KGGSEE V1.1 User Manual

Miaoxin Li, Lin Jiang, Xiangyi Li, Lin Miao

Apr 15, 2022

CONTENTS:

CHAPTER

ONE

SETUP

1.1 System requirements

Operating system	KGGSEE runs in a Java Virtual Machine. It does not matter which operating system it runs in.
Java Runtime Environment	A Java SE Runtime Environment of version 1.8 or higher is needed.
CPU	A CPU with four cores or more is recommended.
Memory	16 GB RAM or higher is recommended.
Free space	KGGSEE and related datasets may take up to 10 GB.

1.2 Setup the Java Runtime Environment (JRE)

KGGSEE needs JRE 1.8 or higher to run. Both Java(TM) SE JRE and OpenJDK JRE are competent for KGGSEE. Please follow the instructions on the websites to complete the installation and also add Java to the system PATH.

Check the JRE by entering `java -version` in a Terminal of Linux or MacOS, or CMD or PowerShell of Windows. If it displays the JRE version like `Java(TM) SE Runtime Environment (build 1.8.0_xxx)` or `OpenJDK Runtime Environment (build 1.8.0_xxx)`, it means the JRE has already been set up. Otherwise, check if JRE has been installed and if Java is in the system PATH.

1.3 Setup KGGSEE

Download the bundled file of `kggsee.jar`, running resource dataset and quick tutorial dataset from the download page and unzip.

The running resource dataset includes:

`resources/hg19/kggseqv1.1_hg19_GEncode.txt.gz`	hg19 GENCODE annotation
`resources/hg19/kggseqv1.1_hg19_refGene.txt.gz`	hg19 RefGene annotation
`resources/hg38/kggseqv1.1_hg38_GEncode.txt.gz`	hg38 GENCODE annotation
`resources/hg38/kggseqv1.1_hg38_refGene.txt.gz`	hg38 RefGene annotation
`resources/HgncGene.txt.gz`	HGNC gene ID
`resources/ENSTGene.gz`	Ensembl gene ID and transcript ID
`resources/gtex.v8.gene.mean.tsv.gz`	The gene-level expression profile of the GTEx v8 tissues
`resources/gtex.v8.transcript.mean.tsv.gz`	The transcript-level expression profile of the GTEx v8 tissues
`resources/HCL_scRNA_cluster_mean.tsv.gz`	The expression profile of cell clusters generated from the scRNA-seq dataset of the Human Cell Landscape
`resources/*.symbols.gmt.gz`	MSigDB gene sets

The tutorial dataset includes:

`tutorials/scz_gwas_eur_chr1.tsv.gz`	Summary statistics of chr1 SNPs for a GWAS of schizophrenia on the European population
`tutorials/1kg_hg19_eur_chr1.vcf.gz`	Genotypes of chr1 SNPs sampled from the 1000 Genome Project European population
`tutorials/GTEx_v8_gene_BrainBA9.eqtl.txt.gz`	Summary statistics of eQTLs calculated from gene-level expression profile of GTEx v8 brain BA9
`tutorials/GTEx_v8_transcript_BrainBA9.eqtl.txt.gz`	Summary statistics of eQTLs calculated from transcript-level expression profile of GTEx v8 brain BA9

For running customized analyses, the following data is needed, refer to *Detailed Document* for descriptions of file formats.

- A file of GWAS summary statistics of the phenotype to be studied.
- VCF files of genotypes sampled from the population of the GWAS to be studied. Genotypes of the 1000 Genomes Project Phase3 v5 can be downloaded from the NCBI FTP site or the 1000 Genomes Project FTP site.
- A file of eQTL summary statistics calculated from target tissues may be used. We provide gene-based and transcript-based eQTL summary statistics for GTEx v8 tissues available for download (refer to the download page).

CHAPTER

TWO

QUICK TUTORIALS

We provide four quick tutorials; each shows one function of KGGSEE. In each tutorial, we provide the command line and a brief explanation of flags and output files. Please refer to *Detailed Document* and *Options* for details. The first tutorial (*Gene-based association tests*) should be done first, then there is no need to do the following in order.

Make sure the KGGSEE Java archive `kggsee.jar`, the running resource data folder `resources/`, and the tutorial data folder `tutorials/` are under the same folder. For convenience, enter the `tutorials/` directory first.

2.1 Gene-based association tests

GATES and ECS are two statistical methods combining the p-values of a group of SNPs into one p-value. This analysis inputs p-values of SNPs and outputs p-values of genes. The command is:

```
java -Xmx4g -jar ../kggsee.jar \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --keep-ref \
  --gene-assoc \
  --out t1
```

Explanation of the flags and input files:

Flag	Description
`--sum-file`	Specifies a whitespace delimitated file of GWAS summary statistics. In this analysis, columns of SNP coordinates and p-values (CHR, BP, and P by default) are needed.
`--vcf-ref`	Specifies a VCF file of genotypes sampled from a reference population. These genotypes are used to estimate LD correlation coefficients among SNPs.
`--keep-ref`	Keep the parsed VCF file (KGGSEE object format) in a folder named `VCFRefhg19` under the output folder. KGGSEE will read these files in the following tutorials, which will be faster than parsing VCF files.
`--gene-assoc`	Triggers gene-based association tests.
`--out`	Specifies the prefix of output files.

Explanation of the output files:

The numeric results of gene-based association tests are saved in `t1.gene.pvalue.txt`. There are seven columns in the file:

Header	Description
Gene	Gene symbol
#Var	Number of variants within the gene
ECSP	p-value of ECS
GATESP	p-value of GATES
Chrom	Chromosome of the gene
Pos	Coordinate of the variant with the lowest p-value within the gene
GWAS_Var_P	p-value of the variant

The columns of `t1.gene.var.pvalue.txt.gz` are the same as `t1.gene.pvalue.txt`. The difference is that, for each gene, in `t1.gene.pvalue.txt`, only the variant with the lowest p-value is output, while in `t1.gene.var.pvalue.txt.gz`, all variants are output.

The Q-Q plots for p-values of inputted GWAS file (inside or outside of gene) and gene-based association tests by GATES or ECS are saved in `t1.qq.png`.

