Table of Contents

Detailed Document

Detailed Document

KGGSum（Knowledge-based Genetic and Genomic analysis platform for GWAS Summary statistics）

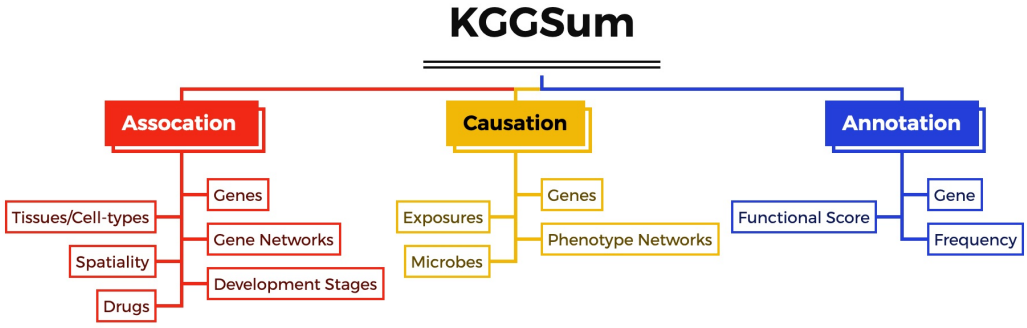
Introduction

KGGSum (*Knowledge-based Genetic and Genomic analysis platform for GWAS Summary statistics*) is **an open-source, Java-based platform tailored for interpreting LARGE-scale GWAS signals through comprehensive integrative analyses of omics data by advanced statistical models**. It provides a cohesive framework that combines GWAS summary statistics with diverse omics layers—ranging from transcriptomic and proteomic data to perturbation expression datasets—across different cell types. Utilizing the robust GTB and CCF file systems, KGGSum effectively manages intermediate and resource data, supporting advanced models for prioritizing genes, cell types, gene networks, life exposures and even microbion driving complex phenotypes. With an intuitive command-line interface, extensive functionality, and thorough documentation, KGGSum is a versatile, one-stop solution for large-scale post-GWAS analyses.

Functionality

KGGSum currently supports the following function modules:

- `association` : a module that links genes, cell types, and gene networks to phenotypes based on the GWAS signals at variants. Various association analyses are driven by the corresponding resource data.
- `causation` : a module that infers causation from genes or exposures to phenotypes, mainly by various Mendelian randomization methods.
- `annotation` : a module that annotates variants with gene features and functional prediction scores.



Unique Advantages of KGGSum

- **Fast and easy screening for multiple LARGE-scale GWAS summary datasets at a time**
- **Comprehensive with numerous high-quality resources for integrative analyses**
- **One-stop platform for a series of professional data mining**

Citations

We kindly ask that you cite the relevant papers associated with the methods you utilize in the KGGSum platform. You can find the corresponding references under each method’s introduction.

KGGSum and Its Running Resources

KGGSum is a Java-based tool distributed as a Java Archive file (kggsum.jar). Depending on the type of analysis, additional resources are required:

- **Gene-based association tests (GATES and ECS)** and **heritability estimations (EHE)** require reference genotypes and gene annotations.
- **Gene-expression causal-effect estimations (EMIC)** require eQTL summary statistics.

The following table provides download links for KGGSum and its resources:

File	Description	Version	Size
kggsum.jar	The KGGSum program	1.0	~21 MB
resources.zip	Resource files (excluding reference genotypes and eQTL statistics)	1.0	~180 MB
tutorials.zip	Toy data for workshop tutorials	1.0	~255 MB

Update History

1. 2025-3-10 Release the first formal version of KGGSum.
2. 2025-3-26 Optimize some internal data flows to speed up analysis and reduce memory usage.
3. 2025-4-2 Update some download links of KGGSum for resource data.

