**Welcome to the runepacts wiki!**

RunEpacts is designed to prepare and/or submit datasets to EPACTS for rare variant association, and to present EPACTS output from said tests in a format that's easy to interpret visually (output described in a later section.)

1. RunEpacts can be used to submit the same set of input files to EPACTS for multiple runs without separate formatting for each run (for example, the same set of input files can be submitted with specifications for different masks/different sets of covariates).
2. RunEpacts generates a PDF file containing QQ plots and Manhattan plots for the group-wise tests, for the group-wise tests with filtered groups (using only those genes that fit criteria of minimum numbers of variants/minor allele counts), and single-marker results for those significant genes (those genes that meet the p-value cutoff--default is 0.05)

RunEpacts consists of 2 scripts: a Python script and an R script (which is called internally by the python script.)

**Dependencies:**

1. tabix
2. python pandas package
3. epacts

**RunEpacts Output:**

**Some features of the RunEpacts output:**

1. A QQ-plot produced using only genes that pass the MINVARS and GENEMINMAC thresholds, along with the original EPACTS qq-plot.

****

1. A graphical look at carriers of variants belonging to the significant genes, with single-marker P-values for these variants**.** In the figure below, the left panel below the histogram shows where carriers for variants of each significant gene (listed on the right) fall on the trait distribution (adjusted trait distribution, if covariates are used in the analysis.) Each line represents 1 variant. Variants are clustered together by their genes (genes are separated by black lines, and rows separating variants within the same gene are separated by dotted lines). Red dots represent homozygous carriers, and blue dots represent hets. The genes are sorted by their P-values. Green vertical lines in each row of the left bottom panel represent the mean phenotype (or adjusted phenotype) of carriers of that variant.

****

**Running Runepacts:**

*Input files:*

1. Compulsory files:

-Config file: Specifies parameters and directories

-VCF file: Can be annotated/unannotated

-PED file: Sample information files

2. Optional files:

-Annotation file: Allows users to create non-standard groups of variants

-Groupfile: Pre-created group file in the EPACTS format, if this exists

-Gene list file: File that restricts analysis to a list of genes/groups

*File formats:*

Config file: There are two columns per line, the first being the parameter name and the second the parameter value. When all parameters have been input, the line 'PROCESS' (with only one column, and no parameter value in the 2nd column) submits a job to the script, which in turn process the parameters and input files into an epacts run.

Every instance of 'PROCESS' must have a unique instance of 'OUTPREFIX'

The following are the parameters that can be used in the config file:

Files:

|  |  |
| --- | --- |
| VCFFILE | VCF file name. If present in the INPUTDIR, just the vcf file name. If not, then directory/vcfname. If option SEPCHR is on, this refers to the chr1 VCF [COMPULSORY] |
| PEDFILE | Tab-delimited file containing sample information. Header compulsory. The first five columns have to be Fam\_ID,IND\_ID,FAT\_ID,MOT\_ID,SEX respectively (they can be named differently but must contain this information). [COMPULSORY] |
| ANNOTFILE | Tab-delimited variant-wise annotation file. Gzipped. [OPTIONAL] |
| ANNOTVARCOL | The column in ANNOTFILE containing the variant name (The first column is 1, and not 0.). Default = 1[OPTIONAL] |
| ANNOTGENECOL | The column in ANNOTFILE containing the gene name (The first column is 1, and not 0.) Default = 5[OPTIONAL] |
| ANNOTPOSCOL | The column in ANNOTFILE containing the variant position (The first column is 1, and not 0.) Position is formatted as chr:position (example: 10:11001248). Default = 2[OPTIONAL] |
| MAFFILE | File containing MAF information. Providing this is not recommended, since setting a sample filter will change the MAF values for each variant for the sample subset being analysed. If not provided, MAF calculations and any MAF-based filtering are performed by the script. [OPTIONAL] |
| KINSHIPFILE | Sample Kinship file used for EMMAX. If not provided, the script will use EPACTS to generate a kinship file. [OPTIONAL] |
| GROUPFILE | Groupfile in EPACTS group format [OPTIONAL] |

Directories:

|  |  |
| --- | --- |
| INPUTDIR | Directory where input files are found. Can be local path from the folder containing the python script, or global path [COMPULSORY] |
| EPACTSDIR | Directory from which epacts should be run. If not set, the script assumes epacts is in the path. [OPTIONAL] |
| OUTPREFIX | Directory + prefix where results and output files are written. Example: outputdir/chr1. The directory (outputdir, in the example) doesn’t have to be created in advance [COMPULSORY] |

Specifying the test options:

|  |  |
| --- | --- |
| MODEL | The phenotype with (if any) covariates to be used. The format is: Phenotype ~ Covariate1+ Covariate2. The ped file should have a column for the phenotype and each of the covariates. [COMPULSORY] |
| TEST | Which test to run. Example: group=skat. For single-marker tests: it would be single=q.linear. All tests allowed by EPACTS can be specified here. [COMPULSORY] |
| SINGLEMARKERTEST | Which single-marker test should be run on the variants of the significant genes. Default q.linear. Can be set to ‘FALSE’.[OPTIONAL] |
| EPACTSCMD | If any additional parameters have to be fed into epacts. They have to be specified exactly as would be added to the epacts command. Example: skat-o |

Filtering parameters:

|  |  |
| --- | --- |
| FILTERPED | An expression that allows you to filter samples using one of the columns in the PEDFILE [OPTIONAL] |
| FILTERMAF | This can either be used as a MAF filter (example 0.05), or an expression that allows filtering of the MAF file (example: 'COLUMN1 < 0.001')[OPTIONAL] |
| FILTERANNOT | [OPTIONAL] |
| MINVARS | This can either be used as a MAF filter (example 0.05), or an expression that allows filtering of the MAF file (example: 'COLUMN1 < 0.001')[OPTIONAL] |
| GENEMINMAC | Specify the minimum number of minor alleles a gene must have [OPTIONAL] |
| PVALUETHRESHOLD | What p-value cut-off from the group-wise test should be used to classify a gene as significant? Default = 0.05 [OPTIONAL] |
|  |  |
|  |  |

To run:

PROCESS

This single-column ‘PROCESS’ line submits preceding parameters to the script for processing and preparing for an EPACTS run.