2.2 DESE

DESE performs phenotype-tissue association tests and conditional gene-based association tests at the same time. This analysis inputs p-values of a GWAS and expression profile of multiple tissues; outputs p-values of phenotype-tissue associations and conditional p-values of genes. The command is:

```
java -Xmx4g -jar ../kggsee.jar \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --saved-ref VCFRefhg19 \
  --expression-file ./GTEx_v8_TMM.gene.meanSE.txt.gz \
  --gene-finemapping \
  --out t2
```

Explanation of the flags and input files:

Flag	Description
`--sum-file`	Specifies a whitespace delimitated file of GWAS summary statistics. In this analysis, columns of SNP coordinates and p-values are needed.
`--saved-ref`	Specifies the folder of genotypes of reference population in KGGSEE object format, which is saved by the `--keep-ref` flag in the first tutorial.
`--expression-file`	Specifies a gene expression file that contains means and standard errors of gene expressions in multiple tissues/cell types. Here `gtex.v8.gene.mean.tsv.gz` is for gene-level DESE. Try `gtex.v8.transcript.mean.tsv.gz` for transcript-level DESE; try `HCL_scRNA_cluster_mean.tsv.gz` for cell-cluster based DESE.
`--gene-finemapping`	Triggers the DESE analysis.
`--out`	Specifies the prefix of output files.

Explanation of the output files:

The three files of `t2.gene.pvalue.txt`, `t2.gene.var.pvalue.txt.gz`, and `t2.qq.png` are the same as their counterparts with the same suffixes of the first tutorial. In addition, the results of conditional gene-based association tests are in `t2.finemapping.gene.ecs.txt` which contains nine columns:

Header	Description
Gene	Gene symbol
Chrom	Chromosome of the gene
StartPos	Start coordinate of the gene
EndPos	End coordinate of the gene
#Var	Number of variants within the gene
Group	LD group number. Conditional ECS tests were performed for genes within a same LD group.
ECSP	p-value of ECS
CondiECSP	p-value of conditional gene-based association tests by conditional ECS
GeneScore	The gene's selective expression score in all tissues. A gene with a high socre will be given higher priority to enter the conditioning procedure.

Results of phenotype-tissue associations are in `t2.celltype.txt`. This is basically a Wilcoxon rank-sum test which tests whether the selective expression median of the phenotype-associated genes is significantly higher than that of other genes in an interrogated tissue. The file contains three columns:

Header	Description
TissueName	Name of the tissue being tested
PValue	p-values of phenotype-tissue associations.
Log(p)	The negative logarithm (base 10) of p-values of phenotype-tissue associations

2.3 EMIC

EMIC inferences gene expressions' causal effect on a complex phenotype with dependent expression quantitative loci by a robust median-based Mendelian randomization. SNPs with effects on both the phenotype and a gene are considered instrumental variables (IVs) of the gene, which can be used to infer the gene's expression effect on the phenotype. This analysis inputs effect sizes of SNPs on the phenotype and genes' expressions; outputs effect sizes and p-values of genes' expression effects on the phenotype. The command is:

```
java -Xmx4g -jar ../kggsee.jar \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --saved-ref VCFRefhg19 \
  --eqtl-file GTEx_v8_gene_BrainBA9.eqtl.txt.gz \
  --beta-col OR \
  --beta-type 2 \
  --emic \
  --out t3
```

Explanation of the flags and input files:

Header	Description
Flag	Description
`--sum-file`	Specifies a whitespace delimitated file of GWAS summary statistics. In this analysis, in addition to the columns of SNP coordinates and p-values, two columns of SNP alleles (named A1 and A2 by default), a column of A1 allele frequency (named FRQ_U by default), and two columns of SNP effect sizes (no default header) and their standard errors (named SE by default) are also needed.
`--saved-ref`	Specifies the folder of genotypes of reference population in KGGSEE object format, which is saved by the `--keep-ref` flag in the first tutorial.
`--eqtl-file`	Specifies a fasta-styled file of SNPs' effects on gene expressions. Here `GTEx_v8_gene_BrainBA9.eqtl.txt.gz` for gene-level EMIC. Try `GTEx_v8_transcript_BrainBA9.eqtl.txt.gz` for transcript-level EMIC.
`--beta-col`	Specifies the column name of effect sizes in the GWAS file.
`--beta-type`	Specifies the type of the effect size; here `2` means that it is the odds ratio for a qualitative phenotype.
`--emic`	Triggers the EMIC analysis.
`--out`	Specifies the prefix of output files.

Explanation of the output files:

The numeric results of EMIC are saved in `t3.emic.gene.txt`. There are nine columns in the file:

Header	Description
Gene	The gene symbol
#Var	Number of IVs within the gene
minP_EMIC	p-value of EMIC. When a transcript-level EMIC is performed, this is the minimum p-value among all transcripts of the gene.
Details_EMIC	Each detailed result has four components in brackets: the number of IVs, the causal effect estimate and its standard error, and the p-value. When a transcript-level EMIC is performed, results for each transcript are listed.
Chrom	Chromosome of the gene
Pos	The coordinate of the IV with the lowest GWAS p-value
GWAS_Var_P	GWAS p-value of the IV
GWAS_Var_Beta	The phenotype association effect size of the IV
GWAS_Var_SE	Standard error of the effect size

The columns of `t3.emic.gene.var.tsv.gz` are the same as `t3.emic.gene.txt`. The difference is that, for each gene, in `t3.emic.gene.txt`, only the eQTL with the lowest GWAS p-value is output, while in `turorial_3.emic.gene.var.tsv.gz`, all eQTLs are output. In this tutorial, the file `t3.emic.gene.PleiotropyFinemapping.txt` is empty, we ignore it here.

File `t3.qq.png` saves the Q-Q plot for GWAS p-values of IVs. File `t3.emic.qq.png` saves the Q-Q plot for EMIC p-values. File `t3.scatterplots.emic.pdf` saves the scatter plots of genetic association with gene expression. Each gene with an EMIC p-value lower than 2.5E-3 (default threshold) is saved on a separate page of the PDF. A filled rectangle on the plots denotes an IV. The red rectangle denotes the most significant GWAS variant among all the IVs of a gene. The slope of the line represents the estimated causal effect. The color of an IV denotes the degree of the LD between the IV and the most significant GWAS variant. The error bars in the rectangles denote the standard errors of the coefficient estimates.