Resource Files in resources.zip

The resources.zip file includes essential files for coordinate conversion and gene annotations:

File	Description
resources/liftover/hg19ToHg38.over.chain.gz	Chain file for converting coordinates from hg19 to hg38
resources/reference/CanonicalTranscript.txt.gz	Canonical transcript details (symbol, Ensembl ID, etc.)
resources/reference/GEncode_hg38_kggseq_v2.txt.gz	GENCODE annotations for hg38
resources/reference/refGene_hg38_kggseq_v2.txt.gz	RefGene annotations for hg38

Setting Up a Java Runtime Environment (JRE)

KGGSum requires a Java Runtime Environment (JRE) version 1.8 or higher. You can use either Java SE JRE or OpenJDK JRE.

Installation Steps

1. Download and Install JRE
    - For Java SE JRE, visit the official Java download page and follow the instructions.
    - For OpenJDK JRE, refer to the OpenJDK installation guide.
2. Verify Installation
    - Open a terminal (Linux/macOS) or Command Prompt/PowerShell (Windows).
    - Run the following command:

```
java -version
```

- If installed correctly, you should see output similar to:

  `Java(TM) SE Runtime Environment (build 1.8.0_XXX)`

  or

  `OpenJDK Runtime Environment (build 1.8.0_XXX)`

- If the command is not recognized, ensure the JRE is installed and the java command is added to your system's PATH.

Setting Up an Environment for Quick Tutorials

To set up an environment for running the tutorial examples, follow these steps:

1. Download Files
   - kggsum.jar
   - resources.zip
   - tutorials.zip
2. Unzip resources.zip and tutorials.zip
   - Extract the contents of both zip files.
3. Organize Files
   - Place kggsum.jar, the extracted resources/ folder, and the extracted tutorials/ folder into a single directory (e.g., kggsum/).
4. Verify KGGSum Installation
   - Open a terminal (Linux/macOS) or Command Prompt/PowerShell (Windows).
   - Navigate to the kggsum/tutorials/ directory:

     ```
     cd path/to/kggsum/tutorials
     ```

   - Run the following command to test KGGSum:

     ```
     java -jar ../kggsum.jar
     ```

   - If successful, KGGSum will display usage information or a help message.

Notes

- The resources/ folder contains files necessary for tutorials, such as liftover chains and gene annotations.
- For advanced analyses, additional resources (e.g., reference genotypes, eQTL summary statistics) may be required and can be downloaded separately.

General Notes

- Ensure your system meets the recommended hardware specifications, particularly for large datasets.
- If you encounter Java-related issues, double-check the JRE installation and PATH configuration.
- KGGSum allows flexibility in resource usage—download only the files needed for your specific analysis.

Resources for Analyses

We have investigated various types of public data that you can download to carry out large-scale knowledge-based data mining analyses on KGGSum. The following are available links for downloading.

