MultiTDS 1.0 (12/08/2013) Manual

Ran Tao (*taor@live.unc.edu*)

Contents

1. GENERAL INFORMATION

MultiTDS is a command-line software program written in C++ to implement the methods described in Tao et al. (2014) for the analysis of sequence data under multivariate trait-dependent sampling. The sampling can depend on multiple quantitative traits in any manner. Quantitative traits are related to genetic variants and covariates through a multivariate linear regression model, while the distributions of genetic variants and covariates are unspecified. Both the maximum likelihood estimation (MLE) and standard least-squares (LS) methods are available. The MLE method properly accounts for multivariate trait-dependent sampling, whereas the LS method does not. Throughout this manual, we assume that there are K traits to be analyzed.

2. OPTIONS

- **-PHENO_SEQ** {*phenofile.seq*}

  Specifies the file that contains the traits and covariates (if any) for the sequenced individuals. The default value is `pheno_seq.tab`.

- **-PHENO_NONSEQ** {*phenofile.nonseq*}

  Specifies the file that contains the traits for the non-sequenced individuals. The default value is `pheno_nonseq.tab`.

- **-GENO** {*genofile*}

  Specifies the genotype file. The default value is `geno.tab`.

- **-MAP** {*mapfile*}

  Specifies the gene-SNP mapping file used in gene-level rare-variant analysis (i.e., when **-RARE_ALL** or **-RARE_TRAIT** is specified). The default value is `map.tab`. This file is not needed in single-variant analysis (i.e., when **-COM** or **-COM_SCORE** is specified).

- **-OUT_MLE** {*outfile.mle*}

  Specifies the output file for the MLE method. The default value is `mle.tab`.

- **-OUT_NAIVE** {*outfile.naive*}

  Specifies the output file for the LS method. The default value is `naïve.tab`.

- **-OUT_WARN** {*outfile.warn*}

  Specifies the output file for the warning texts. The default value is `warning.txt`.

- **-COM**

  Conducts single-variant analysis using Wald statistics. Only one of the four options **-COM**, **-COM_SCORE**, **-RARE_ALL**, or **-RARE_TRAIT** can be specified.

- **-COM_SCORE**

  Conducts single-variant analysis of the k_0th trait using score statistics. Only one of the four options **-COM**, **-COM_SCORE**, **-RARE_ALL**, or **-RARE_TRAIT** can be specified.

- **-RARE_ALL**

  Conducts gene-level rare-variant analysis to test the global null hypothesis that there is no genetic effect on any trait. Only one of the four options **-COM**, **-COM_SCORE**, **-RARE_ALL**, or **-RARE_TRAIT** can be specified.

- **-RARE_TRAIT**

  Conducts gene-level rare-variant analysis of the k_0th trait. Only one of the four options **-COM**, **-COM_SCORE**, **-RARE_ALL**, or **-RARE_TRAIT** can be specified.

- **-MODEL** {*model*}

  Specifies one of the two genetic models (i.e., `Additive` or `Dominant`) in gene-level rare-variant analysis (i.e., when **-RARE_ALL** or **-RARE_TRAIT** is specified). The default value is `Additive`.

- **-TRAIT** $\{k_0\}$

  Specifies the trait to be tested when **-COM_SCORE** or **-RARE_TRAIT** is specified. k_0 should be a number in $\{1, \ldots, K\}$. The default value is `1`.

- **-T1** {*outfile.T1*}

  Conducts the T1 test in gene-level rare-variant analysis (i.e., when **-RARE_ALL** or **-RARE_TRAIT** is specified). Specifies the output file for the T1 test.

- **-T5** {*outfile.T5*}

  Conducts the T5 test in gene-level rare-variant analysis (i.e., when **-RARE_ALL** or **-RARE_TRAIT** is specified). Specifies the output file for the T5 test.

- **-MB** {*outfile.MB*}

  Conducts the MB test in gene-level rare-variant analysis (i.e., when **-RARE_ALL** or **-RARE_TRAIT** is specified). Specifies the output file for the MB test.

- **-SKAT** {*outfile.SKAT*}

  Conducts the SKAT test in gene-level rare-variant analysis (i.e., when **-RARE_ALL** or **-RARE_TRAIT** is specified). Specifies the output file for the SKAT test. In gene-level rare-variant analysis of one trait (i.e., when **-RARE_TRAIT** is specified), both the MLE and LS methods have output files. The output file for the MLE method is *outfile.SKAT*. The output file for the LS method is `naive_`*outfile.SKAT*.

