Reading: Chapter 14 p413-424

Marker-Trait Associations

Recall that alleles at two loci can occur in a gamete MORE (or less) often than expected by chance. When this occurs, we say that the alleles are in *linkage disequilibrium* or in *gametic phase disequilibrium* or are *associated*. Previously, we defined the linkage disequilibrium coefficient as a measure of this lack of independence:

D =

No linkage disequilibrium $\Rightarrow$

Linkage disequilibrium $\Rightarrow$

When such associations occur, we expect the distribution for a trait to differ depending on an individual's marker genotype. This simple idea can be used for mapping QTLs.

Does this imply that the marker is in a causative relationship with the QTL?

Fine Mapping using Linkage Disequilibrium

There are many evolutionary forces that can cause an association between alleles at two loci; however, only linkage disequilibrium is correlated with physical distance between two loci (week 2 lecture notes). With this in mind, it was suggested by Bodmer (1986):

> "...it should be possible to saturate the relevant region with further polymorphic markers and look for the ones that have a population association with the trait, rather than increasing the number of families analyzed to search for closer linkage. This is the most efficient way of finding closely linked markers, since [human] family data are very inefficient at distinguishing between small recombination fractions such as 0.5% versus 5%"

The optimal case:

- "bad" alleles descend from a single ancestral mutation
- age of the ancestral mutation is neither too young nor too old
- expanding population that can be traced back to a small number of founders
- a few examples: Finnish, Alpine isolates, some Jewish populations, Hutterites, Amish

Example 6.1: Hastbacka et al. (1992) studied the autosomal recessive disease Diastrophic Dysplasia (DTD). Using 18 pedigrees, the gene for DTD was localized to within 1.6 cM of a marker locus using multipoint linkage analysis. Fine-mapping thru linkage disequilibrium mapping narrowed the region to a marker within 70kb.

Usage of linkage disequilibrium for mapping genes in more heterogeneous populations has been suggested:

-
-
-

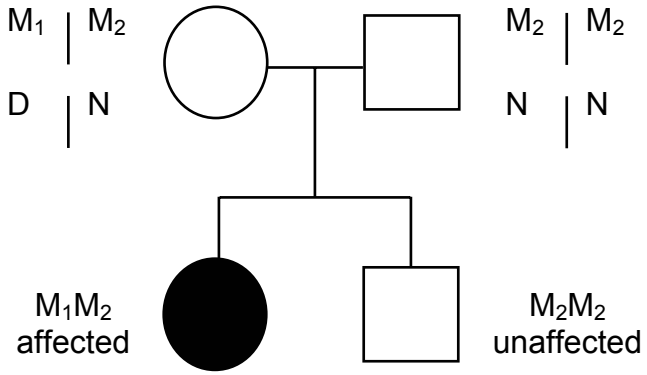
In heterogeneous populations, associations between a marker and quantitative trait can arise due to population substructure. Chapter 14 example 17 of the text provides a nice example of this.

<u>The transmission/disequilibrium test (TDT)</u>

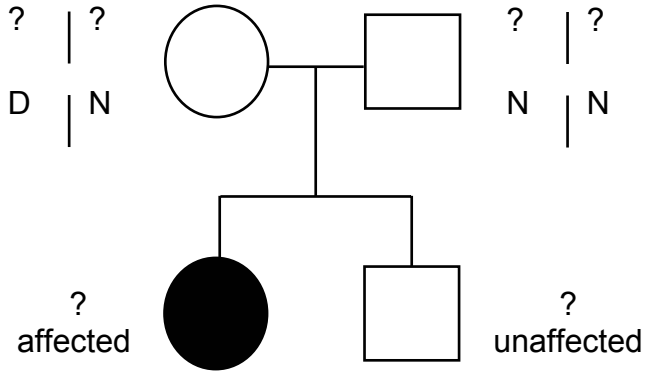
The TDT uses family data to avoid finding associations due strictly to population substructure. The basic idea behind the TDT (and any of its derivatives) is:

- to look for preferential transmission of a parental marker allele to an affected offspring
- using non-transmitted alleles from heterozygous parents as "controls"
- thus providing a **test of linkage and association** for a sample of trios

