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

Introduction

Many population-based rare-variant 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 rare variants, 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) (Spielman et al., 1993), which is robust to population substructure and admixture.

The TDT was extended to analyze rare variants (RV)-TDT using severally commonly used aggregate rare variant association tests which were originally developed to analyzed populations based data. They included the Collapsed Multivariate Collapsing Method (CMC) (Li and Leal, 2008); the Burden of Rare Variants (BRV) (Auer et al., 2013); Weighted Sum Statistic (WSS) (Madsen and Browning, 2009), and Variable Threshold (VT) (Price et al., 2010). Before analysis of the trio data the data must be phased which can be performed using several phasing programs including BEAGLE (Browning and Browning, 2007), Shape-It (Delaneau et al., 2012) and PHASE (Stephens and Scheet, 2005). When the RV-TDT is used to analyze trios empirical p-values must be obtained using haplotype permutation to avoid inflation of type I error. The only exception is that analytical p-values can be obtained for RV-TDT-CMC and type I errors is well controlled.

For more information about RV-TDT, please refer to He et al. 2014.

Variant annotation

In this exercise, we will perform variant annotation and variant selection in Variant Association Tools (VAT) (Wang et al., 2014). First we import the genotype and phenotype data into vtools project.

```
vtools init rvtdt
vtools import --format vcf data/data.vcf --build hg19
vtools phenotype --from_file data/phen.txt
```

The quality control has been performed on this given data, including selecting autosomal SNPs, removing low quality variant call and remove variant with minor allele frequency (MAF) > 1%. In rare variant association, we usually only analyze the functional variants within a gene. Here we use ANNOVAR to annotate the variants and select variants that are nonsynonymous, splicing site, frameshift, or stop gain/loss.

```
vtools execute ANNOVAR  geneanno
vtools select variant "variant.region_type like '%splicing%'or variant.mut_type like
```

```
'nonsynonymous%' or variant.mut_type like 'frameshift%' or variant.mut_type like 'stop%'" -t
func_variant
```

We need to export the following files from the VAT. Three files are exported from VAT (backups of these files can be found in data/ folder):

- vat_raw.tped: This file follows standard plink tped file format, which contains genotype information. One row is one variant/SNP. The first 4 columns are: chromosome, variant identifier, genetic distance (not used in this exercise), and base-pair position. Then all genotypes are listed for all individuals for each particular variant on each line (A/T/G/C coding and 0=missing).

```
vtools export func_variant --format tped --samples 'phenotype is not null'  > vat_raw.tped
# set marker name as chr_pos, needs to avoid duplicate name
sort -k4 -n vat_raw.tped | awk 'BEGIN{OFS="\t";prev="None";copy=1} {$2=$1"_"$4; $3=0;
if($2==prev) {$2=$2"_"copy; copy=copy+1} else {prev=$2; copy=1}; print $0}' > vat_export.tped
```

- vat_raw.tfam: This file follows standard plink tfam file format, which contains individual and family information. One row is an individual. The first six columns are: Family ID, Individual ID, Paternal ID (set to 0 for founder), Maternal ID (set to 0 for founder), Sex (1=male; 2=female; other=unknown), and Phenotype (1=unaffected, 2=affected, 0=unknown).

```
vtools phenotype --out family sample_name pid mid sex phenotype > vat_export.tfam
```

- vat_raw.anno: This file contains variant information. More specially, we need to know which gene each variant belongs since RV-TDT is a gene-based association test. We also need minor allele frequency (MAF) for each variant, which will be used in Variable Threshold method. The MAF of each variant can be annotated by public database (e.g ESP, ExAC), or calculated using the founders' genotype. In this exercise, all of the variants have passed the MAF filter and here we set the MAF of all variants to be 0.001 since this is simulated data, but usually the allele frequencies from a database such as ExAC would be used.

```
vtools use refGene-hg19_20130904
vtools update func_variant --set 'maf=0.001'  # set the maf to be 0.001
vtools select func_variant -o chr pos refGene.name2 maf --header > vat_export.anno
```

Phasing Trio

In this exercise, we will be using BEAGLE (v3.3.2) to phase to trio data. Before we prepare inputs for BEAGLE, Mendelian inconsistencies were identified and removed with the PLINK software (v1.07). The PLINK will also help us to recode the genotype as 0/1/2 from A/T/G/C.

```
plink --noweb --tfile vat_export --recode12 --me 1 1 --set-me-missing --out "recode12_noME"
```

Now we have genotype data that doesn't have any Mendelian errors and the genotypes are coded as 0/1/2. The outputs of previous command include recode12_noME.ped and recode12_noME.map. We need convert these files into BEAGLE input format. One way is to use the java program linkage2beagle.jar (https://faculty.washington.edu/browning/beagle_utilities/utilities.html - linkage2beagle), which is provided as a utility script in BEAGLE. The following commands will convert PLINK format files into BEAGLE format files.

```
sort -n -k1 -k6 -k2 recode12_noME.ped | sed 's/ /\t/g' | cut -f1,3,4,5 --complement > linkage.ped
cut -f2 recode12_noME.map | awk 'BEGIN{OFS="\t";} {print "M",$0}' | sed '1i\I\tid\nA\tDisease' >
linkage.dat
java -Xmx10000m -jar java/linkage2beagle.jar linkage.dat linkage.ped > pre_beagle.bgl
```

