Testing the null hypothesis against dominant or recessive alternatives also leads to tests for 2 × 2 tables – but counting subjects:

Genotype	Cases	Controls
Leu/*	458	306
Pro/Pro	342	266
Total	800	572

Table 27: Subject counting: dominant case

While there may be a case for such tests for a functional variant, indirect associations with markers tend to give alternatives in which heterozygots have intermediate risk.

A general analysis of genotype relative risks involves the $3 \times 2 table$. We need two odds ratios to emasure association. Again taking a/a as reference category, compare A/A vs. a/a and A/a vs. a/a:

We have $\hat{\theta}_{A/A} = \frac{D_{A/A}H_{a/a}}{D_{a/a}H_{A/A}}$, and $\hat{\theta}_{A/a} = \frac{D_{A/a}H_{a/a}}{D_{a/a}H_{A/a}}$. Again, we may test for association using likelihood ratio, score, or Wald tests – difefrent large sample approximations. These are 2-df χ^2 tests (the conventional test is the score test).

Another approach consists in using 2-df tests from logistic regression. We introduce two indicator variabels, x_1 and x_2, indicating genotypes A/a and A/A respectively, and regress disease status of subject against both of these (Table 30).

Genotype	Cases	Controls
Leu/Leu	89	56
Pro/*	711	516
Total	800	572

Table 28: Subject counting: recessive case

Genotype	Cases	Controls
A/A	$D_{A/A}$	$H_{A/A}$
A/a	$D_{A/a}$	$H_{A/a}$
a/a	$D_{a/a}$	$H_{a/a}$

Table 29: Two-df tests in the 3 × 2 table

We have $\log(\text{OR, A/A vs. a/a}) = \beta_2$ and $\log(\text{OR, A/a vs. a/a}) = \beta_1$. Again we can use a logistic regression program to estimate β_1, β_2 and to test the hypothesis of no association ($\beta_2 = \beta_1 = 0$).

A further 1-df test is provided by the multiplicative model – the most convenient model in which the risk for A/a is intermediate between risks for a/a and A/A. Fit using logistic regression using a single indicator, x, taking values 0, 1 or 2 (Table 31).

This time, $\log(\text{OR, A/A vs. a/a}) = 2\beta$ (odds ratio = Φ^2), and $\log(\text{OR, A/a vs. a/a}) = \beta$ (odds ratio = Φ).

The score test for $\beta = 0$ is the Cochran-Armitage test for trend of proportions.

- The score, U, is exactly the same as that for allele counting
- We use a different estimate of its variance, V, which does not assume HWE
- This test is very nearly the same as carrying out 2-sample t-test to compare the means of x between cases and controls.

These tests are preferable to allele counting, since they avoid the need to assume HW equilibrium.

From Table 28, we see that $\hat{\theta}_{Leu/Leu} = \frac{89\times266}{342\times56} = 1.236$ and $\hat{\theta}_{Pro/Leu} = \frac{369\times266}{342\times250} = 1.148$. The relative risks agree quite closely with those predicted by the multiplicative model fitted to the ”chromosomes” Table 26 (1.259 and 1.112). A summary of some of the tests we have discussed is given in Table 32.

The multiplicative model is equivalent to a model, on the log odds scale, in which there is an additive effect of alleles, but no dominance effect. We can parametrize the model in just this way (Table 33).

Here, $\beta_D = \log \frac{\text{Odds}_{A/a}}{\sqrt{\text{Odds}_{a/a}\text{Odds}_{A/A}}}$, $\beta_D = 0$ corresponds to the multiplicative model.

Genotype	x_1	x_2	Probability	Odds	Log Odds
a/a	0	0	$\pi_{a/a}$	$\pi_{a/a}/(1-\pi_{a/a})$	α
A/a	1	0	$\pi_{A/a}$	$\pi_{A/a}/(1-\pi_{A/a})$	$\alpha+\beta_1$
A/A	0	1	$\pi_{A/A}$	$\pi_{A/A}/(1-\pi_{A/A})$	$\alpha+\beta_2$

Table 30: Two-df tests using logistic regression

Genotype	x	Probability	Odds	Log Odds
a/a	0	$\pi_{a/a}$	$\pi_{a/a}/(1-\pi_{a/a})$	α
A/a	1	$\pi_{A/a}$	$\pi_{A/a}/(1-\pi_{A/a})$	$\alpha+\beta$
A/A	2	$\pi_{A/A}$	$\pi_{A/A}/(1-\pi_{A/A})$	$\alpha+2\beta$

Table 31: One-df tests using logistic regression

Type of test	2-df test	1-df test (subj)	1-df test (chrm)
LLR	2.056	1.991	1.954
Score	2.055	1.984	1.949

Table 32: Summary of all tests

The number of genotypes increases with the square of the number of alleles. For markers with $3, 4, 5, \ldots, K$ alleles there are $6, 10, 15, \ldots, K(K+1)/2$ genotypes. Unrestricted tests based on genotype relative risks lack power – it is usually better to consider the multiplicative alterantive model:

$$\theta_{i/j} = \Phi_i \Phi_j$$

Logistic regression: generate $K-1$ indicator variables – one for each (non reference) allele. Each one counts the number of occurences of the allele. This test for association has $K-1$ df.

For example, for a 3-allele locus, indicator variables are shown in Table 34.

A likelihood ratio test (on 2 df) is carried out by comparing the log likelihoods for regression models with and without inclusion of x_1, x_2. The score test is equivalent to Hotelling's T^2 (a multivariate generalization of the t-test) – compare the means of x_1 and x_2 for cases and controls.

Current experience of SNP markers suggests that linkage disequilibrium (LD) exhibits a ″block″ structure – groups of adjacent markers in close LD, separated by ″hot spots″ at which recombination has destroyed LD. LD blocks seem to be, on average, rather smaller than genes. Demonstration that a disease susceptibility locus lies within an LD block can be acrried out by typing a set of ″haplotype tagging″ SNPs (htSNPs), chosen to capture haplotype diversity of blocks. We can look for association in three ways:

- Analysis of each htSNP separately, with correction for multiple testing
- Comparison of htSNP haplotype frequencies in cases and controls
- Compare allele frequencies between cases and controls for all loci

Genotype	x_1	x_D	Probability	Odds	Log Odds
a/a	0	0	$\pi_{a/a}$	$\pi_{a/a}/(1-\pi_{a/a})$	α
A/a	1	1	$\pi_{A/a}$	$\pi_{A/a}/(1-\pi_{A/a})$	$\alpha+\beta_A+\beta_D$
A/A	2	0	$\pi_{A/A}$	$\pi_{A/A}/(1-\pi_{A/A})$	$\alpha+2\beta_A$

Table 33: Alterantive parametrization of the 2-df model

Genotype	x_1	x_2	log Odds	Odds ratio
1/1	0	0	α	1 (reference)
2/1	1	0	$\alpha + \beta_1$	Φ_1
3/1	0	1	$\alpha + \beta_2$	Φ_2
2/2	2	0	$\alpha + 2\beta_1$	$(\Phi_1)^2$
3/2	1	1	$\alpha + \beta_1 + \beta_2$	$\Phi_1\Phi_2$
3/3	0	2	$\alpha + 2\beta_2$	$(\Phi_2)^2$

Table 34: Example of 3-allele locus

We can calculate p-values for any of the test statistics considered thus far using a Monte Carlo appraoch:

- Randomly permute the order of the vector which assigns case/control status to subjects
- Recalculate the test statistic for each random permutation
- In what proportion of random permutations is the observed value of the test statistic exceeded?