- **-MAF_LB** {*MAF_LB*}

  Specifies the minor allele frequency (MAF) lower bound in single-variant analysis (i.e., when **-COM** or **-COM_SCORE** is specified). SNPs with observed MAFs less than *MAF_LB* will not be analyzed. The default value is `0`.

- **-MAC_LB** {*MAC_LB*}

  Specifies the minor allele count (MAC) lower bound in single-variant analysis (i.e., when **-COM** or **-COM_SCORE** is specified). SNPs with observed MACs less than *MAC_LB* will not be analyzed. The default value is `0`.

- **-MAF_UB** {*MAF_UB*}

  Specifies the MAF upper bound in MB and SKAT tests. SNPs with external MAFs in *mapfile* greater than *MAF_UB* will be skipped. The default value is `0.05`.

- **-SNP_MAC_LB** {*SNP_MAC_LB*}

  Specifies the MAC lower bound in gene-level rare-variant analysis (i.e., when **-RARE_ALL** or **-RARE_TRAIT** is specified). SNPs with observed MACs less than *SNP_MAC_LB* will be skipped. The default value is `0`.

3. INPUT FILES

The input files consist of a phenotype file for the sequenced individuals, a phenotype file for the non-sequenced individuals, a genotype file for the sequenced individuals, and a gene-SNP mapping file. In all input files, the field separator character is the tab-delimiter, and missing values are denoted by `-999`.

- Phenotype file for sequenced individuals (*phenofile.seq*)

  The first row contains the headers. The remaining rows represent sequenced individuals. The first K columns pertain to the K traits, and the remaining columns pertain to the covariates (if any). Each individual must have at least one non-missing trait. Covariates are not permitted to have missing values. All covariates must be numeric.

- Phenotype file for non-sequenced individuals (*phenofile.nonseq*)

  The rows represent non-sequenced individuals. A header row is not permitted. There are K columns pertaining to the K traits. The order of the columns is the same as that of the first K columns in *phenofile.seq*. Each individual must have at least one non-missing trait.

- Genotype file (*genofile*)

  The rows represent SNPs. The first column is the SNP ID. The remaining columns are the genotypes for individuals, which take the values 0, 1, or 2. The order of the columns in the genotype file is the same as the order of the rows in *phenofile.seq*. That is to say, the ith column in the genotype file corresponds to the same individual as the ith row in *phenofile.seq*, where $i \geq 2$.

- Mapping file (*mapfile*)

  The rows represent gene-SNP pairs. A header row is not permitted. There are exactly three columns. The first column is the gene ID. The second column is the SNP ID. The third column is the external MAF to be used in gene-level rare-variant analysis.

4. SINGLE-VARIANT ANALYSIS USING WALD STATISTICS

4.1 Synopsis

MultiTDS [**-PHENO_SEQ** *phenofile.seq*] [**-PHENO_NONSEQ** *phenofile.nonseq*] [**-GENO** *genofile*] [**-OUT_MLE** *outfile.mle*] [**-OUT_NAIVE** *outfile.naive*] [**-OUT_WARN** *outfile.warn*] [**-COM**] [**-MAF_LB** *MAF_LB*] [**-MAC_LB** *MAC_LB*]

4.2 Output Files

- Output file for the MLE method (*outfile.mle*)

  The first row contains the headers. The remaining rows represent SNPs. There are $3K + 7$ columns. The first column is the SNP ID. The $(3k - 1)$th, $(3k)$th, and $(3k + 1)$th columns are the estimate, standard error, and p-value, respectively, for the genetic effect on the kth trait, $k = 1, \ldots, K$. The remaining columns are:

  - MAC: MAC in *genofile*.
  - MAF: MAF in *genofile*.
  - N0/N1/N2: number of individuals carrying 0/1/2 minor alleles in *genofile*.
  - NMISS: number of individuals with missing genotypes in *genofile*.

Missing values are denoted by `NA`.

- Output file for the LS method (*outfile.naive*)

  The format of this file is exactly the same as that of *outfile.mle*.

