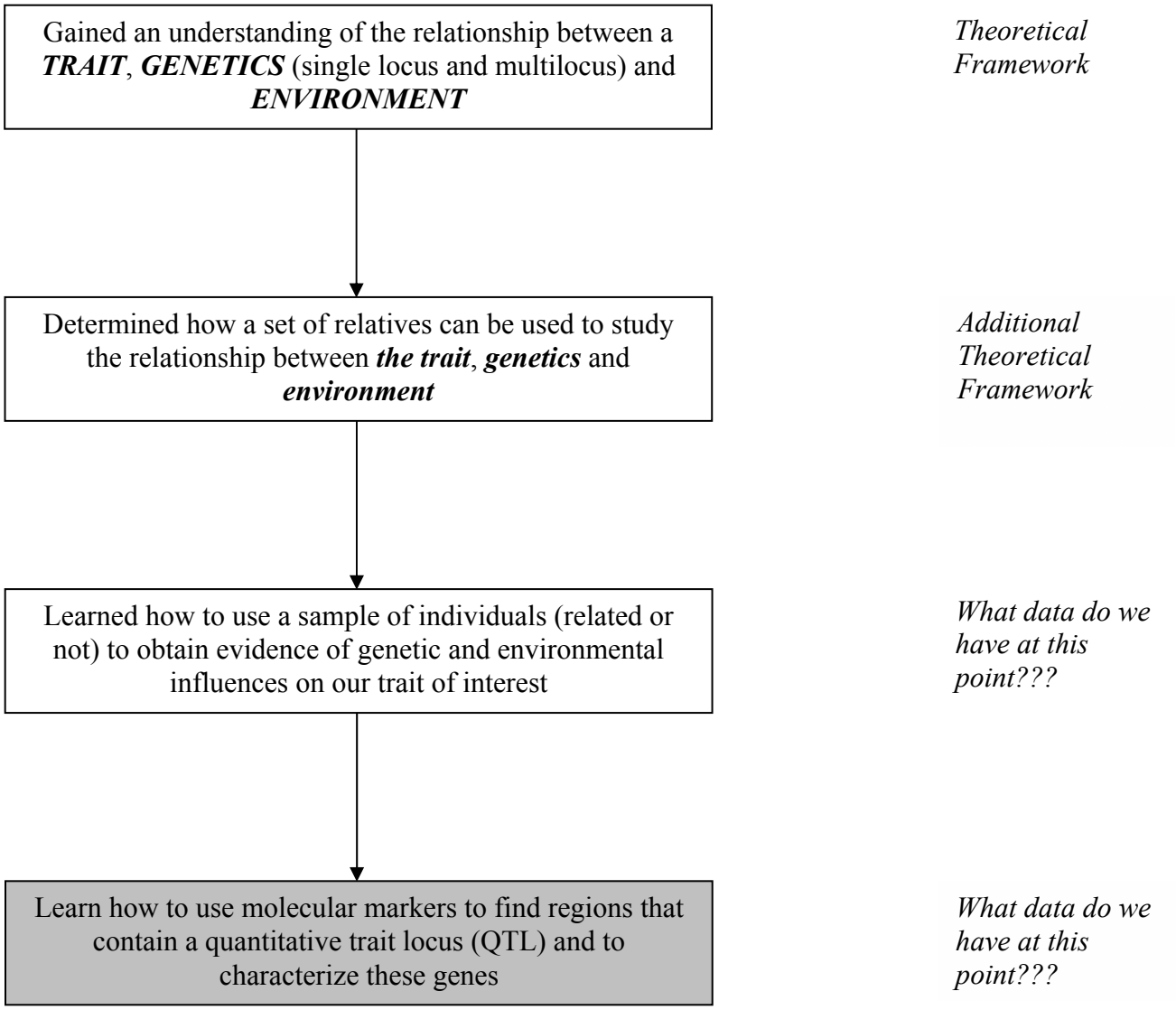
BIOSTAT/STAT551, Statistical Genetics II: Quantitative Traits Mapping and Characterizing QTLS **Winter 2004 Handout 5**

Reading: Chapter 16

ROAD MAP TO THIS POINT:



Molecular Markers

Techniques are available that measure variation at the molecular level. Pages 390-393 of the text provide a nice and clear summary of current methods.

	RFLP	RAPD	Microsatellite DNA	SNPs
Number of alleles/locus	Few	Few (Two)	Many	Two (generally)
Genomic Abundance	High	Very High	Medium	Very High
Marker dominance	Codominant	Dominant	Codominant	Dominant

Marker Informativeness

In order for a parent to provide linkage information, it must be heterozygous at both the trait and the marker.