2.4 Gene-based heritability estimation

Heritability is a measure of how well differences in people's genes account for differences in their phenotypes. This tutorial estimates the heritability of each gene with GWAS summary statistics. The command is:

```
java -Xmx4g -jar ../kggsee.jar \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --saved-ref VCFRefhg19 \
  --case-col Nca \
  --control-col Nco \
  --estimate-heritability \
  --out t4
```

Explanation of the flags and input files:

Flag	Description
`--sum-file`	Specifies a whitespace delimitated file of GWAS summary statistics. In this analysis, in addition to the columns of SNP coordinates and p-values, two columns of case and control sample sizes are also needed.
`--saved-ref`	Specifies the folder of genotypes of reference population in KGGSEE object format, which is saved by the `--keep-ref` flag in the first tutorial.
`--case-col`	Specifies the column name of the case sample size.
`--control-col`	Specifies the column name of the control sample size.
`--estimate-heritability`	Triggers gene-based association tests and estimation of gene heritability.
`--out`	Specifies the prefix of output files.

Explanation of the output files:

The output files are generally the same as the first tutorial, except that, in `t4.gene.pvalue.txt`, `t4.gene.var.pvalue.txt.gz`, there are two more columns named Herit and HeritSE, which are the estimate and its standard error of the gene heritability.

CHAPTER

THREE

DETAILED DOCUMENT

We first describe the general aspects of all analyses and then describe details for each analysis. KGGSEE performs analysis according to the following procedure:

1. Reads genotypes of an ancestrally matched reference population, e.g., a panel of 1000 Genomes Project. The genotypes can be in a VCF file specified by `--vcf-ref`, and if `--keep-ref` is used at the same run, KGGSEE saves the parsed VCF file in KGGSEE object format in the folder of `path/to/outputs/VCFRefhg*/`. For later run with `--keep-ref path/to/outputs/VCFRefhg*/`, KGGSEE reads genotypes from the object format files, which will be faster than parsing VCF files. KGGSEE calculates the minor allele frequency of each SNP and filters out SNPs with a minor allele frequency lower than the threshold specified by `--filter-maf-le` (default: `0.05`). KGGSEE also calculates the p-value of rejecting Hardy-Weinberg equilibrium for each SNP and filters out SNPs with a p-value lower than the threshold specified by `--hwe-all` (default: `1E-5`). Only SNPs with genotypes of the reference population and who have passed the two filters will be considered in the following procedures.
2. Reads GWAS summary statistics from a whitespace delimitated file specified by `--sum-file`. Depending on the analysis performed, this file needs to have different columns, which we will describe separately below. For all analyses, an eQTL summary statistic file specified by `--eqtl-file` may be read. We provide gene-based and transcript-based eQTL summary statistics for GTEx v8 tissues available for downloading on OneDrive.
3. Based on the flag specified, KGGSEE reads more needed files and performs the corresponding analysis.
   - *Gene-based association tests* trigered by `--gene-assoc`;
   - *DESE* trigered by `--gene-finemapping`;
   - *EMIC* trigered by `--emic`;
   - *Gene-based heritability estimation* trigered by `--estimate-heritability`.

The KGGSEE format of eQTL summary statistics is fasta-styled. An example is as follows:

```
#symbol   id    chr     pos     ref     alt     altfreq beta    se      p       neff    r2
>WASH7P ENSG00000227232 1
52238   T       G       0.94    -1.77   0.28    5.1E-9  65      0.38
74681   G       T       0.95    -1.45   0.33    1.1E-5  63      0.23
92638   A       T       0.24    0.54    0.20    7.9E-3  53      0.12
>MIR130 ENSG00000284557 1
52238   T       G       0.94    -1.77   0.28    5.1E-9  65      0.38
74681   G       T       0.95    -1.45   0.33    1.1E-5  63      0.23
```

The first row starting with # is the header line. Then, eQTLs of each gene/transcript are chunked. For each gene/transcript, the first row has three columns of (1) the gene symbol prefixed by >, (2) Ensembl gene/transcript ID, and (3) chromosome; the second and following rows have nine columns of (4) the eQTL coordinate, (5) the reference allele, (6) the alternative allele, (7) the frequency of the alternative allele, (8) the effect size, (9) the standard

error of the effect size, (10) the p-value of nonzero effect size, (11) the effective sample size and (12) coefficient of determination.

3.1 Gene-based association tests

KGGSEE performs the gene-based association analysis by GATES (a rapid and powerful **G**ene-based **A**ssociation **T**est using **E**xtended **S**imes procedure) and ECS (an **E**ffective **C**hi-square **S** tatistics). The `--gene-assoc` flag trigers both.

GATES (the GATES paper) is basically an extension of the Simes procedure to dependent tests, as the individual GWAS tests are dependent due to LD. GATES calculates an effective number of independent p-values which is then used by a Simes procedure.

ECS (the ECS paper) first converts the p-values of a gene to chi-square statistics(one degree of freedom). Then, merges all chi-square statistics of a gene after correcting the redundancy of the statistics due to LD. The merged statistic is called an ECS which is used to calculate the p-value of the gene.

3.1.1 Synopsis

```
java -Xms16g -Xmx16g -jar kggsee.jar
  --gene-assoc
  --out <prefix>
  --vcf-ref <file>
  --sum-file <file>
  --chrom-col <header>  # default: CHR
  --pos-col <header>  # default: BP
  --p-col <header>  # default: P
  --neargene <basepair>  # default: 5000
  --eqtl-file <file>
  --filter-eqtl-p <pval>  # default: 0.01
```

The flag `--gene-assoc` triggers the gene-based association tests. `--sum-file` specifies a white space-delimited GWAS summary statistic file which must have three columns of the chromosome of SNP, coordinate of SNP, and p-value of SNP; headers of the three columns can be specified by `--chrom-col`, `--pos-col` and `--p-col` separately. SNPs belonging to a gene can be defined either by SNPs close to the gene or by eQTLs of the gene. If `--neargene` is specified, KGGSEE reads gene annotations and considers SNPs inside a gene and its adjacent regions at a fixed number of basepairs on both sides to be a test unit. If `--eqtl-file` is specified, KGGSEE reads the eQTL summary statistic file and considers eQTLs of a gene or a transcript to be a test unit, and `--neargene` is overridden. When `--eqtl-file` is specified, `--filter-eqtl-p` can be used to specify a threshold of eQTL p-values. Only eQTLs with a p-value lower than the threshold will be considered. *A description of the eQTL file format* is near the beginning of this chapter.

