User's Guide to TDTae program version 2.04

20 Nov 2012

TABLE OF CONTENTS

1.0 OVERVIEW

This program is compiled to work in UNIX Solaris, LINUX, and Windows (PC) operating systems. All commands are executed from the command line in UNIX and LINUX or CMD (command) prompt in Windows.

This program is designed to perform a likelihood-based transmission disequilibrium test (TDT) on genotype data from families in which there is at least one affected child. Unlike other programs that perform analyses with TDT, our program will allow Mendelian inconsistencies to be present in the data. A two-stage search procedure is implemented (see Section 4.1) to compute the maximum log-likelihoods under a null (H0) and alternative hypothesis (H1), and the difference of these log likelihoods provides the value of the test statistic.

A key assumption in our method is that errors that occur in the data are random and independent. If this assumption is incorrect, or if the genotype data are “cleaned” (Mendelian inconsistencies removed before use of the TDTae program) results of our analysis are likely to be invalid.

1.1 Theory

As mentioned above, the TDTae method performs a likelihood ratio test to test for linkage in the presence of association. Our motivation for developing this method is robustness. Two issues regarding the robustness of the original transmission disequilibrium test (TDT) developed by Spielman et al. (Spielman et al. 1993) are: (i) missing parental genotype data and (ii) the presence of undetected genotype errors. While extensions of the

TDT that are robust to items (i) and (ii) have been developed, there was no single TDT statistic that is robust to both for general pedigrees. We developed a likelihood method (Gordon et al. 2001; Gordon et al. 2004), the TDTae, which is robust to these items in general pedigrees. The TDTae assumes a more general disease model than the traditional TDT, which assumes a multiplicative inheritance model for genotypic relative risk. Our model is based on Weinberg's work (Weinberg 1999). Full details of the TDTae method may be found in our paper (Gordon et al. 2004). The TDTae statistic is given by the formula,

$$\mathrm{TDT}_{ae} = 2\ln[L(G'_p, \hat{R}_1, \hat{R}_2, \hat{p}_{11}, \hat{p}_{12}, \hat{E}) / L(G'_p, 1, 1, \tilde{p}_{11}, \tilde{p}_{12}, \tilde{E})], \qquad (*)$$

Where G'_p is the set of observed (and possibly inconsistent) genotypes for all pedigrees, and $\boldsymbol{R_1}$ = Pr(*aff*|+*d*)/Pr(*aff*|++) and $\boldsymbol{R_2}$ = Pr(*aff*|*dd*)/Pr(*aff*|++) are the genotypic relative risks for an (unobserved) di-allelic trait locus with low-risk (wild-type) allele + and high-risk (disease) allele *d*, p_{11} is the population genotype frequency of the *11* genotype, p_{12} is the population genotype frequency of the *12* genotype, and *E* is the vector of parameters for a given error model (see below for list of error models and their parameters). The null hypothesis is that $R_1 = R_2 = 1$. The likelihood, $L(G'_p, R_1, R_2, p_{11}, p_{12}, E)$, which is a function of the parameters R_1, R_2, p_{11}, p_{12}, *E*, is maximized under the alternative (carat in equation (*)) and under the null (~ in equation (*), with R_1, R_2 set equal to 1.0). Twice the log-difference is asymptotically distributed as a central χ^2 distribution with 2 degrees of freedom. By placing constraints on the genotypic relative risks, the degrees of freedom for the TDTae statistic may be reduced to 1 (see Section 2.1.1). The SNP with alleles *1* and *2* is taken to be the disease variant so that *2* = *d*.

This software program computes the maximum log-likelihoods under the alternative and null hypotheses and thus produces maximum likelihood estimates (MLEs) of R_1, R_2, p_{11}, p_{12}, *E* under each hypothesis. Discussion of the maximization procedure is described below (section 4.1).

2.0 RUNNING THE TDTae PROGRAM

2.1 Input files and usage

With our new version of TDTae (version 2.0), users only need one file to run the TDTae program. It is a pedigree file (described below). Another optional file is a marker file that contains the list of markers being used in the analysis.

