Extensions of transmission/disequilibrium tests for correlated data or how to use your entire dataset for a test of linkage and association

Monks SA[1], Kaplan NL[2]

[1]Department of Biostatistics, University of Washington;
[2]Biostatistics Branch, National Institute of Environmental Health Sciences

1999 American Society of Human Genetics Meeting
San Francisco, CA

Background

Tests of association have become widely used for the localization of genes influencing disease susceptibility. These methods range from the standard case-control test to tests that utilize within family controls. The standard case-control test can yield greater power than family-based tests but can be misleading since it will also detect associations resulting from population heterogeneity, thus, yielding no information with respect to localization.

On the other hand, family-based tests, such as the transmission disequilibrium test (Spielman et al. 1993), only detect association in the presence of linkage. This type of association is expected to occur at very small genetic distances so that the family-based tests ensure that positive results are due to being in close proximity to the susceptibility locus. For this reason, there has been much focus on family-based tests.

Current Family-Based Tests of Linkage and Associaton

For disease susceptibility, Martin et al. (1997) proposed a test of linkage and association that allows for the use of parental transmission information to *all affected* children of marker heterozygous parents.

For a quantitative trait, Monks and Kaplan (1998) introduced a test of linkage and association for data consisting of parental and child genetic information, for data consisting of child genetic information alone as well as for a data set that is a combination of these two types of family data. This work was unique in that *no restrictions were placed on the number of children* that could contribute data to the test.

Unifying Framework

A key observation that must be made, when testing for linkage and association, is that related individuals will have similar marker genotypes in a region around the trait locus. *If children are treated independently, then this excess sharing will elevate the false-positive rate for testing association.*

A general statistical method, **within-cluster resampling (WCR),** has recently been proposed by Hoffman et al. (1998). WCR is defined for data sets that contain independent clustered units. Recognizing that a collection of nuclear families of non-minimal size or a collection of unrelated pedigrees can be defined as a data set of independent clustered units, *WCR can be used to obtain a valid test of linkage and association adjusting for correlation due strictly to linkage*.

Outline

- Introduce the WCR framework and define a general test statistic that results from this framework
- Show that the test of Martin et al. (1997) can be derived using WCR
- Show that the test of Monks and Kaplan (1998) can be derived using WCR
- Study the performance of WCR on three generation pedigree data for a quantitative trait
- *Outline how WCR can be applied to family data consisting of a mixture of pedigree structures (and nuclear families)*

WCR Framework

Consider a collection of random variables

$$X_{11}, X_{12}, \ldots, X_{1n_1},$$
$$X_{21}, X_{22}, \ldots, X_{2n_2},$$
$$\vdots$$
$$X_{N1}, X_{N2}, \ldots, X_{Nn_N}$$

such that (1) observations within a row are correlated while (2) observations between rows are not correlated. Rows will be referred to as clusters so that we have a sample of N independent clusters.

Suppose further that under the null hypothesis:

$E(X_{ij})=0$ for any (i,j).

Given $n_1=n_2=\ldots=n_N=1$, the following statistic is asymptotically distributed as a standard normal random variable under the null hypothesis:

$$\frac{\bar{X}}{\sqrt{\hat{Var}(\bar{X})}} \quad \text{where} \quad \hat{Var}(\bar{X}) = \frac{\sum_{i=1}^{N} X_{i1}^2}{N^2} \qquad \text{Eq. [1]}$$

If any of the n_i are greater than one, then the variance estimate is no longer correct. However, within cluster resampling (WCR) can be used to obtain a correct estimate of the variance.

The WCR estimate of variance is derived from R reduced data sets. These sets are generated by randomly selecting one observation from each cluster. For each reduced data set, the mean, $\bar{X}_r$, and variance, $\hat{Var}(\bar{X}_r)$, can be computed as in equation [1]. These values are then combined as follows:

$$Z = \frac{\bar{\bar{X}}}{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\hat{Var}(\bar{X}_r) - \frac{1}{R}\sum_{r=1}^{R}(\bar{X}_r - \bar{\bar{X}})^2}} \quad \text{where} \quad \bar{\bar{X}} = \frac{1}{R}\sum_{r=1}^{R}\bar{X}_r$$

As more reduced data sets are utilized in the computation of the WCR statistic, i.e. R$\rightarrow\infty$, and using the assumption of independence of observations from different clusters, we have

$$Z \rightarrow \frac{\sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}}{\sqrt{\sum_{i=1}^{N} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \right)^2}} \qquad \text{Eq. [2]}$$