Type	Description
Reference	Variant IDs and genotypes in VCF or GTB formats are used as reference panels to correct the coordinates LD of GWAS variants.
Genotypes	1. 1000 Genomes Project (1KG): the genotypes of 5 different ancestry panels can be downloaded here. 2. The Haplotype Reference Consortium (HRC): It contains high-quality haplotype reference data for approximately 64,940 European individuals. The dataset publishes its sub-reference dataset at https://ega-archive.org/datasets/EGAD00001002729, and you can apply to download the original .vcf file. **Note** that the ancestry of the reference panel should be identical to that of the GWAS sample.
dbSNP IDs	This is required when only dbSNP SNP IDs (rs#), but not genomic coordinates, are available in the GWAS summary data. RSID coordinates can be accessed at NCBI FTP. We provide a GTB version of this dataset, converted from dbSNP's VCF (B157).
GWAS summary	The GWAS summary statistics of phenotypes. Many public domains are providing such data. As long as the data are in text and contain the columns required by KGGSum, it can be used as input.
Phenotypes	The following domains are widely used with downloadable GWAS summary statistics datasets. 1. Open GWAS: A database of genetic associations from GWAS summary datasets for querying or downloading. ==Note==: The GWAS summary data downloaded from Open GWAS in VCF form can be directly used as input of KGGSum without format conversion. 2. GWAS Catalog: The NHGRI-EBI Catalog of human genome-wide association studies 3. Pan-UKBB: Pan-UKBB provides a multi-ancestry analysis of 7,228 phenotypes using a generalized mixed model association testing framework spanning 16,131 genome-wide association studies. 4. FinnGen: is a research project in genomics and personalized medicine. It is a large public-private partnership that has collected and analyzed genome and health data from 500,000 Finnish biobank donors to understand the genetic basis of diseases.
Microbes	A collection of GWAS summary datasets for microbes quantities in the human digestive system. They were curated from three databases: Dutch Microbiome Project, Mibiogen, and Finrisk. The curated datasets can be downloaded from here.
Gene expression	The gene expression profiles are curated from various public domains.
Tissues	The expression profiles of genes in 54 organs of tissues were curated from the GTEx project. It has been included in the downloaded resources.zip file, with the file name, GTEx_v8_TMM_all.gene.meanSE.txt.gz.
Cell-types	Gene expression profiles of 6,598 single-cell types from humans and mice, compiled by PCGA (https://pmglab.top/pcga) from publicly available scRNA-seq datasets. The curated datasets can be downloaded from here.
Spatial and cell-types	The phenotype-relevant Spatial and cell types can be inferred at the online platform, PSC https://pmglab.top/psc/. The backend analysis tool for this online platform is KGGSum.
Drug perturbation	The gene expression perturbation profiles in multiple cell types by various drugs. The preprocessing pipeline and methods can be seen in this Molecular Psychiatry paper. The curated datasets can be downloaded from here.
xQTL	The variants linked to genes by properties of genes such as RNA expression, splicing, protein expression, and DNA methylation.
eQTL	GTEx eQTL: cis-eQTL summary statistics calculated from the gene or transcript-level expression profile of 49 tissues or organs provided by the GTEx project (v8). The curated datasets can be downloaded from here.

Resources for Annotation

The great efforts of international projects and functional genomic studies have led to a rich tapestry of information from diverse biological domains that can be leveraged for variant annotation. We meticulously pre-processed some annotation fields from several commonly used databases to facilitate convenient variant annotation. Users can download the entire or study-needed databases directly from our website and start annotations for a large number of variants. For small-scale annotations (e.g., <10,000 variants), users can access these databases remotely via FTP/HTTP. The program will automatically fetch the specified slices of databases based on the index file of the given variants in a distributed fashion, eliminating the need to download the entire database locally and ensuring a more convenient and expedited process. Given the expansive and dynamic nature of genomic databases, KGGSum also offers users the adaptability to incorporate custom databases for annotation purposes.

Gene feature annotation

We support three well-established gene annotation systems: RefSeq genes (refgene), and GENCODE genes (gencode). In KGGSum, databases used for gene feature annotation should be constructed in **FASTA format**. Additionally, KGGA also accommodates custom gene databases formatted in FASTA, allowing users to integrate their own gene annotations tailored to specific research needs or preferences.

Database Name	FTP/HTTP	Version
refgene	https://ftp.ncbi.nih.gov/refseq/H_sapiens/annotation/	Updated on 2024-08-27
encode	https://www.gencodegenes.org/human/	Release 47

1. Why do I need KGGSum?

   Response: It is fast and convenient, and it has many functions and resources for one-stop large-scale GWAS summary data mining.

2. How do I cite KGGSum?

   Response: KGGSum is a platform. The publication of the specific method(s) you use should be cited to allow readers to understand the principles of your analyses.

3. Where do I report bugs and feedback on KGGSum?

   Response: You can send an email to limx54@163.com for this.