- Output file for the warning texts (*outfile.warn*)

  The rows represent the SNPs. Each row starts with the SNP ID followed by the warning text (if any). Possible warnings include:

  1. *MAF lower than MAF_LB!*
  2. *MAC lower than MAC_LB!*
  3. *No genetic variation in this sample!*
  4. *Covariate design matrix is singular!*

     This warning indicates that the SNP will be skipped in the analysis because the augmented design matrix corresponding to the SNP and the covariates is singular.
  5. *EM algorithm does not converge!*

     The maximum number of iterations is 500.
  6. *The information matrix is not positive definite!*

5. SINGLE-VARIANT ANALYSIS OF ONE TRAIT USING SCORE STATISTICS

5.1 Synopsis

MultiTDS [**-PHENO_SEQ** *phenofile.seq*] [**-PHENO_NONSEQ** *phenofile.nonseq*] [**-GENO** *genofile*] [**-OUT_MLE** *outfile.mle*] [**-OUT_NAIVE** *outfile.naive*] [**-OUT_WARN** *outfile.warn*] [**-COM_SCORE**] [**-TRAIT** k_0] [**-MAF_LB** *MAF_LB*] [**-MAC_LB** *MAC_LB*]

5.2 Output Files

- Output file for the MLE method (*outfile.mle*)

  The first row contains the headers. The remaining rows represent SNPs. There are 10 columns. The first column is the SNP ID. The second, third, and fourth columns are the score statistic, variance of the score statistic, and p-value, respectively, for the genetic effect on the k_0th trait. The remaining columns are:

  - MAC: MAC in *genofile*.
  - MAF: MAF in *genofile*.

  - N0/N1/N2: number of individuals carrying 0/1/2 minor alleles in *genofile*.
  - NMISS: number of individuals with missing genotypes in *genofile*.

  Missing values are denoted by `NA`.

- Output file for the LS method (*outfile.naive*)

  The format of this file is exactly the same as that of *outfile.mle*.

- Output file for the warning texts (*outfile.warn*)

  The rows represent the SNPs. Each row starts with the SNP ID followed by the warning text (if any). Possible warnings include:

  1. *MAF lower than MAF_LB!*
  2. *MAC lower than MAC_LB!*
  3. *No genetic variation in this sample!*
  4. *Covariate design matrix is singular!*

     This warning indicates that the SNP will be skipped in the analysis because the augmented design matrix corresponding to the SNP and the covariates is singular.
  5. *EM algorithm does not converge!*

     The maximum number of iterations is 500.
  6. *The information matrix is not positive definite!*
  7. *V is not positive!*

     V is the variance of the score statistic.

6. GENE-LEVEL RARE-VARIANT ANALYSIS TO TEST THE GLOBAL NULL HYPOTHESIS

6.1 Synopsis

MultiTDS [**-PHENO_SEQ** *phenofile.seq*] [**-PHENO_NONSEQ** *phenofile.nonseq*] [**-GENO** *genofile*] [**-MAP** *mapfile*] [**-OUT_MLE** *outfile.mle*] [**-OUT_WARN** *outfile.warn*] [**-RARE_ALL**] [**-MODEL** *Additive*] [**-T1** *outfile.T1*] [**-T5** *outfile.T5*] [**-MB** *outfile.MB*] [**-SKAT** *outfile.SKAT*] [**-MAF_UB** *MAF_UB*] [**-SNP_MAC_LB** *SNP_MAC_LB*]

6.2 Output Files

- Output file for the MLE method (*outfile.mle*)

  The first row contains the headers. The remaining rows represent genes. The columns are:

- GENE: gene ID.
- T1_MAC: sum of MACs across SNPs with MAFs less than 0.01 and MACs greater than *SNP_MAC_LB*. T1_MAC is present if **-T1** is specified.
- T1_T: test statistic for the T1 test. T1_T is present if **-T1** is specified.
- T1_PVAL: p-value for the T1 test. T1_PVAL is present if **-T1** is specified.
- T5_MAC: sum of MACs across SNPs with MAFs less than 0.05 and MACs greater than *SNP_MAC_LB*. T5_MAC is present if **-T5** is specified.
- T5_T: test statistic for the T5 test. T5_T is present if **-T5** is specified.
- T5_PVAL: p-value for the T5 test. T5_PVAL is present if **-T5** is specified.
- MB_MAC: sum of MACs across SNPs with MAFs less than *MAF_UB* and MACs greater than *SNP_MAC_LB*. MB_MAC is present if **-MB** is specified.
- MB_T: test statistic for the MB test. MB_T is present if **-MB** is specified.
- MB_PVAL: p-value for the MB test. MB_PVAL is present if **-MB** is specified.
- SKAT_MAC: sum of MACs across SNPs with MAFs less than *MAF_UB* and MACs greater than *SNP_MAC_LB*. SKAT_MAC is present if **-SKAT** is specified.
- SKAT_T: test statistic for the SKAT test. SKAT_T is present if **-SKAT** is specified.
- SKAT_PVAL: p-value for the SKAT test. SKAT_PVAL is present if **-SKAT** is specified.