This is easily adapted to correct for multiple testing:

- For each permutation, calculate tests for all markers. Note the largest
- In what proportion of random permutations does this exceed the largest observed value?

With K diallelic markers, there are potentially 2^K different haplotypes. In studies of unrelated individuals, haplotypes may not be assigned to individuals unless they are homozygous at all loci or all loci bar one. Neverthelss, test statistics constructed for the phase-known case may be adapted to the phase-unceratin case (Schaid et al., AJHG, 70: 425–34). But what test statistic?

- Even for multiplicative alternatives, the simple test has up to $2^K - 1$ df.
- Geary-Moran statistics may be constructed based on ”distances” between haplotypes (Clayton and Jones, AJHG, 65:1161–9). But how do we measure distance?

Consider frequencies of 2-locus haplotypes in complete LD ($D' = 1$), as in Table 35.

Locus		
A	B/1	B/2
1	$f_{1,1}$	$f_{1,2}$
2	$f_{2,1}$	$f_{2,2} =$ Small

Table 35: Strong LD

The test for association has 3-df, but one haplotype is rare and may carry little information. Comparing allele frequencies at the two loci in cases and controls gives a test with 2-df. If LD is strong, this carries nearly as much information but uses one less degree of freedom.

In the haplotype test, we construct three indicator variables counting, say, occurences of haplotypes 1.2, 2.1, 2.2 in each subject, and then compare the mean scores for cases and controls. In the 2-locus test, we construct two indicator variables counting occurences of allele 2 at each locus, and then compare the mean scores for cases and controls. When LD is strong, the 2-locus test is more powerful, and does not require phase to be known. The 2-locus test can also be done in logistic regression, using two indicators counting occurence of allele 2 at each locus. This approach generalises to > 2 loci.

Control for confounding: Association may be false, due to

- different risks between different subpopulations, accompanied by
- different allele frequencies between these subpopulations

We control for confounding by comparing cases and controls within strata – is there a significant difference within strata? Effect modification involves gene-gene interaction, and gene-environment interaction. We again make comparisons within strata – does the size of the effect differ between strata?

In the case of a diallelic marker, we have either a 3×2 table of subject counts, or a 2×2 table of chromosome counts for each population stratum (Figure 21).

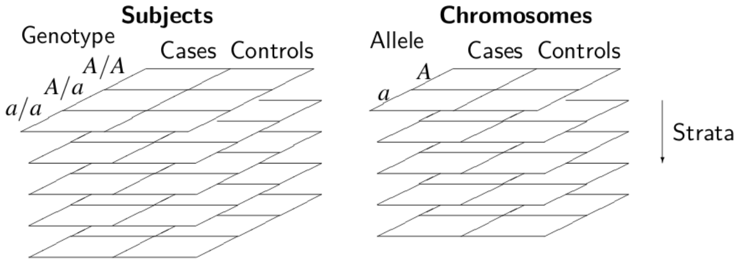


Figure 21: Stratification to control for confounding

But there may be little data in each stratum. Tests which simply add chi-squared values across strata have many df and lack power.

A more parcimonious appraoch consists in using common odds ratios. Consider, as an alternative to the null hypothesis, the model in which the association parameter(s) are constant across strata. Here, the association parameters are either the genotypic odds ratios, $\theta_{A/A}$, $\theta_{A/a}$, or the allelic odds ratios, θ_A. The standard 1 df ”score” tests have been generalized to the stratified case:

- The test for the 2×2 table generalizes to the Mantel-Haenszel test

- The Cochran Armitage test (which we used for the 1 df test in the 3×2 table) generalizes to the Mantel extension test
- The general idea is to take $U = U_1 + U_2 + \ldots$ and $V = V_1 + V_2 + \ldots$.

Assuming constant odds ratios across strata, log odds that a subject in stratum s is a case rather than a control is given by:

$$\log \text{Odds} = \alpha + \beta x + \text{Stratum effect} \quad \text{(1 df model)}$$
$$\log \text{Odds} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \text{Stratum effect} \quad \text{(2 df model)}$$

Table 36: Logistic regression

	1-df model	2-df model	
Genotype	x	x_1	x_2
a/a	0	0	0
A/a	1	1	0
A/A	2	0	1

We simply include stratum in the logistic regression as a series of indicator variables.

We may wish to test the hypothesis that the strength of association as measured, for example, by an odds ratio, differs between strata. This is conveniently carried out using logistic regression. Consider two strata, coded $s = 0$ and $s = 1$. The logistic regression model

$$\log \text{Odds} = \alpha + \beta x + \gamma s + \delta(s, x)$$

becomes

$$\log \text{Odds} = \alpha + \beta x \quad (s = 0)$$
$$\log \text{Odds} = (\alpha + \gamma) + (\beta + \delta)x \quad (s = 1)$$

Testing for $\delta = 0$ is a test for interaction or (better) effect modification.

Regarding inference abount functional variants, consider loci A and B that are both strongly associated with disease and are in LD with one another. If A is the functional variant, stratification by genotype at A will destroy the association with B. However, stratification by B will not destroy the association with A. If neither is functional, both reflecting a third functional variant,

- stratification by each will not entirely destroy association with the other
- there may be ”cis” interaction – an additional effect of the $A.B$ haplotype

If LD between A and B is too strong, there will be little power to detect these patterns.

Little power is gained by having large number of controls. If costs per subject dominate, the most effeicient design has approximately equal numbers of cases and controls. If there is strong confounding due to stratification, we can lose power. We may have $\ll 1$ control per case in some strata, and $\gg 1$ controls per case in others. This can be avoided by matching at the design stage. But another (arguably more) important reason for matched designs is to provide a sampling from controls. Note that, in general, matching at the design stage does not avoid the need to stratify during analysis.