4. How can we use a generative AI to help me use KGGSum?

   You can upload the PDF version of the user manual to the AI and then ask questions. e.g. "Please generate commands for me to do gene-based analyses by KGGSum with my GWAS summary input file (path=abc.tsv.gz), reference genome = hg19".

Input

The main input of KGGSum includes the GWAS summary data in a text file and reference genotypes of GWAS in VCF or GTB formats.

GWAS summary data file

The GWAS summary data file includes variants with GWAS association summaries arranged in rows, with columns separated by whitespace or tabs. The minimal required columns are genomic coordinates and p-values (**CHR**, **BP**, and **P** by default). Additional statistical attributes may be needed for specific analyses; these requirements are detailed in the descriptions of the corresponding analysis functions.

CHR	POS	A1	A2	AF_A1	Beta	SE	P
1	86028	T	C	0.908392	-0.00108	0.00356	0.7626
1	693731	A	G	0.877949	-0.00034	0.00339	0.9209
1	713092	A	G	0.00623653	-0.04053	0.02491	0.1038
1	714596	T	C	0.962227	-0.00045	0.00615	0.9407
1	715205	C	G	0.993317	0.03543	0.02428	0.1445

The corresponding option for the GWAS summary input file is `--sum-file` or `-sf`. More settings about the variants and statistics can be specified in the option.

Format

```
--sum-file file [cp12Cols=CHR,POS,..] [r12Cols=RSID,..] [pbsCols=P,...] [refG=] [sep=] [freqA1Col=] [sampleSizeCols=] [betaType=<0/1/2>] [prevalence=] [exclude=]
```

Example

```
--sum-file ./CAD_UKBIOBANK.gz cp12Cols=chr,bp,a1,a2 pbsCols=pval,beta,se refG=hg19 freqA1Col=AF_A1 exclude=chr6:545554~444555545
```

options

- `file` specifies the path to the GWAS summary statistics. This can be a local file, an internet URL, or an intranet file path accessed via SFTP. **Note:** A LOCAL file path allows wildcards (say, '*.tsv') to specify multiple files as a single input. KGGSum will process the files one by one.
- `cp12Cols` specifies the column name in the summary file: chromosome, positions, effective (VCF alternative, A1) allele, and base (VCF reference, A2) allele.
- `r12Cols` specifies the column names in the summary file: dbSNP rs ID, effective allele (VCF alternative, A1), and base allele (VCF reference, A2). The corresponding coordinates will be retrieved from the dbSNP database. Ensure the database file in GTB format is downloaded, unzipped, and placed in KGGSum's working directory: ./resources/dbsnp/*.gtb. Note that `r12Cols` and `cp12Cols` are mutually exclusive.
- `pbsCols` specifies the column name in the summary file, which are p-values, effect size, and standard errors of effect size.
- `type` specifies the file type of the GWAS summary file. The default one is the TSV format. In addition, there are two alternative formats, VCF and GTB.
- `refG` specifies the reference genome of input variants, . **The default is `refG=hg19`.** Note that incorrect specification of the genome version will lead to the mismatching of GWAS variants with the annotation base variants. All built-in annotation of KGGSum is hg38.

- `sep` specifies the separator of the summary file. By default, it can recognize tabs and spaces. It recognizes four values, . The UNIVERSAL means tabs and spaces or commas. **The default is** `sep=UNIVERSAL`
- `freqA1Col` specifies the column for the value of A1's frequency
- `sampleSizeCols` specifies the columns for the sample sizes of cases and controls. If only one column is specified, it is supposed to be the whole sample size.
- `betaType` specifies the type of effect sizes, <0 1 2>.

  0 means coefficients of linear regression for a quantitative phenotype beta; 1 means coefficients of logistic regression or the logarithms of odds ratio for a qualitative phenotype; 2 means the odds ratio for a binary phenotype. **The default is** `betaType=1` .
- `prevalence` specifies the disease prevalence in a population. This is only required for a GWAS of disease phenotypes.
- `exclude` specifies the genomic regions of variants to be excluded.

