Reading: Chapter 13 and Jarvik GP (1998) Complex segregation analyses: uses and limitations. Am J Hum Genet 63:942-946.

There are a number of reasons why there is interest in the detection of major genes. For us, the two most relevant are:

- the potential to isolate and characterize genes influencing our trait of interest (whether good or bad)
- to be able to adjust for these since a good deal of quantitative genetic theory assumes that a trait is influenced by many genes of small effect

There are various methods available that test for the presence of a major gene:

- for populations in which controlled breeding is possible

- methods that rely on departure from normality to indicate the presence of a major gene

- methods that are based on resemblance between parents and offspring: Bartlett's test, Fain's test, Major-gene Indices (MGI)

Of greater interest to us will be

- mixture models (does not take into account relatedness of individuals)
- complex segregation analysis (used for data consisting of pedigrees)

<u>Mixture Models and Commingling Analysis</u>

In general, we can write the distribution of a trait that is a mixture of *n* underlying distributions as:

It is usual to assume that the population is in Hardy-Weinberg equilibrium at the major locus and that the variances across the underlying distributions are equal. For a diallelic locus that gives us a likelihood for individual *i* of:

and a full likelihood of

Maximum likelihood estimates can then be found for the 5 parameters.

Testing for Specific Modes of Inheritance

The likelihood ratio test can be used to test between various hypotheses:

where L_r corresponds to the likelihood function for which r parameters from the full model have been assigned values.

If the null hypothesis is not nested within the alternative hypothesis, then the large-sample distribution of the LRT statistic is not necessarily distributed chi-squared. One method for comparing nonnested hypotheses is Akaike's information content (AIC) where the model with the smallest AIC is preferable:

As an example of commingling analysis, consider the results of Knoblauch et al. (2000) A Cholesterol-lowering gene maps to chromosome 13q. Am J Hum Genet 66:157-166. They performed a commingling analysis on a subset of individuals within familial hypercholesterolemia families. The trait of interest was low density lipoprotein (LDL) which was adjusted for age effects.

Excerpt from paper:

> In the FH-heterozygous family members, we found a modest, albeit significant ($P <$ .04), effect of age on LDL values, allowing us to correct for the effects of age. Figure 1 shows the frequency distribution of corrected LDL-cholesterol values in FH-heterozygous subjects. Commingling analysis using ILINK gave significant evidence ($P <$ .03) against a unimodal distribution of the corrected LDL values. The mean LDL value of the lowest genotypic distribution in the model was 138 mg/dl; the SD was 29 mg/dl. The other means were close together, at 198 and 203 mg/dl, respectively. This finding was interpreted as indicating a recessive trait, which was in accordance with a segregation analysis that used LOKI (Heath 1997).

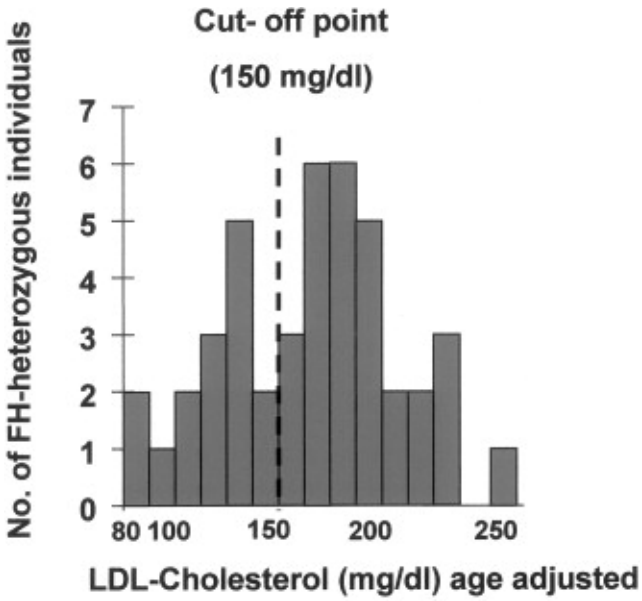


Figure 1 Frequency distribution of the subjects, in terms of their corrected LDL-cholesterol concentrations. Commingling analysis by use of ILINK allowed us to establish phenotypic criteria in terms of being "affected" by a putative cholesterol-lowering gene (LDL cholesterol <150 mg/dl).