3.1.2 Examples

3.1.2.1 Gene-based association tests based on physical distance

In this example, SNPs inside a gene and its 10 kb adjacent regions will be grouped for association tests.

```
java -Xmx4g -jar ../kggsee.jar \
--gene-assoc \
--vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
--sum-file scz_gwas_eur_chr1.tsv.gz \
--neargene 10000 \
--out t1.1
```

3.1.2.2 Gene-based association tests based on eQTLs

In this example, eQTLs of a gene will be grouped for association tests.

```
java -Xmx4g -jar ../kggsee.jar \
  --gene-assoc \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --eqtl-file GTEx_v8_gene_BrainBA9.eqtl.txt.gz \
  --out t1.2
```

3.1.2.3 Transcript-based association tests based on eQTLs

In this example, eQTLs of a transcript will be grouped for association tests.

```
java -Xmx4g -jar ../kggsee.jar \
  --gene-assoc \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --eqtl-file GTEx_v8_transcript_BrainBA9.eqtl.txt.gz \
  --out t1.3
```

3.1.3 Outputs

The file with a suffix of `.gene.pvalue.txt` saves the results of gene-based association tests. Columns of the file are as follow:

Header	Description
Gene	Gene symbol
#Var	Number of variants within the gene
ECSP	p-value of ECS
GATESP	p-value of GATES
Chrom	Chromosome of the gene
Pos	The coordinate of the variant with the lowest p-value within the gene
GWAS_Var_P	p-value of the variant

Columns of the file with the suffix of `.gene.var.pvalue.txt.gz` are the same as `*.gene.pvalue.txt`. The difference is that, for each gene, in `*.gene.pvalue.txt`, only the variant with the lowest p-value is output, while in `*.gene.var.pvalue.txt.gz`, all variants are output. The file with the suffix of `.qq.png` is the Q-Q plots for p-values of GWAS summary statistics and gene-based association tests by GATES and ECS.

3.2 DESE

DESE (**D**river-tissue **E**stimation by **S**elective **E**xpression; the DESE paper) estimates driver tissues by tissue-selective expression of phenotype-associated genes in GWAS. The assumption is that the tissue-selective expression of causal or susceptibility genes indicates the tissues where complex phenotypes happen primarily, which are called driver tissues. Therefore, a driver tissue is very likely to be enriched with selective expression of susceptibility genes of a phenotype.

DESE initially performed the association analysis by mapping SNPs to genes according to their physical distance. We further demonstrated that grouping eQTLs of a gene or a transcript to perform the association analysis could be more powerful. We named the **e**QTL-guided **DESE** eDESE. KGGSEE implements DESE and eDESE with an improved effective chi-squared statistic to control type I error rates and remove redundant associations (the eDESE paper).

3.2.1 Synopsis

```
java -Xms16g -Xmx16g -jar kggsee.jar
  --gene-finemapping
  --out <prefix>
  --vcf-ref <file>
  --sum-file <file>
  --chrom-col <header>  # default: CHR
  --pos-col <header>  # default: BP
  --p-col <header>  # default: P
  --neargene <basepair>  # default: 5000
  --eqtl-file <file>
  --filter-eqtl-p <pval>  # default: 0.01
  --multiple-testing <bonf|benfdr|fixed>  # default: bonf
  --p-value-cutoff <pval>  # default: 0.05
  --top-gene <number>
  --expression-file <file>
  --geneset-db <cura|cgp|cano|cmop|onto|onco|immu>
  --geneset-file <file>
```

The flag `--gene-finemapping` triggers DESE. First, KGGSEE performs gene-based association tests, which is the same as the analyses triggered by `--gene-assoc`. `--sum-file` specifies a white space delimited GWAS summary statistic file which must have three columns of the chromosome of SNP, coordinate of SNP, and p-value of SNP; headers of the three columns can be specified by `--chrom-col`, `--pos-col` and `--p-col` separately. SNPs belonging to a gene can be defined either by SNPs close to the gene or by eQTLs of the gene. If `--neargene` is specified, KGGSEE reads gene annotations and considers SNPs inside a gene and its adjacent regions at a fixed number of basepairs on both sides to be a test unit. If `--eqtl-file` is specified, eDESE is evoked; KGGSEE reads the eQTL summary statistic file and considers eQTLs of a gene or a transcript to be a test unit, and `--neargene` is overridden. When `--eqtl-file` is specified, `--filter-eqtl-p` can be used to specify a threshold of eQTL p-values. Only eQTLs with a p-value lower than the threshold will be considered. *A description of the eQTL file format* is near the beginning of this chapter.

Second, after the gene-based association tests, significant genes by ECS are retained for fine-mapping. `--multiple-testing` specifies the method for multiple testing correction: `bonf` denotes Bonferroni correction; `benfdr` denotes Benjamini–Hochberg FDR; `fixed` denotes no correction. `--p-value-cutoff` specifies the threshold of the adjusted p-value. `--top-gene` specifies the maximum number of genes retained for fine-mapping. So, only

genes (no more than the specified maximum number) with adjusted p-values lower than the specified threshold are retained for fine-mapping. Then, KGGSEE reads the expression file specified by `--expression-file` and performs iterative estimation of driver tissues.

Finally, if `--geneset-db` is specified, KGGSEE tests if the conditional significant genes are enriched in gene sets of MSigDB. The abbreviations of gene sets are as follow:

`cura`: C2. curated gene sets;
`cgp` : C2. chemical and genetic perturbations;
`cano`: C2. canonical pathways;
`cmop`: C4. computational gene sets;
`onto`: C5. ontology gene sets;
`onco`: C6. oncogenic signature gene sets;
`immu`: C7. immunologic signature gene sets.

Customized gene sets for enrichment tests can be specified by `--geneset-file`. Please refer to `resources/*.symbols.gmt.gz` under the KGGSEE directory for file formats.

