

# Combining Family Studies to Improve Genetic Association Tests

K. F. Cheng

Biostatistics Center, and Department of  
Public Health, China Medical University &  
Graduate Institute of Statistics, National Central University

- Association studies of candidate genes or loci have long been popular among human genetics.
- In this paper we consider family-based association study where sample of family trios is obtained for making association test.
- In this study, each family trio consists of two parents and one diseased child.

- We assume the candidate marker has two alleles **S** and **s** and obtain genotypes for affected individuals and their parents.
- Let **S** be the mutant disease allele and let **s** be the normal allele.
- Let **D** be the event that the individual has the disease and risks be define as

$$P(D | ss) = f_0, P(D | Ss) = f_1, P(D | SS) = f_2$$

- These conditional probabilities would be the penetrances of the candidate gene if it is the one and only disease-causing gene.

- We also define the relative-risk parameters,  $\psi_1 = f_1 / f_0$ ,  $\psi_2 = f_2 / f_0$ , which measure the relative increase in disease probabilities for heterozygous **Ss** individuals and homozygous **SS** individuals, respectively, when compared to that probability for homozygous **ss** individuals.

- The null hypothesis of no association is

$$H_0 : \psi_1 = \psi_2 = 1$$

# Definitions

- $G_o$  = the genotype of the disease child; defined as the number of mutant disease allele carried by the child.
- $G_p$  = parental genotype (mating type), such as  $Ss \times SS$  ( $=SS \times Ss$ , the order is irrelevant); defined in the next table

# Table 1: $P(G_o = g \mid G_p = g_p, D)$

Table 1:  $P(G_o = g \mid G_p = g_p, D)$

$g_p$	Parental Mating type	offspring genotype (g)		
		SS(2)	Ss(1)	ss(0)
0	ss×ss	0	0	1
1	Ss×ss	0	$\frac{\psi_1}{\psi_1 + 1}$	$\frac{1}{\psi_1 + 1}$
2	Ss×Ss	$\frac{\psi_2}{\psi_2 + 2\psi_1 + 1}$	$\frac{2\psi_1}{\psi_2 + 2\psi_1 + 1}$	$\frac{1}{\psi_2 + 2\psi_1 + 1}$
3	SS×ss	0	1	0
4	SS×Ss	$\frac{\psi_2}{\psi_1 + \psi_2}$	$\frac{\psi_1}{\psi_1 + \psi_2}$	0
5	SS×SS	1	0	0

Note: This table is from Schaid and Sommer (1993).

- Using the probabilities given in Table 1, Schaid and Sommer (Am J Hum Genet 1993) derived score tests using conditional likelihood inference under different genetic models.

- Under multiplicative model ( $\psi_2 = \psi_1^2$ ) (or additive model,  $\psi_1 = (1 + \psi_2)/2$ ), the corresponding score test is the usual transmission/disequilibrium test (TDT) suggested by Spielman et al. (1993, Am J Hum Genet).
- This test has gained popularity as a method to evaluate the linkage or association of candidate genes with disease and to screen the genome for susceptibility loci.
- The data used in TDT may come from families that are **simplex** (i.e. families with only one affected offspring), **multiplex** (i.e. families with two or more affected sibs), or **multigenerational**; and the population may exhibit structure.

- Consider the same marker that is used to evaluate association with another disease  $D^*$ .
- Similar conditional probabilities are given in Table 2, where  $\phi_i, i = 1, 2$  are relative risk parameters corresponding to the second disease.

# Table 2: $P(G_o = g \mid G_p = g_p, D^*)$

Table 2:  $P(G_o = g \mid G_p = g_p, D^*)$

$g_p$	Parental Mating type	offspring genotype (g)		
		SS(2)	Ss(1)	ss(0)
0	ss×ss	0	0	1
1	Ss×ss	0	$\frac{\phi_1}{\phi_1 + 1}$	$\frac{1}{\phi_1 + 1}$
2	Ss×Ss	$\frac{\phi_2}{\phi_2 + 2\phi_1 + 1}$	$\frac{2\phi_1}{\phi_2 + 2\phi_1 + 1}$	$\frac{1}{\phi_2 + 2\phi_1 + 1}$
3	SS×ss	0	1	0
4	SS×Ss	$\frac{\phi_2}{\phi_1 + \phi_2}$	$\frac{\phi_1}{\phi_1 + \phi_2}$	0
5	SS×SS	1	0	0

# Example:

- Apolipoprotein  $E_4$  ( $ApoE_4$ ) allele has shown to be linked to Alzheimer's disease.
- However, the genetic association between ApoE and age-related macular degeneration (AMD) also suggested a protective effect of the  $E_4$  allele (Thakkinstian et al. (2006), Am J Epidemiology).
- AMD is the leading cause of blindness in the developed world, accounting for half of all new cases of registered blindness.
- With an aging population, the burden of AMD is set to grow, with almost 30 percent of persons aged 75 years or older showing early signs of disease.