Problems with Non-normality

Non-normality can be accommodated within the likelihood function through the use of the Box-Cox transformation:

with each of the underlying distributions having a maximum of 3 parameters. This allows us to test for whether an observed skewed distribution results from

- a single naturally skewed distribution

- a mixture of underlying normals

- a mixture of underlying distributions that can be transformed to normality

Modeling skewness

Given

z = observed phenotype, i.e. observed trait value

the distribution of z WITHIN each of the genotypic classes can be transformed to a normal distribution through the use of the Box-Cox transformation:

For our scenario, it could be assumed that the individual distributions have the same transformation. That is,

Let us suppose that we can find the correct transformation. The distribution on the transformed scale is then:

If we assume that the observed phenotypic distribution is such that $0<y<\infty$ then

Our full likelihood is then

which represents a mixture of three distributions that have separate means, but the same variance and transformation parameter. It would obviously be possible to extend this such that you allow for unequal variances or unequal transformation parameters. The program NOCOM does not allow for different transformation parameters but can allow for separate variances (although it is not recommended). More operational characteristics can be found at:

http://linkage.rockefeller.edu/ott/nocom.htm

The NOCOM program was originally introduced in the following article:

Ott J (1979) Detection of rare major genes in lipid levels. Hum Genet 51:79-91

They used the program to analyze the distribution of triglycerides for 991 unrelated men in the Seattle area. Their results follow:

Table 4. Analysis of normal mixtures for triglycerides of 991 unrelated male index cases in Boman et al (1978), with log likelihood adjusted to zero at its lowest value

	1 component			2 components							P value 1 vs 2 components
	$\hat{\lambda}$[b]	$\log_e L$	P[a]	$\hat{\lambda}$[b]	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{v}$	$\hat{h}$[c]	$\log_e L$	P[a]	
$\lambda = 0$	—	0	0.070	—	4.55	5.36	0.17	0.078	7.455	0.617	0.003
λ est.	−0.15	4.364	0.295	0.15	6.68	8.57	0.62	0.085	8.949	0.588	0.010

[a] Significance level in goodness-of-fit test
[b] Exponent in power transformation (6)
[c] Proportion of component with mean μ_2

They had reason to believe that triglycerides were distributed log-normal within each sub-distribution (one distribution or more). From their results, what would conclusions would you draw concerning the underlying genetics of triglycerides?

Final words on commingling analysis

- the trait distribution could be skewed due to non-genetic reasons
- it is hard to distinguish a skewed distribution from a mixture of normals
- if we had a mixture of a number of different **skewed** distributions, then by assuming a mixture of normals our modeling will be incorrect and so we might expect a loss in power to detect commingling

Through our discussion of commingling analysis, we assumed that we had a sample of *n* random (unrelated) individuals. This gave us a likelihood that was the product over the phenotypic distribution for these *n* individuals.

Obviously, if our sample contains related individuals, then this likelihood will no longer be correct as related individuals will have correlated trait values (recall the information that we covered in Chapter 7).

IN ADDITION, through the use of related individuals, we would expect to have increased power to detect a major gene due to the extra information given by transmissions from parents to offspring.

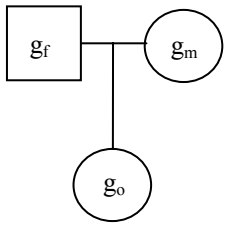
The extension of the simple mixture model of commingling analysis to the use of pedigree information is known as . . .

Complex Segregation Analysis

Consider a diallelic QTL locus with alleles Q and q. Genotypes are encoded by

$$g = \begin{cases} 1 & for \quad QQ \\ 2 & for \quad Qq \\ 3 & for \quad qq \end{cases}$$

For the i^{th} family, we have



The likelihood for the j^{th} offspring of the i^{th} family CONDITIONAL on the parent's genotypes is:

It is important to note that *g* represents the genotype **AT THE QTL** which we cannot actually measure. Also, note that this is **NOT THE SAME AS LOOKING AT A MARKER**.

Example 4.1: What is the likelihood for an individual with (*Qq, Qq*) parents if we assume Mendelian segregation at the QTL?