Expression files should be white space delimitated. The first column is gene/transcript IDs. The following columns are means and standard errors of expression levels of genes or transcripts in multiple tissues. A gene-level expression file looks like this:

Name	Tissue1.mean	Tissue1.SE	Tissue2.mean	Tissue2.SE	...
ENSG00000223972.5	0.0038016	0.00036668	0.0045709	0.00046303	...
ENSG00000227232.5	1.9911	0.030021	1.8841	0.040247	...
ENSG00000278267.1	0.00049215	0.00010645	0.00036466	9.2944E-05	...
ENSG00000243485.5	0.0047772	0.00038018	0.0067897	0.00074318	...
ENSG00000237613.2	0.0030462	0.00027513	0.0030465	0.00031694	...
ENSG00000268020.3	0.011766	0.00061769	0.013409	0.0011429	...
ENSG00000240361.1	0.017913	0.00093294	0.021833	0.001556	...

A transcript-level expression file looks like this:

Name	Tissue1.mean	Tissue1.SE	...
ENST00000373020.8:ENSG00000000003.14	35.06	0.52271	...
ENST00000494424.1:ENSG00000000003.14	0.0034329	0.001209	...
ENST00000496771.5:ENSG00000000003.14	1.0462	0.019697	...
ENST00000612152.4:ENSG00000000003.14	2.5764	0.041124	...
ENST00000614008.4:ENSG00000000003.14	0.42826	0.01346	...
ENST00000373031.4:ENSG00000000005.5	15.215	0.58333	...
ENST00000485971.1:ENSG00000000005.5	1.0715	0.04074	...

3.2.2 Examples

3.2.2.1 DESE based on physical distance

In this example, SNPs inside a gene and its 10 kb adjacent regions will be considered as belonging to a gene. Significant genes by ECS with FDR<0.05 will be retained for fine-mapping.

```
java -Xmx4g -jar ../kggsee.jar \
  --gene-finemapping \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --neargene 10000 \
  --multiple-testing benfdr \
  --p-value-cutoff 0.05 \
  --expression-file ./GTEx_v8_TMM.gene.meanSE.txt.gz \
  --out t2.1
```

3.2.2.2 eDESE based on gene-level eQTLs

In this example, eQTLs of a gene will be considered as a unit for a gene-based association test. The top 100 significant genes by ECS with nominal p<0.05 will be retained for fine-mapping. Significant genes by eDESE will be tested if they are enriched in the C5. ontology gene sets of MSigDB:

```
java -Xmx4g -jar ../kggsee.jar \
  --gene-finemapping \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --eqtl-file GTEx_v8_gene_BrainBA9.eqtl.txt.gz \
  --multiple-testing fixed \
  --p-value-cutoff 0.05 \
  --top-gene 100 \
  --expression-file ./GTEx_v8_TMM.gene.meanSE.txt.gz \
  --geneset-db onto \
  --out t2.2
```

3.2.2.3 SelDP based on gene-level eQTLs

In this example, `--expression-file` specifies a customized file of the drug-induced gene-expression fold-change profile which has the same format as a gene expression file. SelDP estimates the drug selective perturbation effect on the phenotype-associated genes' expression to aid the drug repositioning for complex diseases.

```
java -Xmx4g -jar ../kggsee.jar \
  --gene-finemapping \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --eqtl-file GTEx_v8_genet_BrainBA9.eqtl.txt.gz \
  --expression-file drug-induced_expression_change_profile \
  --out t2.3
```

3.2.3 Outputs

The three files with suffixes of `.gene.pvalue.txt`, `.gene.var.pvalue.txt.gz`, and `.qq.png` are the same as their counterparts output by *Gene-based association tests*.

In addition, results of conditional gene-based association tests are saved in a file with a suffix of `.finemapping.gene.ecs.txt`. Columns of the file are as follow:

Header	Description
Gene	Gene symbol
Chrom	Chromosome of the gene
StartPos	Start position of the gene
EndPos	End position of the gene
#Var	Number of variants within the gene
Group	LD group number. Conditional ECS tests were performed for genes within the same LD group.
ECSP	p-value of ECS
CondiECSP	p-value of conditional gene-based association tests by conditional ECS
GeneScore	The gene's selective expression score in all tissues. A gene with a high score will be given higher priority to enter the conditioning procedure.

Results of phenotype-tissue associations are saved in a file with a suffix of `.celltype.txt`. Columns of the file are as follow:

Header	Description
TissueName	Name of the tissue being tested
PValue	p-values of phenotype-tissue associations. This is basically a Wilcoxon rank-sum test which tests whether the selective expression median of the phenotype-associated genes is significantly higher than that of other genes in an interrogated tissue.
Log(p)	The negative logarithm (base 10) of p-values of phenotype-tissue association

If `--geneset-db` or `--geneset-file` is specified, results of enrichment tests are saved in a file with a suffix of `.geneset.txt`. Columns of the file are as follow:

Header	Description
GeneSet_ID	Gene-set ID in the first column of the gene-set file
Enrichment_PValue_Hypergeometric	p-values of the hypergeometric tests.
IsSignificant_Hypergeometric	If the conditional significant genes are significantly enriched in the gene set.
Total_GeneSet_Gene#	The total number of genes in the gene set.
GeneSet_URL	Gene-set URL in the second column of the gene-set file
Gene_PValue	p-values of conditional significant genes within the gene set.

3.3 EMIC

EMIC (**E**ffective-median-based **M**endelian randomization framework for **I**nferring the **C**ausal genes of complex phenotypes) inferences gene expressions' causal effect on a complex phenotype with dependent expression quantitative loci by a robust median-based Mendelian randomization. The effective-median method solved the high false-positive issue in the existing MR methods due to either correlation among instrumental variables or noises in approximated linkage disequilibrium (LD). EMIC can further perform a pleiotropy fine-mapping analysis to remove possible false-positive estimates (the EMIC paper).