Pedigree file

This file contains information on pedigree structure, affection status, and genotypes. This file is the same as the "pedin.pre" or "pedin.dat" file in the LINKAGE and FASTLINK programs (see Terwilliger and Ott 1994). This file can be in either pre-MAKEPED (default) or post-MAKEPED LINKAGE format, delimited by either spaces or tabs. If there are any observed Mendelian inconsistencies in the dataset, our program requires that such inconsistencies NOT BE REMOVED from the pedigree file. We make this requirement to insure that the results from the TDTae analyses are valid (Gordon et al. 2001).

An example of a pedigree file (in pre-MAKEPED format) is given below:

Pedigree	Ind.	Father	Mother	Gender	Aff	Locus 1 Genotype	Locus 2 Genotype etc
1	1	0	0	1	1	1 1	1 2
1	2	0	0	2	1	2 1	1 2
1	3	1	2	2	2	2 2	1 2
2	1	0	0	1	1	1 1	1 2
2	2	0	0	2	1	2 2	0 0
2	3	1	2	2	2	2 2	1 2
2	4	1	2	1	2	1 1	2 2

…

Note that in this example file, there are Mendelian inconsistencies at the first marker locus for the first and second families. Also, IDs for pedigrees and individuals may be alphanumeric; affection status is 1 for unaffected, 2 for affected, and 0 for unknown.

Marker file

This file is a list of marker names corresponding to the markers in the pedigree file. There is no header line - one just lists the names of each marker, starting with the first locus in your list.

An example of a marker file is given below:

D8S1125
D8S565
GATA3266
SNP1
…

Usage of the program is as follows:

> tdtae [OPTIONS] <input file> <error model> [<locus> ..]

We explain each item:

[OPTIONS]: A list of options that may be used when running TDTae (see Section 2.1.1).

<input file>: The name of the pedigree file.

<error model>: The selected error model for use in the analysis – The options are GLHO, DSB, SPL, or MA. More details on each of these error models is provided below (see Section 2.3.1).

[<locus>..] (optional): The list of markers for which the TDTae analysis will be performed. Note that this list must be a list of positive integers corresponding to ordered markers in the pedigree file. If no list is provided, the program will run on all markers.

It is important to note that the list of options MUST come before the pedigree file and error model chosen on the command line.

Example:
Suppose the name of the pedigree file is “pederr.pre” and the name of the marker file is “markers.txt”. Suppose further that pederr.pre has 5 marker loci that have been genotyped. If the name of our executable code is “tdtae” then we can run the TDTae program by typing:

>tdtae pederr.pre MA 2 3

at the command line. Here, the phrase MA indicates what error will be used when performing the analysis. In this case, it is the Mote-Anderson error model (Mote and Anderson 1965) (see Section 2.3). The numbers following “MA” indicate that only markers 2 and 3 will be analyzed (out of a possible 5 markers).

If we type

>tdtae pederr.pre MA

then all 5 loci will be analyzed.

2.1.1 Command line options

The TDTae program version 2.0 comes with a list of features that are available at the command line. To access these features, type

> tdtae

at the command line (prompt) and hit the Enter (Return) key. You should see the following list of command line features.

Missing arguments
Program TDTAE - Version 2.0 using NR library
Usage: tdtae [OPTIONS] <input file> <error model> [<locus> ..]

OPTIONS:

-a Specify minimum allele percentage (default: 10)
-g Group alleles with low count

-s Calculate support interval
-b Set support bound (default: 2)

-n Specify number of search results to use (default: 5)
-po Input file is in post-MAKEPED format
-o Specify output file
-f Specify file containing marker names
-t Specify file for trimming output
-x Specify maximum number of founders (default: 9)
-v Verbose
-e <outfile> Calculate deviations from Hardy-Weinberg Equilibrium

-d Use dominant model
-r Use recessive model
-m Use multiplicative model
-c Specify number of cuts (default: 5)

Valid Error Models are: DSB, GLHO, SPL, MA.
If no loci are specified all will be analyzed.

We explain each of these options in the order of their appearance above.

-a: The default setting for the minimum minor allele frequency of any allele being tested is 10%. That means, for a SNP with two alleles coded *1* and *2*, unless either allele has a frequency of 10%, TDTae will not analyze that marker. For multi-allelic loci, with coded alleles *1*, 2, *3*, etc, unless an allele *i* or *not i* (i.e., all other alleles) each have frequency of at least 10%, TDTae will not analyze that allele. This option, which requires a positive integer to follow it, allows the user to change the minimum frequency. For example, typing "-a 20" changes the minimum allele frequency requirement to 20%. Our experience with this software is that **the minimum allele frequency should be at least 10%. The maximization method does not perform well when the minor allele frequency is very small** (see also Section 4.2).

