# Transmission Disequilibrium Test in GWAS

Sorin Istrail

Department of Computer Science
Brown University, Providence
sorin@cs.brown.edu

November 10, 2010

# Outline

- The transmission/disequilibrium test (TDT) was intro- duced several years ago by Spielman et al. (1993) as a test for linkage between a complex disease and a genetic marker.
- The original intended use of the TDT was to test for linkage with a marker located near a candidate gene, in cases where disease association already had been found.
- However, even if prior evidence for association is absent, the TDT is valid and can be used to test any marker (or a set of markers) for which data are available from parents and one or more affected offspring.

- The genetic model consists of both a locus D that contributes to disease susceptibility and a (possibly linked) marker locus M.
- The standard way to identify disease loci is to use classical (LOD) or nonparametric (affected-sib-pair [ASP]) methods to test for linkage with such a marker or a set of markers. It has been known for some time, however, that these methods may fail to detect a disease locus linked to a marker, even though the locus may be of biological significance (Cox et al. 1988; Spielman et al. 1989).
- Even when standard linkage tests fail to provide evi- dence, however, a disease locus linked to the marker may be suggested sometimes by the presence of a "dis- ease association," usually established by a case-control study.

- The underlying premise is that, if an association is found, it is likely to be due to linkage disequilibrium. (We use the term strictly to mean the presence of both linkage and association between marker and disease.)
- Since linkage disequilibrium is found only over very small map distances, close linkage between marker and disease susceptibility is implied.

- However, it is a well- known result of population genetics that admixture, het- erogeneity, or stratification in a population can make it impossible to draw valid conclusions from a conven- tional case-control study, since these conditions ("popu- lation structure") can give rise to substantial association even for unlinked loci.
- Accordingly, Spielman and Ewens developed the TDT to test for linkage in the presence of association that is, to distinguish this case (linkage disequilibrium) from associations that arise from population structure in the absence of linkage.

- The TDT grew out of earlier proposals for avoiding incorrect conclusions from disease associations. Recog- nizing the problem, Falk and Rubinstein (1987) pro- posed the haplotype relative risk (HRR) as a family- based test for association, but they did not focus on linkage. Field et al. (1986), Thomson et al. (1989), and Thomson (1995) developed this approach further as a test for association.
- Ott's (1989) analysis of the mathe- matical model for the HRR was the point of departure for Spielman and Ewens development of the TDT, and Parsian et al. (1991) presented and applied a similar test without call- ing attention to its mathematical properties.

- Conventional tests for linkage (e.g., LOD and ASP) require sibships with multiple offspring.
- In the TDT, by contrast, sibships with a single affected offspring can be used to detect linkage between disease and marker, provided that disease association with some particular marker allele is also present.

- We consider a marker locus M, with two alleles M1 and M2, and obtain genotypes for affected individuals and their parents. In the most general form, the data to be analyzed are numbers of "transmissions";
- that is, for parents of each genotype (M1M1, M1M2, and M2M2, we determine the number of times that the M1 allele or the M2 allele was transmit- ted to an affected offspring. Spielman et al. (1993, table 2), denoted these counts as follows:
  - a, number of times that M1MI transmits M1 to af- fected offspring;
  - b, number of times that M1M2 transmits M1 to af- fected offspring;
  - c, number of times that M1M2 transmits M2 to af- fected offspring;
  - d, number of times that M2M2 transmits M2 to af- fected offspring.

- The counts may come from families that are simplex (i.e., data are from only one affected offspring), multi- plex (data are from two or more affected sibs), or multi- generational; and the population may exhibit structure.
- The null hypothesis of interest is that the marker and disease are unlinked.
- For any specified null hypothesis and a given set of data, there is a standard procedure ("Neyman-Pearson"; e.g., see Kendall and Stuart 1979) for obtaining the most powerful statistical test. Applica- tion of this procedure, for the present data and hypothe- sis, yields the TDT as the optimal test.
- When marker and disease are unlinked, data used in the TDT for related individuals are independent.
- It follows that data from a large pedigree may be used (if desired) by applying the TDT to each affected individual separately.

# The TDT is carried out as follows

- The TDT is carried out as follows.
- The result obtained by applying Neyman-Pearson theory dictates that we use data (observations b and c) only from those parents who are heterozygous M1M2; this result holds regardless of whether there is population structure.
- The TDT statistic $\frac{(b-c)^2}{(b+c)}$; it tests for equal numbers of transmissions of M1 and M2 from heterozygous parents to affected offspring.
- If there is linkage between marker and disease, as well as allelic association, b and c will tend to differ in value.

# The TDT is carried out as follows

- The statistical significance of the TDT is tested by X2 ("McNemar Test") or by the exact binomial test (see Spielman et al. 1993); a significant difference provides evidence that the marker is linked to the disease locus.
- Note that, if there is no linkage, alleles of M segregate independently of disease, so the presence of association (e.g., from population structure) will not cause b to dif- fer from c.
- Thus such an association would not lead us to infer linkage incorrectly.

# The TDT is carried out as follows

- Similarly, when there is linkage but no association, there is also, on average, no tendency for b to differ from c.
- Thus the TDT can detect linkage only in the presence of association.
- In contrast, when tests of linkage are carried out by ASP methods, there is no requirement for association, since linkage is then detected as departure from random assortment within families.
- However, ASP methods do require the presence of two or more affected sibs. This restriction, which poses practical difficulties for study of diseases for which multiplex families are rare, does not apply to the TDT.

- The structure of the TDT shows that one should use only heterozygous parents of total number $b + c$.
- The TDT tests whether $\frac{b}{(b+c)}$, $\frac{c}{(b+c)}$ are about the same with the probabilities $(0.5, 0.5)$
- The hypothesis is tested by a binomial (asymptotically chi-square) with one degree of freedom

$$\chi^2 = \frac{(b - \frac{(b+c)}{2})^2}{\frac{(b+c)}{2}} + \frac{(b - \frac{(b+c)}{2})^2}{\frac{(b+c)}{2}} = \frac{(b-c)^2}{(b+c)}$$