Example config file

EPACTSDIR /net/fantasia/home/hmkang/bin/epactsTest/bin

INPUTDIR /net/snowwhite/home/sramdas/final.testing/fullvcf

VCFFILE wave2.clean.vcf.gz

PEDFILE Wave3.ped

ANNOTFILE w2.var.merged.tsv.gz

PEDCOLUMNS FAM\_ID,IND\_ID,FAT\_ID,MOT\_ID,SEX,LDL,C1,C2,C3,C4

OUTPREFIX outputfullvcf.onetestonly.july23/W3.skat.pcs.sex

MODEL LDL ~ SEX + C1 +C2 + C3 + C4

PVALUETHRESHOLD 0.01

FILTERANNOT MOST\_DEL\_SCORE < 3

SINGLEMARKERTEST q.linear

ANNOTCOLUMNS VEP\_TAG

GROUPFILE W3.skat.pcs.sex\_groups.txt

MINVARS 2

GENEMINMAC 5

TEST group=skat

PROCESS

To run RunEpacts:

> python testrunepacts.py --config configfile.txt

Example Runs:

1. To run a single test with a single set of parameters

CONFIG FILE:

EPACTSDIR /net/fantasia/home/hmkang/bin/epactsTest/bin

INPUTDIR /net/snowwhite/home/sramdas/final.testing/fullvcf

VCFFILE wave2.clean.vcf.gz

PEDFILE Wave3.ped

ANNOTFILE w2.var.merged.tsv.gz

PEDCOLUMNS FAM\_ID,IND\_ID,FAT\_ID,MOT\_ID,SEX,LDL,C1,C2,C3,C4

OUTPREFIX outputfullvcf.onetestonly.july23/W3.skat.pcs.sex

MODEL LDL ~ SEX + C1 +C2 + C3 + C4

PVALUETHRESHOLD 0.01

FILTERANNOT MOST\_DEL\_SCORE < 3

SINGLEMARKERTEST q.linear

ANNOTCOLUMNS VEP\_TAG

GROUPFILE W3.skat.pcs.sex\_groups.txt

MINVARS 2

GENEMINMAC 5

TEST group=skat

PROCESS

2. To run multiple tests with the same set of parameters

CONFIG FILE:

EPACTSDIR /net/fantasia/home/hmkang/bin/epactsTest/bin

INPUTDIR /net/snowwhite/home/sramdas/final.testing/fullvcf

VCFFILE wave2.clean.chr1.vcf.gz

SEPCHR ON

PEDFILE Wave3.ped

ANNOTFILE w2.var.merged.tsv.gz

OUTPREFIX outputfullvcf/W3.4pcs

MODEL LDL ~ SEX + C1 +C2 + C3 + C4

PVALUETHRESHOLD 0.01

FILTERANNOT MOST\_DEL\_SCORE < 3

ANNOTCOLUMNS VEP\_TAG

GROUPFILE W3.skat.pcs.sex\_groups.txt

MINVARS 2

GENEMINMAC 5

TEST group=skat,VT

PROCESS

3. To run multiple tests with the same set of input files, but different parameters

EPACTSDIR /net/fantasia/home/hmkang/bin/epactsTest/bin

INPUTDIR /net/snowwhite/home/sramdas/final.testing/fullvcf

VCFFILE wave2.clean.vcf.gz

PEDFILE Wave3.ped

ANNOTFILE w2.var.merged.tsv.gz

OUTPREFIX outputfullvcf/W3.skat

MODEL LDL ~ SEX + C1 +C2 + C3 + C4

PVALUETHRESHOLD 0.01

FILTERANNOT MOST\_DEL\_SCORE < 3

SINGLEMARKERTEST q.linear

ANNOTCOLUMNS VEP\_TAG

GROUPFILE W3.skat.pcs.sex\_groups.txt

MINVARS 2

GENEMINMAC 5

TEST group=skat

PROCESS

FILTERPED COLUMN1 != ‘NA’

TEST group=VT

OUTPREFIX outputfullvcf/W3.VT

PROCESS

4. To get the visualization for two genes only

CONFIG FILE:

EPACTSDIR /net/fantasia/home/hmkang/bin/epactsTest/bin

INPUTDIR /net/snowwhite/home/sramdas/final.testing/fullvcf

VCFFILE wave2.clean.vcf.gz

PEDFILE Wave3.ped

ANNOTFILE w2.var.merged.tsv.gz

OUTPREFIX outputfullvcf/W3.4pcs

MODEL LDL ~ SEX + C1 +C2 + C3 + C4

GENELIST HES2,PSCSK9

GROUPFILE W3.skat.pcs.sex\_groups.txt

PVALUETHRESHOLD 0.01

FILTERANNOT MOST\_DEL\_SCORE < 3

ANNOTCOLUMNS VEP\_TAG

MINVARS 2

GENEMINMAC 5

TEST group=skat,VT

PROCESS

\*When GENELIST is specified, SEPCHR cannot be set to ON, and a groupfile must be specified.