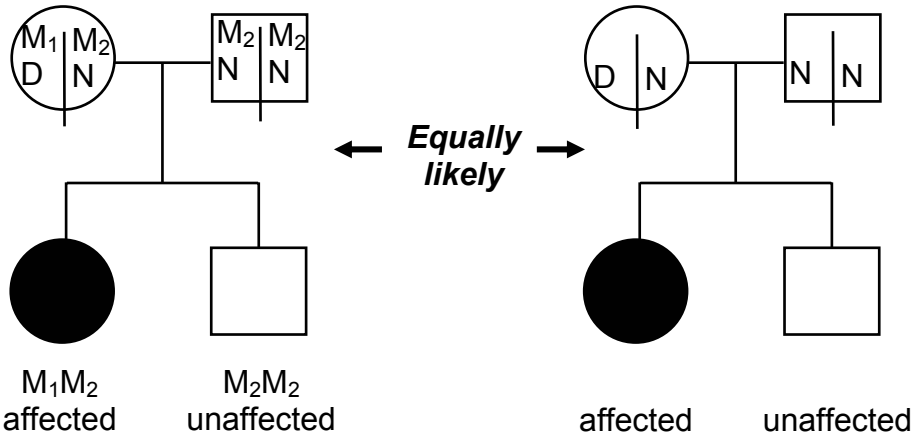
Example 6.2: Consider the following family:



If the marker and susceptibility locus are not linked, what family is equally likely to be observed (conditional on all marker and disease genotypes and the children's disease status):



What if the marker and susceptibility locus are linked, say $\theta=0.001$. What can we say about the observed family?



What allows us to make this statement? How is this different from the scenario of an unlinked marker and susceptibility locus?

The data from such families can be summarized as follows:

		Not Transmitted	
		A	a
Transmitted	A	n_{11}	n_{12}
	a	n_{21}	n_{22}

If we have $h=n_{12}+n_{21}$ heterozygous parents, then the number of transmissions of marker allele A from these parents is distributed:

$n_{12} \sim$

where p=Pr(A transmitted | Aa parent).

Under the null hypothesis of no linkage or no association, p=

Exact tests can be constructed or for large *h:*

$$T = \frac{\hat{p} -}{\sqrt{\quad}}$$

or

$$T^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

Example 6.3: In the paper,

> Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506-516

the TDT was used to test for linkage for the class 1 alleles of the insulin gene 5' VNTR with an IDDM susceptibility locus. Previous case-control tests had found significant association, but linkage studies were not able to find significant linkage for this marker.

	NOT TRANSMITTED	
TRANSMITTED	Class 1	Other
Class 1		78
Other	46	

Find T:

What would your conclusion be?

Segregation Distortion

The TDT can be positive due to segregation distortion. Segregation distortion occurs when heterozygous individuals preferentially transmit one allele (irrespective of the disease status of the child). Care must be taken to rule out this possibility.

Example 6.4: Spielman et al. (1993) also obtained transmission information from parents to their unaffected offspring:

	NOT TRANSMITTED	
TRANSMITTED	Class 1	Other
Class 1		42
Other	62	

Segregation distortion can be ruled out by testing whether the transmission (from heterozygous parents) of the allele to affected children is different from the transmission to their unaffected children:

X^2=

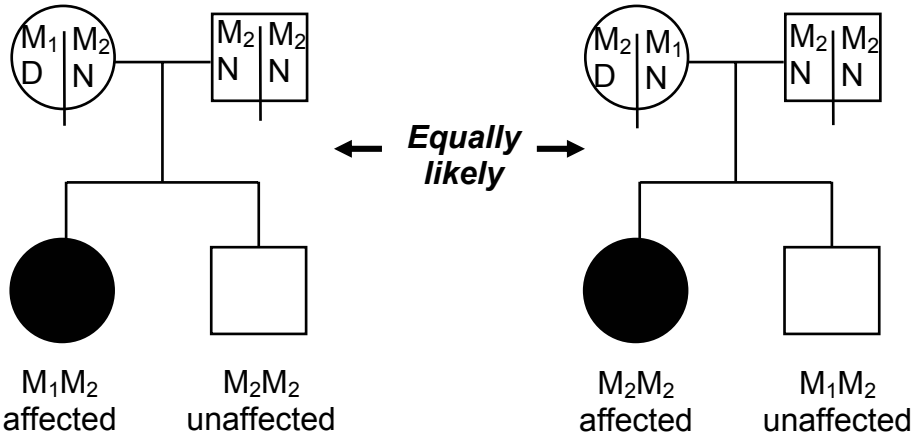
p-value=

Conclusion:

The TDT has been extended to many different scenarios:

- logistic regression framework
- dichotomous trait when parental genotype information is not available
- quantitative trait in a regression framework (for families with and without parental genotype information)
- asymptotic normal test statistic for a quantitative trait
- markers with multiple alleles
- multiple markers
- pedigree data

All of these extensions can be applied to samples of families with multiple children. However, the null hypothesis must be adjusted. To understand why, consider the following family again:



Suppose a third offspring is obtained that is affected and has marker genotype M_1M_2. There are two observations that can be made:

- under the null hypothesis of no linkage, marker transmissions to the each of the children are independent
- under the null hypothesis of no association, marker transmissions to each of the children ARE NOT independent

Why are these statements true?