Missing values are denoted by `NA`.

- Output file for the T1 test (*outfile.T1*)

  Each gene has K rows of output, i.e. one row for each trait. The first column is the gene ID. The second column is the trait name. The third column is the MAC, which is the same as the column “T1_MAC” in *outfile.mle*. The fourth column is the score statistic. The remaining columns constitute a lower diagonal covariance matrix of the score statistics among different traits within the same gene.

- Output file for the T5 test (*outfile.T5*)

  The format of this file is exactly the same as that of the output file for *outfile.T1*.

- Output file for the MB test (*outfile.MB*)

  The format of this file is exactly the same as that of the output file for *outfile.T1*.

- Output file for the SKAT test (*outfile.SKAT*)

  The first column is the gene ID. The second column is the trait name. The third column is the SNP ID. The fourth column is the MAF, which is the same as the third column in *mapfile*. The fifth column is the MAC. The sixth column is the score statistic. The remaining columns constitute a lower diagonal covariance matrix of the score statistics among different trait by SNP combinations within the same gene.

- Output file for the warning texts (*outfile.warn*)

  The rows represent the genes. Each row starts with the gene ID followed by the warning text (if any). Possible warnings include:

  1. *No eligible SNP in this gene!*

     This warning indicates that all SNPs in this gene have MACs less than *SNP_MAC_LB*.
  2. *(T1) EM algorithm for the null model does not converge!*
  3. *(T1) V is not positive definite!*
  4. *(T5) EM algorithm for the null model does not converge!*
  5. *(T5) V is not positive definite!*
  6. *(MB) EM algorithm for the null model does not converge!*
  7. *(MB) V is not positive definite!*
  8. *(SKAT) EM algorithm for the null model does not converge!*
  9. *(SKAT) The asymptotic p-value may not be accurate!*

  Note that the maximum number of iterations for EM algorithms is 500. V is the variance of the score statistic.

7. GENE-LEVEL RARE-VARIANT ANALYSIS OF ONE TRAIT

7.1 Synopsis

MultiTDS [**-PHENO_SEQ** *phenofile.seq*] [**-PHENO_NONSEQ** *phenofile.nonseq*] [**-GENO** *genofile*] [**-MAP** *mapfile*] [**-OUT_MLE** *outfile.mle*] [**-OUT_NAIVE** *outfile.naive*] [**-OUT_WARN** *outfile.warn*] [**-RARE_TRAIT**] [**-TRAIT** k_0] [**-MODEL** *Additive*] [**-T1** *outfile.T1*] [**-T5** *outfile.T5*] [**-MB** *outfile.MB*] [**-SKAT** *outfile.SKAT*] [**-MAF_UB** *MAF_UB*] [**-SNP_MAC_LB** *SNP_MAC_LB*]

7.2 Output Files

- Output file for the MLE method (*outfile.mle*)

  The first row contains the headers. The remaining rows represent genes. The columns are:

- GENE: gene ID.
- T1_MAC: sum of MACs across SNPs with MAFs less than 0.01 and MACs greater than *SNP_MAC_LB*. T1_MAC is present if **-T1** is specified.
- T1_T: test statistic for the T1 test. T1_T is present if **-T1** is specified.
- T1_PVAL: *p*-value for the T1 test. T1_PVAL is present if **-T1** is specified.
- T5_MAC: sum of MACs across SNPs with MAFs less than 0.05 and MACs greater than *SNP_MAC_LB*. T5_MAC is present if **-T5** is specified.
- T5_T: test statistic for the T5 test. T5_T is present if **-T5** is specified.
- T5_PVAL: *p*-value for the T5 test. T5_PVAL is present if **-T5** is specified.
- MB_MAC: sum of MACs across SNPs with MAFs less than *MAF_UB* and MACs greater than *SNP_MAC_LB*. MB_MAC is present if **-MB** is specified.
- MB_T: test statistic for the MB test. MB_T is present if **-MB** is specified.
- MB_PVAL: *p*-value for the MB test. MB_PVAL is present if **-MB** is specified.
- SKAT_MAC: sum of MACs across SNPs with MAFs less than *MAF_UB* and MACs greater than *SNP_MAC_LB*. SKAT_MAC is present if **-SKAT** is specified.
- SKAT_T: test statistic for the SKAT test. SKAT_T is present if **-SKAT** is specified.
- SKAT_PVAL: *p*-value for the SKAT test. SKAT_PVAL is present if **-SKAT** is specified.

