RV-TDT: Rare Variant Extensions of the Transmission Disequilibrium Test

Zong-Xiao He & Suzanne M. Leal

March 16, 2015

1 Introduction

Many population-based rare-variant (RV) association tests, which aggregate variants across a region, have been developed to analyze sequence data. A drawback of analyzing population-based data is that it is difficult to adequately control for population substructure and admixture, and spurious associations can occur. For RVs, this problem can be substantial, because the spectrum of rare variation can differ greatly between populations. A solution is to analyze parent-child trio data, by using the transmission disequilibrium test (TDT), which is robust to population substructure and admixture.

We extended the TDT to test for RV associations using four commonly used methods. We demonstrate that for all RV-TDT methods, using proper analysis strategies, type I error is well-controlled even when there are high levels of population substructure or admixture. For trio data, unlike for population-based data, RV allele-counting association methods will lead to inflated type I errors. However type I errors can be properly controlled by obtaining p values empirically through haplotype permutation. The methods that RV-TDT provides include:

- RV-TDT-CMC-Analytical: extension of Combined Multivariate and Collapsing (CMC, Li and Leal 2008), analytical p value is evaluated;
- RV-TDT-CMC-Haplotype: extension of CMC, haplotype permutation is used to evaluate statistical significance;
- RV-TDT-BRV-Haplotype: extension of Burden of Rare Variants (BRV, Auer et al 2013), with haplotype permutation;
- RV-TDT-VT-BRV-Haplotype: extension of Variable Threshold (VT, Price et al 2010) with BRV coding and haplotype permutation;
- RV-TDT-VT-CMC-Haplotype: extension of VT with CMC coding and haplotype permutation;
- RV-TDT-WSS-Haplotype: extension of Weighted Sum Statistic (WSS, Madsen 2009).

For more information about RV-TDT, please refer to:

He, Z., O'Roak, B. J., Smith, J. D., Wang, G., Hooker, S., Santos-Cortez, R. L. P., ... & Leal, S. M. (2014). Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data.
The American Journal of Human Genetics, 94(1), 33-46.

2 Installation

RV-TDT can be installed and run on Linux and Mac OS. Download the source code from `http://bioinformatics.org/rv-tdt/` and run the following shell commands:

```
make rvTDT
```

An executable file *rvTDT* will be compiled in the current folder.

3 Input Format

The RV-TDT requires three input files: a tped file, a map file, and a phenotype information file.

3.1 tped file

The tped file provides the genotype information. Each line presents the genotypes for a variants.

```
1042809 0 1 1 0 0 1 0 1 0 0 0 0 1 0 0 1 1 0 1 1 1 0 1 1 0 1 ...
```

The first column is SNP/variant id, and followed by the genotype on every individual. Since RV-TDT only takes phased data, every two columns present the haplotypes of one individual, that is, every six columns present one trio.

3.2 map file

This file provides the gene-variant map information. The first two columns are the gene and variant id. The variant id must matches with the variant id in tped file.

```
ZAN 100331837 0.013496
ZAN 100334516 0.033025
ZAN 100346012 0.000452
```

The third column is used for filtering, and this information can be used to determine whether the variant should be included in the analysis. For example, in above example, this column contain population minor allele frequencies (MAF) from public available database, and later we can use *-u 0.05* option to exclude variants with MAF > 0.05.

3.3 phenotype file

The phenotype contains six columns in total: sample ID, family ID, father ID, mother ID, sex (1 for male, 0 for female), case(1) / control(0).

```
5252    11006   0    0    0    0
5284    11006   0    0    1    0
5316    11006   5284    5252    0    1
5253    11009   0    0    0    0
5285    11009   0    0    1    0
5317    11009   5285    5253    0    1
```

The order of the sample ID must match the order of individuals in tped file.

4 Output Format

The output includes two folder

- ${proj}_pval: contains a .pval file for each gene in your dataset, which lists the p-values for the all RV-TDT tests.

```
#gene   CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo
ABCA7   0.006688    0.166417    0.006997    0.024988    0.000500    0.028986
```

- ${proj}_rvTDT: contains a .rvTDT file for each gene in your dataset, which lists the detailed information about the gene, such as transmission counts for each variants, WSS weights etc. This file is in the json format.

```
{
        "basicInformation" :
        {
                "variants" : []
        },
        "tdtStatic" :
        {
                "Transmitted" : [],
                "denovoCount" : [],
                "effectiveTrioNum" : [],
                "singleSitePval" : [],
                "siteBeAnalyzed" : [],
                "unTransmitted" : [],
                "wssWeight" : []
        },
        "tdtTest" :
        {
                "pvalues" : {},
                "tests" : []
        },
        "variantStatic" :
        {
                "missingRatio" : [],
                "populationMafs" : [],
                "sampleMafs" : [],
                "varFoundInKid" : [],
                "varFoundInParent" : []
        }
}
```

