



## A New *S*-test for Haplotype Analysis: Concordance and Discordance

Mahnaz Khattak,<sup>†</sup> Shuhrat Shah,<sup>†</sup> and Salahuddin<sup>‡,\*</sup>

<sup>†</sup> Jinnah College for Women

<sup>‡</sup> University of Peshawar

**Abstract.** A new test of inheritance, *S*, is proposed, which uses information from affected as well as unaffected siblings in the family. The siblings are analyzed in terms of similarities of haplotypes. The distribution of the proposed *S*-test is derived under the null hypothesis of random inheritance. Mean and variance are obtained for the distribution. The test is then applied to data sets published in the literature. The results suggest some sort of linkage between haplotypes and disease genes.

**Keywords.** genetics; haplotypes; *S*-test.

### 1 Introduction

To demonstrate the heritability of a trait, one way is to provide evidence for its linkage with a known genetic marker that shows the two traits tend to be inherited together more often than would be expected by chance alone. This implies that the loci with alleles determining the two traits are located at the same chromosome with recombination frequency less than  $1/2$ , which means they are linked.

Penrose (1937, 1953) first discussed the problem of detecting linkage between a quantitative trait and a marker locus. His test uses data on inde-

\* Corresponding author

pendent pairs. Many workers have contributed a lot to the development and generalization of sib-pair methods in different situations.

The affected sib-pairs (AS) methods assume the presence of a tightly linked disease susceptibility locus (DS) in the vicinity of HLA region. Sib-pairs from different families are categorized according to whether they share 0, 1 or two haplotypes identical by descent (IBD).

Sib-pair methods can be extended to include any relative that share at least one haplotype in common and this could be generalized to apply to extended pedigrees (Cantor and Rotter, 1987; Cantor, 1989). De Vries et al. (1976) used the criterion

$$F = \{( \text{maximum} - \text{minimum} ) \text{ number of haplotypes from one parent} \} \\ + \{( \text{maximum} - \text{minimum} ) \text{ number of haplotypes from the other parent} \}.$$

Green et al. (1983) improved the *F*-test of de Vries et al. (1976) by using criterion *N* which takes account of the family size distribution as well. It is given by

$$N = \{ \text{maximum haplotype frequency from one parent} \} + \{ \text{maximum haplotype frequency from the other parent} \}.$$

Criterion *N* is essentially equivalent to *F*. That is *F* = *N* - *s*, where *s* is the number of minimum HLA haplotypes from both parents. HLA is the abbreviation of human histocompatibility system — a blood group like system detected on the white cells of the blood.

All these tests use information from affected sibs only, but to fully utilize the data available on genotypes of unaffecteds as well, tests involving unaffecteds have been suggested (for example, Spielman et al., 1980; Clerget-Darpoux et al., 1980; Rubinstein et al., 1981).

Green and Montasser (1988) proposed another *T*-test based on haplotype discordance. This is defined as  $T = n_1 + n_2 - n_3$ , Where  $n_1$  is the *N*-measure for the affected sibs in a sibship,  $n_2$  is the *N*-value for the unaffected sibs in the same sibship, and  $n_3$  is the *N*-measure of the whole sibship, with *N* as defined earlier. One advantage of the *T*-test and others using unaffecteds, is that sibships with only one affected and only one unaffected can contribute. A recent survey of these tests, their powers, relative merits and demerits, various extensions and generalizations is given by Green and Shah (1993), and Shah and Green (1993, 1994).

A new method for differentiating between groups of patients according to severity of disease proposed by Shah et al. (1995) is found to be an effective tool in analyzing data sets with respect to disease severity. Khattak et al.

(2005) suggested a simpler and easier method to distinguish between recessive and dominant mode of inheritance of HLA associated diseases and also provides measures to estimate probability of the disease under consideration in the population.

## 2 The Proposed *S*-test

On the same pattern, we will present a new test, which is based upon the proband and his/her affected as well as non-affected siblings. Here the siblings are analyzed in terms of similarities of haplotypes. The hypothesis is to test that the disease has random pattern of inheritance against the alternative hypothesis that it has non-random pattern of inheritance.