WCR for Disease Susceptibility

Consider a marker with alleles M_1 and M_2. The transmission disequilibrium test can be derived from the random variable $X_{ij}=T_{ij}-0.5$ where T_{ij} is 1 if the i^{th} heterozygous parent transmitted marker allele M_1 to their j^{th} child and 0 otherwise. Thus, the clusters are composed of the $\{X_{ij}\}$ for heterozygous parent i and their n_i children. For a sample of affected sib-pairs with N heterozygous parents, the statistic from equation [2] becomes

$$\frac{\sum_{i=1}^{N}\frac{1}{2}\sum_{j=1}^{2}(T_{ij}-0.5)}{\sqrt{\sum_{i=1}^{N}\left(\frac{1}{2}\sum_{j=1}^{2}(T_{ij}-0.5)\right)^2}}=\frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{2}T_{ij}-1\right)}{\sqrt{\sum_{i=1}^{N}\left(\sum_{j=1}^{2}T_{ij}-1\right)^2}}$$

which is equal to the extension of the TDT from Martin et al. (1997).

WCR for a Quantitative Trait

For a quantitative trait, the transmission disequilibrium test proposed by Rabinowitz (1997) can be derived from the random variable

$$X_{ij} = (Y_{ij} - \bar{Y})[T^*_{iM}(T_{ijM} - 0.5) + T^*_{iF}(T_{ijF} - 0.5)]$$

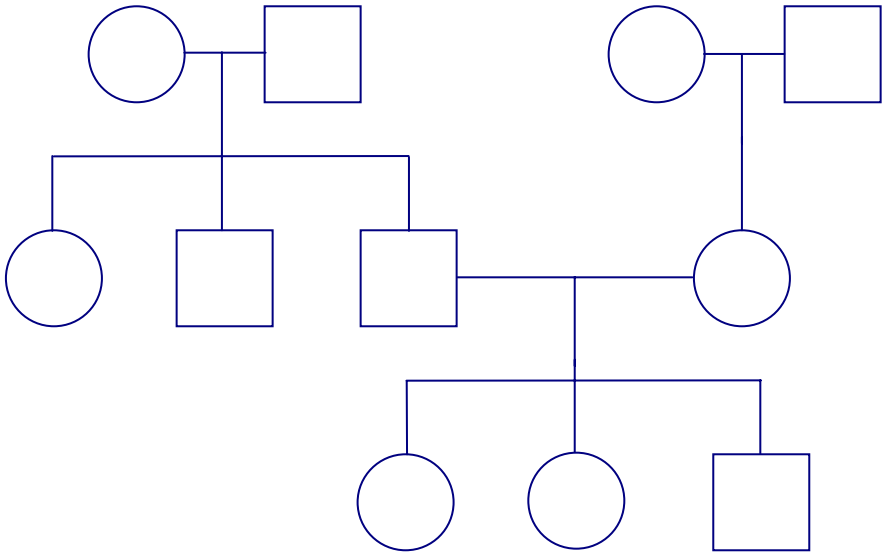
where T^*_{iM} (T^*_{iF}) is 1 if the mother (father) of the i^{th} family is heterozygous and 0 otherwise and T_{ijM} (T_{ijF}) is 1 if the mother (father) of the i^{th} family transmitted marker allele M_1 to their j^{th} child and 0 otherwise. Thus we have clusters composed of the $\{X_{ij}\}$ for the i^{th} family. Using this random variable, with the statistic from equation [2], we get

$$\frac{\sum_{i=1}^{N} \frac{1}{t_i} \sum_{j=1}^{t_i} (Y_{ij} - \bar{Y})[T^*_{iM}(T_{ijM} - 0.5) + T^*_{iF}(T_{ijF} - 0.5)]}{\sqrt{\sum_{i=1}^{N} \left(\frac{1}{t_i} \sum_{j=1}^{t_i} (Y_{ij} - \bar{Y})[T^*_{iM}(T_{ijM} - 0.5) + T^*_{iF}(T_{ijF} - 0.5)] \right)^2}}$$

which is equal to the extension of Monks and Kaplan (1998).

Pedigree Data

For a sample of pedigrees, WCR can be used for disease susceptibility (i^{th} cluster consists of all heterozygous parent-child pairs from pedigree i) and a quantitative trait (i^{th} cluster consists of all mother-father-child trios contained in pedigree i)



We now demonstrate the validity of the procedure for a quantitative trait and provide sample size calculations for a number of genetic models. The reader is referred to poster 2462 for further discussion of the procedure for a susceptibility locus.