The following table describes the hypotheses tested given the trait and data types:

Trait	Data	Linkage	Association[1]
Dichotomous	Independent Trios	Yes	Yes
	Families with marker genotypes for more than one affected child and for parents	Yes	No
	Sibships with marker genotypes for a single affected and unaffected child (no parental genotypes)	Yes	Yes
	Sibships with marker genotypes beyond a single affected and unaffected child (no parental genotypes)	Yes	No
Quantitative	Independent Trios	Yes	Yes
	Families with marker genotypes for more than one child and on parents	Yes	No
	Sibships with marker genotypes for two children (no parental genotypes)	Yes	Yes
	Sibships with marker genotypes for more than two children (no parental genotypes)	Yes	No

[1]With the exception of methods that explicitly adjust for the correlation among siblings due to linkage

Extensions of TDT to Quantitative Traits

The use of transmission disequilibrium tests has become a popular design for evaluating linkage and association for candidate genes. A recent search of Pubmed yielded 638 citations since 1993. In particular,

- the TDT has been extended to quantitative traits
- quantitative TDTs are based on whether parental transmissions are associated with the magnitude of the child's trait.
- We'll cover two primary extensions: a permutation based test and a test based on a variance component model. Each of these tests use heterozygous parents to avoid spurious associations due to population sub-structure.

But first an example of association testing in a random sample of unrelated individuals:

Example 6.5: Suppose we are studying a disease for which there is a quantitative trait, Y, that indicates progression to that disease. Further, suppose we have collected genotype data for a diallelic candidate gene with allele A_1 and A_2 with the following properties: $\theta=0$, Q_H and A_1associated and $Pr(A_1)=Pr(Q_H)=0.5$.

A_1A_1	A_1A_2	A_2A_2
9.141	8.370	5.932
10.877	7.215	5.027
8.867	8.373	5.755
11.692	7.854	4.498
10.054	7.373	4.668

Propose a test for association between the quantitative trait and the marker alleles. What are some of the potential causes of association that will be detected by your test?

Example 6.6: Illustration of spurious associated due to population sub-structure

Now suppose that, unknown to you, the population is composed of an equal mixture of two sub-populations with:

- $Pr(A_1)=Pr(Q_H)=0.5$ in population 1
- $Pr(A_1)=Pr(Q_H)=0.1$ in population 2
- $D=0$ and $\theta=0.5$ in each population

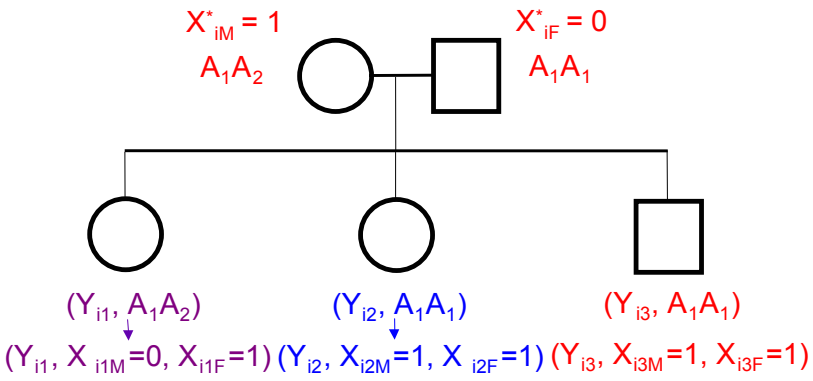
If this was tested using an F-test from the corresponding ANOVA, F=9.058 (p=0.00017).

Ignoring sub-structure, is there association?

As with the original TDT, these types of spurious associations can be avoided by using transmissions from heterozygous parents to their offspring.

A Permutation-based Transmission Disequilibrium for Quantitative Traits

Consider the i^{th} family with s_i children:



In 1997, Rabinowitz developed a test of linkage in the presence of association based on a score test for the regression of a child's trait value on the sum of their transmission variables from heterozygous parents:

$$Y_{ij} = \beta_0 + \beta_1\left[X^*_{iM}(X_{ijM} - \frac{1}{2}) + X^*_{iF}(X_{ijF} - \frac{1}{2})\right] + \varepsilon_{ij}$$

with the test statistic based on F families

$$T_R = \frac{\sum_{i=1}^{F}\sum_{j=1}^{s_i}(Y_{ij}-\bar{Y})\left[X^*_{iM}(X_{ijM}-0.5)+X^*_{iF}(X_{ijF}-0.5)\right]}{\sqrt{\sum_{i=1}^{F}\sum_{j=1}^{s_i}(Y_{ij}-\bar{Y})^2\left[X^*_{iM}(X_{ijM}-0.5)^2+X^*_{iF}(X_{ijF}-0.5)^2\right]}}$$

T_R is asymptotically N(0,1) under the null hypothesis of no linkage. If s_i=1 for all families, then T_R is also asymptotically N(0,1) under the null hypothesis of no association.