3.3.1 Synopsis

```
java -Xms16g -Xmx16g -jar kggsee.jar
  --emic
  --out <prefix>
  --vcf-ref <file>
  --sum-file <file>
  --chrom-col <header>  # default: CHR
  --pos-col <header>  # default: BP
  --a1-col <header>  # default: A1
  --a2-col <header>  # default: A2
  --freq-a1-col <header>  # default: FRQ_U
  --beta-col <header>
  --beta-type <0|1|2>
  --se-col <header>  # default: SE
  --eqtl-file <file>
  --filter-eqtl-p <pval>  # default: 1E-4
  --ld-pruning-mr  <r2>  # default: 0.5
  --emic-pfm-p <pval>  # default: 2.5E-6
  --emic-plot-p <pval>  # default: 2.5E-3
```

When performing EMIC (triggered by `--emic`), a GWAS summary statistic file (specified by `--sum-file`) and an eQTL summary statistic file (specified by `eqtl-file`) are needed. The GWAS summary statistic file must have columns of SNP coordinates (specified by `--chrom-col` and `--pos-col`), the two alleles (specified by `--a1-col` and `--a2-col`), frequencies of the allele specified by `--a1-col` (specified by `--freq-a1-col`), the effect sizes and its standard errors (specified by `--beta-col` and `--se-col`). The type of effect sizes is specified by `--beta-type` (`0` for linear regression coefficient of a quantitative phenotype; `1` for the logarithm of odds ratio or logistic regression coefficient of a qualitative phenotype; `2` for an odds ratio of a qualitative phenotype). `--filter-eqtl-p` specifies the p-value threshold of eQTLs; only eQTLs with a p-value lower than the threshold will be considered; we note here that the default value is `1E-4` for EMIC, which is different from the other analyses. `--ld-pruning-mr` specifies the threshold of LD coefficient when pruning variants; for each gene or transcript, eQTLs with LD coefficients higher than the threshold will be pruned. `--emic-pfm-p` specifies the p-value threshold to further perform an EMIC pleiotropy fine-mapping (EMIC-PFM) analysis; if the EMIC p-value of a gene is lower than the threshold, an EMIC-PFM will be performed to control the false-positive caused by pleiotropy. `--emic-plot-p` specifies the p-value threshold for plotting a scatter plot; genes with an EMIC p-value lower than the threshold will be plotted. *A description of the eQTL file format* is near the beginning of this chapter.

3.3.2 Examples

3.3.2.1 EMIC based on gene-level eQTL

This is an example of gene-level EMIC. Only eQTLs with a p-value lower than 1E-6 will be considered IVs. Genes with a p-value of EMIC lower than 0.05 will also undergo EMIC-PFM. Genes with a p-value of EMIC lower than 0.01 will be plotted.

```
java -Xmx4g -jar ../kggsee.jar \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --eqtl-file GTEx_v8_gene_BrainBA9.eqtl.txt.gz \
  --beta-col OR \
  --beta-type 2 \
  --emic \
  --filter-eqtl-p 1e-6 \
  --emic-pfm-p 0.05 \
  --emic-plot-p 0.01 \
  --out t3.1
```

3.3.2.2 EMIC based on transcript-level eQTL

This is an example of transcript-level EMIC. Only eQTLs with a p-value lower than 1E-6 will be considered IVs. Transcripts with a p-value of EMIC lower than 0.05 will also undergo EMIC-PFM. Transcripts with a p-value of EMIC lower than 0.01 will be plotted.

```
java -Xmx4g -jar ../kggsee.jar \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --eqtl-file GTEx_v8_transcript_BrainBA9.eqtl.txt.gz \
  --beta-col OR \
  --beta-type 2 \
  --emic \
  --filter-eqtl-p 1e-6 \
  --emic-pfm-p 0.05 \
  --emic-plot-p 0.01 \
  --out t3.2
```

3.3.3 Outputs

The numeric results of EMIC are saved in a file with a suffix of `.emic.gene.txt`. There are nine columns in the file:

Header	Description
Gene	The gene symbol
#Var	Number of IVs within the gene
minP_EMIC	p-value of EMIC. When a transcript-level EMIC is performed, this is the minimum p-value among all transcripts of the gene.
Details_EMIC	Detailed results of EMIC-PFM separated by semicolons. Each result has four components in brackets: the number of IVs, the causal effect estimate and its standard error, and the p-value. When a transcript-level EMIC is performed, results for each transcript are listed.
Chrom	Chromosome of the gene
Pos	The coordinate of the IV with the lowest GWAS p-value
GWAS_Var_P	GWAS p-value of the IV
GWAS_Var_Beta	The phenotype association effect size of the IV
GWAS_Var_SE	Standard error of the effect size

The numeric results of EMIC-PFM are saved in a file with a suffix of `.emic.gene.PleiotropyFinemapping.txt`. Only genes with a p-value lower than the threshold specified by `--emic-pfm-p` are saved. The file has thirteen columns, in which nine are the same as columns of `*.emic.gene.txt`. The other four columns are:

Header	Description
Group	IDs of a group of genes that share eQTLs.
minP_EMIC_PFM	p-value of EMIC-PFM. When a transcript-level EMIC-PFM is performed, this is the minimum p-value among all transcripts of the gene.
DetailsEMIC_PFM	Detailed results of EMIC-PFM separated by semicolons. Each result has four components in brackets: the number of IVs, the causal effect estimate and its standard error, and the p-value. When a transcript-level EMIC-PFM is performed, results for each transcript are listed.
CochransQ	The p-value of an extended Cochran's Q test. The significance (p<1E-3) means that the causal effect is more likely to be false-positive. At this point, KGGSEE excludes its eQTLs which are also the eQTLs of other significant genes, and redoes EMIC. In this case, results in the columns of minP_EMIC_PFM and DetailsEMIC_PFM will be different from in the columns of minP_EMIC and Details_EMIC.

Columns of the file with a suffix of `.emic.gene.var.tsv.gz` are the same as `*.emic.gene.txt`. The difference is that, for each gene, in `*.emic.gene.txt`, only the eQTL with the lowest GWAS p-value is output, while in `*.emic.gene.var.tsv.gz`, all eQTLs are output. The file with a suffix of `.qq.png` saves the Q-Q plot for GWAS p-values of IVs. The file with a suffix of `.emic.qq.png` saves the Q-Q plot for EMIC p-values. The file with a suffix of `.scatterplots.emic.pdf` saves the scatter plots of genetic association with gene expression. Each gene with an EMIC p-value lower than the threshold specified by `--emic-plot-p` is saved on a separate page of the PDF. A filled rectangle on the plots denotes an IV. The red rectangle denotes the most significant GWAS variant among all the IVs of a gene. The slope of the line represents the estimated causal effect. The color of an IV denotes the degree of the LD between the IV and the most significant GWAS variant. The error bars in the rectangles denote the standard errors of the coefficient estimates.