Thus for the i^{th} family we have the CONDITIONAL likelihood:

Recall, that the parental genotypes are unknown and so we must sum over all possible parental genotypes. This gives us the UNCONDITIONAL likelihood:

If it is assumed that the QTL is in Hardy-Weinberg Equilibrium, then the frequencies for the parental genotypes can be written as our well-known functions of p=Pr(Q). The overall likelihood is then a function of five unknown parameters:

To lead to a more robust approach, it was suggested by Elston et al. (1975) that

$\Pr(g_o \mid g_m, g_f)$

be modeled as unknown parameters:

τ_x=Pr(genotype x transmits a Q allele)

Example 4.2: Given these additional parameters, what is the conditional likelihood for the trait value of a child given (*Qq,QQ*) parents?

In order to feel strongly about the presence of a major-gene, it is recommended that

1. there is a significantly better overall fit of a mixture model compared with a single normal
2. the hypothesis of Mendelian segregation is NOT rejected
3. while the hypothesis of equal transmission for all genotypes is rejected

1⇒

2⇒

3⇒

In addition to the genetic effects due to the major locus, polygenic background can be added to the model. This is done by assuming the background polygenes are completely additive and that the background genetic value *A* is normally distributed with mean 0 and variance σ_A^2.

Example 4.3: The following table is from a segregation analysis of Radiosensitivity. It can be found in Roberts SA (1999) Heritability of cellular radiosensitivity: a marker of low-penetrance predisposition genes in breast cancer? Am J Hum Genet 65: 784-794

Table 3: **Model Parameters from Segregation Analysis of G_2 Radiosensitivity in 20 Families**

	Model 1: General Transmission and Polygenes	Model 2: Major-Gene Only	Model 3: Sporadic	Model 4: Polygene Only	Model 5: Major-Environmental Only	Model 6: Major Gene + Polygene	Model 7: Major-Gene-Only, Non-Mendelian	Model 8: General-Transmission Only
Allele frequency, *p*	.96	.96	NA	NA	.77	.96	.98	.94
Means:								
nn	4.48	4.48	4.63	4.67	4.48	4.48	4.48	4.47
ns	4.85	4.85	NA	NA	4.81	4.84	4.85	4.81
ss	5.27	5.27	NA	NA	5.17	5.30	5.28	5.18
Residual SD	.101	.101	.23	.21	.077	.10	.10	.077
Polygene heritability, *H*	0[a]	[0]	NA	.79	[0]	.30	[0]	[0]
Genotype-transmission probabilities:[b]								
$_1$	.96	[1]	NA	NA	[*p*]	[1]	[1]	.96
$_2$	.37	[.5]	NA	NA	[*p*]	[.5]	.35	.37
$_3$	.29	[0]	NA	NA	[*p*]	[0]	[0]	.29
2 × Log likelihood	76.6	68.5	9.8	18.1	50.8	71.2	71.4	76.6
No. of fitted parameters	9	5	2	3	5	6	6	8
Likelihood-ratio ^{2}-test *P* values:								
Compared with model 1	NA	.090	<.001	<.001	<.001	.15	.16	1.0
Compared with model 2	.090	NA	<.001	NA	NA	.11	.087	.045

NOTE.NA = not applicable. Parameters in square brackets were held fixed in the model.
[a] Fitted parameter value at its lowest boundary.
[b] $_1$, $_2$, and $_3$ take values of 1, .5, and 0, respectively, for Mendelian inheritance (see Subjects and Methods).

Other extensions of CSA

CSA has been extended in several ways. A good review is given in the following paper:

Jarvik GP (1998) Complex segregation analyses: uses and limitations. Am J Hum Genet 63:942-946

CSA has been extended to allow for:

- multivariate traits
- genotype $\times$ environment interaction
- general pedigrees
- various computational procedures for the likelihood function
- nonrandom ascertainment

Complex Segregation Analysis of Discrete Characters

Consider a dichotomous trait which is coded as follows:

$$y = \begin{cases} 0 & normal \\ 1 & diseased \end{cases}$$

Define the penetrance, ψ_g, of a genotype to be the probability that an individual with genotype *g* is diseased:

ψ_g=Pr(*y=1* | *g*).

