Gene Mapping: Introduction

Gene mapping studies the relation of genotypes and phenotypes. Its goal is to identify, as precisely as possible, genomic regions affecting particular phenotypes of interest and to estimate the importance of those regions to phenotypic variability of the trait.

Phenotypes can include disease status (usually coded 0 or 1), quantitative measurements associated with an individual (blood pressure, fasting glucose), transient molecular measurements associated with organ function (RNA transcript abundance), etc.

Basic Logic of Genetic Mapping

1. **Experimental crosses of inbred strains.** Because we can know the original genetic makeup of the inbred strains and can control breeding and other environmental factors, we can study under controlled conditions the correlation of genotype and phenotype. Peak is relatively broad, so a few hundred markers suffice to cover an entire genome.

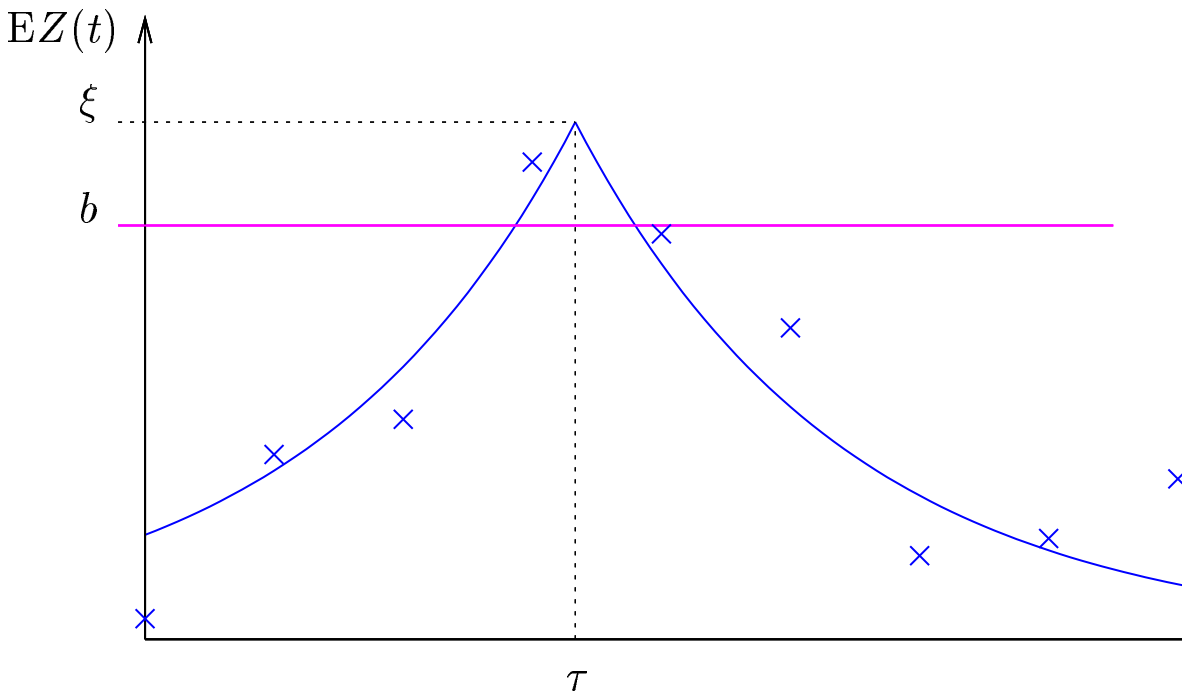
2. **Association Mapping in Humans.** This is, in principle, similar to 1, but is complicated by the uncontrolled features of population history and inadequate knowledge of the mode of inheritance of the trait. Success of this method depends on *association* (correlation) between genetic markers and the phenotype of interest. Peak is very narrow, so many, perhaps 500000, genetic markers are required to cover a genome.

3. **Recombination (linkage) mapping in Humans.** Relatives have similar traits because they have similar genotypes at those loci where there are genes that influences those traits. *Identity by descent* (IBD) at a locus t measures genetic relatedness at t. Hence we should expect to find the gene(s) that influence a trait that is shared by related individuals to be found in regions where those individuals are IBD. This involves analysis of covariance (of genotype and phenotype) and generates a signal that is usually weaker, but much wider than for association mapping. Since one utilizes family relationships going back only a few generations, there are fewer difficulties due to population history.

3. **Association mapping with family controls** has advantages and disadvantages of both 2 and 3.

4. **Admixture mapping** utilizes historical phenotype and genotype differences between populations that for the most part have intermarried only relatively recently (e.g., Asians and Europeans). One studies the correlation between phenotypes and the population origin of genetic markers.