For the later case, statistical significance can also be assessed by permuting the transmissions from heterozygous parents.

- Why is this true?

- Why is this not true for families with multiple offspring?

Correcting for Dependent Transmissions

If the marker and QTL are linked, then transmissions among offspring are no longer independent. However, consider the numerator of the test statistic:

$$U_i = \frac{1}{s_i}\sum_{j=1}^{s_i}(Y_{ij}-\bar{Y})\left[X^*_{iM}(X_{ijM}-\frac{1}{2})+X^*_{iF}(X_{ijF}-\frac{1}{2})\right]$$

The expectation of U_i is

$$E(U_i) = D(1-2\theta)\left(\frac{\mu_{HH}+\mu_{HL}(1-2q_H)}{4p_1p_2(1-p_1p_2)}\right)$$

Under no linkage, what is $E(U_i)$?

Under no association, what is $E(U_i)$?

Hence, the statistic

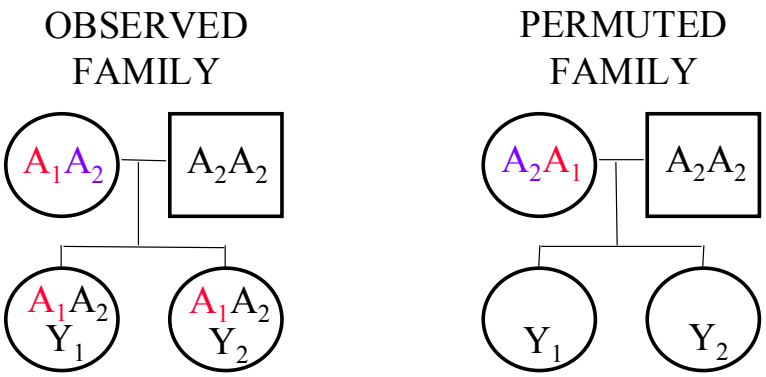
$$T_{QP} = \frac{\sum_{i=1}^{F} U_i}{\sqrt{\sum_{i=1}^{F} U_i^2}}$$

H_o: no linkage or no association

H_a: linkage AND association

T_{QP} is asymptotically N(0,1) under the null hypothesis of no linkage or no association regardless of the number of children per family.

Example 6.7: Conditional on trait values for the children and genotypes for the family, what genotype configuration is equally likely under the null hypothesis of no association?



Compute the value of U for each family. How do they compare? How could this be used to assess significance in a permutation framework?

The T_{QP} test provides a model-free framework for testing for linkage and association. Due to the Central Limit Theorem, the distribution of the test statistic is N(0,1) under the null hypothesis. Note that

no assumptions are necessary regarding the mode of inheritance of the underlying QTL.

While this provides a robust testing procedure, if information on the underlying mode of inheritance is available, then a more powerful test could be formed by using this information in a likelihood framework.

Transmission Disequilibrium Test for Quantitative Traits in the Variance Components Framework

General idea: the trait mean is modeled as a function of allelic effects for the marker under study while maintaining the same parameterization for the variance matrix.

Set-up:

For the j^{th} offspring of the i^{th} family, define the genotypes as

$$g_{ij} = \begin{cases} 1 \ if \ AA \\ 0 \ if \ Aa \\ -1 if \ aa \end{cases}$$

Then the trait mean can be modeled as

$E(y_{ij}) =$

Recall, the variance matrix is

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

Assuming the vector of phenotypes, y_i, for family i has a multivariate normal distribution, the likelihood for N_f families is:

$$L = \prod_{i=1}^{N_f} \frac{(2\pi)^{-n_i/2}}{\sqrt{|V_i|}} \exp\left\{-1/2[y_i - E(y_i)]' V_i^{-1}[y_i - E(y_i)]\right\}$$

To test for association, the likelihood ratio test can be used to evaluate whether β_a=0. This test will be affected by population substructure.

Extending Variance Components Analysis to Account for Population Stratification

Fulker et al. (1999) proposed the following modification:

Abecasis et al (2000) proved that

$$\begin{bmatrix} \beta_b \\ \beta_w \end{bmatrix} = \begin{bmatrix} \dfrac{\sum_i n_i (p_i - q_i)\mu_i}{NV_b} + a \\ \\ a \end{bmatrix}$$

where μ_i, p_i, q_i are the phenotypic mean and the marker-allele frequencies for the population from which family i was drawn. N is the total number of individuals with n_i equal to the number of families from population i. Note that V_b is the component of V due to between family allele effects.

Hence, we can test for association in the presence of population stratification by evaluating whether β_w=0. It is of note that β_b is a biased estimator of a and with this bias being a function of the unknown population stratification.