The likelihood function for the trait of an individual with genotype *g* is then:

$\ell(y \mid g) =$

Arguments for constructing the likelihood for traits measured on a collection of siblings are the same as for a continuous trait. The conditional likelihood for a child's trait given their parental genotypes is:

$\ell(y \mid g_m, g_f) =$

The unconditional likelihood is then:

$\ell(y) =$

A polygenic background can also be added to the model. This can be done first by allowing the penetrance to be a function of the polygenic backround, $\psi(g, A)$.

The likelihood for an individual's trait given their genotype, g, and polygenetic background, A, is then:

$\ell(y \mid g, A) =$

This can be used to compute the likelihood for a child' trait conditional on their parental genotypes and polygenic backgrounds:

$\ell(y \mid g_m, g_f, A_m, A_f) =$

Relating a discrete and quantitative traits

The penetrance functions, $\psi(g, A)$, can be modeled assuming an underlying LIABILITY MODEL.

For instance, suppose disease occurs once the liability exceeds some threshold T. The normal condition then corresponds to individuals with liability under T:

$$\begin{aligned}\psi(g,A) &= \Pr(z > T \mid g, A) \\ &= \int_T^{\infty} \varphi(z; \mu_G + A, 1)dz\end{aligned}$$

Defining the cumulative distribution function for a unit normal U as $\Phi(x) = \Pr(U < x)$ then:

Example 4.4: Use this relationship to write the likelihood for a child's trait value conditional on their parental genotypes being $g_m=1$ and $g_f=2$.

That is, derive $\ell(y \mid g_m = 1, g_f = 2, A_m, A_f)$.

More information regarding threshold characters is given in Chapter 25.

Before we move on to methods for linkage analysis of quantitative traits, we're going to cover a type of aggregation/segregation analysis for dichotomous traits that relies heavily on the biometrical model.

The ***biometrical model*** corresponds to Fisher's decomposition of the genotypic value using least squares to fit additive effects for each allele with dominance effects being the residuals to this fit.

The following reference for this material has been placed in the class file:

> Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222-228

Recurrence risk of disease in relatives

The **relative risk ratio** represents the increased risk of disease given an individual is related to a diseased person. This is often denoted by λ_R and is equal to:

$$\lambda_R = \frac{\Pr(\text{relative of type R is diseased} \mid \text{diseased individual})}{\Pr(\text{diseased individual})}$$

The numerator is often called the relative recurrence risk.

- We will use the biometrical model to develop a general framework for computing the relative recurrence risk of disease as a function of the underlying genetic parameters and the relative type.

- This framework has proven useful for studying patterns of risk across relatives as an indicator for the underlying mode of inheritance. In particular, it has proved useful for separating additive and multiplicative risk models.

For example, Risch[1] studied Schizophrenia and compared observed relative recurrence risk across many types of relatives with what is expected from seven different disease models:

Table 1

Multilocus Multiplicative Models for Schizophrenia

		Model Prediction[b]						
Risk Ratio[a]	Observed	I	II	III	IV	V	VI	VII
λ_O	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
λ_S	8.6							
λ_M	52.1	19.0	100.0	75.0	55.6	43.8	56.3	42.2
λ_D	14.2							
λ_H	3.5							
λ_N	3.1							
λ_G	3.3							
(λ_2).	3.2	5.50	3.16	3.35	3.65	3.95	3.56	3.77
λ_C.	1.8	3.25	1.78	1.87	2.03	2.20	1.96	2.07

[a] Definitions of subscripts: O = offspring; D = DZ twins; H = half-sibs; N = niece/nephew; G = grandchild; C = first cousins. All other subscripts are as defined in the text.

[b] Definitions of models: I—one locus, $\lambda_{10} = 10.0$; II—infinite loci, each with small effect; III—$\lambda_{10} = 2.0$, infinite other loci; IV—$\lambda_{10} = 3.0$, infinite other loci; V—$\lambda_{10} = 4.0$, infinite other loci; VI—$\lambda_{10} = \lambda_{20} = 2.0$, infinite other loci; VII—$\lambda_{10} = \lambda_{20} = \lambda_{30} = 2.0$, infinite other loci.

Let's see how this is done for a single-locus model.