Why?

Obviously only a small portion of our families will be informative. While we do not have control over heterozygosity at our trait locus, we can control heterozygosity at our marker locus.

Types of marker-informative matings:

- Fully informative

- Backcross

- Intercross

A common measure of marker-informativeness is the polymorphism information content (PIC). The PIC is the probability that the transmitted marker allele from **a parent** can be deduced for all offspring:

PIC = $\leq \frac{(n-1)^2(n+1)}{n^3}$

Another measure of informativeness is the proportion of fully informative matings (PFIM) or the probability that transmitted marker alleles can be distinguished for both of the parents in all offspring:

PFIM= $\leq \frac{(n-1)(n-2)(n+1)}{n^3}$

Figure 16.1 depicts the relationship between PIC, PFIM and heterozygosity:

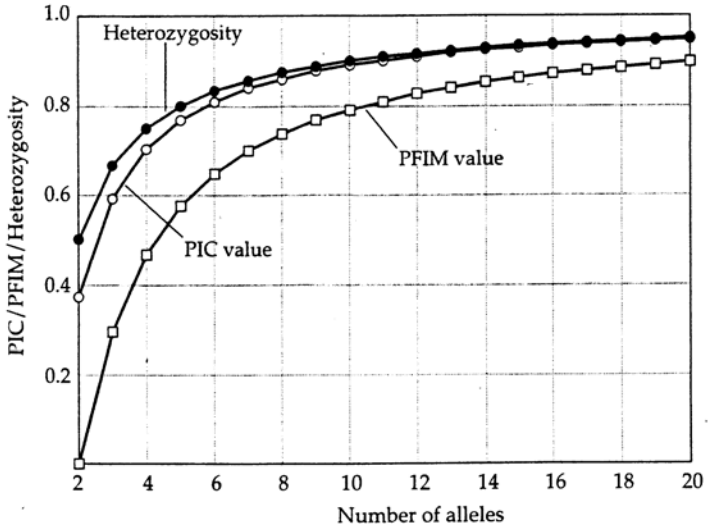


Figure 16.1 The relationship between three measures of marker information: heterozygosity, PIC, and PFIM. Plotted are the maximum values for each measure as a function of the number of marker alleles under the assumption of an even allele-frequency distribution ($p_i = 1/n$).

We already stated that for a parent to be informative for linkage that they must be heterozygous at both the marker and the QTL.

Suppose q_i=Pr(Q_i), i=1,...,n_Q. What is the probability that a parent is heterozygous at the QTL?

If we assume that the marker and QTL are in linkage equilibrium, then what is

- Pr(at least on parent fully informative)=

- Pr(both parents fully informative) =

Recall that it is possible to construct a likelihood for our quantitative trait parameters given trait information on a pedigree (complex segregation analysis):

$$\Pr(z_o \mid Q_M, Q_F) = \sum_{Q_o} \Pr(z_o \mid Q_o) \Pr(Q_o \mid Q_M, Q_F)$$

and

$$\Pr(z_o) = \sum_{Q_M} \sum_{Q_F} \Pr(z_o \mid Q_M, Q_F) \Pr(Q_M) \Pr(Q_F)$$

We now would like to add information for marker genotypes so that we can evaluate whether the recombination fraction, θ, between the marker and QTL differs from 0.5.

Consider the trait density for an offspring given marker information on the mother, father, and offspring in addition to QTL genotypes for the parents:

$$\Pr(z_o \mid M_o, M_M, M_F, Q_M, Q_F) =$$

What is $\Pr(Q_o \mid M_o, M_M, M_F, Q_M, Q_F)$?

Example 5.1: Suppose that you have the following information: M_o=*Aa*, M_M=*Aa*, M_F=*aa*, Q_M=*Bb* and Q_F=*bb*. Find $\Pr(Q_o = Bb \mid M_o, M_M, M_F, Q_M, Q_F)$.

The density for the offspring's trait can be obtained by averaging over all possible parental genotypes for the QTL.

$$\Pr(z_o \mid M_o, M_M, M_F) =$$

What if we have multiple offspring?

The likelihood ratio test can then be used to test for linkage between the marker and a QTL.

Various comments:

- single marker analysis for a quantitative trait is not very powerful (specifically referring to outbred populations)
- using additional markers will increase power (usage of flanking markers, multipoint analysis)
- choosing highly polymorphic marker loci will further increase power

The LOD score method explicitly models the transmissions from parents to offspring and requires the estimation of QTL allele frequencies. The likelihood could be extended to use data on multigenerational pedigrees.