5 Example

In this example, the genotype and phenotype data for 20 genes and 2,000 trios are randomly generated. The tped and map files are located in *rvtdt_exercise_data* folder, and you can find the phenotype file *rvtdt_exercise.phen* in current folder. Use the following script to run the RV-TDT analysis for all these genes (see the script in *rv_tdt_cmds.sh*, and run it by typing *sh rv_tdt_cmds.sh* in terminal).

```
for g in `ls rvtdt_exercise_data | grep tped | cut -d"." -f1`
do
echo "runing rvTDT on gene "${g}
./rvTDT exercise_proj -G ./rvtdt_exercise_data/${g}.tped -P ./rvtdt_exercise.phen \
-M ./rvtdt_exercise_data/${g}.map \
--adapt 500 --alpha 0.00001 --permut 2000 \
--lower_cutoff 0 --upper_cutoff 100 \
--minVariants 3 \
--maxMissRatio 1
done
```

For demonstration purposes most arguments are written out including some that use default values:

- *-G*: tped file location;
- *-P*: phenotype file location;
- *-M*: map file location;
- *–adapt*: To reduce computational time, adaptive permutation is used in *rvTDT*. Every *$adapt* permutations (default: 500 permutations), the program will check if we should keep doing permutation (which means this gene looks promising to reach the desired α level), or we should give up on this gene (which means this gene will not reach the desired α level based on the permutations we have done so far, or we have done enough permutations);
- *–alpha*: The α level in adaptive permutation;
- *–permut*: The maximum number of permutations;
- *–lower_cutoff* and *–upper_cutoff*: The cutoffs to determine which variants we should include in the analysis. In this example, the third column of map file is the number of minor allele counts, and here we only include the variants who have minor allele counts less than 100;
- *–minVariants*: The minimum number of variant sites for a gene. Genes with variant site number less than *$minVariants* will be excluded from analysis (after check missing);
- *–maxMissRation*: The max missing ratio allowed for a variant. The variants with missing ratio greater than *$maxMissRatio* will be excluded from analysis. In this example, we generated the genetic data file without any missing genotypes, so *–maxMissRation 1* is used here.

It takes about 2 minutes to finish above commands. The pvalues for these genes are as follows:

```
#gene   CMC-Analytical  BRV-Haplo   CMC-Haplo   VT-BRV-Haplo    VT-CMC-Haplo    WSS-Haplo
gene20  0.041632    0.053892    0.061876    0.071856    0.055888    0.053892
gene9   0.051235    0.065868    0.063872    0.061876    0.077844    0.091816
gene17  0.176580    0.187625    0.185629    0.381238    0.381238    0.255489
gene5   0.369441    0.349301    0.343313    0.489022    0.475050    0.377246
gene16  0.500000    0.504990    0.528942    0.508982    0.536926    0.530938
gene14  0.841345    0.756487    0.744511    0.750499    0.770459    0.778443
gene1   0.948765    0.948104    0.944112    0.968064    0.968064    0.966068
gene13  0.958368    0.944112    0.960080    0.948104    0.948104    0.946108
gene19  0.987326    0.988024    0.990020    0.952096    0.956088    0.968064
```

Please note the number of pvalue files in *exercise_proj_pval* is less than 20, because some genes have variant sites less than 3 and are not included in the analysis.

6 rvTDT Command Options

```
Usage: rvTdt projName
(-h to list usage, --help to list all arguments)

All arguments:
 projName (currently set to: rareTdtTest)
        Project name. (STRING)
        * set output folder name

-G/--genoFile <string> (currently set to: NoInputFile)
        Genotype file, with folder path.
        | Format (no header): SNP_ID 0 1 1 0 ....
        * missing site marked as -9

-M/--mapFile <string> (currently set to: NoInputFile)
        mapping file, with folder path.
        | Format (no header): Gene SNP_ID Annotation

-P/--phenoFile <string> (currently set to: NoInputFile)
        Phenotype file, with folder path.
        | Format (no header): Ind_ID Fam_ID Father_ID Mother_ID Sex Status ...
        * Fahter_ID or Mother_ID = 0, if not available
        * SEX = (male)?1:0
        * Status = (affected)?1:0

-a/--adapt <int> (currently set to: 500)
        Number of permutations for adaptive check
        * only applicable to permutation based methods

-e/--minVariants <int> (currently set to: 4)
        The minimum variant sites for a gene to be analyzed
        * genes with variant site No. < $minVariants will be excluede from analysis
          (after check missing).

-l/--lower_cutoff <float> (currently set to: 0)
        The lower bound of variants to be included in analysis
```

```
        * Will compared with the third column of map file, only loci having annotated
          value > $lower_cutoff will be analyzed

-m/--maxMissRatio <float> (currently set to: 1)
        The max missing ratio allowed on each site
        * A variant site will be excluded from analysis, if its missing ratio > $maxMissRatio

-n/--nopermut (currently set to: 0)
        Only run CMC-analytical test

-p/--permut <int> (currently set to: 200000)
        Number of permutations
        * only applicable to permutation based methods

-s/--alpha <float> (currently set to: 1e-05)
        Alpha level for adaptive check.

-u/--upper_cutoff <float> (currently set to: 0.05)
        The upper bound of variants to be included in analysis
        * Will compared with the third column of map file, only loci having annotated
          value <= $upper_cutoff will be analyzed
```