PreMeta

GENERAL INFORMATION

PreMeta is an R program that reformats summary statistics for four meta-analysis pipelines (MASS, RAREMETAL, MetaSKAT, and seqMeta). PreMeta requires the output files from the study-level analysis of any of these pipelines. In particular, MASS uses one text file to contain all the summary statistics. RAREMETAL uses two text files: one contains score statistics and SNP-level information, and the other contains between-SNP covariance by sliding windows. MetaSKAT uses .MSSD and .MInfo files: the .MSSD is a binary file with between-SNP information matrices, and the .MInfo is a text file with information on studies and SNP sets. seqMeta uses an R object to contain all the summary statistics. In addition, PreMeta normalizes the score statistics from RAREMETALWORKER by the estimated residual variance. Thus, PreMeta allows the summary statistics from any operators to be used in any other packages to perform meta-analysis. When PreMeta reformats gene-based summary statistics from other operators for the RAREMETAL pipeline, the between-SNP covariance is set to 0 if the two SNPs do not belong to the same gene. This workaround will produce the correct covariance information in meta-analysis if the same gene annotation is used in generating gene-based summary statistics and in performing meta-analysis (since the covariances between different genes are not used anyway).

SYNOPSIS

PreMeta(scriptFile=*script.txt*, software=*meta_software*, version=*version_num*)

The option **scriptFile** is to specify a script file including a list of the output files from the study-level analysis, as well as the name and the version number of the pipeline within which the study-level analysis were performed. The option **software** is to specify the meta-analysis software and the option **version** is to specify version of the software. We currently support version 5.1 for MASS, version 0.4.0 for RAREMETAL, version 0.40 for MetaSKAT, and version 1.5 for seqMeta.

INPUT FILES

The following is an example of the PreMeta script file. The line starting with # is treated as a comment and is ignored. The keyword **SOFTWARE** indicates which software was used to generate the file(s) for the study and the keyword **VERSION** indicates the version of the software. The name of the file is specified by the file keywords **FILE_*****. *Note that for each study the keywords* ***SOFTWARE*** *and* ***VERSION*** *should appear prior to the file keywords.*

If "**SOFTWARE** = MASS", one file keyword **FILE** should follow. For the detailed description of the file, refer to the documentation of SCORE-Seq (http://dlin.web.unc.edu/software/score-seq/) or SCORE-SeqTDS (http://dlin.web.unc.edu/software/score-seqtds/).

If "**SOFTWARE** = RAREMETAL", at least two file keywords **FILE_SCORE** and **FILE_COV** should follow. For the detailed description of the two files, refer to the documentation of RAREMETALWORKER (http://genome.sph.umich.edu/wiki/RAREMETALWORKER). If the study-level analysis were performed within the **RAREMETAL** pipeline but the meta-analysis will not be performed by **RAREMETAL**, PreMeta need to convert the sliding-window summary statistics to the gene-based summary statistics. To this end, a group file need to be provided to specify the grouping of the SNPs. The format of the group file is described in the documentation for **RAREMETAL** (http://genome.sph.umich.edu/wiki/RAREMETAL_Documentation).

If "**SOFTWARE** = MetaSKAT", two file keywords **FILE_MSSD** and **FILE_MInfo** should follow.

If "**SOFTWARE** = seqMeta", one file keyword **FILE_RDATA** should follow.

One need to make sure that the format of the SNP IDs (rs#, or chr:pos) are consistent across studies. The SNP ID in the RAREMETAL pipeline takes the form "chr:pos". So, if any one of those files is generated from the RAREMETAL pipeline, then the SNP ID format should be "chr:pos" across all studies.