- number of possible combinations of genotypes for individuals in the entire pedigree increases exponentially with the number of pedigree members
- an alternative is to construct likelihood functions using the variance components associated with a QTL (not specifying the details of the QTL model)

Basic idea behind variance components:

Notation:

Why R_{ij} for σ^2_A BUT $2\Theta_{ij}$ for σ^2_{A*}?

If we have a pedigree with *n* individuals, then for the vector of trait values $\underset{\rightarrow}{z}$

What is $\underset{\rightarrow}{R}$?

What is $\underset{\rightarrow}{A}$?

Assuming that $\underset{\rightarrow}{z}$ is multivariate normal:

Unknown parameters:

Example 5.2: In the following paper:

> Duggirala R et al. (1996) Quantitative variation in obesity-related traits and insulin precursors linked to the OB gene region on human chromosome 7. Am J Hum Genet 59:694-703

Variance components analysis was used to test for linkage for 15 markers on chromosome 7 for various obesity-related traits as well as associated metabolic traits.

Recall that our variance-covariance matrix is

$$V = \underset{\rightarrow}{R}\sigma_A^2 + \underset{\rightarrow}{A}\sigma_{A^*}^2 + \underset{\rightarrow}{I}\sigma_e^2$$

so that our hypotheses for a test for linkage are

H_o: $\sigma_A^2 = 0$
H_A: $\sigma_A^2 > 0$.

The likelihood ratio test can be used to test for linkage with the LRT statistic being approximately χ^2 with 1 degree of freedom.

Following are two figures from the paper that provide LOD scores for skinfold measurement and waist circumference, respectively.

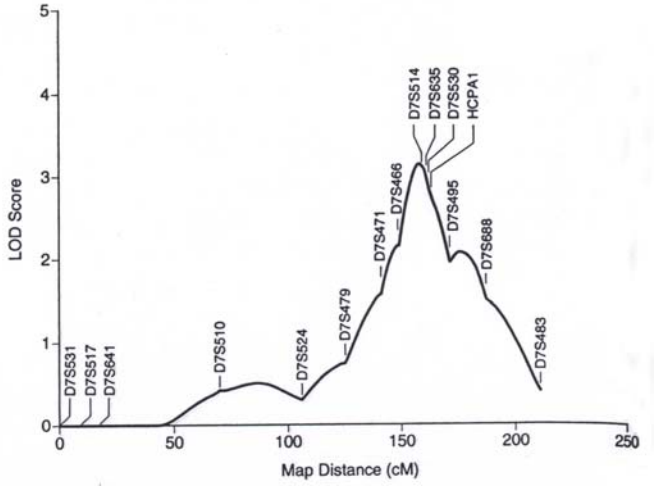


Figure 1 Linkage of microsatellite marker D7S514 locus in the *OB* gene region, with variation in extremity skinfolds: LOD scores vs. map positions on human chromosome 7.

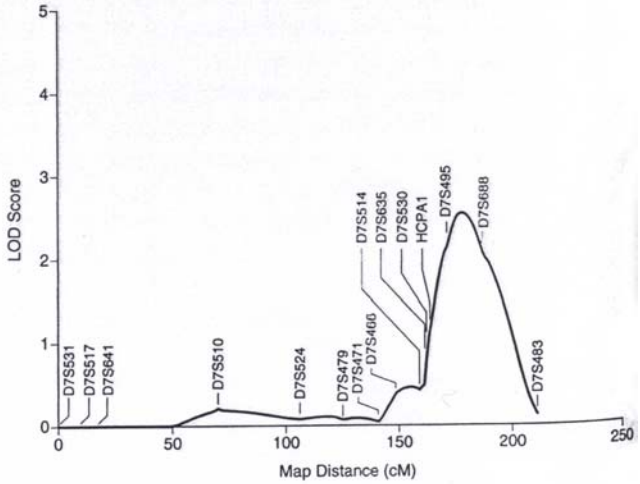


Figure 2 Linkage of a genetic location near microsatellite marker D7S495, with variation in waist circumference: LOD scores vs. map positions on human chromosome 7.