We'll represent our disease trait by X where X=1 indicates a diseased individual and X=0 indicates a non-diseased individual. Further, let X_1 be the indicator of disease for an individual with X_2 the indicator for their relative. With this notation,

$K = Pr(X_i=1) = Pr(\text{individual is diseased}) = \text{population prevalence}$

and

$K_R = Pr(X_2=1 \mid X_1=1) = \text{relative recurrence risk.}$

This leads to the relative risk ratio being equal to

$$\lambda_R = \frac{K_R}{K}$$

[1] Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222-228

What is K equal to? For the single-locus model, assume there are two alleles, D and N, that are in Hardy-Weinberg Equilibrium with p=Pr(D).

$$\begin{aligned} K &= \Pr(X_1 = 1) \\ &= \sum_{g \in \{genotypes\}} \Pr(X_1 = 1 \text{ and individual 1 has genotype g}) \\ &= \sum_{g \in \{genotypes\}} \Pr(X_1 = 1 \mid \text{individual 1 has genotype g}) \times \Pr(\text{individual 1 has genotype g}) \\ &= \end{aligned}$$

Before we continue with the derivation of K_R, let's reconsider the **biometrical model** for the genotypic contribution to a trait. This will allow us to derive

$$\begin{aligned} K_R * K &= \Pr(X_2=1, X_1=1) \\ &= E(X_1 * X_2) \\ &= \mathrm{Cov}(X_1, X_2) + K^2 \end{aligned}$$

Consider the phenotype Y where Y represents a standardized version of X as described below. Let $\boldsymbol{f_{kl}}$ be the **penetrance** for genotype *(k,l)*. We will treating the penetrance values as we previously used genotypic values.

Let p_k and p_l represent allele frequencies for the k^{th} and l^{th} allele. We will also assume (for simplicity) that

$$E(Y) = \sum_k \sum_l f_{kl} p_k p_l = 0$$

Recall, the biometrical model breaks the genetic contribution into an effect due to each allele and the departure from these effects. We now apply this concept to the penetrances:

$$f_{kl} = \alpha_k + \alpha_l$$

The α_k and α_l are chosen to minimize the deviations $\delta_{kl} = \mu_{kl} - \alpha_k - \alpha_l$ using least squares. That is, they minimize $SS = \sum_k \sum_l (f_{kl} - \alpha_k - \alpha_l)^2 p_k p_l$. From our previous work, we then have:

$$\alpha_k = \sum_l f_{kl} p_l \ ,$$

and the variance in the genotypic contribution to the trait Y due to the additive effects of alleles or the **additive genetic variance** is

$$\sigma_a^2 = 2\sum_k \alpha_k^2 p_k$$

The variance due to departures from the additive effects of alleles or the **dominance genetic variance** is then

$$\sigma_d^2 = \sum_k \sum_l \delta_{kl}^2 p_k p_l$$

So that $\mathrm{cov}(Y_i, Y_j) = 2\Phi_{ij}\sigma_a^2 + \Delta_{7ij}\sigma_d^2$ where Φ_{ij} is the kinship coefficient and Δ_{7ij} is the coefficient of fraternity.

We can now complete our derivation:

$$\begin{aligned}
\lambda_R &= \frac{K_R}{K} \\
&= \frac{\Pr(X_2 = 1 \mid X_1 = 1)}{K} \\
&= \frac{\Pr(X_2 = 1, X_1 = 1)/K}{K} \\
&= \frac{\Pr(X_2 = 1, X_1 = 1)}{K^2} \qquad \text{(equation 4.1)} \\
&= \frac{E(X_1 X_2)}{K^2} \\
&= \frac{Cov(X_1, X_2) + K^2}{K^2} \\
&= 1 + \frac{2\Phi_{ij}\sigma_a^2 + \Delta_{7ij}\sigma_d^2}{K^2}
\end{aligned}$$

The last step was due to Y being a standardized version of X.

Example 4.5: Relatives of n degree are separated by n meioses. Parent-offspring pairs are first-degree relatives. Grandparent-grandchild and uncle-niece are second-degree relatives and first cousins are typical third-degree relatives. The following table summarizes the adjusted relative risk ratios for first, second and third-degree relatives.

R	Relative Type	Adjusted Risk Ratio (λ_R-1)
1	First-degree	$\frac{\sigma_a^2}{2K^2}$
2	Second-degree	$\frac{\sigma_a^2}{4K^2}$
3	Third-degree	$\frac{\sigma_a^2}{8K^2}$

Summarize how the adjusted relative risk ratios for first, second and third degree relatives compare. How does this information help in study planning?