Missing values are denoted by `NA`.

- Output file for the LS method (*outfile.naive*)

  The first row contains the headers. The remaining rows represent genes. The columns are:

  - GENE: gene ID.
  - T1_MAC: sum of MACs across SNPs with MAFs less than 0.01 and MACs greater than *SNP_MAC_LB*. T1_MAC is present if **-T1** is specified.
  - T1_U: score statistic for the T1 test. T1_U is present if **-T1** is specified.
  - T1_V: variance of the score statistic for the T1 test. T1_V is present if **-T1** is specified.
  - T1_PVAL: *p*-value for the T1 test. T1_PVAL is present if **-T1** is specified.
  - T5_MAC: sum of MACs across SNPs with MAFs less than 0.05 and MACs greater than *SNP_MAC_LB*. T5_MAC is present if **-T5** is specified.
  - T5_U: score statistic for the T5 test. T5_U is present if **-T5** is specified.

  - T5_V: variance of the score statistic for the T5 test. T5_V is present if **-T5** is specified.
  - T5_PVAL: p-value for the T5 test. T5_PVAL is present if **-T5** is specified.
  - MB_MAC: sum of MACs across SNPs with MAFs less than *MAF_UB* and MACs greater than *SNP_MAC_LB*. MB_MAC is present if **-MB** is specified.
  - MB_U: score statistic for the MB test. MB_U is present if **-MB** is specified.
  - MB_V: variance of the score statistic for the MB test. MB_V is present if **-MB** is specified.
  - MB_PVAL: p-value for the MB test. MB_PVAL is present if **-MB** is specified.
  - SKAT_MAC: sum of MACs across SNPs with MAFs less than *MAF_UB* and MACs greater than *SNP_MAC_LB*. SKAT_MAC is present if **-SKAT** is specified.
  - SKAT_T: test statistic for the SKAT test. SKAT_T is present if **-SKAT** is specified.
  - SKAT_PVAL: p-value for the SKAT test. SKAT_PVAL is present if **-SKAT** is specified.

  Missing values are denoted by `NA`.

- Output file for the T1 test (*outfile.T1*)

  The rows represent genes. The first column is the gene ID. The second column is the MAC, which is the same as the column "T1_MAC" in *outfile.mle*. The third column is the score statistic. The fourth column is the variance of the score statistic. Missing values are denoted by `NA`.

- Output file for the T5 test (*outfile.T5*)

  The format of this file is exactly the same as that of *outfile.T1*.

- Output file for the MB test (*outfile.MB*)

  The format of this file is exactly the same as that of *outfile.T1*.

- Output file for the SKAT test using the MLE method (*outfile.SKAT*)

  The first column is the gene ID. The second column is the variant ID. The third column is the MAF, which is the same as the third column in *mapfile*. The fourth column is the MAC. The fifth column is the score statistic. The remaining columns constitute a lower diagonal covariance matrix of the score statistics among different SNPs within the same gene.

- Output file for the SKAT test using the LS method (`naive_`*outfile.SKAT*)

  The format of this file is exactly the same as that of *outfile.SKAT*.

- Output file for the warning texts

  The rows represent the genes. Each row starts with the gene ID and followed by the warning text (if any). Possible warnings include:

  1. *No eligible SNP in this gene!*

     This warning indicates that all SNPs in this gene have MACs less than *SNP_MAC_LB*.
  2. *(T1) The design matrix is singular!*
  3. *(T1) EM algorithm does not converge!*
  4. *(T1) V is not positive!*
  5. *(T5) The design matrix is singular!*
  6. *(T5) EM algorithm does not converge!*
  7. *(T5) V is not positive!*
  8. *(MB) The design matrix is singular!*
  9. *(MB) EM algorithm does not converge!*
  10. *(MB) V is not positive!*
  11. *(SKAT) The design matrix is singular!*
  12. *(SKAT) EM algorithm does not converge!*
  13. *(SKAT) The asymptotic p-value of the MLE method may not be accurate!*
  14. *(SKAT) The asymptotic p-value of the LS method may not be accurate!*

  Note that the maximum number of iterations for EM algorithms is 500. V is the variance of the score statistic.