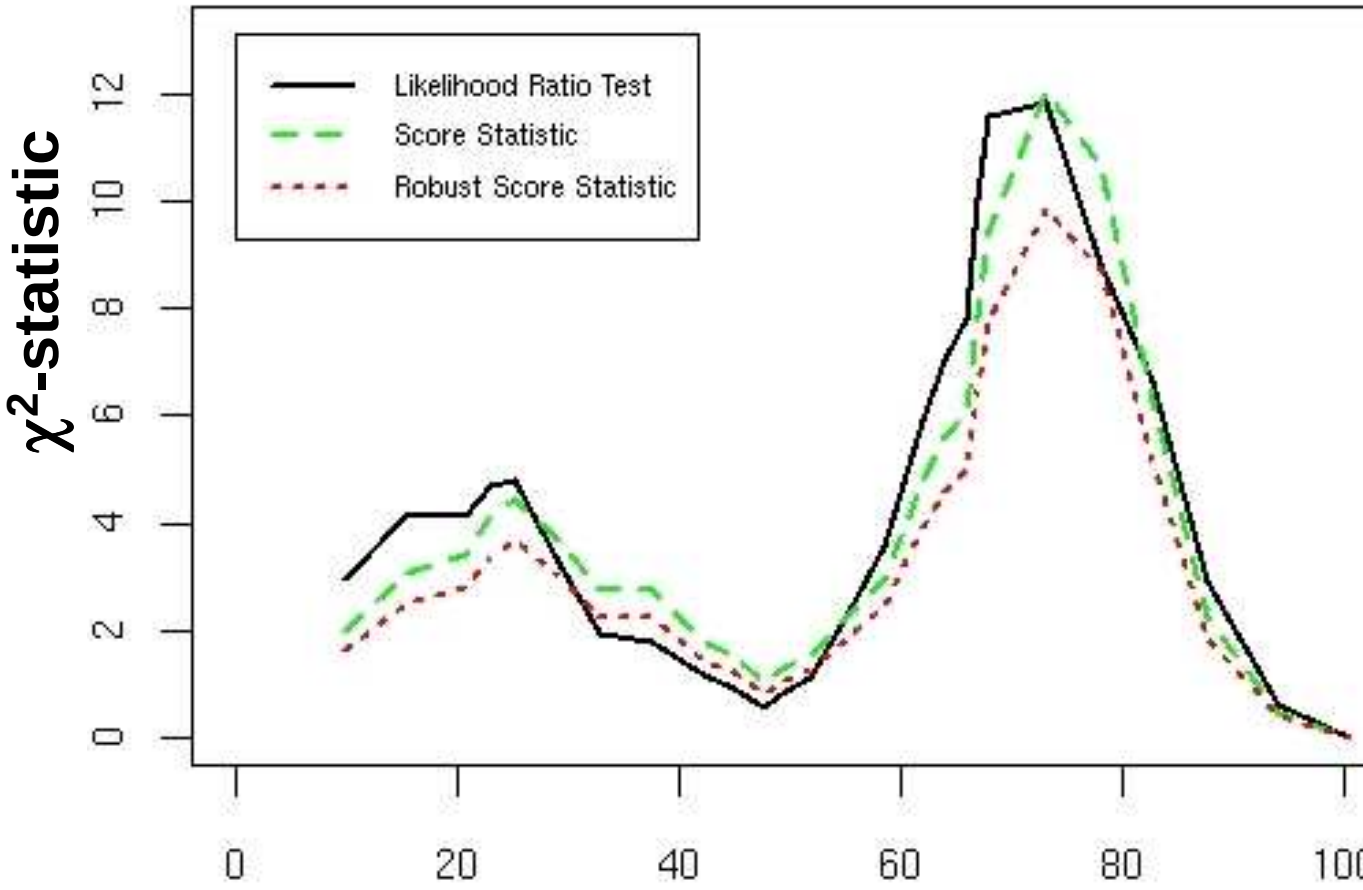
Generic Data



Questions: (1) How should the threshold b be determined in order to account properly for multiple comparisons and control the false positive error rate? (2) How can one increase the power to detect genes affecting the phenotype? (3) How can one estimate the location τ and the genetic effect ξ?

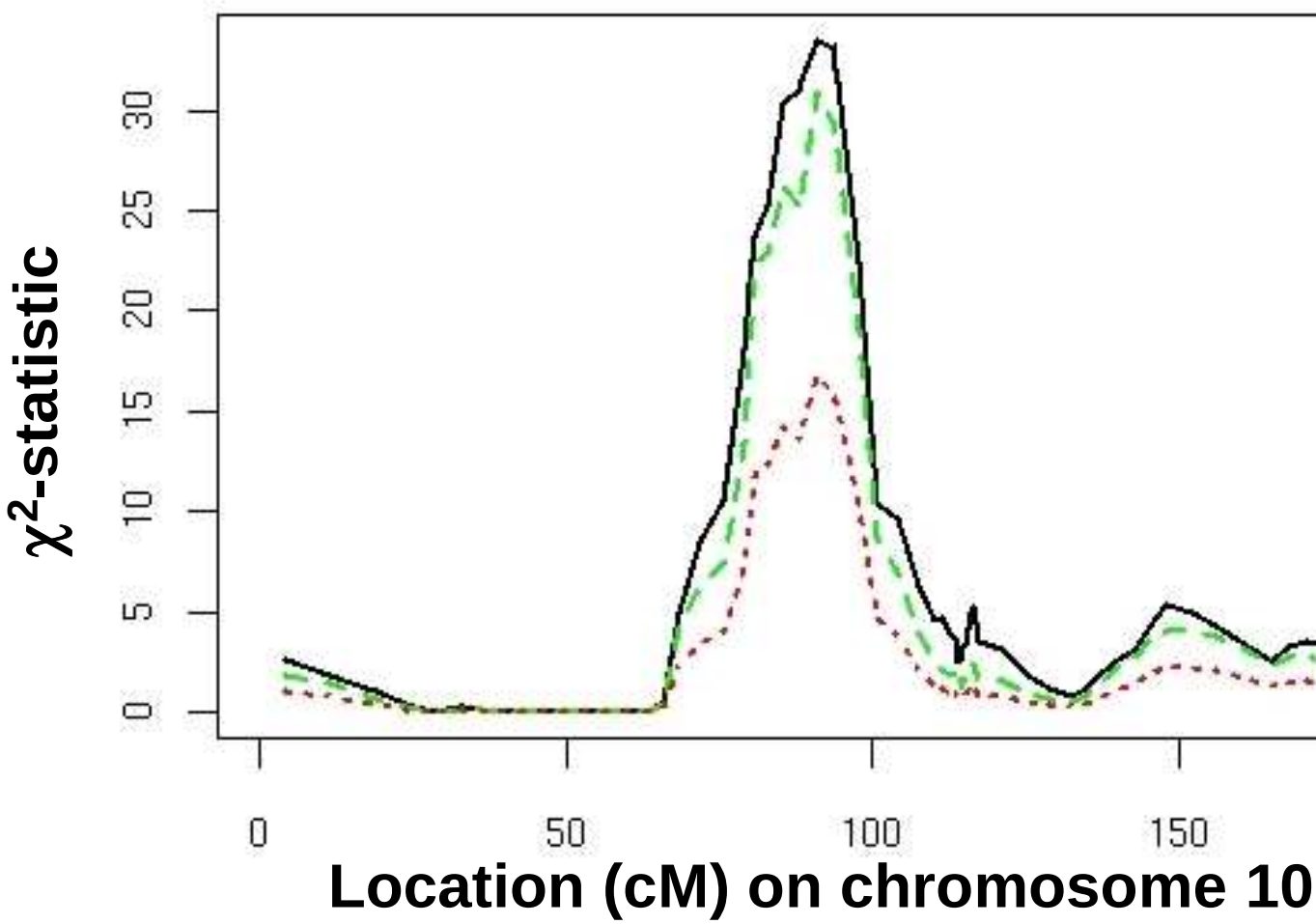
Note irregularity as a statistical problem: if $\xi = 0$, then τ is unidentifiable; also the cusp.