The new *S*-test is a modification of *M*-test of haplotype concordance in which affected as well as unaffected sibs in a sibship are considered. The *M*-test is based on the proband and his/her affected sibs only and uses haplotype concordance among all affected children in a family. The *M*-test is defined as

$$M = \{\text{Sum of haplotypes from both parents in the affected siblings}\} \\ - \{\text{the number of distinct haplotypes in the affected siblings}\}.$$

That is, if  $m$  is the number of affected siblings and  $k$  is the number of distinct haplotypes, then  $M = 2m - k$ , where  $k = 2, 3$ , or  $4$  as the sibs may share  $2, 3$ , or  $4$  distinct genes from two heterozygous parents having  $a, b, c$ , and  $d$  genes. The *S*-test is applicable only when there is at least one unaffected sib in a sibship. However, in case all sibs are affected in a sibship, *S*-test is essentially similar to *M*-test.

The *S*-test is stated as

$$S = \{\text{Sum of haplotypes from both parents in the affected siblings}\} \\ - \{\text{the number of distinct haplotypes in the affected siblings}\} \\ - \{\text{Sum of haplotypes from both parents in the unaffected siblings}\} \\ - \{\text{the number of distinct haplotypes in the unaffected siblings}\}.$$

That is,  $S = (2m - k_1) - (2r - k_2)$ , where  $m$  is the number of affected siblings and  $r$  is the number of unaffected siblings.  $k_1$  and  $k_2$  take values  $2, 3$ , or  $4$  as the siblings may share  $2, 3$ , or  $4$  different genes from two parents having  $a, b, c$  and  $d$  genes in affected and non-affected sibs. We assume here that the two parents are heterozygous.

If we proceed with  $m = 2$  and  $r = 1$ , and  $m = 3$  and  $r = 2$ , then the possible values of *S* will be calculated as in Table 1. Similary, if we proceed with  $m = 4$  and  $r = 1, 2$ , and  $3$ , then the possible values of *S* will be calculated

**Table 1.** Possible combinations of haplotype in 2 to 3 affected and 1 to 2 unaffected sibs with their *S*-scores

$m = 2$	$r = 1$	$S_{21}$	$m = 3$	$r = 2$	$S_{31}$	$S_{32}$
ac ac	ac	2	ac ac ac	ac ac	4	2
ac ad	ac	1	ac ac ad	ac ac	3	1
ac bc	ac	1	ac ac bc	ac ac	3	1
ac bd	ac	0	ac ac bd	ac ac	2	0
ac ac	ad	2	ac ac ac	ac ad	4	3
ac ad	ad	1	ac ac ad	ac ad	3	2
ac bc	ad	1	ac ac bc	ac ad	3	2
ac bd	ad	0	ac ac bd	ac ad	2	1
ac ac	bc	2	ac ac ac	ac bc	4	3
ac ad	bc	1	ac ac ad	ac bc	3	2
ac bc	bc	1	ac ac bc	ac bc	3	2
ac bd	bc	0	ac ac bd	ac bc	2	1
ac ac	bd	2	ac ac ac	ac bd	4	3
ac ad	bd	1	ac ac ad	ac bd	3	2
ac bc	bd	1	ac ac bc	ac bd	3	2
ac bd	bd	0	ac ac bd	ac bd	2	1

**Table 2.** Possible combinations of haplotype in 4 affected and 1 to 3 unaffected sibs with their *S*-scores

$m = 4$	$r = 1$	$r = 2$	$r = 3$	$S_{41}$	$S_{42}$	$S_{43}$
ac ac ac ac	ac	ac ac	ac ac ac	6	4	2
ac ac ac ad	ac	ac ac	ac ac ac	5	3	1
ac ac ac bc	ac	ac ac	ac ac ac	5	3	1
ac ac ac bd	ac	ac ac	ac ac ac	4	2	0
ac ac ac ac	ad	ac ad	ac ac ad	6	5	3
ac ac ac ad	ad	ac ad	ac ac ad	5	4	2
ac ac ac bc	ad	ac ad	ac ac ad	5	4	2
ac ac ac bd	ad	ac ad	ac ac ad	4	3	1
ac ac ac ac	bc	ac bc	ac ac bc	6	5	3
ac ac ac ad	bc	ac bc	ac ac bc	5	4	2
ac ac ac bc	bc	ac bc	ac ac bc	5	4	2
ac ac ac bd	bc	ac bc	ac ac bc	4	3	1
ac ac ac ac	bd	ac bd	ac ac bd	6	6	4
ac ac ac ad	bd	ac bd	ac ac bd	5	5	3
ac ac ac bc	bd	ac bd	ac ac bd	5	5	3
ac ac ac bd	bd	ac bd	ac ac bd	4	4	2