- Note that the joint probability  $G_o$  of  $G_p$  and can be expressed by combining Table 1 (or Table 2) with the mating type probabilities. (Unknown, 5 parameters)
- Conceptually, suppose the samples from two studies are obtained from two homogeneous populations but their mating type probabilities are identical, then one can combine two samples to make more efficient inference about  $\psi_i$  (for example, testing  $H_0 : \psi_1 = \psi_2 = 1$  )

- Under this case, the log of the profiled (mating type probabilities were profiled out) likelihood is

$$\ell_p = \sum_{g_o=0}^2 \sum_{g_p=0}^5 N(g_o, g_p) \log f(g_o, g_p) + \sum_{g_o=0}^2 \sum_{g_p=0}^5 N^*(g_o, g_p) \log f^*(g_o, g_p)$$

where  $N(g_o, g_p)$  is the count corresponding to  $G_o = g_o$  and  $G_p = g_p$  from the combined sample and  $N^*(g_o, g_p)$  is similarly defined but from the “first” sample.

# Table 3: $f(g_o, g_p)$

Table 3:  $f(g_o, g_p)$

		$g_o$		
		0	1	2
$g_p$				
0		$\frac{c_0}{1 + \exp(\alpha)}$	0	0
1		$\frac{c_1}{(1 + \phi_1) + (1 + \psi_1) \exp(\alpha)}$	$\frac{\phi_1 c_1}{(1 + \phi_1) + (1 + \psi_1) \exp(\alpha)}$	0
2		$\frac{c_1}{(1 + 2\phi_1 + \phi_2) + (1 + 2\psi_1 + \psi_2) \exp(\alpha)}$	$\frac{2c_2 \phi_1}{(1 + 2\phi_1 + \phi_2) + (1 + 2\psi_1 + \psi_2) \exp(\alpha)}$	$\frac{c_2 \phi_2}{(1 + 2\phi_1 + \phi_2) + (1 + 2\psi_1 + \psi_2) \exp(\alpha)}$
3		0	$\frac{c_3 \phi_1}{\phi_1 + \psi_1 \exp(\alpha)}$	0
4		0	$\frac{2c_4 \phi_1}{(\phi_1 + \phi_2) + (\psi_1 + \psi_2) \exp(\alpha)}$	$\frac{c_4 \phi_2}{(\phi_1 + \phi_2) + (\psi_1 + \psi_2) \exp(\alpha)}$
5		0	0	$\frac{c_5 \phi_2}{\phi_2 + \psi_2 \exp(\alpha)}$

$$c_{g_p} = N(+, g_p) / \{N(+, +) - N^*(+, +)\}$$

# Table 4: $f^*(g_o, g_p)$

Table 4:  $f^*(g_o, g_p)$

$g_p$	$g_o$		
	0	1	2
0	$\exp(\alpha)$	0	0
1	$\exp(\alpha)$	$\frac{\psi_1}{\phi_1} \exp(\alpha)$	0
2	$\exp(\alpha)$	$\frac{\psi_1}{\phi_1} \exp(\alpha)$	$\frac{\psi_2}{\phi_2} \exp(\alpha)$
3	0	$\frac{\psi_1}{\phi_1} \exp(\alpha)$	0
4	0	$\frac{\psi_1}{\phi_1} \exp(\alpha)$	$\frac{\psi_2}{\phi_2} \exp(\alpha)$
5	0	0	$\frac{\psi_2}{\phi_2} \exp(\alpha)$