-g: This option groups together all alleles whose number of appearances is below the minimum count (default = 30; also see -a command above) into one allele.

-s: This option allows for calculation of support intervals (Edwards 1992) for each of the maximum likelihood parameters under H0 and H1. The default setting is 2; that is, the endpoints of the 2-unit support interval (i.e., 100:1 odds) of the MLEs of each parameter are provided. The default setting can be changed by using the "-b" option (see directly below).

-b: This option enables the user to specify the length of the support interval when calculating using the "-s" option above. This option must be followed by a positive integer. For example, typing "-b 3" produces a 3-unit support interval instead of the default 2-unit interval.

-n: When performing maximization under H0 or H1, a two-stage procedure is employed (see Section 4.1 below on maximization). Once the grid search (1^{st} stage) is finished, parameter values corresponding to the largest *n* log-likelihoods are used as starting points for the Powell maximization method (Acton 1970; Brent 1973; Jacobs 1977). This option allows the user to specify the number of largest *n* log-likelihoods that will be followed up (default is 5). When using this option, the user must specify a positive integer indicating the number of largest log-likelihoods that will be followed up.

-po: This option instructs the program that the format of the pedigree file is post-MAKEPED format. If this option is not used, then the program will assume that the format is pre-MAKEPED.

-o: This option enables the user to specify the name of the output file. It is followed by the user-specified name of the file. If this option is not used, the results will be written to the screen only.

-f: Invoking this option enables the user to specify marker names that will be used when reporting results. If no such file of marker names is provided, then the output file will label each of the markers "Locus #1, Locus #2, etc".

-t: With this option, the user can view what individuals were "trimmed" by the TDTae program to decrease the computational load. Also see the "-x" option (next).

-x: In its present formulation, TDTae's computational time to produce results increases with the number of individuals in a pedigree. This option enables the user to trim the number of founders from the pedigree so that the size of the pedigree that is analyzed is reduced. This option must be followed by a positive integer. The default maximum number of founders in a pedigree is 9.

-v: This option allows the user to view progress of the maximizations for each marker locus and allele. It also provides an estimation of the time till completion for each TDTae analysis with a given allele.

-e: With this option the user can test whether recoded genotypes on founders are in Hardy-Weinberg proportions. This option uses the method employed in the HWE program (see UTIL, Statistical Utility programs). This option requires that the user specify an output file for the results.

The next three options are related. The default analysis for TDTae involves maximization over the parameters R_1 and R_2 under the alternative, with no constraints placed on the relationship of these parameters. As such, the resulting test statistic has 2 degrees of freedom (df). The following options allow the user to perform a 1 df test, subject to constraints on the parameters R_1 and R_2. These tests might be used if the user has some prior knowledge of the mode of inheritance of the trait being studied and wishes to potentially increase power by reducing the degrees of freedom. The following constraints are invoked when using the three options:

-d: $R_1 = R_2$ (dominant mode of inheritance)
-r: $R_1 = 1$ (recessive mode of inheritance)
-m: $R_1^2 = R_2$ (multiplicative mode of inheritance); see graph further down.

It is interesting to note that using the "-m" option **(multiplicative mode) is equivalent to performing a TDT analysis with the original TDT statistic** (Weinberg 1999), but this does not mean that the TDTae with option –m is identical to the original TDT (see section 3.2).

-c: This option allows the user to specify the number of "cuts" *c* that are used in the first stage of the search procedure (see Section 4.1 below). This option must be followed by a positive integer greater than 1. The default number of cuts *c* used is 5.

2.2 Example files with this distribution

Example pedigree files, marker files, and output files are provided with the distribution of this software. The pedigree files are: pedsim-err.pre (simulated data), psor17.pre (real data from a study of psoriasis pedigrees on chromosome 17 (Helms et al. 2003)) and sito.pre (real data from a study of sitosterolemia pedigrees on chromosome 2 (Lee et al. 2001)). The corresponding marker files are: markers-pedsim.txt and markers-sito.txt (there is no marker file for the psoriasis data).