3.4 Gene-based heritability estimation

This analysis estimates the heritability of each gene and performs gene-based association tests at the same time.

3.4.1 Synopsis

```
java -Xms16g -Xmx16g -jar kggsee.jar
  --estimate-heritability
  --out <prefix>
  --vcf-ref <file>
  --sum-file <file>
  --chrom-col <header>  # default: CHR
  --pos-col <header>  # default: BP
  --p-col <header>  # default: P
  --nmiss-col <header>  # default: Neff
  --case-col <header>
  --control-col <header>
  --neargene <basepair>  # default: 5000
  --eqtl-file <file>
  --filter-eqtl-p <pval>  # default: 0.01
```

`--estimate-heritability` triggers gene-based association tests and estimation of gene heritability. `--sum-file` specifies a white space delimited GWAS summary statistic file which must have three columns of the chromosome of SNP, coordinate of SNP, and p-value of SNP; headers of the three columns can be specified by `--chrom-col`, `--pos-col` and `--p-col` separately. In addition, for quantitative phenotype, a column of sample sizes is needed, and its header is specified by `--nmiss-col`; for qualitative phenotype, two columns of case sample sizes and control sample sizes are needed, and their header is specified by `--case-col` and `--control-col` separately. SNPs belonging to a gene can be defined either by SNPs close to the gene or by eQTLs of the gene. If `--neargene` is specified, KGGSEE reads gene annotations and considers SNPs inside a gene and its adjacent regions at a fixed number of basepairs on both sides to be a test unit. If `--eqtl-file` is specified, KGGSEE reads the eQTL summary statistic file and considers eQTLs of a gene or a transcript to be a test unit, and `--neargene` is overridden. When `--eqtl-file` is specified, `--filter-eqtl-p` can be used to specify a threshold of eQTL p-values. Only eQTLs with a p-value lower than the threshold will be considered. *A description of the eQTL file format* is near the beginning of this chapter.

3.4.2 Examples

3.4.2.1 Gene heritability based on physical distance

In this example, SNPs inside a gene and its 10 kb adjacent regions will be grouped to estimate heritability.

```
java -Xmx4g -jar ../kggsee.jar \
  --estimate-heritability \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --case-col Nca \
  --control-col Nco \
  --neargene 10000 \
  --out t4.1
```

3.4.2.2 Gene heritability based on eQTLs

In this example, eQTLs of a gene will be grouped to estimate heritability.

```
java -Xmx4g -jar ../kggsee.jar \
  --estimate-heritability \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --case-col Nca \
  --control-col Nco \
  --eqtl-file GTEx_v8_gene_BrainBA9.eqtl.txt.gz \
  --out t4.2
```

3.4.2.3 Transcript heritability based on eQTLs

In this example, eQTLs of a transcript will be grouped to estimate heritability.

```
java -Xmx4g -jar ../kggsee.jar \
  --estimate-heritability \
  --vcf-ref 1kg_hg19_eur_chr1.vcf.gz \
  --sum-file scz_gwas_eur_chr1.tsv.gz \
  --case-col Nca \
  --control-col Nco \
  --eqtl-file GTEx_v8_transcript_BrainBA9.eqtl.txt.gz \
  --out t4.3
```

3.4.3 Outputs

The file with a suffix of `.gene.pvalue.txt` saves the results of gene-based heritability estimates and association tests. Columns of the file are as follow:

Header	Description
Gene	Gene symbol
#Var	Number of variants within the gene
ECSP	p-value of ECS
GATESP	p-value of GATES
Herit	Heritability estimate
HeritSE	Standard error of the heritability estimate
Chrom	Chromosome of the gene
Pos	The coordinate of the variant with the lowest p-value within the gene
GWAS_Var_P	p-value of the variant

Columns of the file with the suffix of `.gene.var.pvalue.txt.gz` are the same as `*.gene.pvalue.txt`. The difference is that, for each gene, in `*.gene.pvalue.txt`, only the variant with the lowest p-value is output, while in `*.gene.var.pvalue.txt.gz`, all variants are output. The file with the suffix of `.qq.png` is the Q-Q plots for p-values of GWAS summary statistics and gene-based association tests by GATES and ECS.

CHAPTER

FOUR

OPTIONS

The options for *Reference population genotypes*, *GWAS summary statistics*, and *Miscellaneous global options* act on all analyses. For clarity, we have categorized the other parameters by *Gene-based association and heritability*, *DESE* and *EMIC*, although this has resulted in some duplication of parameters.

In the "Default" columns of the following tables, "null" denotes that the flag works with an argument but there is no default value; "n/a" denotes that the flag works without an argument.

4.1 Reference population genotypes

These options work on the VCF file of reference population genotypes. Only SNPs that pass the filters will be used for subsequent analyses. These options act on all analyses.

Flag	Description	Default
`--vcf-ref`	Specifies a VCF file of genotypes sampled from a reference population. These genotypes are used to estimate LD correlation coefficients among SNPs. For VCF files of separated chromosomes, use wildcards with quotes like `"chr*.vcf.gz"`.	null
`--keep-ref`	Keep the parsed VCF files as KGGSEE object format in a folder named `VCFRefhg*` under the output folder.	n/a
`--saved-ref`	Specifies the folder of genotypes of reference population in KGGSEE object format, which is saved by the `--keep-ref`. Reading KGGSEE object format files is faster than parsing VCF files.	null
`--filter-maf-le`	Filter SNPs with a minor allele frequency lower than the setting.	`0.05`
`--hwe-all`	Filter SNPs with a p-value of rejecting Hardy-Weinberg equilibrium lower than the setting.	`1E-5`

4.2 GWAS summary statistics

These options work on the GWAS summary statistics and act on all analyses.