Reference genotypes for linkage-disequilibrium calculation

In the analysis of some methods, genotype data from GWAS samples is required to perform linkage disequilibrium correction. However, such genotypes are often unavailable. In these cases, ancestry-matched reference genotypes can be used as a substitute for KGGSum. Suitable reference datasets include genotypes from the 1000 Genomes Project or the UK Biobank. These references are primarily used to estimate LD for common variants. Ideally, the reference dataset should include between 500 and 5000 subjects; larger datasets, while more comprehensive, may increase computation time. KGGSum supports two genotype file formats: VCF and [GenoType Block (GTB)], with the option `--ref-gty-file` .

Option	Description	Default
`--ref-gty-file`	Specify the input file. It is a combination of multiple parameters. `<type>` is used to specify the format of the input file. `<refG>` is used to specify the reference genome of input variants. Format: `--input <file> type=[AUTO/VCF/GTB] refG=[hg18/hg19/hg38]` Eaxmple: `--input ./example.vcf.gz type=VCF refG=hg38`	type=AUTO refG=hg19

Gene Score Profiles

The gene score profile contains various values representing different contexts or conditions, such as RNA or protein expression levels or perturbation effects. Each row corresponds to a gene, and the columns, separated by tabs, represent different contexts or conditions. For each context or condition, two columns can be provided: one for the mean (labeled with .mean) and another for the standard error (SE, labeled with .SE). While the SE column is optional, including it can enhance the accuracy of the analysis. Below is an example of the file format.

Gene	Adipose-Subcutaneous.mean	Adipose-Subcutaneous.SE	Adipose-VisceralOmentum.mean	Adipose-VisceralOmentum.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	9.29E-05
ENSG00000243485.5	0.0047772	0.00038018	0.0067897	0.00074318
ENSG00000237613.2	0.0030462	0.00027513	0.0030465	0.00031694

The path and relevant settings can be specified by the option

Option	Description
`--`	The scores can represent various attributes, such as RNA expression, protein expression, epigenetic markers, or perturbation profiles at genes. Each row corresponds to a gene, and each column (except the first) represents a condition. The first column should contain the gene symbols. This is a combination parameter with the following options: `file` : Specifies the file path of the gene score file, which can be a local path or a remote path accessed via a network. ==**NOTE**== For a LOCAL file path, it allows wildcards (say, 'brain*.tsv') to specify multiple files as a

`--gene-score-file`	single input. `calcSpecificity` : Triggers the calculation of the specificity of gene scores for each condition. The default is "y(es)". `noDirection` : Instructs KGGSum to ignore the directionality of specificity. The default is "y(es)". Format: `--gene-score-file file=file/path calcSpecifity=<y/n> noDirection=<y/n>` Example: `--gene-score-file file/path` \ calcSpecify=y

xQTL summary data

This dataset is used to link variants to their target genes, typically using each gene's eQTL summary statistics. Each row represents an eQTL and must include the following nine columns: gene symbol, gene ID, chromosome, position, p-value, effective (alternative) allele, base (reference) allele, effect size, and standard error. Below is an example of the file format.

symbol	id	chr	pos	ref	alt	altfreq	beta	se	p
LINC00115	ENSG00000225880	1	796375	T	C	0.149	-0.223	0.081	5.87E-03
LINC00115	ENSG00000225880	1	797440	T	C	0.159	-0.24	0.078	2.28E-03
LINC00115	ENSG00000225880	1	802496	C	T	0.146	-0.247	0.083	2.95E-03
LINC00115	ENSG00000225880	1	812743	C	T	0.17	-0.19	0.073	9.57E-03
LINC01128	ENSG00000228794	1	693731	A	G	0.118	-0.258	0.094	6.31E-03
LINC01128	ENSG00000228794	1	731718	T	C	0.151	-0.293	0.084	4.50E-04