2.3 Error models

To run the TDTae program with your data, you will need (as mentioned above) a pedigree file in either pre- or post-MAKEPED format. You must also specify the particular error model that you will use when performing the TDTae analyses. The choices are:

GLHO	(Gordon Liu Heath Ott)	(Gordon et al. 2001)
DSB	(Douglas Skol Boehnke)	(Douglas et al. 2002)
SPL	(Sobel Papp Lange)	(Sobel et al. 2002)
MA	(Mote Anderson)	(Mote and Anderson 1965)

A brief description of each error model is provided here. Notationally, we assume that all markers have two (possibly down-coded) alleles labeled 1 and 2. The parameter list for each error model is provided below this description. Also see our website, PAWE at http://linkage.rockefeller.edu/pawe/. The GLHO model introduces errors into alleles as opposed to genotypes. It is described by 2 parameters. The DSB model introduces errors into genotypes, and is the only model for which it is not possible for a homozygous 11 genotype to be incorrectly recoded as a homozygous 22 genotype, or vice versa. It is described by 2 parameters. The SPL model is, for di-allelic loci, described by 3 parameters. It is the most general error model possible for di-allelic loci, under the constraint that errors are independent of the particular allele. The MA model, which is the most general error model possible in the sense that it can describe all other error models, is described by 6 parameters. The GLHO, SPL, and MA error models all allow for errors in which one homozygote is incorrectly miscoded as another homozygote.

Gordon Heath Liu Ott (GHLO) error model parameters
The parameter settings for this error model are:
E1 = Pr(*1* allele incorrectly coded as *2* allele)
E2 = Pr(*2* allele incorrectly coded as *1* allele)

Both entries must be positive real numbers less than 1.0.

Douglas Skol Boehnke (DSB) error model
The parameter settings for this error model are:
Gamma = Pr(homozygous *11* or *22* genotype incorrectly coded as heterozygote *12*)
Eta = Pr(heterozygote *12* genotype incorrectly coded as homozygote *11* or *22*)

Both entries must be positive real numbers less than 1.0.

Note: for the Eta parameter, it is assumed that the *12* genotype has an equal probability (0.5) of being incorrectly coded as *11* or *22*. Also, the notation used here comes from the Gordon et al. (2002) reference.

Sobel Papp Lange (SPL) error model
The parameter settings for this error model are:
V_1 = Pr(true homozygote incorrectly coded as heterozygote)
V_2 = Pr(one homozygote incorrectly coded as another homozygote)
V_3 = Pr(true heterozygote incorrectly coded as a homozygote)

Note: This parameterization of the SPL error model is an improvement over the parameterization previously used (Gordon et al. 2002) in that it only requires three parameter settings. The author gratefully acknowledges S. Seaman and P. Holmans for the improvement.

All entries must be positive real numbers less than 1.0, subject to the following constraints:
$V_1 + V_2 < 1.0$
$V_3 < 0.5$

Mote and Anderson (MA) error model
The parameter settings for this error model are:
e_{21} = Pr(*12* genotype observed | *11* true)
e_{31} = Pr(*22* genotype observed | *11* true)
e_{12} = Pr(*11* genotype observed | *12* true)
e_{32} = Pr(*22* genotype observed | *12* true)
e_{13} = Pr(*11* genotype observed | *22* true)
e_{23} = Pr(*12* genotype observed | *22* true)

The following constraints are needed for the MA error model:
$e_{21} + e_{31} < 1.0$
$e_{12} + e_{32} < 1.0$
$e_{13} + e_{23} < 1.0$

The MA error model is the most robust error model in that it completely characterizes all other error models given certain constraints. Therefore, it is the "best" error model to use. However, it comes with a computational price. It requires three more parameters to be maximized than the SPL model, and four more than the GLHO and DSB error models.

