

GENETIC COIN FLIPS: FAMILY-BASED ASSOCIATION STUDIES

ŚAUNAK SEN*

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

sen@biostat.ucsf.edu

INTRODUCTION

Family-based association tests such as the TDT (transmission disequilibrium test)[†] are robust to confounding by population stratification – a common problem for genetic association studies. The basic idea behind family-based association studies is best illustrated by the TDT. It may be viewed as a naturally randomized experiment, therefore providing strong evidence in favor of causation. After reviewing the TDT which is based on parent-child trios, we will consider its generalization to quantitative traits, and to sibships.

TDT: TRANSMISSION DISEQUILIBRIUM TEST

Let us consider the setting of the case-parent trio of the TDT. Suppose the marker of interest has two alleles, M and m . Then we can tabulate the allele frequencies of the cases and their parents as follows.

Parental genotype	Case genotype			Total
	mm	mM	MM	
* mm x mm	22	–	–	22
mm x mM	17	25	–	42
* mm x MM	–	7	–	7
mM x mM	1	11	13	25
mM x MM	–	1	1	2
* MM x MM	–	–	2	2
Total				100

Note that when both parents are homozygous, the mating is not informative, since there is no genotypic variation at that locus in the progeny. Since our goal is to detect the association between genotype and phenotype variation, if no genotype variation is generated, it is not of interest.[‡] Assume for a moment that the marker allele we are testing is tightly linked with the causative disease allele. Then the above data takes the form of a randomized trial, where we only observe the treatment assignments in the cases (and infer those in the controls).

Under random assortment, the expected frequencies of the data can be found very easily and are given by the table below:

Parental genotype	Case genotype						Total
	Observed			Expected			
	mm	mM	MM	mm	mM	MM	
mm x mM	17	25	–	21	21	–	42
mM x mM	1	11	13	6.25	12.5	6.25	25
mM x MM	–	1	1	–	1	1	2
Total							69

*© 2006 Śaunak Sen; Last updated April 18, 2006.

[†]Sometimes also called transmission distortion test

[‡]This is analogous to the situation in model organisms where the F_1 individuals are genetically identical, and hence not directly usable for genetic mapping.

We can use a χ^2 test (which has 1+2+1=4 degrees of freedom) to test for departure from the null hypothesis of 'no distortion'. The value of the χ^2 statistic is 13.4 which has a p-value of 0.0095. However, the p-value is suspect because some cells had small counts and the χ^2 approximation is not always good under those scenarios.

TDT Using some simplifying assumptions we can try to collapse the cells in the original table. One popular summary is as follows:

Transmitted	Non-transmitted		
	m	M	Total
m	93	31	124
M	63	13	76
Total	156	44	200

We can use two tests of the null hypothesis. One is the test of "symmetry" (which is the TDT), and the other is a test of "marginal homogeneity" (which is the test for haplotype relative risk).

```
. symmi 93 31 \ 63 13, contrib mh
```

```
-----
      |      col
      |      1      2      Total
-----+-----
      |
  1 |      93      31      124
  2 |      63      13      76
      |
Total |     156      44      200
-----
```

```
-----
Cells          Contribution
              to symmetry
              chi-squared
-----
n1_2 & n2_1          10.8936
```

```
-----
                                chi2      df      Prob>chi2
-----+-----
Symmetry (asymptotic)          |      10.89      1      0.0010
Marginal homogeneity (Stuart-Maxwell) |      10.89      1      0.0010
Marginal homogeneity (Bickenboller) |      12.19      1      0.0005
Marginal homogeneity (no diagonals) |      10.89      1      0.0010
-----
```

Both tests are used in practice, but the test for symmetry (TDT) is generally preferred since it takes into account the paired nature of the transmission.

Multi-allelic case The multi-allelic extension is simple. Suppose the data are arrayed as follows:

Untransmitted	Transmitted			Total
	a	b	c	
a	47	56	38	141
b	28	61	31	120
c	26	47	10	83
Total	101	164	79	344

We can use the same command as before to test for marginal homogeneity or symmetry, the latter being the extension of the TDT to the multi-allelic case. If the counts in the table are small, one may have to use permutation tests instead of relying on the χ^2 approximation to get the right p-value.

```
. symmi 47 56 38 \ 28 61 31 \ 26 47 10 , contrib mh
```

```
-----
      |
      |           col
      |         1     2     3     Total
-----+-----
      |
      | 1 | 47     56     38     141
      | 2 | 28     61     31     120
      | 3 | 26     47     10     83
      |   |
      | Total | 101    164    79    344
-----
```

```

      |           Contribution
      |         to symmetry
      |       chi-squared
-----+-----
      |
      | n1_2 & n2_1           9.3333
      | n1_3 & n3_1           2.2500
      | n2_3 & n3_2           3.2821
-----
```

```

                                     chi2      df      Prob>chi2
-----+-----
      | Symmetry (asymptotic) | 14.87      3      0.0019
      | Marginal homogeneity (Stuart-Maxwell) | 14.78      2      0.0006
      | Marginal homogeneity (Bickenboller) | 13.53      2      0.0012
      | Marginal homogeneity (no diagonals) | 15.25      2      0.0005
-----
```

SIB TDT

For late-onset diseases, parental genotypes may be difficult to obtain. However, affected and unaffected sibling genotypes may be more easily ascertained. Consider the simplest case where we collect discordant sib pairs. We can view this study as a matched case-control design, where the stratum is the family.

We use the RNASEL data[§] as our example where the data would look like this.