The path and relevant settings can be specified by the option

Option	Description
`--xqtl-file`	Specify the xQTL summary file. This is a combination of parameters. In the file, one row represents a genetic variant with its association summary to a gene. The association can be calculated based on various gene characteristics, including RNA expression (eQTL), RNA splicing (sQTL), protein expression (pQTL), and methylation (mQTL). The first column should contain the gene symbols. This is a combination parameter with the following options: `file` specifies the path to the xQTL summary statistics. This can be a local file, an internet URL, or an intranet file path accessed via SFTP. ==**NOTE**==: For a LOCAL file path, it allows wildcards (say, a*b?.qtl.tsv.gz) to specify multiple files as a single input. `cp12Cols` specifies the column names in the summary file, which are chromosome, positions, effective (VCF alternative, A1) allele, and base (VCF reference, A2) allele. `pbsCols` specifies the column names in the summary file, which are p-values, effect size, and standard errors of effect size. `giCols` specifies the column names in the summary file, which are gene symbols, and gene ID. `freqA1Col` specifies the column for the value of A1's frequency `sampleSizeCols` specifies the columns for the sample sizes for the xQTL. `refG` specifies the reference genome of input variants. The default value is hg19. `sep` specifies the separator of the summary file. By default, it can recognize Tabs, spaces and commas, and the corresponding tag is UNIVERSAL `pCut` specifies the p-value threshold for selecting significant xQTL for subsequent analyses. The default value is `pCut=1E-6<` br /> `ldCut` specifies the LD r^2^ to prune highly redundant xQTLs. The default value is `ldCut=0.8` Format: `--xqtl-file file=file/path [cp12Cols=chr,pos,alt,ref] [pbsCols=p,beta,se] [giCols=symbol,id] [refG=hg19] [sep=TAB] [freqA1Col=altfreq] [sampleSizeCols=neff] [pCut=1E-6][ldCut=0.8]`

Output

KGGSum operates using a task-by-task model for all analyses. Each task generates an efficient binary file named variants.annot.hg38.gtb, which stores variants along with harmonized statistics, intermediate results and annotations. This design allows users to seamlessly resume interrupted workflows or re-run analyses with adjusted parameters, starting from the previous breakpoints or creating branched workflows.

Option	Description	Default
`--output`	Specify the output folder path. All data from each task will be put under the specified folder. That preserve intermediate files and can avoid duplicate tasks. Format: `--output <dir>` Example: `--output ./out/test`	./kggsum
`--clean-intermediate-data`	Clean the all intermediate data of the analysis, reducing memory usage. Format: `--clean-intermediate-data`	[OFF]

While each analysis has a unique task, some common tasks are listed below.

File	Description
ConvertVCF2GTBTask*.gtb	The gtb format file converted from the input VCF file by `--ref-gty-file`
GenerateRootVariantSetTask\variants.annot.hg38.gtb	The variants extracted from the reference genotype file specified by `--ref-gty-file` in gtb format will be used as the base for the following analysis.
AppendVariants2RootVariantSetTask\variants.annot.hg38.gtb	The base variants appended with GWAS summary statistics specified by the `--sum-file`
GeneFeatureAnnotationTask\variants.annot.hg38.gtb	The variants annotated with gene features subsequently
OutputVariants2TSVTask\variants.hg38.tsv.gz	The GWAS summary and annotations of variants retained for analysis in TSV format.

General Setting

There are also some options for resource file paths, template data, and parallel tasks.