3.0 INTERPRETING RESULTS FROM TDTae OUTPUT

A critical ingredient in running the TDTae analysis is interpretation of the outcome. The program produces MLEs of parameter estimates, values for the TDTae statistic, and uncorrected and corrected (for multiple testing) *p*-values. Headings for each of the parameters are as follows:

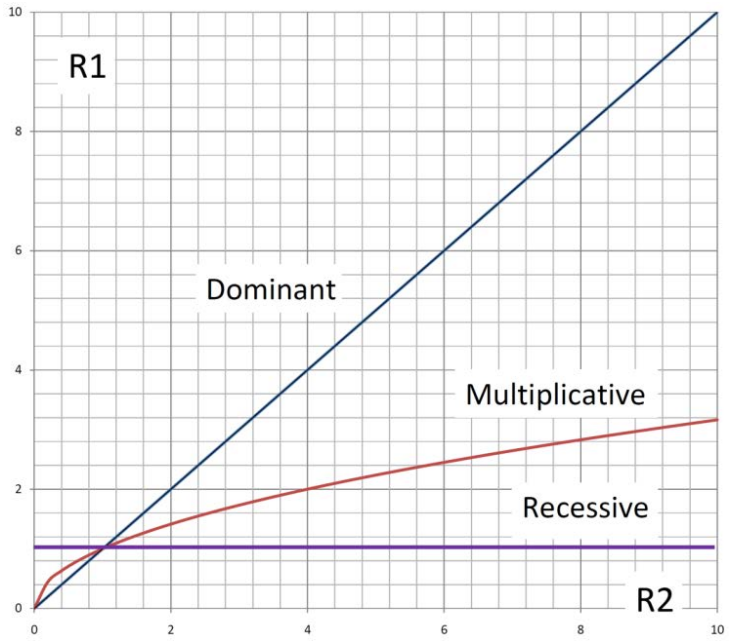
r1: MLE of the genotypic relative risk R_1 under alternative (H1) and null (H0) hypotheses.

r2: MLE of the genotypic relative risk R_2 under alternative (H1) and null (H0) hypotheses.

p11: MLE of the genotype frequency p_{11} under alternative (H1) and null (H0) hypotheses – note that the allele being tested is considered the "2" allele for estimation purposes.

p12: MLE of the genotype frequency p_{12} under alternative (H1) and null (H0) hypotheses – note that the allele being tested is considered the "2" allele for estimation purposes.

LogLike: Maximum log-likelihood estimates of the data under alternative (H1) and null (H0) hypotheses using two-stage search procedure (for purposes of programming, (-LogLike is minimized rather than LogLike being maximized).



LRT: The TDTae statistic - this quantity is given by the formula -2[LogLike(H1)-LogLike(H0)].

P: *p*-value (uncorrected for multiple testing) corresponding to the LRT statistic for allele being tested.

Corrected: P-value corrected for multiple testing. The correction is done as follows: if *k* alleles at marker locus are tested, and *p* is the uncorrected *p*-value corresponding to a particular allele, then the corrected *p*-value is given by $1 - (1 - p)^{k-1}$. See our paper (Gordon et al. 2004) for more details. Note that for SNPs, no correction for multiple testing is performed.

Also provided are MLEs for all error model parameters. See Section 2.3.1 above for the list of different error model parameters.

Maximizing the log likelihood in the space (plane) spanned by the R_1 and R_2 coordinates above could lead to estimates specifying $R_1 > R_2$, which would mean that the *+d* heterozygote has a larger relative risk than the *dd* homozygote. This does not generally make sense, so one may want to view with caution estimates outside the space in the above graph bounded by dominant and recessive inheritance. In practice, one may want to run TDTae three times, one time each for dominant, multiplicative, and recessive inheritance and retain the largest of the resulting three test statistics, each with 1 df. While the test statistic for (R_1, R_2) has 2 df, this maximized 1 df statistic has a number of df somewhere between 1 and 2. Assuming 2 df is conservative yet simple (the *p*-value associated with chi-square of 2 df is known analytically). If X^2 is the maximized test statistic, then the associated 2 df *p*-value is given by $p = \exp(-\frac{1}{2} X^2)$.

3.1 Example runs

We present here the results of some example runs. The first example uses the simulated data provided in this distribution (*pedsim-err.txt* and *markers-pedsim.txt*).

We comment that the results file below was created by typing:

>tdtae -f marker-pedsim.txt –n 20 –o pedsim-tdtae.out pedsim-err.pre GLHO

Note that we chose the error model of Gordon et al. (Gordon et al. 2001) for this analysis, because we simulated the data according to that error model.