Fasting Insulin

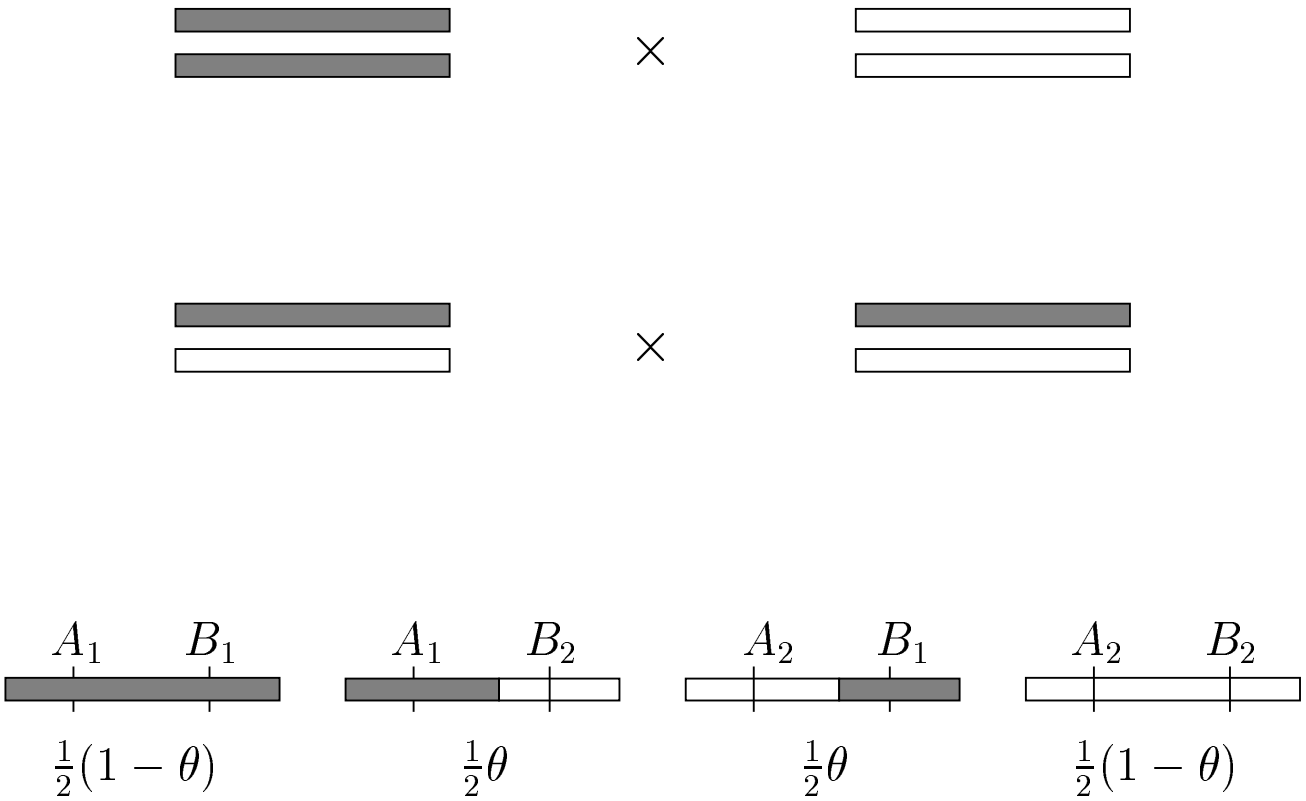


Location (cM) on chromosome 19

Fasting Blood Glucose



Intercross



Given two genetic markers, the recombination fraction between them is denoted by θ. If we assume the Haldane model of no interference, the genetic distance D between two markers, in units of Morgans(M), is defined by solving the equation (explained below) $\theta = [1-\exp(-2D)]/2$. One M is the distance in which the expected number of crossovers is one.

The mouse genome consists of 20 pairs of chromosomes of total length 16 M, while the human genome consists of 23 pairs of chromosomes of approximate total length 33 M. The genome of baker's yeast consists of 16 chromosomes of total length 44 M.

Crossovers and Recombination

At any position t on a child's (for example) maternally inherited chromosome, the allele is equally likely to come from either of the mother's two chromosomes. A *crossover* occurs at position t if the inherited DNA switches from one of the mother's chromosomes to the other.

According to the model suggested by Haldane (1919), crossovers occur in random positions, according to a Poisson process. This means that the number of crossovers between positions s and t has a Poisson distribution, so the probability of k crossovers is given by

$$Q_k = [\lambda(s,t)]^k \exp[-\lambda(s,t)]/k!, \quad k = 0, 1, \ldots.$$

Here $\lambda(s,t)$ is the mean number of crossovers. In addition the numbers of crossovers in non-overlapping intervals are independent random quantities. By changing the scale in which "distance" along a chromosome is is measured, it is possible to assume that $\lambda(s,t) = |t-s|$. This new distance scale is called "genetic distance" and the units of distance are Morgans (after Thomas Hunt Morgan). Sometimes one uses centi-Morgans (cM). 1 cM equals 0.01 M.

Remark. It is known that Haldane's model is a very rough approximation, and better models have been suggested. But the mathematical convenience of using Haldane's model often outweighs the improvement in accuracy of using a more precise model.

A *recombination* occurs between the two genetic loci s and t if there is an *odd* number of recombinations between these loci. If θ denotes the probability of recombination, then

$$\theta = \sum_{k\text{odd}} |t-s|^k \exp(-|t-s|)/k! = [1 - \exp(-2|t-s|)]/2. \qquad (*)$$

Exercise. Recall that for every real number x, $\exp(x) = \sum_0^\infty x^k/k!$. Hence $[\exp(x) - \exp(-x)]/2 = \sum_{k \text{ odd}} x^k/k!$. Use this information to derive (*).

Natural Populations: Hardy-Weinberg and Linkage Equilibrium

A genetic marker with alleles $A_1, \ldots, A_m$ that have frequencies in the population of $p_1, \ldots, p_m$ is said to be in Hardy-Weinberg equilibrium (HWE) if the *genotypes* A_iA_i have frequencies p_i^2, while the genotypes A_iA_j for $i \neq j$ have frequencies $2p_ip_j$. In an infinitely large randomly mating population, HWE is achieved in one generation of random mating.