For missing genotype data, BEAGLE imputes missing data and only provides the most likely genotype. Analyzing the most likely genotype will increase false-positive rates for trio data. In this exercise, we replace the missing genotype as wildtype to bypass the problem. If any member within a trio has missing genotype, we mark the genotypes of all members within the trio to be wildtype.

```
python script/pre_phase.py -i pre_beagle.bgl -a pre_beagle_withMissing.bgl
```

Now we call BEAGLE to phase the trios.

```
java -Xmx10000m -jar java/beagle.jar missing=0 trios=pre_beagle.bgl out=bgl_phased
verbose=false redundant=true
```

The output file is a zipped file (bgl_phased.pre_beagle.bgl.phased.gz) and the following command will unzip it.

```
gunzip bgl_phased.pre_beagle.bgl.phased.gz
```

RV-TDT Analysis

RV-TDT is a gene-based rare variant association test. For each gene, we need to select its genotype data (in file *.bgl.phased) and corresponding variant annotations (vat_export.anno). The following python script will put each gene's genotype and annotation information into single files (under the folder genes/).

```
python script/post_phase.py -a vat_export.anno -b bgl_phased.pre_beagle.bgl.phased -o genes/
```

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

- **tped file**: This file provides the genotype information. Each line presents the genotypes for a variant. The first column is variant id, and followed by the genotype on every individual. Every two columns present the genotypes of one individual abd every six columns present one trio (genotype of father/mother/child).
- **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. The third column is the MAF of the variant.
- **phenotype file**: This file contains six columns: sample ID, family ID, father ID, mother ID, sex (1=male; 0=female), disease status (1=affected; 0=unaffected).

The tped and map files for each gene are in genes/ folder, and the phenotype file can be found in data/ folder. Now we are ready to run RV-TDT analysis for each gene. To save time, we will run the test over 20 genes in this exercise.

```
for g in `ls genes | grep tped | cut -d"." -f1 | head -20`
do
      echo "runing rvTDT on gene "${g}
      rvTDT exercise_proj -G ./genes/${g}.tped -P ./data/rvtdt.phen -M ./genes/${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. The descriptions of each argument are:

- -*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.

You can use command ‘./rvTDT –help’ to see the all options of rvTDT. The output includes two folder

- exercise_proj_pval: This folder contains .pval files, which lists the p-values of the all RV-TDT tests for each gene analyzed. For example,

```
#gene   CMC-Analytical   BRV-Haplo    CMC-Haplo    VT-BRV-Haplo   VT-CMC-Haplo   WSS-Haplo
BTG3    0.028890         0.037924     0.031936     0.045908       0.039920       0.041916
```

- exercise_proj_rvTDT: This folder contains .rvTDT files, which lists the detailed information about for each gene analyzed, such as transmission counts for each variants, WSS weights etc. This file is in the json format.

References

Auer, P.L., Wang, G., and Leal, S.M. (2013). Testing for rare variant associations in the presence of missing data. Genet Epidemiol *37*, 529–538.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet *81*, 1084–1097.

Delaneau, O., Marchini, J., and Zagury, J.F. (2012). A linear complexity phasing method for thousands of genomes. Nat Methods *9*, 179–181.

He, Z., O'Roak, B.J., Smith, J.D., Wang, G., Hooker, S., Santos-Cortez, R.L.P., Li, B., Kan, M., Krumm, N., Nickerson, D.A., et al. (2014). Rare-Variant Extensions of the Transmission Disequilibrium Test: Application to Autism Exome Sequence Data. Am. J. Hum. Genet. *94*, 33–46.

Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet *83*, 311–321.

Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet *5*, e1000384.

Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet *86*, 832–838.

Spielman, R.S., McGinnis, R.E., and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet *52*, 506–516.

Stephens, M., and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet *76*, 449–462.

Wang, G.T., Peng, B., and Leal, S.M. (2014). Variant association tools for quality control and analysis of large-scale sequence and genotyping array data. Am. J. Hum. Genet. *94*, 770–783.

Questions

1. How many trios are imported?
2. How many variants are imported from vcf file?
    a. How many of them are functional variants?
3. How many Mendel errors detected in total by PLINK?
    a. How many markers remaining?
4. How many genes are there after we split the genotype data by gene?
5. List the gene with smallest p-value on CMC-Analytical method.
    a. What is the p-value?
6. If the alpha level is 0.05, how many genes are significant with BRV-Haplo method?

Answers

1. 1000 (vtools show tables)
2. 8905 (vtools show tables)
    a. 3410 (vtools show tables)
3. 130 (output of PLINK)
    a. 3410 (output of PLINK)
4. 172 (ls genes/ | grep tped | wc)
5. *BTG3*, 0.02889 (cat exercise_proj_pval/*.pval | grep -v "^#" | sort –k2)
6. 1: *BTG3* (cat exercise_proj_pval/*.pval | grep -v "^#" | sort –k3) [Note: the results may be different, since it's a permutation based method]