Results from program TDTAE Version 2.01 using NR library
Written By Chad Haynes and Derek Gordon
Please email tdtae@linkage.rockefeller.edu with any bugs or problems

Locus #1 SNP1
Allele #1 (875 occurrences - 29.2%)
MLE r1 r2 p11 p12 E1 E2 LogLike LRT P Corrected
H1: 1.048007 1.508604 0.583394 0.378448 0.079448 0.000010 1287.559815 1.568215 0.456527 0.456527
H0: 1.000000 1.000000 0.575082 0.385035 0.075341 0.000010 1288.343922
Locus #2 SNP2
Allele #1 (770 occurrences - 25.7%)
MLE r1 r2 p11 p12 E1 E2 LogLike LRT P Corrected

H1: 0.860711 0.338686 0.599816 0.358233 0.055971 0.000010 1214.811515 5.065659 0.079434 0.079434
H0: 1.000000 1.000000 0.448872 0.466392 0.000010 0.180059 1217.344344
Locus #3 SNP3
Allele #1 (883 occurrences - 29.6%)
MLE r1 r2 p11 p12 E1 E2 LogLike LRT P Corrected
H1: 0.942563 1.324528 0.580013 0.345977 0.091583 0.084099 1330.769422 0.683097 0.710669 0.710669
H0: 1.000000 1.000000 0.608122 0.328371 0.101303 0.047451 1331.110970
Locus #4 SNP4
Allele #1 (662 occurrences - 22.4%)
MLE r1 r2 p11 p12 E1 E2 LogLike LRT P Corrected
H1: 0.775813 0.123752 0.559178 0.362844 0.032243 0.186968 1159.912489 15.579787 0.000414 0.000414
H0: 1.000000 1.000000 0.716827 0.251595 0.085297 0.014342 1167.702383

These data were simulated so that the first three markers (SNPs 1-3) are null and the last SNP is in both linkage and linkage disequilibrium with a trait locus. Also, approximately 25% of the parents in this file were not genotyped, and genotyping error was simulated according the GLHO model with each error parameter being set to 0.10. Because all data were simulated independently, a Bonferroni correction is appropriate. Thus, we see that, even in the presence of missing parental data and genotyping errors, the TDTae method provides accurate information in that it indicates that the trait locus is located near SNP 4. The TDTae statistic is not significant at the 5% level for any other marker after the Bonferroni correction.

We also note that, despite the relatively large sample size (500 trios), error parameter estimation is not consistent from marker to marker. Thus, error parameter estimation should be used with caution when considering data from trios.

Note that for two of the loci (#2 and #4), MLEs for genotypic relative risk values R_1 and R_2 for allele *1* are both less than 1. These values can be converted to genotypic relative risks for the "non-*1*" allele using the formulas $R'_1 = R_1/R_2$, $R'_2 = 1/R_2$, where the prime superscript indicates genotypic relative risk for the "non-*1*" allele.

In the next example, we present results for selected markers from the Sitosterolemia data (Lee et al. 2001) provided in this distribution. The pedigree file is *sito-ped.txt* and the marker file is *markers-sito.txt*. We choose the DSB model for our error model, although, as the output file indicates, there are no observed genotyping errors in this data set. Also, because we know that the disease is inherited in a recessive fashion, we chose the "-r" option when running our analyses (Section 2.1.1 – Command Line Options). The advantage of using this option is that there is only one degree of freedom for the corresponding TDTae (LRT) statistic. Also, we chose the "-a 20" option to allow testing for alleles whose minimal number of occurrences is 20.

We comment that the results file below was created by typing:

>tdtae –a 20 -f markers-sito.txt –n 20 –v –r -c 10 –o sito-tdtae.out sito-ped.txt DSB 13 15 20

Results from program TDTAE Version 2.01 using NR library
Written By Chad Haynes and Derek Gordon
Please email tdtae@linkage.rockefeller.edu with any bugs or problems

Locus #13 D2S4009
Allele #2 (39 occurrences - 23.5%)
MLE r1 r2 p11 p12 Gamma Eta LogLike LRT P Corrected
H1: 1.000000 8.987662 0.606503 0.305744 0.000000 0.000000 52.606985 6.274741 0.012267 0.012267
H0: 1.000000 1.000000 0.592418 0.318584 0.000000 0.000000 55.744355
Locus #15 D2S2298
Allele #2 (98 occurrences - 57.6%)
MLE r1 r2 p11 p12 Gamma Eta LogLike LRT P Corrected
H1: 1.000000 22.562454 0.245505 0.554387 0.000000 0.000000 59.997096 35.35158 0.000000 0.000000
H0: 1.000000 1.000000 0.228886 0.552130 0.000000 0.000000 77.672887
Locus #20 D2S2174
Allele #1 (37 occurrences - 22.3%)
MLE r1 r2 p11 p12 Gamma Eta LogLike LRT P Corrected
H1: 1.000000 1.479179 0.575724 0.327053 0.000000 0.000000 55.185774 0.158508 0.690551 0.904241
H0: 1.000000 1.000000 0.574890 0.326453 0.000000 0.000000 55.265027
Allele #2 (38 occurrences - 22.9%)
MLE r1 r2 p11 p12 Gamma Eta LogLike LRT P Corrected
H1: 1.000000 10000.00 0.691292 0.270247 0.000000 0.000000 41.537543 21.31348 0.000004 0.000008