as in Table 2. In these tables,  $S_{mr}$  indicates the  $S$ -score for  $m$ , number of affected sibs, and  $r$ , number of unaffected sibs in a sibship. These are some possible combinations for variable values of  $m$  and  $r$  ( $m > r$ ) and their respective  $S$ -values, where  $S = (2m - k_1) - (2r - k_2) = 2(m - r) - (k_1 - k_2)$ .

## 2.1 Probability Distribution of $S$

We start with a sibship of size two with both affected sibs. We assume heterozygous parents. The variable  $S$  will take values 0, 1, or 2 when the numbers of distinct haplotypes are 4, 3, or 2. Consider the following possible combinations for two affected sibs,

		Haplotype Score $S$
$ac$	$ac$	$2(2 - 0) - 2 = 2$
$ac$	$ad$	$2(2 - 0) - 3 = 1$
$ac$	$bc$	$2(2 - 0) - 3 = 1$
$ac$	$bd$	$2(2 - 0) - 4 = 0$

where  $S = 2(m - r) - (k_1 - k_2)$ . Here  $m = 2$  and  $r = 0$  (no unaffected).  $k_1$  is the number of distinct haplotypes in the affected sibs (2,3,3,4) and  $k_2$  is the number of distinct haplotypes in the unaffected siblings.

The scores 0, 1, and 2 occur with probabilities  $1/4$ ,  $1/2$ , and  $1/4$ , respectively. If a sibship of size  $m$  is taken with no unaffected sib, the score will be  $(2m - 2)$ ,  $(2m - 3)$ , and  $(2m - 4)$  with probabilities  $\{2^{-(m=1)}\}^2$ ,  $2\{2^{-(m=1)}\}[1 - \{2^{-(m=1)}\}]$ , and  $[1 - \{2^{-(m=1)}\}]^2$ , respectively. If we generalize it to a sibship of size  $(m + r)$  with  $m$  affected and  $r$  unaffected sibs, the variable  $S$  takes values  $2(m - r) - 2$ ,  $2(m - r) - 1$ ,  $2(m - r)$ ,  $2(m - r) + 1$ , and  $2(m - r) + 2$ .

For  $k_1$  and  $k_2$  taking values only 2, 3, or 4, the variable  $S$  takes only the following five possible values.

$2(m - r) - 2$	$k_1 = 4, k_2 = 2$
$2(m - r) - 1$	$k_1 = 4, k_2 = 3$ or $k_1 = 3, k_2 = 2$
$2(m - r)$	$k_1 = k_2 = 2$ or 3 or 4
$2(m - r) + 1$	$k_1 = 2, k_2 = 3$ or $k_1 = 3, k_2 = 4$
$2(m - r) + 2$	$k_1 = 2, k_2 = 4$

Now we can present this new variable  $S$  along with their probabilities as in

**Table 3.** Probability distribution of *S* for  $m + r$  size sibships

<i>S</i>	$P(S = s)$
$2(m - r) - 2$	$(1 - 2^{-m+1})^2(2^{-r+1})^2$
$2(m - r) - 1$	$(1 - 2^{-m+1})^2 2(2^{-r+1})(1 - 2^{-r+1}) + 2(2^{-m+1})(1 - 2^{-m+1})(2^{-r+1})$
$2(m - r)$	$(2^{-m+1})^2(2^{-r+1})^2 + 4(4^{-m+1})(2^{-r+1})(1 - 2^{-m+1})(1 - 2^{-r+1})$ $+ (1 - 2^{-m+1})^2(1 - 2^{-r+1})^2$
$2(m - r) + 1$	$2(2^{-m+1})^2(2^{-r+1})(1 - 2^{-r+1}) + 2(2^{-m+1})(1 - 2^{-m+1})(1 - 2^{-r+1})^2$

Table 3. The sum of probabilities is equal to one, hence it shows that it is a complete probability distribution, and we can derive its mean and variance easily.