# Simulation conditions:

- $p = P(S \text{ allele}) = 0.2, 0.3, 0.4, 0.5$  assuming that the population is in Hardy-Weinberg equilibrium.
- Genetic model:
  1. Recessive model:  $\psi_1 = 1, \psi_2 = \psi$
  2. Dominant model:  $\psi_1 = \psi_2 = \psi$
  3. Multiplicative model:  $\psi_1 = \psi, \psi_2 = \psi^2$
- Null hypothesis  $H_0 : \psi = 1$  , significance level=0.05  
Alternative hypothesis  $H_a : \psi = 1.25, 1.50, 1.75, 2.0$
- Working model
  - = true model, assuming genetic model is known;
  - = general model with two parameters ,  $\psi_1, \psi_2$  assuming genetic model is not known.
- The relative risks for the second disease are  $\phi_1 = \phi_2 = 1$  (the risk allele has no effect on the second disease)
- Sample sizes: 150 obs for the first study; 150 obs for the second study
- Number of replications=5000

# Table 5: Empirical type I error rates/powers (Recessive model)

Table 5: Empirical type I error rates/powers (Recessive model)

$p$	$\psi$	True model		general model	
		Score test	New test	Score test	New test
0.2	1.00	.051	.054	.052	.052
	1.25	.084	.085	.078	.078
	1.50	.147	.162	.119	.126
	1.75	.236	.275	.180	.211
	2.00	.358	.417	.285	.331
0.3	1.00	.048	.049	.049	.047
	1.25	.108	.122	.094	.097
	1.50	.237	.279	.187	.209
	1.75	.420	.491	.345	.402
	2.00	.618	.691	.512	.593
0.4	1.00	.051	.049	.047	.044
	1.25	.128	.149	.114	.120
	1.50	.324	.388	.255	.311
	1.75	.572	.660	.475	.561
	2.00	.774	.855	.678	.779
0.5	1.00	.055	.053	.051	.053
	1.25	.153	.190	.123	.139
	1.50	.406	.484	.325	.398
	1.75	.669	.769	.563	.673
	2.00	.861	.926	.779	.871

# Table 6: Empirical type I error rates/powers (Dominant model)

Table 6: Empirical type I error rates/powers (Dominant model)

$p$	$\psi$	Ture model		general model	
		Score test	New test	Score test	New test
0.2	1.00	.051	.046	.052	.052
	1.25	.154	.194	.119	.148
	1.50	.411	.530	.336	.433
	1.75	.685	.802	.585	.715
	2.00	.854	.929	.776	.886
0.3	1.00	.052	.054	.049	.047
	1.25	.172	.207	.136	.164
	1.50	.445	.551	.347	.443
	1.75	.711	.820	.598	.728
	2.00	.852	.931	.773	.881
0.4	1.00	.050	.047	.047	.044
	1.25	.166	.185	.129	.144
	1.50	.416	.492	.324	.392
	1.75	.645	.759	.539	.653
	2.00	.815	.903	.731	.833
0.5	1.00	.046	.050	.051	.053
	1.25	.141	.159	.113	.134
	1.50	.342	.408	.264	.322
	1.75	.552	.648	.449	.549
	2.00	.725	.813	.629	.720

# Table 7: Empirical type I error rates/powers (Multiplicative model)

Table 7: Empirical type I error rates/powers (Multiplicative model)

$p$	$\psi$	True model		general model	
		Score test	New test	Score test	New test
0.2	1.00	.052	.051	.052	.052
	1.25	.259	.258	.208	.207
	1.50	.682	.680	.578	.576
	1.75	.934	.933	.881	.879
	2.00	.987	.987	.977	.977
0.3	1.00	.049	.049	.047	.047
	1.25	.320	.318	.250	.249
	1.50	.786	.785	.689	.688
	1.75	.968	.968	.930	.929
	2.00	.998	.998	.995	.994
0.4	1.00	.048	.048	.044	.044
	1.25	.351	.349	.273	.274
	1.50	.815	.811	.716	.715
	1.75	.977	.975	.954	.954
	2.00	.997	.996	.994	.994
0.5	1.00	.048	.048	.053	.053
	1.25	.353	.351	.277	.276
	1.50	.811	.806	.729	.728
	1.75	.974	.964	.945	.945
	2.00	.998	.982	.992	.992

# Conclusions

- Both tests maintain the correct type I error rate.
- Tests based on the general model are inferior to those based on the true genetic model. However, the former is more robust.
- Under the recessive and dominant genetic models, the new test shows significant improvement over the score test. In contrast, under the multiplicative model, we do not find any noticeable differences between the two tests. However, if we reduce the two sample sizes to 100, then the difference becomes clearer. In this case, the new test is significantly better than the score test.