[§]Casey G, Neville PJ, Plummer SJ, Xiang Y, Krumroy LM, Klein EA, Catalona WJ, Nupponen N, Carpten JD, Trent JM, Silverman RH, Witte JS (2002) RNASEL Arg462Gln variant is implicated in up to 13% of prostate cancer cases. *Nature Genetics*. 32:581-583

RNASEL genotype		
FamilyID	Case	Control
34	GG	AG
349	GG	AG
364	AG	GG
378	GG	GG
1312	AG	AG
1325	AG	GG
1340	GG	GG

Thus, we can analyze this data as a matched case-control design which can be accomplished by either using a Mantel-Haenszel test or by using conditional logistic regression. For the RNASEL data this is accomplished by the following steps.

```

/* read in data */
. insheet using rnasel.csv

/* create prostate cancer yes/no outcome variable */
. gen outcome=1 if cap=="Y" & cap ~=.
. replace outcome=0 if cap=="N"

/* generate SNP variables according to genotype */

. gen snp=0 if rnasel_1385_a_g=="GG" & rnasel_1385_a_g ~=.
. replace snp=1 if rnasel_1385_a_g=="AG"
. replace snp=2 if rnasel_1385_a_g=="AA"

/* conditional logistic (test for trend) */

. clogit outcome snp, group(familyid)
note: multiple positive outcomes within groups encountered.
note: 103 groups (202 obs) dropped due to all positive or
      all negative outcomes.

Iteration 0:  log likelihood = -297.52889
Iteration 1:  log likelihood = -297.11546
Iteration 2:  log likelihood = -297.11541
Iteration 3:  log likelihood = -297.11541

Conditional (fixed-effects) logistic regression      Number of obs   =           844
                                                    LR chi2(1)      =           4.31
                                                    Prob > chi2     =           0.0378
Log likelihood = -297.11541                        Pseudo R2       =           0.0072

```

outcome	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
snp	.2973011	.1444359	2.06	0.040	.014212 .5803903

Notice that we have coded the `snp` variable as counting the number of A alleles. Thus, it gives us a single degree of freedom “test for trend” in the conditional logistic regression. The p-value for the test is 0.04. Alternatively, we may wish to test for the SNP genotype effect as a two degree of freedom test, treating the SNP genotype as factors. We can do this as follows.

```

/* code two dummy variables for SNP genotypes */

. gen snp1=0 if snp==0 & snp~=1.
. replace snp1=1 if snp==1
. replace snp1=0 if snp==2

. gen snp2=0 if snp==0 & snp~=1.
. replace snp2=0 if snp==1
. replace snp2=1 if snp==2

/* conditional logistic with snps as two-level factor */

. clogit outcome snp1 snp2, group(familyid)
note: multiple positive outcomes within groups encountered.
note: 103 groups (202 obs) dropped due to all positive or
      all negative outcomes.

Iteration 0:   log likelihood = -297.53412
Iteration 1:   log likelihood = -297.03991
Iteration 2:   log likelihood = -297.03975
Iteration 3:   log likelihood = -297.03975

Conditional (fixed-effects) logistic regression   Number of obs   =           844
                                                    LR chi2(2)      =           4.46
                                                    Prob > chi2     =           0.1073
Log likelihood = -297.03975                       Pseudo R2      =           0.0075

-----
      outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      snp1 |   .2552311   .1803085     1.42   0.157    - .098167   .6086293
      snp2 |   .6512237   .3246906     2.01   0.045    .0148419   1.287606
-----

```

This test gives a p-value of 0.11, which would not be significant at the 5% level of significance. This is the price of the additional degree of freedom. The conditional logistic regression formulation makes it easy for us to adjust for confounding factors (such as age). We can simply enter them as predictor variables.

DISCUSSION

The TDT is a test for linkage and linkage disequilibrium. The TDT is a valid test for linkage under all situations. However, the extent to which it is sensitive to linkage disequilibrium in the population depends on the sampling scheme. For example, if all the trios come from a large pedigree which share the disease allele from a single founder, the TDT will be a test of linkage only. However, with additional copies of the number of ancestral disease alleles, the TDT will test for both linkage and linkage disequilibrium.

If there is reason to suppose that there is segregation distortion, then we will have to collect controls (control trios, or sib controls) in addition to cases. If environmental exposures are correlated with genotypes, then it is better to stratify the data by parental genotype rather than collapsing the data into transmitted and untransmitted alleles.

The TDT analysis conditions on the genotype of the parents and is therefore robust to the effects of population stratification (which would show up in the differential allele frequencies of the parents). This argument extends to other family-based association tests. The randomized trial analogy for the TDT helps us understand why family-based association tests provide stronger evidence of causation than population-based association studies. If there are genotyping errors, or if families with one missing parental genotype are included in the TDT, the results may be anti-conservative. Thus, it is important to check the fidelity of the data to ensure that the conclusions are robust.

Our discussion above suggests how we can extend the TDT when the trait of interest is quantitative. For the parent-child trio case, we can use conditional logistic (or multinomial) regression conditioning on the trait. Alternatively, one can use a prospective approach where one models the distribution of the trait conditional on the parental genotypes. The latter approach may be more powerful when the model assumptions are correct; the former is more robust. Analogous extensions for the sibship case are possible.

Family-based association test, by conditioning on the family structure are able to remove potential bias due to population stratification and environmental effects. However, this comes at the price of reduced power relative to population-based association studies. Thus, study design choice has to be made according to study goals, and available resources.