## 2.2 Mean and Variance of *S* Under $H_0$

$$\begin{aligned} E(S) &= 2(m - r) + 2^{-m+1} - 2^{-r+1} \\ &= 2\{(m - r) + 2^{-m} - 2^{-r}\}, \end{aligned}$$

$$\text{var}(S) = 2\{2^{-m+1}(1 - 2^{-m+1}) + 2^{-r+1}(1 - 2^{-r+1})\}.$$

Thus mean and variance of *S* are  $2\{(m - r) + 2^{-m} - 2^{-r}\}$  and  $2\{2^{-m+1}(1 - 2^{-m+1}) + 2^{-r+1}(1 - 2^{-r+1})\}$ , respectively.

Now for any number of affected (*m*) and non-affected sibs (*r*) in a sibship of size  $m + r$ , we can find the expected mean and variance of *S*. We can present different numbers of affected and non-affected siblings using the *S*-test along with their probabilities and expected means and variances in a tabular form in Table 4. We will assume that the parents are heterozygous, i.e. they have no common haplotypes and the number of affected sibs is greater than the number of unaffected sibs, i.e.  $m \geq r$ .

## 3 Application to Data Sets

To illustrate the application of *S*, we analyze data relating to rheumatoid arthritis which consist of 22 families with at least one member affected by the disease, which was collected by Cudworth and Woodworth (1975) and was found in the literature. The families were typed for HLA-A, B, Cw and DR alleles by the hematology department. The analysis of the rest of the families is shown in Table 5.

**Table 4.** Probability distribution of  $S$  for  $m + r$  size sibships along with their means and variances

	$S_{21}$	$P(S_{21})$	$S_{31}$	$P(S_{31})$	$S_{32}$	$P(S_{32})$	$S_{41}$	$P(S_{41})$	$S_{42}$	$P(S_{42})$	$S_{43}$	$P(S_{43})$	$S_{mr}$	$P(S_{mr})$
	0	$\frac{1}{4}$	2	$\frac{9}{16}$	0	$\frac{9}{64}$	4	$\frac{49}{64}$	2	$\frac{49}{256}$	0	$\frac{49}{1024}$	$2(m - r) - 2$	$(1 - 2^{-m+1})^2 (2^{-r+1})^2$
	1	$\frac{2}{4}$	3	$\frac{6}{16}$	1	$\frac{24}{64}$	5	$\frac{14}{64}$	3	$\frac{112}{256}$	1	$\frac{308}{1024}$	$2(m - r) - 1$	$(1 - 2^{-m+1})^2 2(2^{-r+1})(1 - 2^{-r+1})$ $+ 2(2^{-m+1})(1 - 2^{-m+1})(2^{-r+1})$
	2	$\frac{1}{4}$	4	$\frac{1}{6}$	2	$\frac{22}{64}$	6	$\frac{1}{64}$	4	$\frac{78}{256}$	2	$\frac{526}{1024}$	$2(m - r)$	$(2^{-m+1})^2 (2^{-r+1})^2 + 4(4^{-m+1})$ $(2^{-r+1})(1 - 2^{-m+1})(1 - 2^{-r+1})$ $+ (1 - 2^{-m+1})^2 (1 - 2^{-r+1})^2$
					3	$\frac{8}{64}$			5	$\frac{16}{256}$	3	$\frac{132}{1024}$	$2(m - r) + 1$	$(1 - 2^{-m+1})^2 (1 - 2^{-m+1})^2$ $+ 4(2^{m+1})(1 - 2^{-m+1})(2^{-r+1})$ $(1 - 2^{-r}) + (2^{-m+1})^2 (2^{-r+1})^2$
mean	1				4	$\frac{1}{64}$			6	$\frac{1}{256}$	4	$\frac{9}{1024}$	$2(m - r) + 2$	$(2^{-m+1})^2 (1 - 2^{-r+1})^2$ $2\{(m - r) + 2^{-m} - 2^{-r}\}$
variance	$\frac{1}{2}$				$\frac{5}{2}$	$\frac{3}{2}$			$\frac{17}{4}$	$\frac{13}{4}$	$\frac{7}{4}$			$2\{2^{-m+1}(1 - 2^{-m+1})$ $+ 2^{-r+1}(1 - 2^{-r+1})\}$