H0: 1.000000 1.000000 0.682654 0.278131 0.000000 0.000000 52.194284
Allele #4 (45 occurrences - 27.1%)
MLE r1 r2 p11 p12 Gamma Eta LogLike LRT P Corrected H1: 1.000000 4.452777 0.532816 0.397743 0.000000 0.000000 57.419163 4.988331 0.025541 0.050429
H0: 1.000000 1.000000 0.527557 0.397515 0.000000 0.000000 59.913329

There are a few interesting things to note about this output. First, the TDTae statistic is performed for several alleles at each marker. As can be noted by studying the maximum LRT value for each marker and the corresponding minimal corrected *p*-value, the results are highly significant. We comment that genotype relative risk estimates for R_2 are large for each marker.

We comment that the location of the "Sitosterolemia" genes, ABCG5 and ABCG8, are approximately 20,000 base pairs from marker D2S2298 (Lee et al. 2001; Lu et al. 2001).

3.2 A note about robustness to population stratification

We comment that our likelihood method as presently designed may not be robust to population stratification, the original reason why the TDT and other statistics were developed. We quote from our European Journal of Human Genetics paper: "*We note that we assumed that the mating type frequencies of founders is given by the product of the individual genotype frequencies, unlike Weinberg (Weinberg 1999). We make this simplification to reduce the number of parameters that we must maximize in finding the maximum log-likelihood of the data. While it may be more powerful to use the six mating types, we comment that our simplification reduces the number of parameters to be estimated by three. However, our assumption does make* ***our statistic potentially non-robust to population stratification****, the original condition for which the TDT and other statistics were developed (Falk and Rubinstein 1987; Spielman et al. 1993). We plan to extend our method to handle the more general mating-type frequencies proposed in Weinberg's work (Weinberg 1999).*"

We therefore caution researchers regarding interpretation of results when using TDTae version 2.0 on data that is potentially stratified due to population admixture.

4.0 MAXIMIZATION PROCEDURES

4.1 Maximization method used

When applying our test statistic, we perform a two-stage maximization procedure. We first compute the log-likelihood under the null and alternative hypotheses using a lattice of points from a multi-dimensional rectangle. We "cut" the cube into a pre-specified number of intervals, and compute the log-likelihoods for the endpoints of each of the intervals. The number of cuts can be user-specified (see "-c" option – Section 2.1.1) and the default setting is five cuts. For example, if we consider the SPL error model, and specify four cuts, then the parameters p_{11}, p_{12}, and V_1 through V_3, all of which have values in the interval [0,1], will be tested at 4 + 1 = 5 values: 0.0, 0.25, 0.5, 0.75, and 1.0. For the relative risk parameters, R_1 and R_2, we initially consider the interval [0, 20]. Thus, in the first stage of our maximization, the log-likelihood is computed for $(c + 1)^{4 + e}$ values under the alternative hypothesis, and for $(c + 1)^{2 + e}$ values under the null, where c is the number of cuts specified, and e is the number of error model parameters for a given error model. The parameter e is equal to 2, 3, and 6 for the DSB, SPL, and MA error models, respectively.

Once the log-likelihoods are computed in the first stage, the parameter values that provide the top n log-likelihoods under each hypothesis are then used as starting values for the Powell maximization procedure (Acton 1970; Brent 1973; Jacobs 1977). The value n may be user specified (see "-n" option – Section 2.1.1). The default setting for this number is five. We use the Powell procedure as implemented in the "Numerical Recipes in C" text (Press et al. 2002). The largest log-likelihood from each set of n runs is then chosen as the maximum log-likelihood for each hypothesis.

