

# COMPUTER EXERCISE 2

## STATISTICS IN GENETICS

### ASSOCIATION ANALYSIS

PETER ALMGREN  
PÄR-OLA BENDAHL  
HENRIK BENGTTSSON  
OLA HÖSSJER  
AZRA KURBASIC  
ROLAND PERFEKT

12th January 2007



LUND UNIVERSITY

Lund Institute of Technology  
Centre for Mathematical Sciences  
Mathematical Statistics



# Contents

<b>1 Association analysis</b>	<b>2</b>
1.1 TDT in Genehunter . . . . .	3

## Introduction

In this exercise we will use *Genehunter* to explore how the TDT test works. All the pedigrees will be trios, i.e. families with two parents and one offspring.

Navigate to the bnf073-directory:

```
cd ~/bnf073
```

Copy the files you need for this computer exercise to the current directory:

```
cp -r /export/local/lib/bnf073/lab2/ ~/bnf073/
```

Navigate to the subdirectory lab2:

```
cd lab2
```

## 1 Association analysis

There are two primary approaches for mapping genes that either cause or increase susceptibility to human diseases. The first is the linkage approach, either a parametric lod score analysis when the disease model is known or a model independent method when the genetic model is unknown. The second approach is allelic association studies, which is another nonparametric approach for disease gene mapping. Particularly in complex diseases, association analysis has turned out to be an important tool in identifying disease gene loci.

There are two major types of association studies. *Case-control* studies compare allele frequencies in a set of unrelated affected subjects to those in a set of matched controls. One drawback in case-control studies may be that they are sensitive to *population stratification*, i.e. multiple population subtypes. Population stratification may be due to recent admixture or selection of controls that match the cases poorly. The other kind of association study is the *family-based* design. The basic idea is that the transmitted allele is considered the 'case' and the untransmitted is considered the 'control'. One of the most popular family based tests is the *Transmission Disequilibrium Test*, *TDT*. The only samples necessary are those from an affected individual and its two parents. It is quite easy to count the transmitted and the untransmitted alleles manually and then perform a McNemar test to evaluate the statistical significance.

**Example 1 (A simple TDT example)** Suppose we have a nuclear family consisting of two parents and their affected child, see Figure 1. The marker in question is a SNP (a diallelic marker). The number of transmitted and untransmitted alleles can then be put into a 2x2-table like Table 1 for further analysis with McNemar's test or an exact binomial test.

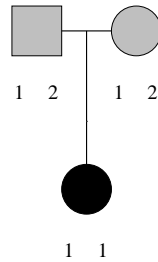


Figure 1: A TDT triad.

<i>Transmitted</i>	<i>Non-transmitted</i>	
	Allele 1	Allele 2
Allele 1	0	2
Allele 2	0	0

Table 1: Number of transmitted and non-transmitted marker alleles 1 and 2.

## 1.1 TDT in Genehunter

Genehunter has a standard implementation of the transmission disequilibrium test (TDT) along with several extensions for using missing data and estimating significance via simulation (permutation procedures). In addition there are commands for haplotype analysis, `tdt2`, `tdt3` and `tdt4`. The haplotype methods look at more than one locus at a time to estimate transmitted and untransmitted allele combinations. The version of TDT that is implemented in Genehunter gives you an output telling how many times a certain allele was transmitted and not transmitted from parents to affected individuals. Note, it only uses transmissions from *heterozygous* parents.

The pedigree in Figure 1 can be found in the pedigree file `tdt00.pre`. List the content of this file and make sure you understand the coding scheme. The corresponding parameter file is called `tdt00.dat`. Start a Genehunter session and perform a TDT-test:

```
load tdt00.dat
tdt tdt00.pre
```

Calculate the TDT-statistic by hand (using the formula presented in the association analysis lecture) and compare with the output from the TDT test in Genehunter. Do they agree?

.....

Note that Genehunter presents two identical hypothesis tests. The second line is completely redundant for markers with only two alleles (SNPs). The output from Genehunter looks like a two-by-two table, but it is NOT the kind of table presented in Table 1. It is the counts in the two off-diagonal cells (North East and South West if you think of a compass).

The pedigree file tdt0.pre has data for one trio and two markers. The TDT-score does not depend on the genetic model when all the individuals have been genotyped. Nevertheless you must present a parameter file to Genehunter each time you run a TDT test. Use tdt.dat this time. Exit Genehunter and start a new Genehunter session.

```
load tdt.dat
tdt tdt0.pre
```

The first marker is not informative so Genehunter reports no TDT-score for this locus. Why is it not informative? Hint: look at the pedigree file.

.....

The pedigree file tdt.pre holds simulated data at two marker loci for 100 trios. Exit and restart Genehunter again.

```
load tdt.dat
tdt tdt.pre
```

Look at the output for marker locus 1. The two numbers (trans and untrans) for allele 1 correspond to the number of times allele 1 was transmitted but not allele 2 (trans) and the number of times allele 1 was not transmitted and allele 2 was transmitted (untrans). Transmissions from homozygous parents hold no information. Construct a 2-by-2 table of the kind presented in Table 1 and carry out the TDT-test by hand:

.....

.....

.....

Was one of the two marker alleles at locus 1 preferentially transmitted from parents to affected offspring? If so, which allele?

.....

Marker locus 2 has four alleles (it is a micro satellite not a SNP). For this marker Genehunter reports four tests - one for each allele. The first row corresponds to trans and untrans of marker allele 1. Note that only parents heterozygous (1,x) at the marker locus will be counted in this analysis. The x-allele could be any of the three other alternatives (2, 3, or 4). The tests for the other three alleles are defined analogously. Do we have significant evidence for association between disease inheritance and inheritance of a specific marker allele at locus number 2?

.....

Exit and restart Genehunter. TDT can also be used to see if a specific haplotype is preferentially transferred from parents to affected offspring. The simplest form of haplotype TDT is two-locus TDT `tdt2`, but a standard one-locus TDT is necessary as a pre processing step for `tdt2`.

```
load tdt2.dat
tdt tdt2.pre
```

That was the one-locus analysis. Try to summarize the Genehunter output in a few sentences:

.....

.....

.....

The two-locus analysis is now carried out with the command `tdt2` at the Genehunter prompt (no arguments).

```
tdt2
```

Try to interpret the output:

.....

.....

.....

Try a three-locus TDT-analysis:

tdt3

and a four-locus analysis:

tdt4

Find the smallest p-values and the corresponding haplotypes in each of the four analyzes above (tdt, tdt2, tdt3, and tdt4).

.....  
.....  
.....  
.....

Does the level of evidence for association increase as we take more and more loci into account simultaneously?

.....  
.....  
.....

As a final comment it may be mentioned that many of the diseases not yet explained genetically are the late-onset ones. As you may guess, association techniques like TDT may be hard to employ. A good alternative to standard implementations of the TDT may be to look at methods taking siblings into account, i.e. instead of counting transmissions from parents it may be wise to compare alleles in siblings. A software available is S-TDT, which uses TDT triads if possible, otherwise it looks at siblings.