**Table 5.** Cudworth's family data on rheumatoid arthritis and HLA typing

Family	Affected Sibs With Haplotypes	Unaffected Sibs With Haplotypes	<i>S</i>	<i>E(S)</i>	<i>var(S)</i>
1	$\frac{a}{b}, \frac{c}{b}$	$\frac{c}{b}$	1	1	$\frac{1}{2}$
2	$\frac{a}{d}, \frac{a}{b}$	—	1	1	$\frac{1}{2}$
3	$\frac{a}{b}, \frac{a}{b}$	—	2	1	$\frac{1}{2}$
4	$\frac{a}{b}, \frac{a}{b}$	$\frac{c}{d}$	2	1	$\frac{1}{2}$
5	$\frac{a}{b}, \frac{a}{d}$	—	1	1	$\frac{1}{2}$
6	$\frac{a}{b}, \frac{c}{b}$	—	1	1	$\frac{1}{2}$
7	$\frac{a}{b}, \frac{a}{d}, \frac{a}{b}$	$\frac{c}{d}$	3	$\frac{5}{2}$	$\frac{3}{8}$
8	$\frac{c}{b}, \frac{c}{b}$	—	2	1	$\frac{1}{2}$
9	$\frac{a}{b}, \frac{c}{d}$	—	0	1	$\frac{1}{2}$
10	$\frac{a}{d}, \frac{a}{d}$	—	2	1	$\frac{1}{2}$
11	$\frac{a}{b}, \frac{c}{b}$	—	1	1	$\frac{1}{2}$
12	$\frac{a}{b}, \frac{c}{b}, \frac{c}{b}$	—	3	$\frac{5}{2}$	$\frac{3}{8}$
13	$\frac{a}{b}, \frac{a}{b}$	$\frac{a}{d}$	2	1	$\frac{1}{2}$
14	$\frac{a}{d}, \frac{c}{d}$	—	1	1	$\frac{1}{2}$
15	$\frac{a}{d}$	$\frac{a}{d}$	0	0	$\frac{1}{2}$
16	$\frac{a}{b}$	$\frac{a}{b}$	0	0	$\frac{1}{2}$
17	$\frac{a}{d}, \frac{c}{d}, \frac{a}{b}$	$\frac{c}{d}$	2	$\frac{5}{2}$	0
18	$\frac{c}{b}, \frac{c}{d}, \frac{c}{d}$	$\frac{a}{d}, \frac{c}{d}$	2	$\frac{3}{2}$	0
19	$\frac{c}{b}$	—	0	0	$\frac{3}{8}$
20	$\frac{a}{b}, \frac{c}{d}$	—	0	1	$\frac{7}{8}$
21	$\frac{a}{b}, \frac{a}{b}$	—	2	1	0
22	$\frac{a}{b}$	$\frac{c}{b}$	0	0	$\frac{1}{2}$
Total			28	23	9

hematology department. The analysis of the rest of the families is shown in Table 5.

The test criterion yielded by *S* is

$$T = \frac{\sum S - \sum E(S)}{\{\sum V(S)\}^{1/2}},$$

that is,  $T = (28 - 0.05 - 23)/3 = 1.667$  and  $p = 0.0475$ , which is just significant at 5% level. Hence it reveals some sort of linkage which is perhaps sufficient to suggest that the possibility of association between HLA and disease genes will be worth exploring when further samples are taken and of course the sample size is enlarged and the data are well randomized.

## 4 Conclusion

The test gives expected means and variance for any number of affected and non-affected siblings. If the number of unaffecteds is greater than the number of affecteds, that is, if for example we have  $S_{13}$  instead of  $S_{31}$  then the variance will remain the same but mean will have the same value with a negative sign. Further, for equal numbers of affected and non-affected siblings, that is,  $m = r$ , mean will always be zero and variance will be  $4\{2^{-m+1}(1 - 2^{-m+1})\}$ . The new  $S$ -test is simpler and easily applicable than the already established test for disease association with genes and their non-random inheritance.