Consider the case $m = 2$. Let $u, 2v, w$ denote the initial frequencies of the genotypes A_1A_1, A_1A_2, and A_2A_2, respectively. Then the allelic frequencies are $p_1 = u + v$, and $p_2 = v + w$. After a single generation of random mating the genotypic frequency of A_1A_1 is $u' = u^2+4uv/2+v^2 = p_1^2$. Similarly $w' = p_2^2$ and by subtraction $2v' = 2p_1p_2$, so $p_i' = p_i$ for $i = 1, 2$.

Remark. This result validates the interpretation of random mating that says we sample a maternal allele and a paternal allele independently from the population of all alleles according to the frequencies p_i.

Exercise Write out details of the preceding argument in the case of m alleles.

In real populations genetic loci frequently are close to HWE, but there are often small departures. Since real populations are finite, some apparently unrelated indviduals may actually have a common ancestor in their family histories. In this case the genotype formed by selecting one allele from two randomly selected parents is of the form $P(A_1A_1) = p_1^2 + Fp_1p_2, P(A_1A_2) = 2p_1p_2(1-F)$, etc., where $F > 0$ is the *coefficient of relatedness* of the parents. Note that the probability of a homozygote is increased while that of a heterozygote is decreased, relative to HWE. Although this departure from HWE is often very small, it has important consequences for population based association studies, especially case control studies, since the coefficient of relatedness may differ in cases and controls.

Alleles at different loci that are both inherited from the mother (or father) are said to be in **linkage equilibrium** if their joint frequency is the product of their marginal frequencies. Symbolically, suppose A, a and B, b are alleles at two loci, A and B have frequencies p_1 and p_2,

respectively. Let $P(AB)$ denote the joint frequency of A at the first locus and B at the second locus on chromosomes inherited from the same parent. Then the alleles are in inkage equilibrium if $P(AB) = p_1 p_2$.

Unlike the case of HWE, linkage equilibrium only becomes established over several generations of random mating. To see this, let $P_t(AB)$ denote the frequency of AB after t generations of random mating, with $P_0(AB)$ the initial frequency. Let θ denote the recombination fraction between the two loci. For recombinant chromosomes, by HWE the frequency of AB is $p_1 p_2$, and hence $P_t(AB) = (1-\theta)P_{t-1}(AB) + \theta p_1 p_2$. This can be re-written $P_t(AB) - p_1 p_2 = (1-\theta)(P_{t-1} - p_1 p_2)$, so $P_t(AB) - p_1 p_2 = (1-\theta)^t (P_0(AB) - p_1 p_2) \to 0$ as $t \to \infty$.

Note that although linkage equilibrium becomes established with repeated random mating, the rate depends on the recombination fraction; and even for loci on different chromosomes, where $\theta = 1/2$, linkage disequilibrium can persist for a few generations of random mating.

Suppose that a population can be divided into two parts (strata) of relative sizes w_1 and w_2, where $w_1 + w_2 = 1$. Assume there is random mating within each stratum, so eventually there is linkage equilibrium in each stratum, i.e., in obvious notation $P_i(AB) = P_i(A)P_i(B)$ for $i = 1, 2$. For a person drawn at random from the entire population there is not linkage equilibrium, since (for example) $w_1 P_1(AB) + w_2 P_2(AB) \neq [w_1 P_1(A) + w_2 P_2(A)][w_1 P_1(B) + w_2 P_2(B)]$. A similar remark applies to Hardy-Weinberg equilibrium, for exactly the same mathematical reason.

Statistical Models for a Phenotype; Fisher, 1918

Quantitative Traits: On a very general level, the phenotype Y is given by $Y = f(G, E)$, where G denotes genotype and E denotes environment. As a very special case assume (Fisher, 1918; Weinberg, 1910) that this function is additive, so $Y = G+E$. In general G may involve many genes, but suppose we focus on a single genetic locus, denoted by its genomic location, τ. There may be additional genes contributing to the trait (all on different chromosomes). Assume that the population is randomly mating and at locus τ let A_1 and A_2 be alleles with frequencies of p and $q = 1 - p$. (An intercross, which occurs when two inbred strains that are homozygous, say A_1A_1 or A_2A_2, at all loci are crossed and their progeny (inter)crossed, can be regarded as a special case with $p = 1/2$.)

Let $g = g(\tau) \in \{0, 1, 2\}$ be the number of A_1 alleles.

We assume Y is given by

$$Y = m + ag(\tau) + d\mathbf{1}_{\{g(\tau)=1\}} + e,$$

where $\mathbf{1}_{\{\cdot\}}$ is equal to 1 if $\{\cdot\}$ is true and 0 otherwise; and e stands for random environmental effects *and* the effects of other genes that do not explicitly appear in the model. After some algebra this equation can be rewritten in the form

$$Y = \mu + \alpha[g(\tau) - 2p] + \delta[\mathbf{1}_{\{g(\tau)=1\}} - 2pq - (q - p)\{g(\tau) - 2p\}] + e. \quad (*)$$

In this equation $\alpha = a + (q - p)d$, $\delta = d$. An equivalent expression is

$$Y = \mu + \alpha[g(\tau) - 2p] - 2\delta[g_M(\tau) - p][g_F(\tau) - p] + e,$$

where $g_M(t)$ is 1 or 0 according as the allele inherited from the mother at the locus t is A, and similarly for $g_F(t)$.

We assume that e has mean 0 and is independent of $g(\tau)$. This important assumption has both advantages and disadvantages. Advantages: It allows us to write the variance of Y as $\sigma_y^2 = \sigma_A^2 + \sigma_D^2 + \sigma_e^2$ in terms of the *variance components*

$$\sigma_A^2 = 2pq\alpha^2, \;\; \sigma_D^2 = (2pq\delta)^2, \;\; \sigma_e^2$$

Disadvantages: We have assumed there is no gene-environment interaction, and that the loci of unmodeled genes are unlinked to the locus τ. With a more complex model one can get around these disadvantages, but that possibility is not discussed here.

Qualitative Traits: Assume that the observed phenotype of an individual is ϕ, which equals 0 or 1 according as the individual does not have or does have a particular disease. For the purpose of constructing a test, ϕ replaces Y in the definition of the statistic. But for constructing a model we put $Y = E(\phi \mid G, E)$ and use a model of the form given above in (*). We could also use a logistic model, where $Y = \log[E(\phi \mid G, E)/\{1-E(\phi \mid G, E)\}]$. The score statistic would be the same, but the likelihood ratio statistic would be substantially more complicated.