Each single case has its own set of controls – each case defines a stratum. Conventional logistic regression fails because we would have to introduce an extra parameter for each stratum. Conditional logistic regression fits the same model but avoids the need for these extra parameters by using a likelihood based on an ingenious conditional probability argument. The score test based on this argument has the same U for each stratum but V is multiplied by $N/(N-1)$.

For some years, matching in case-control studies was considered to be entirely beneficial, but it later emerged that matching for a variable which, while not a confounder, is related to the factor of interest, causes loss of power.

Example: use of sibling controls (”sib TDT” study)

- If there is minimal linkage in the region, the probability that two siblings share 0, 1, or 2 genes IBD are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$ respectively.
- On average, half of the alleles of a case and a sibling control would be expected to be identical; only the remaining half would contribute information about association.
- This design will need twice the sample size required for a study with unrelated controls.

Admixture and/or stratification may confound genetic association when risk differs between subpopulations.

- Any locus whose allele frequencies will vary between subpopulations will have non-zero association parameters (log odds ratios)
- As a result, test statistics will be overdispersed and there will be false positive findings.

Given data concerning a large number of loci, two ways of tackling this have been proposed:

1. Estimate the amount of overdispersion empirically (Devlin and Roeder, Biometrics, 2000)

2. Given ancestry-informative markers, latent stratification/admixture can be estimated – compute a posterior probability that each chromosome segment derives from each ancestral population (Protchard et al., AJHG 67:170, 2000; Hoggart et al., AJHG bf 72:1492, 2003; Patterson et al., AJHG 74:979, 2004).

   A made-up example: Devlin and Roeder's "genome-wide control"

- We carry out 1 df tests for 200 SNP markers, sufficiently well spaced that we can safely assume linkage equilibrium within populations.
- We find 20 tests are significant at $p < 0.05$ and 6 at $p < 0.01$.
- Overdispersion of chi-squared tests is caused by unobserved stratification and random differences in allele frequencies between strata.
- Devlin and Roeder suggest that the true distribution of test statistics can be approximated by a simple multiple of chi-squared.

  Figure 22 show idealised results.

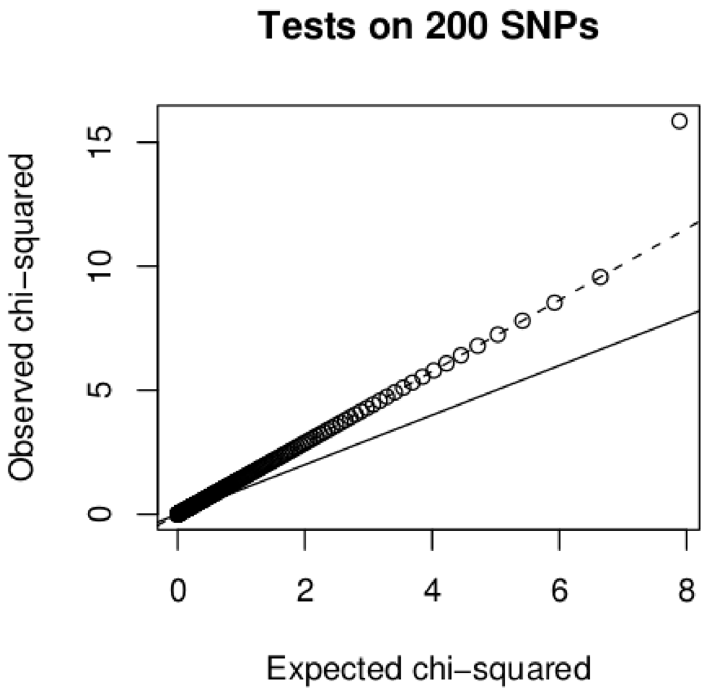


Figure 22: Stratification to control for confounding

Unbroken line represents the line of equality – tests distributed as chi-squared on 1-df. Broken line has slope 1.44, representing overdispersion. Using the overdispersed distribution, there are 10 markers with $p < 0.05$ and 2 with $p < 0.01$. Devlin and Roeder discuss estimation of overdispersion parameter (here 1.44). Problems: (1) in real life, is the line really straight? (2) inefficiency.

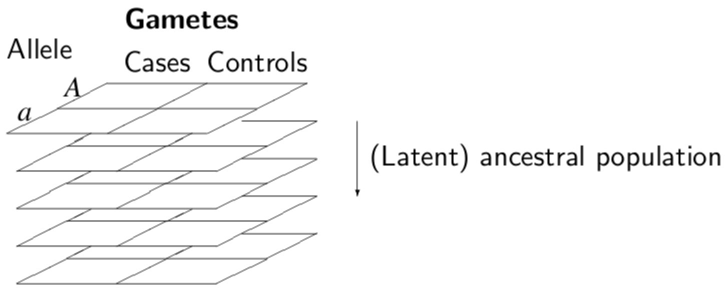


Figure 23: Latent classes

We cannot assign gametes exactly to ancestral populations. Instead we calculate a posterior probability, given ancestry-informative markers, and divide each data point between strata. Null distribution of test statistic can be computed by random permutation of case/control indicators, or by a large-sample approximation to it.

A simple model assumes that trait, Y, is normally distributed with variance σ^2 around a mean that varies with genotype G – μ_G say. Following Fisher (1918), we can decompose the variation of trait with genotype into additive and dominance components. For $G = i/j$,

$$\mu_{i/j} = \mu + \alpha_i + \alpha_j + \delta_{ij}$$

α_i is the additive effect of allele i, while $\{\delta_{ij}\}$ are dominance effects. Additive and dominance terms correspond to ”main effects” of alleles and ”interaction” between alleles in the analysis of variance.

Population-based studies:

- Cross-sectional study: the simplest type of study is of a sample of subjects from the general population, we relate trait value to genotype at one or more loci.
- Two-stage study: an efficient alternative (only sometimes feasible) is to select subjects with extreme trait values: (1) more efficient use of expensive genotyping, (2) analysis is more difficult: ignoring selection results in biased estimates of effects.

Regression analysis of unselected study: Regress trait value on genotype. For a diallelic marker, the additive model gives a 1-df test and the full model gives a 2-df test:

$$Y = \alpha + \beta_A x_A + \beta_D x_D + \text{Residual}$$

Indicator	1/1	1/2	2/2
additive, x_A	0	1	2
dominant, x_D	1	0	1

Table 37: Regression analysis of unselected study

For multiallelic marker, introduce an x_A indicator for every allele (except the reference). There are many indicators for dominance – usually ompitted.

Main references:

1. Breslow and Day (1980) Statistical Methods in Cancer Epidemiology. Vol 1: The Analysis of Case–Control Studies, IARC Publications, Lyon.