The new test not only takes haplotype concordance among the affected siblings into consideration, but it also considers haplotype discordance in the whole sibship. Usually the information given by non-affected siblings of diseased person were ignored in the past. The new  $S$ -test provides room for the incorporation of information contained in unaffected siblings. The distribution of this test under  $H_0$  was the only way to calculate its mean and variance, and then apply it to data set for detecting linkage disequilibrium. The results suggest some sort of linkage between haplotypes and disease genes.

## References

Cantor, R.M. (1989). A linkage test with identity-by-descent marker data from pairs of affected relatives. *Prog. Clin. Biol. Res.* **329**, 111-116.

Cantor, R.M.; Rotter, J.I. (1987). Marker concordance in pairs of distant relatives: a new method of linkage analysis for common diseases. *Ann. Hum. Genet.* **51**, A252.

Clerget-Darpoux, F.; Bonaiti-Pellie, C.; Hors, J.; Deschamps, I.; Fein-gold, N. (1980). Application of the Lod score method to detection of linkage between HLA and juvenile insulin-dependent diabetes. *Clin. Genet.* **18**, 51-57.

Cudworth, A.G.; Woodwrth, J.C. (1975). Evidence for HLA linked gens in "Juvenile diabetes". *Br. Med. J.* **3**, 133-135.

Green, J.R.; Montasser, M. (1988). HLA haplotype discordance. *Biometrics* **44**, 941-950.

Green, J.R.; Woodrow, J.C. (1977). Sibling Method for detecting HLA-linked genes and disease. *Tissue Antigens* **9**, 31-35.

Green, J.R.; Low, H.C.; Woodrow, J.C. (1983). Inference on the inheritance of disease using repetitions of HLA haplotypes in affected siblings. *Ann. Hum. Genet.* **47**, 73-82.

Green, J.R.; Shah, S. (1993). Power comparison of various sibships tests of association. *Ann. Hum. Genet.* **57**, 151-158.

Khattak, M.; Shah, S.; Salahuddin (2005). Mode of inheritance of HLA-associated diseases. *Pak. J. Stat.* **21**, 203-208.

Penrose, L.S. (1937). Genetic linkage in graded human characters. *Ann. Eugen.* **8**, 233-237.

Penrose, L.S. (1953). The general purpose sib-pair linkage test. *Ann. Eugen.* **18**, 120-124.

Rubinstein, P.; Ginsberg-Fellner, F.; Falk, C. (1981). Genetics of type I diabetes mellitus: a single, recessive predisposition gene mapping between HLA-B and GLO. With an appendix on the estimation of selection bias. *Am. J. Hum. Genet.* **33**, 865-882.

Shah, S.; Green, J.R. (1993). Testing for a haplotype concordance with incomplete parental data-modified Q and correction for modified N. *Ann. Hum. Genet.* **57**, 239-247.

Shah, S.; Green, J.R. (1994). The distribution of Q: A powerful sibship test of association. *Ann. Hum. Genet.* **58**, 163-173.

Shah, S.; Khattak, M.; Qazi, U. (1995). A test of inheritance for disease severity. *Pak. J. Hist. Phil.* **1**, 29-36.

Spielman, R.S.; Baker, L.; Zmijewski, C.M. (1980). Gene dosage and susceptibility to insulin dependent diabetes. *Ann. Hum. Genet.* **44**, 135-150.

de Vries, R.R.; Fat, R.F.; Nijenhuis, L.E.; Van Rood, J.J. (1976). HLA-linked genetic control of host response to *Mycobacterium Leprae*. *Lancet* **2**, 1328-1330.

**Mahnaz Khattak**  
Jinnah College for Women,  
Peshawar, Pakistan.

**Shuhrat Shah**  
Jinnah College for Women,  
Peshawar, Pakistan.

**Salahuddin**  
Department of Statistics,  
University of Peshawar.  
Peshawar, Pakestan.  
e-mail: *salahuddin\_90@hotmail.com*