Exercise. Assuming Hardy-Weinberg equilibrium, show that the terms multiplying α and δ in (*) are uncorrelated.

Testing for Linkage in an Intercross

Model: $Y = \mu + \alpha[g(\tau) - 1] + e$ (additive trait, no dominance deviation). Phenotypic variance is $\sigma_y^2 = \sigma_A^2 + \sigma_e^2 = \alpha^2/2 + \sigma_e^2$. Locus specific heritability is $h^2 = \sigma_A^2/\sigma_y^2$.

Data: Phenotypes $Y_1, \ldots, Y_n$, Genotypes $g_1(t_i), \ldots, g_n(t_i)$ at markers $t_1, \ldots, t_m$ spread evenly throughout the genome. Note τ may or may not be one of the markers.

Statistic: $Z_t = \sum_1^n (Y_i - \bar{Y})[g_i(t) - 1]/[\hat{\sigma}_y (n/2)^{1/2}]$, where $\hat{\sigma}_y^2 = n^{-1} \sum_1^n (Y_i - \bar{Y})^2$.

Properties: At unlinked loci $E(Z_t) = 0$, $\mathrm{Var}(Z_t) = 1$, Z_t is approximately normally distributed. For two loci separated by a recombination distance θ, we have $E[g_M(t) g_M(s)] = (1 - \theta)/2$, so $\mathrm{Cov}[g_M(t), g_M(s)] = (1 - \theta)/2 - 1/4 = (1 - 2\theta)/4$, and hence

$$\mathrm{Corr}(Z_s, Z_t) = (1 - 2\theta) = \exp(-2|t - s|) \qquad (**)$$

. Note that these statements are true conditional on the Y's, i.e., without assuming any model for the phenotypes as a function of genotypes.

Now assume the phenotypic model suggested above is valid and there is linkage of the marker at locus t to a QTL at τ. Then from (**) it follows

$$E(Z_t) \approx n^{1/2} \sigma_A \exp(-2|t - \tau|)/\sigma_y.$$

(Exercise: Derive this result.)

Strategy: Since we do not know the location of τ, we scan the entire genome using $\max_t |Z_t|$, where the maximum is taken over all markers $t = t_j, j = 1, \ldots, m$.

Significance levels: For testing a single marker, we use the normal distribution, e.g, $P_0\{|Z_t| \geq 2\} \approx 0.05$, to determine the threshold for significance. For testing k markers, we can use the Bonferroni inequality:

$$P_0\{\max_t |Z_t| \geq z\} \leq \sum_t P_0\{|Z_t| \geq z\}. \qquad (*)$$

Then for a 0.05 genomewide significance level we must choose c so that $P_0\{|Z_t| \geq z\} \leq 0.05/m$. If the markers are widely spaced (say 0.1 M apart or more), this inequality provides a reasonable approximation. But if they are close together, one can use a more complicated formula that gives a good approximation to the probability on the left hand side of (*). The result is that instead of $z = 2$, which would be the 0.05 level threshold for a single test, one should take $z \in (3.3, 4.0)$ (depending on the marker density and the genetic length of the genome being scanned).
Approximation. For equally spaced markers at distance Δ, a very good approximation to the false positive rate is given by

$$P_0\{\max_t |Z_t| \geq z\} \approx 1-\exp\{-2C[1-\Phi(z)]-\sum_c \ell_c[2\beta z\varphi(z)\nu[z(2\beta\Delta)^{1/2}].$$

In this approximation φ denotes the standard normal probability density function, Φ is the standard normal distribution function, ν is a special function, which to a very good approximation is given by

$$\nu(y) \approx \frac{(2/y)[\Phi(y/2) - 1/2]}{(y/2)\Phi(y/2) + \varphi(y/2)},$$

C denotes the number of chromosomes and ℓ_c the genetic length of the cth chromosome in M. The parameter β can vary according to the problem. For an intercross $\beta = 2/\mathrm{M}$.

The preceding analysis ignores the possibility that there is an important dominance deviation (parameterized previously by δ. One can also define and analyse a two-dimensional statistic to test a marker for linkage to a gene having either an additive or a dominance effect (or both).

For a numerical example, we consider an idealized mouse genome consisting of 20 chromosomes of 0.8 M each, for a total genetic length

of 16 M. Assuming that markers are spaced at $\Delta = 1$ cM intervals, the genome wide 0.05 significance threshold would be $z_{\max} = 3.79$. The Bonferroni bound for this threshold would be 0.24, which is unacceptably conservative. At the other extreme, recall that for a single marker, the 0.05 threshold would be $z \approx 2$, so we must compensate for the large number of markers tested with a considerably higher significance threshold.

For the two-degree of freedom statistic, the genome-wide 0.05 significance threshold would be $||z||_{\max} = 4.28$.

Permutation Method. Randomly assign phenotypes to genomes, determine $\max_t Z_t$, repeat a large number of times, and determine the relative frequency of tail events.

Power

Suppose that there is one trait locus at τ, which is also a marker locus. Then the power of a genome scan is

$$P\{\max_t |Z_t| \geq z\} = P\{|Z_\tau| \geq z\} + P\{|Z_\tau| < z, \ \max_{t \neq \tau} |Z_t| \geq z\}. \quad (***)$$

Recalling that for at a marker locus t linked to τ, $E(Z_t) = \exp(-\beta|t-\tau|)N^{1/2}\sigma_A/\sigma_Y$, we see that the most important term on the right hand side of (***) is the first one, which is also easy to evaluate.

In the case that markers are equally spaced at distance Δ, the less important second term can be approximated by using the observation that conditional on Z_τ, the process $z(Z_t - Z_\tau)$ behaves like a (two-sided) random walk with each step having a mean value of $-\beta z^2\Delta$ and a variance of $2\beta z^2\Delta$. Using the notation $\xi = N^{1/2}\sigma_A/\sigma_Y$, we find that the power is approximately

$$1 - \Phi(z - |\xi|) + \varphi(z - |\xi|)[2\nu/|\xi| - \nu^2/(z + |\xi|),$$

where $\nu = \nu[z(2\beta\Delta)]$.