Flag	Description	Default
`--sum-file`	Specifies a whitespace delimitated file of GWAS summary statistics.	null
`--chrom-col`	Specifies the column of chromosomes.	`CHR`
`--pos-col`	Specifies the column of coordinates.	`BP`
`--p-col`	Specifies the column of p-values.	`P`
`--a1-col`	Specifies the column of the reference allele to calculate effect sizes.	`A1`
`--a2-col`	Specifies the column of the other allele.	`A2`
`--freq-a1-col`	Specifies the column of the frequency of the allele specified by `--a1-col`.	`FRQ_U`
`--beta-col`	Specifies the column of effect sizes.	null
`--beta-type`	Specifies the type of effect sizes: `0` for the linear regression coefficient of a quantitative phenotype; `1` for the logarithm of odds ratio or logistic regression coefficient of a qualitative phenotype; `2` for an odds ratio of a qualitative phenotype.	null
`--se-col`	Specifies the column of standard errors of effect sizes. Note: even if the effect size is provided as an odds ratio, this is still the standard error of the logarithm (base e) of the odds ratio.	`SE`
`--nmiss-col`	Specifies the column of sample sizes for a quantitative phenotype.	`Neff`
`--case-col`	Specifies the column of case sample sizes for a qualitative phenotype.	null
`--control-col`	Specifies the column of control sample sizes for a qualitative phenotype.	null

4.3 Gene-based association and heritability

Flag	Description	Default
`--gene-assoc`	Triggers gene-based association tests.	n/a
`--estimate-heritability`	Triggers gene-based association tests and estimation of gene heritability.	n/a
`--prevalence`	Specifies the proportion of cases in the population when estimating the heritability of a qualitative phenotype.	0.01
`--neargene`	Specifies the number of basepairs to extend at both ends of a gene, when considering SNPs belonging to the gene.	`5000`
`--eqtl-file`	Specifies a fasta-styled file of eQTL summary statistics. If this flag is used, `--neargene` is overridden, and eQTLs of a gene or transcript will be grouped and tested.	null
`--filter-eqtl-p`	Specifies the threshold of eQTL p-values. Only eQTLs with a p-value lower than the threshold will be used. The default is `0.01` when performing gene-based association tests and heritability estimating.	`0.01`

4.4 DESE

Flag	Description	Default
`--gene-finemapping`	Trigers the DESE, eDESE or SelDP.	n/a
`--expression-file`	Specifies a gene expression file that contains means and standard errors of gene expressions in multiple tissues.	null
`--multiple-testing`	Specifies the method for multiple testing correction. `bonf` denotes performing Bonferroni correction; `benfdr` denotes controlling false discovery rate by the Benjamini–Hochberg method; `fixed` denotes no correction.	`bonf`
`--p-value-cutoff`	Specifies the threshold of the adjusted p-value for fine-mapping. Only genes with an adjusted p-value lower than the threshold will be retained for fine-mapping.	0.05
`--top-gene`	Specifies the maximum number of genes with the smallest p-values that will be retained for fine-mapping.	null
`--geneset-db`	Specifies MSigDB gene sets for enrichment analysis: `cura`: C2. curated gene sets; `cgp`: C2. chemical and genetic perturbations; `cano`: C2. canonical pathways; `cmop`: C4. computational gene sets; `onto`: C5. ontology gene sets; `onco`: C6. oncogenic signature gene sets; `immu`: C7. immunologic signature gene sets.	null
`--geneset-file`	Specifies a user-defined file of gene sets for enrichment analysis.	null
`--neargene`	Specifies the number of basepairs to extend at both ends of a gene, when considering SNPs belonging to the gene.	`5000`
`--eqtl-file`	Specifies a fasta-styled file of eQTL summary statistics. If this flag is used, `--neargene` is overridden, and eQTLs of a gene or transcript will be grouped and tested.	null
`--filter-eqtl-p`	Specifies the threshold of eQTL p-values. Only eQTLs with a p-value lower than the threshold will be used. The default is `0.01` when performing DESE.	`0.01`

4.5 EMIC

Flag	Description	Default
`--emic`	Triggers the EMIC.	n/a
`--eqtl-file`	Specifies a fasta-styled file of eQTL summary statistics.	null
`--filter-eqtl-p`	Specifies the threshold of eQTL p-values. Only eQTLs with a p-value lower than the threshold will be used. The default is `1E-4` when performing EMIC.	`1E-4`
`--ld-pruning-mr`	Specifies the threshold of LD coefficient when pruning variants. For each gene or transcript, eQTLs with LD coefficients higher than the threshold will be pruned.	0.5
`--emic-pfm-p`	Specifies the p-value threshold to further perform an EMIC pleiotropy fine-mapping (EMIC-PFM) analysis. If the EMIC p-value of a gene is lower than the threshold, an EMIC-PFM will be performed to control the false-positive caused by pleiotropy.	`2.5E-6`
`--emic-plot-p`	Specifies the p-value threshold for plotting a scatter plot. Genes with an EMIC p-value lower than the threshold will be plotted.	`2.5E-3`

4.6 Miscellaneous global options

These options act on all analyses.

Flag	Description	Default
`--nt`	Specifies the number of threads.	`4`
`--buildver`	Specifies the reference genome version of the coordinates. The supported versions are `hg19` and `hg38`.	`hg19`
`--db-gene`	Specifies the database of gene annotations. `refgene` for RefSeq Genes; `gencode` for GENCODE; `refgene,gencode` for both.	`gencode`
`--excel`	Output results in Excel format.	n/a
`--only-hgnc-gene`	Only genes with an HGNC-approved gene symbol are considered in analyses.	n/a
`--out`	Specifies the output prefix of results.	null
`--regions-bed`	Specifies a BED file to define customized gene coordinates instead of the annotation from RefSeqGene or GENCODE. The first three columns of the BED file define gene coordinate and are mandatory; the fourth column defines gene names and is optional. When the fourth column is absent, a gene name of the format like `chr1:100-200` will be allocated.	null
`--regions-out`	Specifies genomic regions to be excluded in analyses, e.g. `chr1,chr2:2323-34434,chr2:43455-345555`.	null
`--resource`	Specifies the path KGGSEE running resource data.	`resources/` under the folder of `kggsee.jar`