Recall that heritability is the fraction of the total phenotypic variation attributable to the additive genetic differences among individuals, i.e. the additive variance. Duggirala et al. provided estimates of heritability due to the variation corresponding to the markers of interest:

Table 4

Multipoint Variance-Components Analysis: Proportion of Total Phenotypic Variance in Anthropometric and Metabolic Traits That Is Attributable to the Susceptibility Loci for Which Maximum LOD Scores Were Obtained

Phenotype	Genetic Variance h_m^2 Explained[a] (%)	*P*-Value
Anthropometric traits:		
BMI	46.9 ± 13.0	.003
Fat mass	43.2 ± 18.2	.017
Trunk skinfolds	46.5 ± 9.7	.003
Extremity skinfolds	55.7 ± 10.1	.00014
Sum of skinfolds	45.8 ± 16.9	.003
Waist circumference	60.8 ± 8.5	.00063
Subscapular/triceps ratio	31.0 ± 10.4	.012
Metabolic traits:		
Glucose	0	
2-h glucose	0	
Fasting insulin	29.7 ± 16.2	.086
2-h insulin	44.9 ± 15.8	.019
Proinsulin	64.0 ± 14.4	.00013
Split proinsulin	73.2 ± 20.3	.000010

[a] By linkage to D7S514 in the case of extremity skinfolds, by linkage to a genetic location near D7S495 in the case of the other anthropometric traits, and by linkage to HCPA1 in the case of the metabolic traits excluding the two glucose phenotypes.

A few additional words:

- the variance components analysis was performed for locations without direct IBD information
- IBD information for untyped regions was inferred using marker IBD information
- criteria used to judge significance was from the following article

  Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241-247

  In short, the article recommended the term “suggestive linkage” for LOD scores of 1.9-2.4 and “significant linkage” for LOD scores of 3.3-3.8.

- a number of the traits under study are highly correlated so that some type of multivariate analysis could seem appropriate

The variance-components method discussed to this point does not implicitly model the location of a QTL. In fact, the method does not require the assumption of a single QTL in the area under study. However, the presence of a single QTL might be concluded from the results.

Incorporating the recombination fraction

In the paper,

> C.I. Amos (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535-543

the additional assumption of a single QTL in the region of study is made. Information regarding recombination between the marker and QTL can then be incorporated into the likelihood.

The following table contains the covariance for an additive major-gene component for several relationships:

Table I

Monogenic Components of Variance

Relative Pair	Component of Variance
Sibs	$[{}^1/_2+(1-2\theta)^2(\pi_{ij}-{}^1/_2)]\sigma_a^2$
Half-sibs	$[{}^1/_4+(1-2\theta)^2(\pi_{ij}-{}^1/_4)]\sigma_a^2$
Avuncular	$[{}^1/_4+(1-2\theta)^2(1-\theta)(\pi_{ij}-{}^1/_4)]\sigma_a^2$
Grandparental	$[{}^1/_4+(1-2\theta)(\pi_{ij}-{}^1/_4)]\sigma_a^2$
First cousin	$[{}^1/_8+(1-2\theta)^2(1-{}^4/_3\theta+{}^2/_3\theta^2)(\pi_{ij}-{}^1/_8)]\sigma_a^2$

We are now able to test for linkage and obtain an estimate of the location of the QTL relative to the marker:

Amos also performed simulations to estimate Type I error and power for samples of 40 nuclear families consisting of six sibs and their parents (100 replications were used for each estimate.

Table 3

Power and Significance of Simulation Studies

Simulation Experiment	Generating Model Trait Means					% of Tests Rejected at Nominal Level of			Median Parameter Estimates Under Full Model			Median Parameter Estimates Under Reduced Model		Estimated Genetic Variance
	μ_{AA}	μ_{AB}	μ_{BB}	σ_e^2	θ	10%	5%	1%	σ_a^2	σ_G^2	σ_e^2	σ_G^2	σ_e^2	
1	2	4	6	0	.5	8	7	2	0.00	1.74	.00	1.80	.00	1.80
2	2	4	6	0	.0	100	100	100	1.98	.00	.00	1.64	.10	1.98
3	2	4	6	1	.5	5	2	0	.03	1.68	1.07	1.83	1.07	1.82
4	2	4	6	1	.0	100	100	100	1.69	.00	.90	1.88	1.03	1.99
5	2	4	6	2	.5	5	3	1	.00	1.45	1.90	1.57	1.90	1.57
6	2	4	6	2	.0	94	92	81	1.46	.00	1.85	1.50	1.99	1.92
7	2	6	6	2	.5	7	4	0	.00	1.67	2.43	1.84	2.17	1.88
8	2	6	6	2	.0	96	93	86	1.96	.00	2.21	1.71	2.59	2.27
9	4	6	4	2	.5	2	1	0	.00	.29	2.51	.38	2.51	.38
10	4	6	4	2	.0	41	29	11	.30	.00	2.67	.00	2.95	.30
11	2	6	−6	2	.5	5	4	1	.04	8.01	.00	8.27	.00	8.28
12	2	4	−6	2	.0	96	96	89	4.84	3.90	.00	8.55	.00	8.77
13	2	4	6	0	.1	100	100	100	1.50	.00	.33	.00	1.95	1.79
14	2	4	6	0	.0[a]	100	100	100	1.99	.00	.00	1.64	.10	1.99
15	2	4	6	1	.0[a]	100	100	100	1.74	.00	.93	1.88	1.03	2.04

[a] Parental alleles are unique.