Gene-gene and Gene-environment Interactions

The simple model described above can be modified in a straightforward way to provide models for gene-gene and gene-environment interactions. For gene-gene interaction, suppose that loci τ and $\tilde{\tau}$ contribute to a trait and put

$$Y = \mu + \alpha[g(\tau) - 2p] + \tilde{\alpha}[g(\tilde{\tau}) - 2\tilde{p}] + \gamma[g(\tau) - 2p][g(\tilde{\tau}) - 2\tilde{p}] + e.$$

The term involving γ is referred to as an additive-additive interaction term. A saturated model would have eight parameters: two additive effects, two dominance effects, and four interactions terms, additive-additive, additive-dominant, dominant additive and dominant-dominant. In addition to assuming for convenience that there are no dominance terms, we also assume that the loci τ and $\tilde{\tau}$ are on different chromosomes. This makes the various genetic contributions uncorrelated, hence simplifies numerous calculations. For example, the phenotypic variance can be written

$$\sigma_Y^2 = \alpha^2 2pq + \tilde{\alpha}^2 2\tilde{p}\tilde{q} + \gamma^2 2pq 2\tilde{p}\tilde{q} + \sigma_e^2 = \sigma_A^2 + \sigma_{\tilde{A}}^2 + \sigma_{A\tilde{A}}^2 + \sigma_e^2.$$

To test for these genetic effects we use a multiple regression statistic consisting of regression of the phenotype on the genotype at a putative trait locus s and at a putative trait locus at t, combined with regression on the product (interaction) $[g(s) - 2p_s][g(t) - 2p_t]$. For each fixed s and t, we could write this statistic asymptotically as a chi-square statistic with three degrees of freedom of the form $Z_{1,s}^2 + Z_{2,t}^2 + Z_{3,s,t}^2$, and for a genome scan we would have to maximize this over $s < t$. It is possible to obtain an approximation to the false positive rate (significance level). The mathematical expression is complicated, but easy to evaluate numerically. Since we are searching over two indices $s < t$, the genome-wide significance threshold is much larger than in the case of a standard genome scan.

For a numerical example we consider again an idealized mouse genome consisting of 20 chromosomes of genetic length 0.8 M for a total genome length of 16 M. For an intercross, where we test only for additive effects, or a backcross, $\beta = 2/\mathrm{M}$. Assume there are markers placed at 1 cM intervals.

As noted above, for a single locus genome scan the threshold for a genome-wide false positive rate of 0.05 is $z_{\max} \approx 3.79$ (or 14.1 on the chi-square scale). The comparable threshold for a two locus scan with no interaction permitted is 5.06 (or 25.6) and for a two locus scan with interaction is 5.58 (or 31.1). Note that the increase in threshold for a two locus scan is quite considerable without the statistic to test for interaction, while the third degree of freedom to test for interaction adds comparatively little.

Remark. The Bonferroni bound for the threshold of 3.79 is 0.24. If we choose our threshold based on the Bonferroni bound, it would be 4.16. In this case the Bonferroni bound is very conservative.

An advantage of experimental genetics is the possibility of maintaining a controlled environment, so differences in phenotype especially humans, it may be important to model environmental effects, so phenotypes that appear to arise from genetic variability do not simply reflect environmental variability (or some combination of both). If certain environmental conditions are thought to play an important role in the trait, one can introduce them into a genetic model.

To study the effect of an environmental variable w, which in experimental genetics can be controlled to take only a small number, perhaps only two, values, one can consider a multiple regression model that contains both a purely environmental effect and an additive gene-environment interaction, by writing

$$Y = \mu + \beta(w - \bar{w}) + \alpha[g(\tau) - 2p] + \gamma[g(\tau) - 2p](w - \bar{w}) + e,$$

where $\bar{w}$ is the sample average of w. We consider w to be non-random, either because it is chosen by the experimenter in experimental genetics, or we assume that we condition on its value in samples from natural populations.

The variance of Y is

$$\sigma^2_{Y,w} = E[Y - \mu - \beta(w - \bar{w})]^2 = 2pq(\alpha^2 + 2\alpha\gamma(w - \bar{w}) + \gamma^2(w - \bar{w})^2 + \sigma^2_e.$$

Assuming that we regard the purely environmental effect measured by β as a nuisance parameter, we can create a two-dimensional statistic

by regressing the phenotype on $[g(t) - 2p_t]$ and on the product term $[g(t)-2p_t](w-\bar{w})$. To estimate the variability of the estimated regression coefficients, we would also estimate μ and β (under the null hypothesis that there are no genetic effects) and use the appropriate residual sum of squares. After standardization, the result at marker locus t is the vector $Z_t = (Z_{1,t}, Z_{2,t})'$, which under the null hypothesis of no linkage has asymptotically independent standard normal entries. The squared norm of this vector would be chi-square on two degrees of freedom, and $\max_t ||Z_t||$ can be used in a genome scan. An approximation for the false positive rate is omitted.

Remark. For studying disease traits in humans a case-control design is common. This amounts to sampling individuals based on the value of their phenotypes. When there are no environmental covariates, we can develop a model for the conditional distribution of the genotypes given the phenotypes and define define the false positive rate in terms of that conditional distribution, so it makes essentially no difference how the phenotypes are selected. The situation is more complicated when there are environmental covariates, since sampling based on the values of the phenotypes will in general also affect the values of the covariates.

Estimation of Genetic Effects

Inverting Tests. Suppose X has distribution P_ξ, which is stochastically increasing in ξ. Given x, define $\tilde{\xi}$ to be the inf of all ξ such that $P_\xi\{X \geq x\} \geq \alpha$. Then $P\{\tilde{\xi} \leq \xi\} = 1-\alpha$, i.e., the set of all values ξ NOT rejected at significance level α give a $1-\alpha$ confidence bound for ξ.

Lower Confidence Bound for a Genetic Effect. Hypothesize that there exists at most one trait locus, of effect at most ξ, and consider the statistic $\max_t Z_t$. Let $G_0(b, C)$ and $G_1(b, \xi)$ denote approximations for the significance level and power, respectively. Then a $1-\alpha$ lower confidence bound for ξ is the solution of $1-[1-G_1(b,\xi)][1-G_0(b, C-1)] = \alpha$. In particular, for $\alpha = 0.5$, this equation defines a median unbiased estimator.