2. Clayton and Hills (1993) Statistical Models in Epidemiology, Oxford University Press, Oxford.

3. Clayton (2001) Population Association. Handbook of Statistical Genetics, ed. Balding, Bishop and Cannings, Wiley, Chichester.

4. Cordell and Clayton (2005) Genetic association studies. The Lancet

Gene-environment interaction in complex disease

Questions for epidemiology:

- What is ”gene-environment interaction”?
- Can we test for its existence in a biologically meaningful way using epidemiological methods?
- Does knowledge of environmental risk factors aid the search for genes?
- Does knowledge of genes aid the search for (causal) environmental risk factors?
- Is gene-environment interaction relevant to designing public health interventions?

What is G-E interaction? The word ”interaction” can mean different things to different people. Many of the more extravagant claims for its importance fail to define it.

> The word ”interaction” means different things to statisticians and biologists. Some epidemiologists prefer the word ”synergism”, but. . . while most would agree that epidemiological synergism among exposures exist, definint it is problematic. — Weinberg CR. On Synergism, Encyclopedia of Biostatistics, Wiley

> The notion of interaction and indeed the very word itself are widely used in scientific discussion. This is largely because of the realtion between interaction and causal connexion. Interaction in the statistical sense has, however, a more specialized meaning related, although often in only a rather vague way, to the more general notion. — Cox DR. Int. Statist. Rev:52, 1984

> The term ”interaction” is widely used in statistics. Although it has a concrete arithmetical meaning in all statistical models. . . it often appears than this term is used when something unusual, something non-specific, is described without an attempt to derive a deeper understanding of the phenomenon. — Wahrendorf J. Perspectives in Medical Statistics, 1981

> A decade ago the concept of interaction among causes of disease was at the center of a lively debate. Since that time, controversy over the nature of interaction has largely subsided, although there seems never to have been an adequate resolution of the conceptual and pragrmatic issues that had been raised. Thompson WD. Clinical Epidemiology:44, 1991

What do biologists mean? What is interesting to biologists – it being self-evident that genes and environment interact? A specific mode of interaction is when a gene and an environmental factor act on the same pathway. The promise of being able to study G-E interaction in epidemiological studies seems to offer the potential for epidemiology to elucidate mechanism.

What do statisticians mean? Statisticians know what they mean by ”interaction”, but does anyone else?

> I have also been pondering over these matters and found that there is only one basic definition of non-interaction between x and y:
>
> $$\text{expected response} = f\left[a.h_1(x) + b.h_2(y)\right],$$
>
> a and b being parameters and f, h_1, h_2 arbitrary functions. If no such functions exist in a given situation it is clear that x and y inextricably interact. — Hilden J. Disucssion of paper by Wahrendorf, 1981

That is, interaction is absent when the effects of the two factors are additive with respect to some quantitative measure of response: Response = Gene effect + Environmental effect. Statistical interaction describes lack of fit of this simple statistical model for joint action.

Epistasis: Similar issues are involved in the study of gene-gene interaction.

> I am not clear as to the exact sense in which he uses the term ”epistacy” but as it has already been used biologically in a sense which is evidently not the one in which it is used here, I think that it should be made quite clear how the new sense proposed differs from that already in use. — Punnett RC. Review of Fisher’s 1918 paper to the Royal Society

Table 38: Diabetes and two loci in two strains of mice

				IDD10		
IDD3	R1.NN	R1.NB	R1.BB	R2.NN	R2.NB	R2.BB
NN	63/81	48/78	50/152	52/68	64/101	95/193
NB	58/73	48/118	9/61	54/93	34/95	–
BB	23/81	6/57	2/159	19/85	5/92	–

There is statistical interaction on some scales of measuremnt of risk, but not on others. Is there epistasis? If one locus blocked or reversed the effects of the other, there would be clear implications for mechanism (”qualitative interaction”).

Relevance to biology:

> Unfortunately, choice among theories of pathogenisis is enhanced hardly at all by the epidemiological assessment of interaction... What few causal systems can be rejected on the basis of observed results would provide decidedly limited etiological insight. — Thompson WD, 1991

Measuremnt error introduces a further difficulty in interpreting statistical interaction or the lack of it. If two factors are measured with error, the precise form of their joint action is distorded. When studying disease risk, the distorsion is towards the multiplicative model: Risk × Effect of A × Effect of B. Is this why multiplicative models tend to fit rather well in practice?

> We focus here on statistical interaction which does not necessarily imply interaction on the biological or mechanistic level. — Witte JS. On Gene-Environment interaction, Encyclopedia of Biostatistics, Wiley

There seems to be little purpose in designing studies whose purpose is to detect statistical interaction... Yet, if we are to believe the power calculations often presented, such studies are advocated:

- Biobank UK: the need for a chort study of 500,000 middle-aged subjects was initially justified on the basis of such calculations
- Drug-gene interaction studies

Elston et al. (Statistics in Medicine:18, 1999) describe a statistical test, and sample size calculations, for drug-gene interaction studies. ”No interaction” means constant additive effect of genotype at every dose of drug upon the probability of the desired response. Is this a realistic model for what a biologist would mean by ”no interaction”?

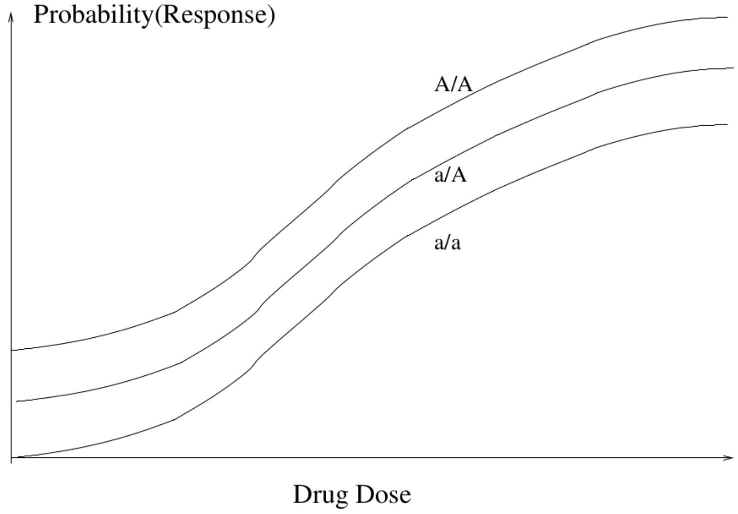


Figure 24: A realistic model?

The statistician’s caveat: Elston et al. remark that interaction could be defined in terms of additivity on a different scale of measurement, and indicate how the test and power calculations would be modified. But they fail to comment on the fact that the very existence of this possibility undermines the purpose of the exercise.