4.2 Potential maximization issues

We have performed extensive analyses with this program. Occasionally, we have seen LRT values for certain alleles that are less than 0, indicating that the maximum log-likelihoods were not found under H1.Typically, this result is caused by using alleles whose minor allele frequency is small (less than 10%). One potential solution is to re-run the analysis using a sufficiently large minor allele frequency (i.e., choose the command line option "-a *x*") where *x* is an integer greater than or equal to 20. If the problem persists, please contact us at the e-mail listed below (Section 5.0).

4. 3 A note about computation time

We comment that the time necessary to complete maximization and therefore to determine the LRT value for a given marker allele is dependent upon the number of individuals in a pedigree and also the number of parameters being maximized (see also Section 4.1). As the number of individuals in a pedigree grows, the computation time for our method also increases. Therefore, our method at present is most efficient for small nuclear families in which there are no genotype errors. We are presently researching approximate likelihood solutions to reduce the computational time necessary to compute the LRT values for *any* pedigree.

5.0 PROBLEMS? COMMENTS?

If there are problems in the execution or compilation of this program or if you would like to provide some feedback, please e-mail Gordon@dls.rutgers.edu.

6.0 ACKNOWLEDGEMENTS

The authors of this software gratefully acknowledge grant K01-HG00055 from the National Institutes of Health. The psoriasis study for which example data are provided is funded in part by NIH grant AR049049.

7.0 REFERENCES

Below are references for this User's Guide. Please cite:

Gordon D, Haynes C, Johnnidis C, Patel S, Bowcock AM, Ott J (2004) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. European Journal of Human Genetics 12:752-61

Gordon D, Heath SC, Liu X, Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. American Journal of Human Genetics 69:371-380

when reporting results obtained by using the TDTae program.

Acton FS (1970) Numerical methods that work. Mathematical Association of America, Washington, DC

Brent RP (1973) Chapter 7. In: Algorithms for minimization without derivatives. Prentice-Hall, Englewood Cliffs, NJ

Douglas JA, Skol AD, Boehnke M (2002) Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. Am J Hum Genet 70:487-495

Edwards AWF (1992) Likelihood. The Johns Hopkins University Press, Baltimore

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Annals of Human Genetics 51:227-233

Gordon D, Haynes C, Johnnidis C, Patel S, Bowcock AM, Ott J (2004) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. European Journal of Human Genetics 12:752-61

Gordon D, Heath SC, Liu X, Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. American Journal of Human Genetics 69:371-380

Helms C, Cao L, Krueger JG, Wijsman EM, Chamian F, Gordon D, Heffernan M, Daw JA, Robarge J, Ott J, Kwok PY, Menter A, Bowcock AM (2003) A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis. Nat Genet 35:349-356

Jacobs DAH (1977) The state of the art in numerical analysis. Academic Press, London

Lee MH, Gordon D, Ott J, Lu K, Ose L, Miettinen T, Gylling H, Stalenhoef AF, Pandya A, Hidaka H, Brewer B, Jr., Kojima H, Sakuma N, Pegoraro R, Salen G, Patel SB (2001) Fine mapping of a gene responsible for regulating dietary cholesterol absorption; founder effects underlie cases of phytosterolaemia in multiple communities. European Journal of Human Genetics 9:375-384

Lu K, Lee MH, Hazard S, Brooks-Wilson A, Hidaka H, Kojima H, Ose L, Stalenhoef AF, Mietinnen T, Bjorkhem I, Bruckert E, Pandya A, Brewer HB, Jr., Salen G, Dean M, Srivastava A, Patel SB (2001) Two genes that map to the STSL locus cause sitosterolemia: genomic structure and spectrum of mutations involving sterolin-1 and sterolin-2, encoded by ABCG5 and ABCG8, respectively. American Journal of Human Genetics 69:278-290

Mote VL, Anderson RL (1965) An investigation of the effect of misclassification on the properties of chisquare-tests in the analysis of categorical data. Biometrika 52:95-109

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2002) Numerical Recipes in C. The art of scientific computing. Cambridge University Press, Cambridge

Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. American Journal of Human Genetics 70:496-508

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). American Journal of Human Genetics 52:506-516

Weinberg CR (1999) Allowing for missing parents in genetic studies of case-parent triads. American Journal of Human Genetics 64:1186-1193