For a numerical example, suppose $Z_{\max} = 3.4$, which has an associated genome-wide p-value of 0.17. The suggested median unbiased estimator of ξ is 2.77; and a lower 0.8 confidence bound is 1.1. For $Z_{\max} = 5$ the median unbiased estimator is 4.8; for $Z_{\max} = 3.1$, it is 1.8. The 0.8 lower confidence bounds are 3.95 and 0, respectively.

Recombinant Inbred Lines and Other Breeding Designs

Experimental genetics provides the possibility of a wide variety of breeding designs, in order to increase the power of gene mapping experiments. One example is provided by recombinant inbred lines (RILs). In mouse genetics, these may be obtained by the following process. After an intercross of two inbred mouse strains, a (preferably large) numbers of brother-sister pairs are mice are selected to become the first generation of the recombinant inbred lines and are repeatedly crossed for a large number of generations. Eventually the mice in each line lose all heterozygosity, so at each locus they are equally likely to be A_1A_2 or A_2A_2, and within each line all mice are genetically identical. This has several consequences.

1. Since mice are genetically identical within each line, we can use, say, the mean phenotype of k mice from each line, so the residual variance σ_e^2 becomes σ_e^2/k.

2. Because there are no heterozygotes, one cannot test for a dominance effect. However the additive variance increases from $\alpha^2/2$ for an intercross to α^2 for the recombinant inbreds.

3. Since genotypes are fixed within lines, we need genotype only one mouse from each line. These genotypes can be stored and used again and again for mapping different phenotypes.

4. The recombination parameter $\beta = 4/\mathrm{M}$, so the peaks identifying the genomic locations of trait loci are roughly half as wide as in an intercross.

5. Production of a reasonably large number of RILs is very expensive, so very few have been produced until recently.

Exercise. For the model $Y = \mu + \alpha[g(\tau) - 1] + e$, where for each marker locus t, $g(t)$ is equally likely to be 0 or 2, what is the asymptotic noncentrality parameter of a regression statistic based on a sample of size N (only one observation per RIL) at a marker t linked to τ in terms of α and $\sigma_e^2 = E(e^2)$?

Population Based Association Studies

For a sample from a human populaton, assumed to be in Hardy-Weinberg equilibrium (more later about this assumption), we consider the same simple model for a phenotype:

$$Y = \mu + \alpha[g(\tau) - 2p_\tau] + \delta[g_M(\tau) - p_\tau][g_F(\tau) - p_\tau] + e.$$

Again we assume that e is independent of the genetic effect at τ, although it can implicitly contain genetic effects of other loci.

Again we can define regresson statistics, either a one dimensional statistic where we assume that $\delta = 0$ and test only for an additive effect, or a two dimensional statistic for the general model. For each marker locus t, the standardized statistics will be either univariate standard normal or uncorrelatd bivariate standard normal. For simplicity, suppose that we test for an additive effect. Let

$$Z_t = \sum_i (Y_i - \bar{Y})[g_i(t) - \bar{g}(t)] / \{\hat{\sigma}_Y^2 \sum_i [g_i(t) - \bar{g}(t)]^2\}^{1/2},$$

which is asymptotically standard normal when there is no association (i.e., correlation due to linkage disequilibrium) between a trait locus τ and the marker locus t.

An important difference with experimental genetics is that there is no mathematical model for the correlation between Z_t and Z_s. The asymptotic noncentrality parameter for the additive statistic based on a sample of size N at a marker t having correlation r, due to linkage disequilibrium, with the trait locus τ is

$$rN^{1/2}\sigma_A/\sigma_Y.$$

This is similar to what we found earlier, but the usually unknown parameter r has replaced the simpler $\exp(-\beta\Delta)$, which gives an analogous correlation (due to linkage) in terms of the genetic distance Δ between marker and trait locus.

Since there will usually be many generations during which linkage disequilibrium is breaking down, the correlation r between marker and

trait locus in natural popoulations is usually much smaller than that in breeding experiments, where there are usually only a few generations to break down this correlation. This has two important consequences: (i) enormous numbers of markers are necessary to scan an entire genome, since peaks indicating association are very narrow, and (ii) the genomic location of trait loci can be estimated relatively precisely.

Recall also that there is no automatic correspondence between genetic distance and the correlation due to linkage equilibrium. Although it is rarely the case, in some populations, for example, those that are mixtures of two or more subpopulations, the correlation r might be quite large between the trait locus and marker loci that are nowhere near the trait locus, hence may give a false impression about the location of the trait locus.

Association Mapping with Family Controls

The analysis of the preceding section depends critically on the assumption that the genotypes in supposedly unrelated individuals are actually independent. The test statistic given here does not presuppose this assumption. We use the same model as before. We assume that we have a sample of parent-child trios, i.e., two parents and one child. At the dialleleic marker t, let $M_{i,t}$ denote the indicator that the mother is heterozygous, and let $F_{i,t}$ be the indicator that the father is heterozygous. At a marker locus where a parent is heterozygous, the dhild is equally likely to inherit either of the parental alleles. To test for an additive genetic effect, we consider the regression statistic

$$Z_t = \frac{\sum_i \{(Y_i - \bar{Y})[M_{i,t}(g_{Mi}(t) - 1/2) + F_{i,t}(g_{Fi}(t) - 1/2)]\}}{\{\hat{\sigma}_Y^2 \sum_i [M_{i,t} + F_{i,t}]/4\}^{1/2}}.$$

A lengthy calculation shows that the asymptotic noncentrality parameter of this statistic is

$$(1 - 2\theta)(N/2)^{1/2} r \sigma_A / \sigma_Y,$$

where θ is the recombination fraction between the trait and marker loci, and r, as before is the correlation due to linkage disequilibrium.

The differences between this noncentrality parameter and that for population based mapping are two-fold: (1) the factor $(1 - 2\theta)$ that appears here, and (2) the factor $1/2^{1/2}$. The first factor shows that there is no signal to detect association between the genotype and phenotype unless there is linkage between the two. This eliminates the concern over spurious correlation that exists in population based studies. Moreover, because the correlation r is usually very small unless the marker is very close to the trait locus, usually θ will be close to 0, so the factor $1-2\theta$ in effect does not reduce the noncentrality parameter. However, the factor $1/2^{1/2}$ does reduce the noncentrality parameter, and means that this statistic will be considerably less powerful than the population based statistic—unless the apparent power of a population baed statistic is bought at the price of an uncontrolled significance level.