Is E relevant to finding G’s? In the presence of strong (statistical) G-E interaction, there can be a gain in power to detect genetic linkage/association...

- but, for discrete outcomes, gains in power are modest except when effects (and their interaction) are strong
- and, in the absence of strong a priori knowledge, gains mau be dissipated in multiple testing

- in the presence of exposure measurement error, effects and their interaction are not strong

This is because nature tends to deliver a balanced design, by ”Mendelian randomization” (Yougman et al., Circulation:102, 2000). It will usually ensure independence of genotype and exogenous exposures. Unless direction of effects is modified, marginal effects of gene and environment are weaker, but not destroyed. For quantitative responses and strong effects, there can be power gain by reduction of residual error, but this is not so pronounced for discrete outcomes.

Is G relevant to finding E’s? As before, power gain is modest or nonexistent, except when strong a priori knowledge guides analysis or there is reversal of direction of effect. In the former case, the answer is already known and studies are essentially confirmatory. The latter case seems implausible in most cases.

Guiding the search for environment determinants:

- Modern epidemiology is bedeviled by a multiplicity of possible environmental causes, often difficult to measure (Taube H, Science 269, 1995)
- Better understanding of genetic effects may provide focus

Confirming causality:

- Stronger effects in genetically prone subgroups are less likely to be due to bias or confounding
- Mendelian randomization ensures that genetic association studies are less susceptible to the problems of observational studies
- When the effect of a gentic polymorphism mimics the effect of an environmental exposure, this can provide convincing evidence that the environmental exposure is causal

Case-control and family-based studies are the method of choice for demonstrating genetic associations. Cohort studies are usually preferred for the study of environmental factors. For studies of G-E interaction, the advantages of cohort studies are not as clear:

- Cases are in short supply and may be phenotypically hetergoneous
- Because of Mendelian randomization, G-E interaction effects may be less affected by biased exposure measurement

Case-only studies: Mendelian randomization also leads to the possibility of gaining information simply by looking at cases:

- G and E should be independent in the population

- Certain forms of joint action will create association between G and E in cases

At the very least this offers the possibility of extracting further information from existing designs.

Reagrding public health relevance, the development, in the late 60's, of statistical methods for studying multiple risk factors led to suggestions to target intervention upon subgroups with high risk scores. For multifactorial disease, it doesn't work – most cases come from low to medium risk groups (Rose G, The strategy of preventive medicine, 1992). The same suggestion has been made in relation to genetic factors. But, for genetically complex diseases, why should the situation be any different?

Genes and environment must interact and how they act together to cause disease is an important field of study. Tests for statistical interaction do not necessarily have any strong biological interpretation. Design of studies explicitly to test for statistical interaction has no scientific jsutification. In the search for genetic and envionmental determinants, consideration of factors together will increase power by little if at all, and requires well founded prior hypotheses. For diseases if complex geentic aetiology, targeting interventions at genetically susceptible groups may not be an effective public health strategy.

Multiple testing and "significance" in genetic epidemiology

The Neyman-Pearson theory cosntitutes an approach to scientific inference based on decision theory. The aim of an investigation is seen as being to refute a null hypothesis, H_0. We deine a test statistic, $T(\text{Data})$, sensitive to departures from H_0, and reject H_0 if the observed T, $T_{\text{obs}} \geq C$ – a critical value. We choose C so that the type 1 error probability $\alpha = \Pr(T \geq C \mid H_0)$ is small. This is the significance level.

This strict decision theoretic approach is now rarely used. Instead, we calculate the p-value, defined by

$$p = \Pr(T \geq T_{\text{obs}} \mid H_0)$$

as a measure of the strength of evidence agaisnt H_0. Small p-values indicate that data are improbable under H_0 and cast doubt upon it. The p-value is a statistic, calculated from the data, with the property that under H_0, its probability distribution is uniform over the interval $(0,1)$.

Returning to the decision theoretic approach, if we do several tests we increase our chances of making errors. We can rely on Bonferroni correction: if α_1 is the type 1 error probability in a single test, the type 1 error probability in N independent tests is

$$\begin{aligned}\alpha_N &= 1 - (1-\alpha_1)^N \\ &\approx N\alpha_1\end{aligned}$$

We need to choose a smaller α_1 to maintain the same overall type 1 error probability. The same correction is advocated for p-values.

Decision theory: we calculate independent tests $T_1, T_2, \ldots, T_N$ and reject H_0 if *any* $T > C$. P-value approach: we calculate independent p-values $p_1, p_2, \ldots, p_N$ and record the smallest, p_{min}. The Bonferroni-corrected value $1 - (1 - p_{\text{min}})^N \approx Np_{\text{min}}$ is uniformly distributed over $(0,1)$. It is a measure of the strength of evidence against H_0 from all the tests taken together. In both cases, there is only one H_0 under consideration.

For the sake argument, assume that we can consider the genome as 46 independent segments. A genome screen can be considered as a test of 46 H_0's – not 46 tests of a single H_0. Does it still make sense to correct for multiple testing? It is generally accepted that it does, and that we need to demonstrate "genome-wide significance". But why?

Consider two scientists. Professor A is cautious and unambiguous. He writes a grant application every year for 46 years and each year tests a separate segment of the genomes for linkage (in random order). Professor B, however, is more ambitious. He writes a big grant and tests all 46 regions in one study. Which scientist should apply a multiple testing correction? Or, perhaps neither? Or both? If Professor A should carry out a multiple testing correction, what correction should be applied? What is N when he carries out his first study?

Both scientists should do the same, and (more controversially) both scientists should apply a multiple testing correction. But, clearly, the reason for a multiple testing correction does not follow from the conventional theory:

- the number of tests carried out doesn't seem relevant
- rather, in some sense, we need to correct for the multiplicity of hypotheses

The Neyman-Pearson theory is concerned with probability calculations of the form $\Pr(\text{Data} \mid \text{Hypothesis is true})$ while what the scientist wants to know is $\Pr(\text{Hypothesis is true} \mid \text{Data})$. This second question raises deep philosophical problems about the nature of probability: hypotheses are either True or Falsen unless there are degrees of truth.

Attempts to develop a coherent theory of probability as degree of truth failed. But a theory has been developed in terms of degree of belief. Unfortunately this loses the important characteristic of objectivity – different scientists are permitted to have different degrees of belief (as they do in the real world). But probability theory provides a theory for modifying beliefs in the light of evidence so that, given the same evidence, beliefs of different scientists will converge.