Comments:

Variance-components analysis is an attractive technique for mapping QTLs. However, it is not necessarily the most intuitive/easy method available. As an alternative or a first attempt at analyzing quantitative trait data, Haseman-Elston regression is often used.

<u>Basic idea behind Haseman-Elston regression</u>

- suppose that the marker under study is linked to a QTL
- pairs of relatives that share marker alleles IBD will share QTL alleles IBD
- the quantitative traits for these relatives should be more similar than those relatives that are not IBD

$\Rightarrow$ the difference in trait values between two relatives is expected to decrease as they share more marker alleles IBD

- methods based on this idea are often called *allele sharing methods*

Notation

Again we assume an additive QTL model:

Thus the expected value of the squared difference for the j^{th} pair is

with the assumptions that e_{1j}-e_{2j}

- has mean zero
- variance σ_e^2
- is uncorrelated with A_{1j}-A_{2j}

So what is $\sigma(A_{1j}, A_{2j})$?

Denote $(z_{1j}-z_{2j})^2$ by Y_j. We have the following relationship for Y_j:

where π_{jt} is the proportion of alleles IBD at the QTL.

However, π_{jt} is unknown! IBD information at the marker must be used to infer IBD status at the QTL.

Regardless of the relationship between the two individuals, we have

$$E(Y_j \mid \pi_{jm}) = \alpha + \beta\pi_{jm}$$

Chapter 16 of the text gives expected slopes for a number of relationships. As an example, for full sibs,

$$\beta = -2(1-2\theta)\sigma_A^2$$

Testing procedure

For a sample of *n* pairs of the same type of relationship,

- regress the squared difference of the trait against the fraction of marker alleles IBD
- perform a one-sided test of H_0: β=0 versus H_A: β<0
- a significant negative slope indicates that the marker is linked to a QTL

Parting Words

- different types of relative pairs cannot be used in the same regression
- residuals tend to be heteroscedatic
- cannot use parent-offspring pairs
- θ and σ_A^2 are confounded and cannot be estimated separately from β
- dominance at the QTL does not bias our TESTING results

Determining whether marker alleles are IBD

For a highly polymorphic marker, it is usually possible to determine the IBD status between full sibs:

If marker information for the corresponding pedigree is available, then it is possible to estimate the probability that the pair of relatives share 0, 1 or 2 alleles IBD.

What would this involve?

What if there is no pedigree information available?

The following general estimator of π_{jm} has been recommended for any pair of relatives:

$$p_{jm} = \frac{f_2 \Pr(M \mid i=2) + (f_1/2)\Pr(M \mid i=1)}{f_2 \Pr(M \mid i=2) + f_1 \Pr(M \mid i=1) + f_0 \Pr(M \mid i=0)}$$

The f_i correspond to the prior probability that the relatives share *i* alleles IBD (these correspond to Δ_7, Δ_8 and Δ_9 for i=2,1,0 from Chapter 7). The Pr(*M* | *i*) are found as follows:

Extensions of Haseman-Elston regression

- large-sample approximations for power have been developed
- power can be poor especially if the marker and QTL are separated by much (even as little as θ=0.1)
- power can be increased through selective genotyping
- interval mapping and multipoint extensions provide an additional increase in power and further allow for estimates of θ and σ_A^2 (recall that they are confounded in single-marker analysis)

Additional Methods for Mapping QTLs

- Within the past few years, there have been several extensions of Haseman-Elston regression to incorporate more than just the squared trait difference

  - Extensions center around strategies for incorporating the relationship between IBD sharing and the mean corrected squared trait sum

  - For a review of regression-based methods, see

    Feingold E (2002) Regression-based quantitative-trait-locus mapping in the 21st century. Am J Hum Genet 71:217-222

- Bayesian methods that combine segregation and linkage analyses

  - The method uses a sampling scheme such that a Markov chain can jump between models with different numbers of QTLs

  - Avoids problems that can arise from mis-specification of the underlying genetic model

  - There will be two class projects focusing on these methods

  - For a starting point, see

    Heath SC (1997) Markov chain monte carlo segretation and linkage aalysis for oligogenic models. Am J Hum Genet 61:748-760