Note that the above adjusted risk ratios are defined by equation 4.1 and the coefficients of kinship and fraternity. Below is a summary of these coefficients for reference.

Table 2.1: Coefficients of kinship and fraternity for common relationships in a non-inbred population. Adapted from Lynch and Walsh[2]

Relationship	Φ_{ij}	Δ_{7ij}
Parent-offspring	1/4	0
Grandparent-grandchild	1/8	0
Great grandparent-great grandchild	1/16	0
Half sibs	1/8	0
Full sibs, dizygotic twins	1/4	1/4
Uncle(aunt) – nephew(niece)	1/8	0
First cousins	1/16	0
Double first cousins	1/8	1/16
Second cousins	1/64	0
Monozygotic twins	1/2	1

[2] Lynch M and Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland

Extension to multiple loci

There are two general extensions that are used for relating susceptibility loci to disease risk: additive and multiplicative.

- For an additive model, the Pr(disease given genotype i at locus 1 and genotype j at locus 2) = Pr(disease given genotype 1 at locus 1) + Pr(disease given genotype 2 at locus 2). In this case,

  - $\lambda_R - 1 = \left(\frac{K_1}{K}\right)^2 (\lambda_{1R} - 1) + \left(\frac{K_2}{K}\right)^2 (\lambda_{2R} - 1)$ where K_i is the prevalence of disease due to locus *i* and λ_{iR} is the relative risk ratio for the i^{th} locus

  - λ_1-1 = 2(λ_2-1)=4(λ_3-1) where λ_i is the risk ratio for relatives of degree I (this is the same relationship as for a single locus)

  - we cannot distinguish a single locus from two (or multiple additive) loci

- For a multiplicative model, Pr(disease given genotype i at locus 1 and genotype j at locus 2) = Pr(disease given genotype 1 at locus 1) $\times$ Pr(disease given genotype 2 at locus 2). In this case,

  - $\lambda_R = \lambda_{1R}\lambda_{2R}$ for relatives of type R

  - $\lambda_2 = \lambda_{12}\lambda_{22} = 1/4(\lambda_{11}+1) \times (\lambda_{21}+1)$ where λ_{ij} is the relative recurrence risk for j degree relatives for locus i

  - $\lambda_3 = 1/16 \times (\lambda_{11}+3) \times (\lambda_{21}+3)$

  - note that the decay is greater than what is expected under the additive model

What does this provide us? Let's answer this in the context of the example from Risch[3]. Consider the following in your discussion:

- Based on the observed data, does an additive model fit the data well?
- Note that $\lambda_D > \lambda_S$. What does this tell you?
- Is there any evidence of dominance? The table on p18 might be useful here.
- The risk to monozygotic twins is highly underestimated when fitting an additive model. What does this imply?
- What appears to be the most consistent model(s)?

Table 1

Multilocus Multiplicative Models for Schizophrenia

		Model Prediction[b]						
Risk Ratio[a]	Observed	I	II	III	IV	V	VI	VII
λ_O	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
λ_S	8.6							
λ_M	52.1	19.0	100.0	75.0	55.6	43.8	56.3	42.2
λ_D	14.2							
λ_H	3.5							
λ_N	3.1							
λ_G	3.3							
(λ_2)	3.2	5.50	3.16	3.35	3.65	3.95	3.56	3.77
λ_C	1.8	3.25	1.78	1.87	2.03	2.20	1.96	2.07

[a] Definitions of subscripts: O = offspring; D = DZ twins; H = half-sibs; N = niece/nephew; G = grandchild; C = first cousins. All other subscripts are as defined in the text.

[b] Definitions of models: I—one locus, $\lambda_{10} = 10.0$; II—infinite loci, each with small effect; III—$\lambda_{10} = 2.0$, infinite other loci; IV—$\lambda_{10} = 3.0$, infinite other loci; V—$\lambda_{10} = 4.0$, infinite other loci; VI—$\lambda_{10} = \lambda_{20} = 2.0$, infinite other loci; VII—$\lambda_{10} = \lambda_{20} = \lambda_{30} = 2.0$, infinite other loci.

[3] Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222-228