This is an uncontroversial theorem of probability theory. Its application in the current (controversial) context is as follows. Given two alternative hypotheses, H_0 and H_1,

$$\underbrace{\frac{\Pr(H_1 \mid \text{Data}}{\Pr(H_0 \mid \text{Data}}}_{\text{Posterior odds}} = \underbrace{\frac{\Pr(\text{Data} \mid H_1}{\Pr(\text{Data} \mid H_0}}_{\text{Likelihood ratio}} \times \underbrace{\frac{\Pr(H_1)}{\Pr(H_0)}}_{\text{Prior odds}}$$

For example, H_1 may be "there is linkage of the region to disease", and H_0 that "there is no linkage". But what about the strength of linkage?

A hypothesis such as "there is linkage" or "there is association" is really composed of infinitely many hypotheses. Even given that there is linkage (or association), its strength (measured by some parameter, θ say) is uncertain. We must also specify a prior distribution for θ. The multiplier which transforms prior odds for H_1 to posterior odds now involves averaging over this distribution of θ:

$$\text{Bayes factor} = \frac{\text{Average}_\theta \left[\Pr(\text{Data} \mid H_1, \theta)\right]}{\Pr(\text{Data} \mid H_0}$$

When testing for linkage we expect most null hypotheses to be true – the prior odds are strongly against each H_1. This is even so in the case of association studies. To overcome strong prior odds against linkage or association, we need a large Bayes factor, corresponding (loosely) to a small p-value. This correspondence is not simple and has been the source of some controversy.

Table 39 show posterior odds for $\$\mathrm{H}_0$ versus H_1 at two significance leveles. Prior mean size of effect is ocnstant for all N. The paradox may be resolved if we let Effect size $\propto 1/\sqrt{N}$ – or $N \propto 1/(\text{Effect size})^2$.

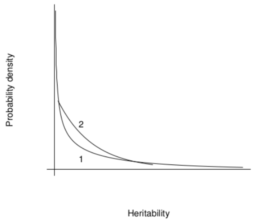
Sample size	Significance level	
N	0.1	0.01
1	0.365	0.0513
10	0.283	0.00797
10^2	0.690	0.0141
10^4	6.70	0.132
10^6	66.8	1.31
10^8	668	13.1

Table 39: A paradox (Cox and Hinkley, 1974, p397)

Small effects need large samples. In well-designed studies, the sample size will be realistic given the expected size of effect (if present) and the desired significance level. For any p-value, take sample size as that required to achieve a given power for a given size of effect. Holding power constant we can plot Bayes factor agaisnt p-value. Assumptions for association studies:

- Prior distribution of heritability attributable to an associated gene is χ_1^2
- We use χ^2 tests for association; degree of freedom determined by number of htSNPs needed

Two prior distributions for the heritability attributed to a causal variant (figure in margin):



- distribution 2 is exponential, as suggested by Sewell Wright
- distribution 1 is χ_1^2, as recently suggested by Rudan et al.

In Figure 25, $P = 10^{-6}$ gives a Bayes factor $> 10^4$. Power is irrelevant, except when derisory (seriously underpowered studies require smaller p-values to convince).

Figure 26 show almost the same curve. $P = 10^{-4}$ gives a Bayes factor $> 10^2$. We will get the same curve for 10 df.

In the case of microarrays and genome screens, how are things chnaged if we test many hypotheses simultaneously? Aim not proof of association, but the identification of a list of good candidates. We can rank genes from the smallest p-value to the largest, and draw a line someway down the list. Can we estimate the proportion of false

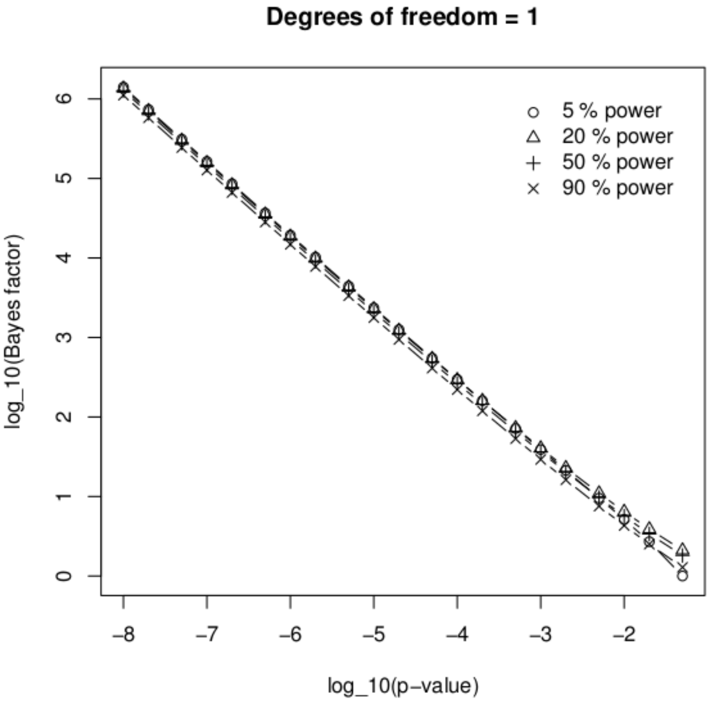


Figure 25: One-df test

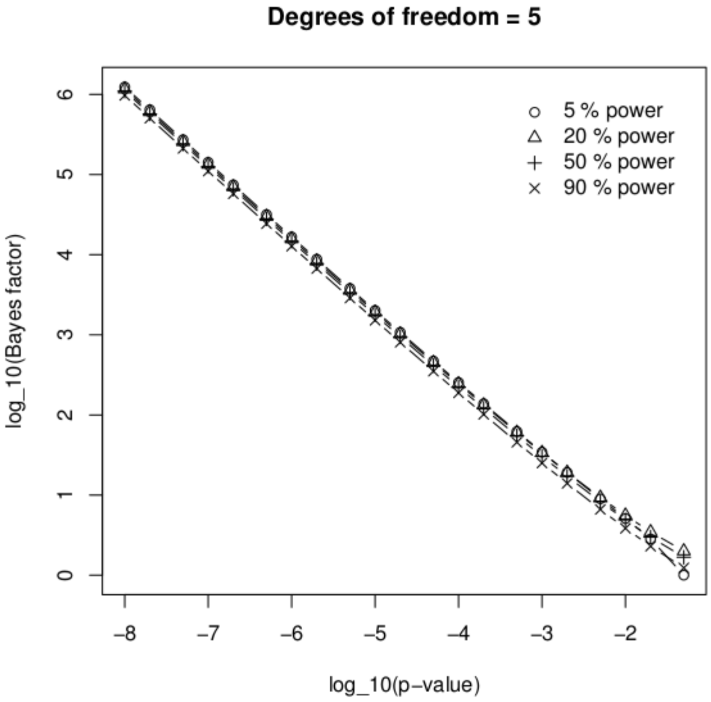
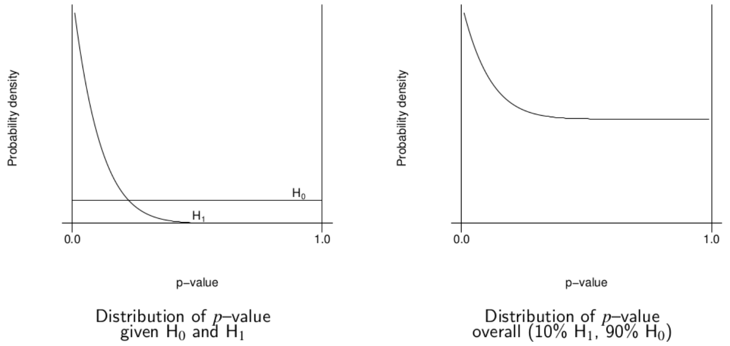