Simulation models for a quantitative trait

- diallelic QTL with the allele causing a higher trait value having frequency $Pr(Q_H) \in \{0.1, 0.5\}$
- heritability, $H^2 \in \{0.1, 0.3, 0.5\}$
- additive, dominant and recessive modes of inheritance considered
- diallelic marker with marker allele, M_1, having frequency $Pr(M_1) \in \{0.5, 0.8\}$
- disequilibrium coefficient, D, set equal to 0 under the null hypothesis and its maximum possible value under the alternative; the recombination fraction between the QTL and marker was 0.0
- one-sided test with significance level, α, of 0.01
- given a QTL genotype, the trait was normally distributed with variance 1(note that given the above information the distribution means are uniquely determined)

Estimates of Significance Level

			MODE OF INHERITANCE		
H^2	$Pr(Q_H)$	$Pr(M_1)$	Additive	Dominant	Recessive
0.1	0.1	0.5	0.0098	0.0109	0.0091
0.1	0.1	0.8	0.0087	0.0087	0.0095
0.1	0.5	0.5	0.0085	0.0109	0.0092
0.1	0.5	0.8	0.0105	0.0098	0.0101
0.3	0.1	0.5	0.0086	0.0089	0.0081
0.3	0.1	0.8	0.0093	0.0091	0.0069
0.3	0.5	0.5	0.0098	0.0098	0.0095
0.3	0.5	0.8	0.0103	0.0107	0.0090
0.5	0.1	0.5	0.0094	0.0086	0.0085
0.5	0.1	0.8	0.0081	0.0082	0.0050
0.5	0.5	0.5	0.0096	0.0100	0.0089
0.5	0.5	0.8	0.0104	0.0108	0.0100

Estimates are based on 10000 samples of 250 pedigrees. Estimates satisfactorily correspond to the alpha level of 0.01. Two estimates are significantly different from 0.01 (two-sided test with type I error 0.05; this corresponds well to the expected number of false positives 2/36=0.05556).

No. of Pedigrees Required for 80% Power

			MODE OF INHERITANCE		
H^2	$Pr(Q_H)$	$Pr(M_1)$	Additive	Dominant	Recessive
0.1	0.1	0.5	335	348	1987
0.1	0.1	0.8	1636	1665	8735
0.1	0.5	0.5	37	56	56
0.1	0.5	0.8	185	269	290
0.3	0.1	0.5	118	118	772
0.3	0.1	0.8	542	558	3200
0.3	0.5	0.5	14	21	21
0.3	0.5	0.8	64	94	100
0.5	0.1	0.5	73	73	499
0.5	0.1	0.8	319	333	1917
0.5	0.5	0.5	10	15	15
0.5	0.5	0.8	41	62	62

Monks and Kaplan. ASHG 1999

Other Extensions

- The same framework can be used for testing for a susceptibility locus using pedigree data (see poster 2462)
- The procedure can be used for a mixture of family structures
  - e.g. nuclear families with varying number of children, pedigrees with different numbers of generations, collections of siblings with no parental information
  - each structure is then viewed as a cluster of minimal units and a mean 0 random variable is defined for each minimal unit
  - once the minimal unit and random variable are defined, the statistic of equation [2] can be used to measure statistical significance
- *WCR can be used in more complicated situations. In particular, Hoffman (1998) showed that WCR could be used for generalized linear models. This encompasses many forms of analysis such as normal and logistic regression.*

Last words (and shameless plug)

- The statistics discussed here could be derived by noting that given the sum of the random variables for each cluster, denote these by $U_i, \ldots, U_N$,
  - the U_i have mean 0
  - an estimate of variance is given by $\frac{1}{N}\sum_{i=1}^{N} U_i^2$
  - and thus a statistic equal to that of equation [2] can be used to test the null hypothesis of no linkage or no association.

This approach is discussed in the talk "Family-based tests of association for a quantitative trait that use families with an arbitrary number of children" by N.L. Kaplan in SESSION 28 (TDT and Other Tests for Linkage Disequilibrium)

References

- Hoffman EB (1998) Within cluster resampling. Thesis: University of North Carolina at Chapel Hill, Department of Biostatistics
- Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61:439-448
- Monks SA, Kaplan NL (1998) Tests of association for QTLs using sibships of unrestricted size, Meeting of the American Society of Human Genetics. Am J Hum Genet Suppl 63:A302
- Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Hum Hered 47:342-350
- 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