Document for HS-TDT software

Shuanglin Zhang

Department of Mathematical Sciences
Michigan Technological University
1400 Townsend Drive
Houghton MI 49931 USA

INTRODUCTION

The program HS-TDT implements a haplotype frequency estimation method based on nuclear family data (Chen and Zhang 2003) and Haplotype Sharing TDT (HS-TDT) test for tightly linked markers and nuclear family (Zhang et al. 2003, 2004, Sha et al. 2004). The method implemented in this version of HS-TDT uses Partition-Ligation EM algorithm to estimate haplotype frequencies by incorporating nuclear family information. The program allows missing genotype, missing parents, and can deal very large number of markers. The detail methods used are introduced in the above mentioned papers.

We strongly encourage users to send me a brief email at shuzhang@mtu.edu, saying that you have download the program and what sort of the questions the program applied to. This will help us to assess the amount of interest in HS-TDT, and how much effort we should put into maintaining and improving it.

GETTING STARTED

You will need an executable file: *HS-TDT.exe* (both for windows and unix), one file with name *parameters*, and one input data files (the name of the input data file given in the file *parameters*). We strongly suggest that you put that executable file, parameters file and input data files in a one folder. To analyze your own data, you need to prepare the input data file in the appropriate format.

PREPARE THE "PARAMETERS" FILE

In the parameter file *parameters* (see example file), the lines beginning with ## are description lines. Do not change the description lines. You can change the number or words of non-description line.

For the term "how many markers in each partition", this is the number of markers in one partition. For example, there are 38 markers. You put 10 markers for each partition, the program will divided the total region into 4 parts. Each of the first three parts contains 10 markers, and the fourth part contains 8 markers. For computational consideration, we suggest that number of markers in each partition <=30 for the case of not many missing, <=15 if there are more than 10% genotype are missing, and <=6 if there are many parents are missing. for each and number lines.

The purpose of merging rare haplotype is to control the false-position results of HS-TDT caused by genotyping errors. The simulation results in Sha et al (2004) suggest that choosing the cutoff value α between 1% to %2 can well control the false-positive for a wide range of genotyping error rates. The meaning of cutoff value α is that the haplotype with frequency<= α is considered as rare haplotype and will be merge to the most similar common haplotype. For each term, only has one description line, the number following the description line can take one line or more than one line.

For the data file, the first line (a description line) is followed by a distinct diseased haplotype tables. The last column of the table is the number of this haplotype. For example, the first haplotype is in the second row. The first, second, … column of the second row is the marker allele of the first haplotype at first, second, … marker.

FORMAT FOR INPUT DATA FILE

The format of the input data file is similar to the input file of linkage analysis. A sample input file is given below

Family_ID	Ind_ID	Fa_ID	Mo_ID	Trait_value	marker_1	marker_2	marker_3	marker_4	marker_5
1	1	0	0	-1	0 0	0 0	0 0	0 0	1 1
1	2	0	0	-1	0 0	0 0	0 0	0 0	1 1
1	3	1	2	2.296042	0 0	0 0	0 0	0 0	1 1
1	4	1	2	1.386646	0 0	0 0	0 0	0 0	1 1
2	5	0	0	-1	0 0	0 0	0 1	0 0	1 1
2	6	0	0	-1	0 0	0 0	-1 1	0 0	1 0
2	7	5	6	-0.906672	0 0	0 0	0 1	0 0	1 0
3	8	0	0	-1	-1 0	1 1	1 1	0 0	1 1
3	9	0	0	-1	-1 0	1 1	0 1	0 0	1 0
3	10	8	9	0.221613	-1 0	1 1	0 1	0 0	1 0
4	11	0	0	-1	0 0	0 0	0 1	0 0	1 0
4	12	0	0	-1	1 0	1 0	0 0	0 0	1 1
4	13	11	12	-0.368437	1 0	1 0	0 0	0 0	1 1

The first line is a description line. Following the description line, the data for each individual takes one line. -1 denotes missing value. For each nuclear family, the first two individuals are parents. If you only estimate haplotype frequencies and do not have trait value. You can put an arbitrary value for trait.

FORMAT FOR OUT DATA FILE

There is one output file (the name of the output file is given in the *parameters* file). A sample output file is given below

==========

No. of total haplotypes:
7
Haplotype No. frequency
1 0 0 0 0 1 0.492187
2 0 0 1 0 1 0.070313
3 0 0 1 0 0 0.125000
4 0 1 1 0 1 0.122919
5 0 1 0 0 0 0.062500
6 1 1 1 0 1 0.064581
7 1 1 0 0 1 0.062500
Then following are the family No. and the number of compatible haplotype groups;
Next line: the haplotype numbers of father and mother in each compatible group
1 1
1 1 1 1
2 2
1 2 3 1
1 2 3 2
3 6
4 4 5 4
4 4 5 6
4 6 5 4
6 4 5 4
4 6 5 6
6 4 5 6
4 1
1 3 7 1

The pvalues of the tests

1th marker 0.274500
2th marker 0.609000
3th marker 0.338000
4th marker 0.377000
5th marker 0.442000
Overall 0.428500

If you merge the haplotypes, the total number of haplotypes is the number of after merging. For each family, we first give the family number and the number of compatible haplotype groups, followed by the haplotypes of father and mother for each compatible haplotype group. For example, the first family has 1 compatible haplotype group, the second family has 2, the third family has 6, and fourth family has 1. For the second family, for example, in the first compatible haplotype group, father has haplotypes 1 and 2 (the number of the haplotype given at the upper part of the file, haplotypes 1 and 2 are the haplotypes 0 0 0 0 1 and 0 0 1 0 1) mother has haplotypes 3 and 1, and in the second compatible haplotype group, father has haplotypes 1 and 2 and mother has haplotypes 3 and 2,

References

Qin ZS, Niu T, Liu JS (2002) Partition-Ligation–Expectation-Maximization Algorithm for Haplotype Inference with Single-Nucleotide Polymorphisms. Am J Hum Genet 71:1242-1247

Zhang SL, Sha Q, Chen HS, Dong J, Jiang R (2004) Impact of genotyping error on type I error rate of the haplotype-sharing transmission /disequilibrium test (HS-TDT): reply to Knapp and Becker. *Am J Hum Genet 74*: 591-593.

Zhang SL, Sha Q, Chen HS, Dong J, Jiang R (2003) Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. *Am J Hum Genet 73:566–579*

Sha Q, Dong J, Jiang R Chen HS, Zhang S (2004) Haplotype Sharing Transmission/Disequilibrium Tests That Allow for Genotyping Errors. submitted