Figure 26: Five-df test

positives identified? Here we are somewhat better of because we can estimate the prior distribution of effect sizes and the proportion of true positives.

We can estimate proportion of genes for which H_1 is true, and the distribution of p given H_1 is true (Figure 27).



Distribution of p–value given H_0 and H_1

Distribution of p–value overall (10% H_1, 90% H_0)

Figure 27: Empirical priors

Consider the rate of false positives if we choose all genes with p-value $\leq P_c$, a threshold. Bayes theorem gives

$$\Pr(H_0 \mid p \leq P_c) = \frac{\Pr(p \leq P_c \mid H_0}{\Pr(p \leq P_c} \times \Pr(H_0)$$

$\Pr(p \leq P_c \mid H_0)$ is simply P_c, and $\Pr(p \leq P_c)$ can be estimated by the proportion of p-values less than or equal to P_c.

As an example, consider testing 10,000 genes and finding that 20 have p-values $\leq$ our chosen threshold, $P_c = 10^{-3}$:

$$\frac{\Pr(p \leq P_c \mid H_0)}{\Pr(p \leq P_c)} = \frac{10^{-3}}{20/10,000} = 0.5$$

If we estimate that the true negative rate, $\Pr(H_0)$ is 0.995, the estimated proportion of false positives if we take $P_c = 10^{-3}$ is $0.5 \times 0.995 = 0.4975$. If true positives are expected to be rare we can tale $\Pr(H_0) \approx 1$; this provides an upper bound for the false positive rate.

Use of the conservative approximation gives the false discovery rate of Benjamini and Hochberg. If we select only the smallest of N p-values, the false disccovery rate is $\frac{p_{\text{min}}}{1/N} = Np_{\text{min}}$ – equivalent to the Bonferroni correction. Storey et al. developed a method for estimating the true negative rate. They called the estimates of false positive rate for each threashold Q-values.

Whole genome association studies

Prospects for whole-genome screens, estimated numbers of single nucleotide polymorphisms (SNPs):

- Direct studies of common nsSNPs (MAF $>$ 1%): ˜ 30,000–50,000 SNPs
- Indirect studies of genes: ˜ 300,000–500,000 SNPs
- "Nearly" whole genome: 500,000–1,000,000 SNPs
- Whole genome: ˜ 2,000,000-4,000,000 SNPs
- Initial model-based estimate (Kruglyak, 1999): 500,000 SNPs

Mutliplicity of tests, or rather *a priori implausibility* of each hypothesis, means that we require at least $p < 10^{-6}$ for "significance". Common variants are likely to have small effects – odds ratio ¡ 1.5. This militates for case/control studies with sample sizes of 5,000–20,000 (although few such collections currently exist).

Current "gene-chip" technologies deliver 250,000–500,000 SNPs on a study subject in a single determination, at a cost of $500–$1000. Next year we expect double the number of SNPs for approximately the same price. Nevertheless, whole genome screening studies will be expensive undertakings.

We might need 8-10,000 subjects to achieve adequate power to detect associations at $p < 10^{-6}$. But it is more efficient to use a multiphase design. For example:

1. 2,000 cases + 2,000 controls with 500,000 SNP chip
2. Further 2,000 cases + 2,000 for best 100,000 SNPs
3. Further 4,000 cases + 4,000 for best 10,000 SNPs

These designs are in current use in candidate gene studies, and will be essential in whole genome studies, although constrained by available levels of mutliplexing. Computation of the characteristics of such designs requires Monte Carlo integration – optimization is computationally intensive.

Optimal choice of markers requires detailed mapping of LD, e.g. based on HapMap data. Understanding LD depends on two aspects:

- Mapping physical extend of LD along the chromosome
- Understanding the structure of the haplotype phylogeny at each point

Truly optimal solutions will be computationally intensive. Current chip designers are using single marker r^2 cluster-based algorithms.

The large number of markers typed in whole-genome studies will allow us to assess the impact of population substructure and to apply ”genomic control”. As ancestry-informative markers are discovered, they will be incorporated into gene chips.

The Wellcome Trust Case-Control Collaboration arose out of independent proposals from Universities of Cambridge and Oxford and from the Sanger Institute, but now it has been extended to a wider consortium. Funding of £8.6m announced by the Wellcome Trust in April 2005. In total, there are 700,000 SNPs, including all known common non-synonymous coding SNPs, and tagging SNPs for as much as possible of the remaining genome. Eight case groups covering a range of pathologies (cardiovascular, cancer, autoimmune, psychiatric) and two control groups (blood donors + 1958 birth cohort subjects).

Phased design:

1. 1,000 DNA’s from each group typed for all SNPs
2. Current funding is to type SNPs/regions for which $p < 0.01$ in a further 1,000 cases and controls for each disease – but it is hoped that falling costs will allow more SNPs to be typed in this phase.

Unfortunately 2,000 cases + 2,000 controls is barely adequate. Some groups have more. We hope that the study will provide an impetus for further case collections and greater cooperation amongst clinical groups.

Initial design for a study of nsNSPs and type 1 diabetes:

- Stage 1: ˜ 900 cases and ˜ 900 controls and ˜ 7,500 nsSNPs with MAF $> 1\%$
- Stage 2: 3,000 nsSNPs in further 1,000 cases and 1,000 controls
- Stage 3: 1,000 nsSNPs in further 4,000 cases and 4,000 controls

Cases are ˜ 50% of all the juvenile-onset diabetic cases in GB. Controls are drawn from the national 1958 birth cohort. Confirmation of positive results in case-parent family studies (˜ 3,000 trios).