In this example, we have files from four studies. The one text file for the first study is generated by the MASS pipeline. The two text files for the second study are generated by the RAREMETAL pipeline, the .MSSD and .MInfo files for the third study is generated by the MetaSKAT pipeline, and the .Rdata file for the fourth study is generated by the seqMeta pipeline.

```
## === THE FIRST STUDY: SUMMARY STAT === ##
SOFTWARE = MASS
VERSION = 5.1
FILE = path/study1.txt

## === THE SECOND STUDY: SUMMARY STAT === ##
SOFTWARE = RAREMETAL
VERSION = 0.4.0
FILE_SCORE = path/study2_score.txt
FILE_COV = path/study2_cov.txt
FILE_GROUP = path/group.txt

## === THE THIRD STUDY: SUMMARY STAT === ##
SOFTWARE = MetaSKAT
VERSION = 0.40
FILE_MSSD = path/study3.MSSD
FILE_MInfo = path/study3.MInfo

## === THE FOURTH STUDY: SUMMARY STAT === ##
SOFTWARE = seqMeta
VERSION = 1.5
FILE_RDATA = path/study4.Rdata
```

OUTPUT FILES FOR MASS PIPELINE

For each study, PreMeta generates one text file that can be loaded by MASS for meta-analysis. PreMeta also prepares the MASS script file that will be used as MASS input. For detailed description of the MASS script file, please refer to MASS documentation (http://dlin.web.unc.edu/software/mass/). The following is an example of the MASS script file.

```
## MASS script file
## === THE FIRST STUDY: INPUT FILE AND COLUMN SPECIFICATION === ##
FILE = path/MASS_STUDY1.txt
GENE_ID_COLUMN = 1
GVAR_ID_COLUMN = 2
MAC_COLUMN = 3
N_OBS_COLUMN = 4
SCORE_COLUMN = 5

## === THE SECOND STUDY: INPUT FILE AND COLUMN SPECIFICATION === ##
FILE = path/MASS_STUDY2.txt
GENE_ID_COLUMN = 1
GVAR_ID_COLUMN = 2
MAC_COLUMN = 3
```

```
N_OBS_COLUMN = 4
SCORE_COLUMN = 5

## === THE THIRD STUDY: INPUT FILE AND COLUMN SPECIFICATION === ##
FILE = path/MASS_STUDY3.txt
GENE_ID_COLUMN = 1
GVAR_ID_COLUMN = 2
MAC_COLUMN = 3
N_OBS_COLUMN = 4
SCORE_COLUMN = 5

## === THE FOURTH STUDY: INPUT FILE AND COLUMN SPECIFICATION === ##
FILE = path/MASS_STUDY4.txt
GENE_ID_COLUMN = 1
GVAR_ID_COLUMN = 2
MAC_COLUMN = 3
N_OBS_COLUMN = 4
SCORE_COLUMN = 5
```

OUTPUT FILES FOR RAREMETAL PIPELINE

For each study, PreMeta generates two text files that can be loaded by RAREMETAL for meta-analysis. PreMeta also prepares two lists that summarize the two sets of the text files across studies. The two list can be directly used by RAREMETAL. For detailed description of the lists, please refer to RAREMETAL documentation (http://genome.sph.umich.edu/wiki/RAREMETAL). The following is an example.

```
## SCORE FILES
## === THE FIRST STUDY === ##
STUDY1_score.txt
## === THE SECOND STUDY === ##
STUDY2_score.txt
## === THE THIRD STUDY === ##
STUDY3_score.txt
## === THE FOURTH STUDY === ##
STUDY4_score.txt

## COV FILES
## === THE FIRST STUDY === ##
STUDY1_cov.txt
## === THE SECOND STUDY === ##
STUDY2_cov.txt
## === THE THIRD STUDY === ##
STUDY3_cov.txt
## === THE FOURTH STUDY === ##
STUDY4_cov.txt
```

OUTPUT FILES FOR MetaSKAT PIPELINE

For each study, PreMeta generates .MSSD and .MInfo files that can be loaded by MetaSKAT for meta-analysis. For detailed description of the files, please refer to MetaSKAT manual (http://cran.r-project.org/web/packages/MetaSKAT/index.html).

OUTPUT FILES FOR seqMeta PIPELINE

For each study, PreMeta generates a .Rdata file that can be loaded by seqMeta for meta-analysis. For detailed description of the files, please refer to seqMeta manual (http://cran.r-project.org/web/packages/seqMeta/index.html).