Note also that there will be considerably greater cost in obtaining phenotype and genotype data from an individual *and* the parents of that individual.

A Numerical Example

Suppose $\sigma_A^2 = 0.25, \sigma_e^2 = 0.75$, so $\sigma_y^2 = 1$, and assume the heritability is 0.5. Assume that markers are equally spaced at intermarker distance $\Delta = 0.01$ M. We detect linkage in the neighborhood of any genomic position t, where $Z_t \geq b$. Assume that the trait locus τ is itself a marker, so there is no deterioration in the signal strength between trait locus and marker locus.

Intercross in mice: 20 chromosome pairs of average length 0.8 M. For $b \approx 3.79$ the genome wide significance level is 0.05. To have 90% power to detect a QTL having noncentrality ξ, one needs $\xi \approx 4.8$, where

$$\xi = N^{1/2}\sigma_A/\sigma_Y.$$

A sample size of about 93 mice is required for 90% power.

Association in humans: To detect association with 500K SNP markers in humans (hence an average intermarker distance of 0.000066 M), a threshold of about $b = 5.25$, is required. Assuming an average correlation of 0.7 between the functional polymorphism and some SNP marker, a sample size of about $N = 350$ is required for 90% power.

Admixture Mapping

Admixture occurs when two or more populations merge to form a new population. A classical example in humans is the African-American population. When this occurs, it may be possible to map a phenotype that has different (average) values in the two populations by regressing the individuals' phenotypes on the population origin of their genotypes.

For a model we assume population I is the origin of a proportion π of the genes in an admixed population, while population II is the origin of $1-\pi$ of the genes. Within population I the allele A at a particular genetic marker has frequency p_1, while that same allele has frequency p_2 in population II. Then $\bar{p} = \pi p_1 + (1-\pi)p_2$ is the frequency of A in a random selection from the population. We use the same model for the phenotype as before, but we are now interested not directly in g_M and g_F, but in (for example) $E(g_M|\chi_M)$,, where χ_M is the indicator that the maternal allele came from population I. It may be shown (Exercise: prove this) that $E(g_M - \bar{p}|\chi_M) = (p_1 - p_2)(\chi_M - \pi)$, and consequently that at a QTL

$$E(Y|\chi_M, \chi_F) = \mu + \alpha(p_1 - p_2)[\chi_F + \chi_M - 2\pi] + \delta(p_1 - p_2)^2[(\chi_M - \pi)(\chi_F - \pi)].$$

Hence we can define our statistic to test for an additive effect at marker locus t by

$$Z_t = \frac{\sum_i [(Y_i - \bar{Y})(\chi_{M,i}(t) + \chi_{F,i}(t) - 2\pi)]}{[n\hat{\sigma}_Y^2 2\pi(1-\pi)]^{1/2}}.$$

Its noncentralilty when the marker t equals the QTL is

$$n^{1/2}\alpha(p_1 - p_2)[2\pi(1-\pi)]^{1/2}/\sigma_Y.$$

Note that the noncentrality parameter depends on the difference between the allele frequencies in the two populations, but otherwise it is of the same form as the noncentrality parameter for an intercross.

As above we can assess significance thresholds for $\max_t Z_t$ if we know the covariance function, which depends on the the length of time and the nature of the admixture. Simple models suggest that for t and s separated by recombination fraction θ, $\mathrm{Cov}(Z_s, Z_t) \approx (1-\theta)^g$, where g is the number of generations since the initiation of admixture.

Remark. A natural question is whether one can combine information from an association study with admixture data.

An Example: eQTL Mapping

The study of Morley *et al.* (2004) involved the expression levels of 3000 genes as phenotypes and 2800 markers in 14 CEPH familes consisting of sibships of size 8. To control the number of false positives, Morley *et al.* used a threshold of $b = 4.94$, which they said was appropriate for a genome-wide signficance level of 0.001, hence about three false positives.

However, for sibships of size 8, the skewness of the score statistic is approximately $\gamma = 1.6$. For $b = 4.94$, a more refined approximation, which corrects for skewness, gives a significance level of about 0.016, so one should expect about 48 false positives in 3000 phenotypes. (Simulations gave the value of 0.0195 ± 0.0035.) To achieve a significance level of about 0.001, one should take $b \approx 5.8$.

Refined p-value approximation

With markers equally spaced at distance Δ, an approximation for the significance level, which accounts for skewness in large pedigrees, is

$$P_0\{\max_{0 \le i\Delta < L} Z_{i\Delta} > b\}$$

$$\approx [2\pi(1+\gamma\theta)]^{-1/2}\{1/\theta + \nu\beta L b\} \exp[-\theta^2(1+2\gamma\theta/3)/2].$$

Here γ is a measure of skewness, $\theta = [-1+(1+2b\gamma)^{1/2}]/\gamma$, β depends on the pedigree structure ($= 4/\mathrm{M}$ for sibships), and $\nu = \nu[b(2\beta\Delta)^{1/2})$, where $\nu(2x) \approx x^{-1}[\Phi(x) - 1/2]/[x\Phi(x) + \varphi(x)]$.

References

Abecasis,G.R., Cherny,S.S., Cookson,W.O., Cardon,L.R., (2002). Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97-101.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees, *Am. J. Hum. Genet.* **62** 1198-1211.

Balding, D. (2006). A tutorial on statistical methods for population association studies, *Nature Reviews: Genetics* **7** 781-791.

Churchill, G.A. and Doerge, R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-971.

Dupuis, J., Shi, Jianxin, Manning, A., Benjamin, E., Meigs, J., Cupples, L.A., and Siegmund, D. (2009). Mapping quantitative traits in unselected families: algorithms and examples, *Genetic Epidemiology* **32**, 1-9.

Feingold, E., Brown, P.O., Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent, *Am. J. Hum. Genetics*, **53**, 234-251.

Fisher, R.A. (1918). The correlation of relatives on the assumption of Mendelian inheritance, *Proc. Roy. Soc. Edinburgh.*

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics* **121**, 185-199.

Lander, E. S. and Schork, N.J. (1994). Genetic Dissection of complex traits, Science **265:** 2037-2048.

Siegmund D. and Yakir B. (2007). *The Statistics of Gene Mapping*, Springer, New York.

Tang, Siegmund, D., H. Johnson, N., Romieu, I., London, S. (2010) Joint testing of genotype and ancestry association in admixed families, *Genet. Epidemiol.*