Most extreme results are in the HLA region. There is at least one large known association in the HLA region and strong LD. There is also evidence of general overdispersion.

Most extreme finding is another knwn association (at least it is known now). General overdispersion remains.

Repeating these analyses using highly reliable single-marker typing revealed:

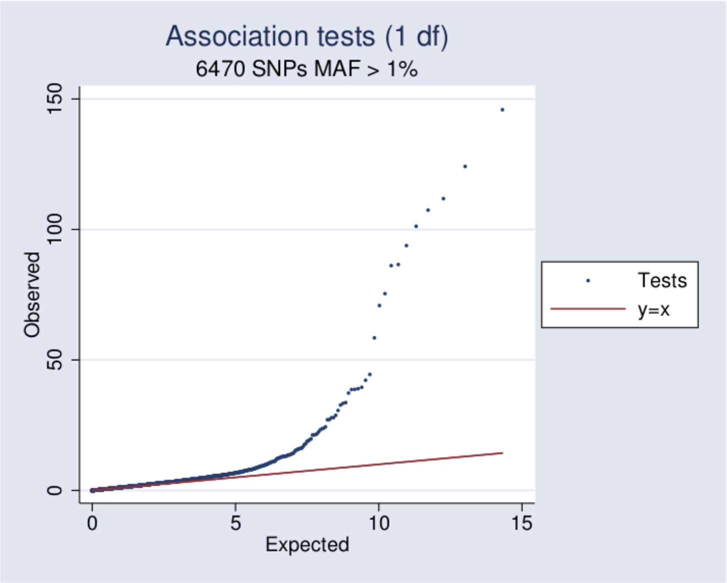


Figure 28: QQ plot: all SNPs with MAF > 1%

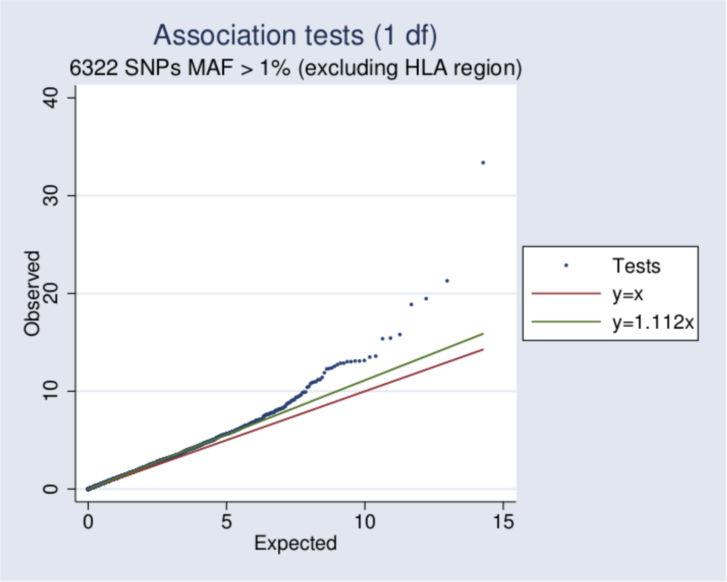


Figure 29: QQ plot: omitting the HLA region

- genotyping errors, acting differentially between cases and controls
- ”informative missingness”, again acting differentially

How? Typing was done ”blind” to case/control status and in random order. Yet closer inspection showed different characteristics of the genotype scoring – presumably due to the fact that cell lines were created and DNA extracted in different laboratories.

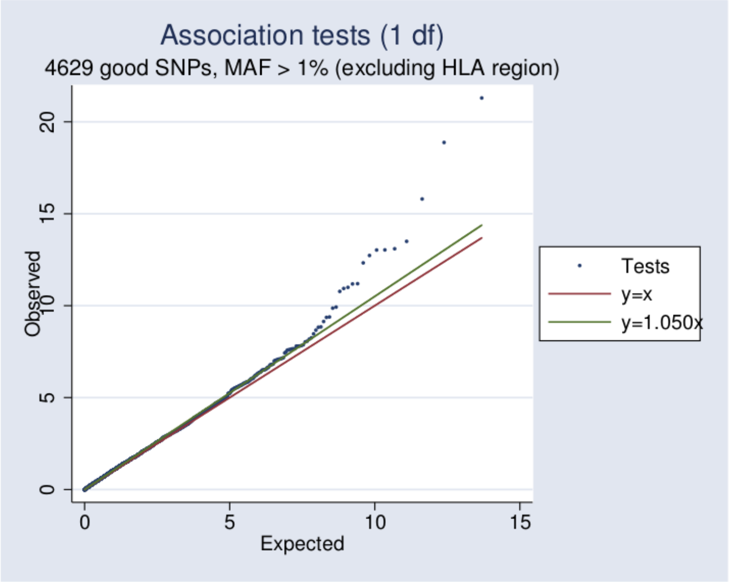


Figure 30: QQ plot: omitting the HLA region

As can be seen in Figure 30, there is still some overdispersion. Controlling for geography, little overdispersion remains (Mantel-extension test, stratified by 12 broad regions).

In this study, differential misclassification of ”exposure” (here genotype) between cases and controls was more of a problem than unmeasured confounding by population substructure. Great care will be necessary – particularly with very high throughput technologies and automated genotype scoring procedures.

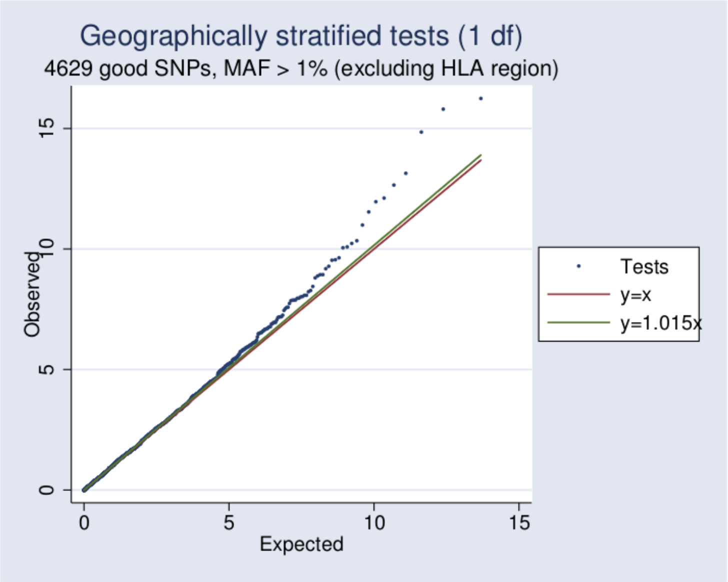


Figure 31: QQ plot: omitting the HLA region