Option	Description	Default
`--threads`	Specify the number of threads on which the program will be running. Format: `--threads <int>` Example: `--threads 8` **NOTE**: As a rule of thumb, please do not give a thread number larger than the number of CPU cores. Too many threads may even slow down the analysis.	4
--channel	Specify the path of resource data files. It could be a local system file path, an intranet file path, or even an internet one. Format: `--channel path/to/file` Example: `--channel /public1/resources`	./resources https://pmglab.top/kggsum/resources
--cache	Specify the path of the temple data files. It must be a local file path on the Operating System. Format: `--cache path/to/file` Example: `--cache ./tmp/`	[--output]/tmp/
--chromosome-tag	Specify the recognizable chromosomes. By default, it only considers the standard chromosomes, 1, ..., 22, X, Y, and M.	1,...,22,X,Y,M
update	Check and update the local kggsum.jar file to the latest version available on the website. Example: `java -jar kggsum.jar update` **Note:** The update option does **not** work on Windows. On Windows, you must manually download the latest kggsum.jar file and replace the existing one.	-

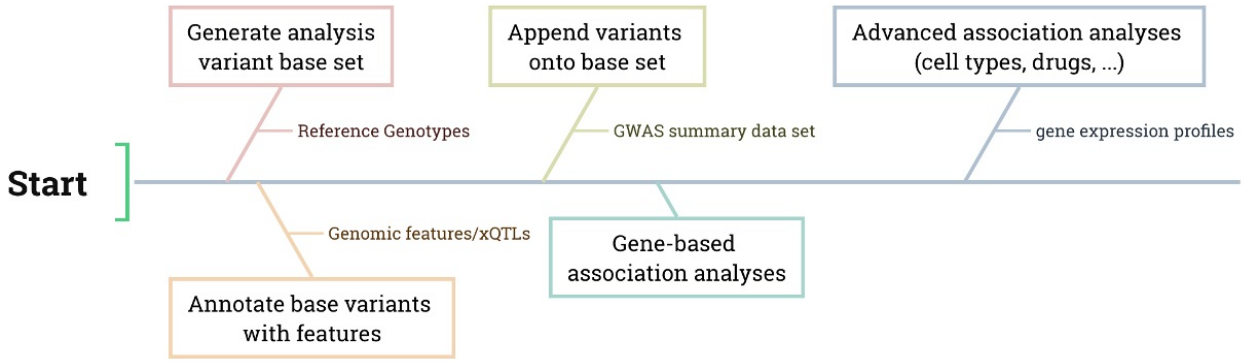
Association

About

The association module in KGGSum offers various functions to link genes, cell types, gene networks, and even drugs to phenotypes using GWAS signals from variants. A common step across all analyses involves aggregating association signals from multiple variants to derive gene-level associations using GATES (Li et al., 2011) and ECS (Li et al., 2019). Advanced association analyses build upon these gene-level associations. The rationale behind each type of association analysis is detailed in the respective option descriptions and relevant publications.

Main Workflow of the Association Module

1. **Generation**: Extract variant coordinates and frequencies from the VCF or GBC file to create a root variant set for further analysis.
2. **Annotation**: Annotate the root variant with gene features or xQTLs.
3. **Append**: Integrate GWAS variants and their summary statistics into the annotated root variant set.
4. **Gene-based association**: Conduct gene-based association analyses based on the gene feature annotations.
5. **Advanced association**: Conduct gene-based association analyses based on the gene feature annotations.
6. **…**: *(Additional association analysis as specified).*



Basic Usage

```
java -jar kggsum.jar assoc --sum-file <input1> --ref-gty-file <input2> --output <output> [options]
```

By default, the program performs gene-based association tests using GATES and ECS. Additional association analyses in this module can be initiated by specifying relevant input options (e.g., `--gene-score-file`).

Genes

GATES and ECS

Performs the gene-based association analysis using 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).

GATES (Li et al. 2011) 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 (Li et al. 2019) first converts the p-values of a gene to chi-square statistics(one degree of freedom). Then, it 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 and is used to calculate the p-value of the gene.

Citations

1. **For gene-based association analysis:** Miaoxin Li, Hong-Sheng Gui, Johnny S.H. Kwan, et al. GATES: a rapid and powerful gene-based association test using extended Simes procedure. The American Journal of Human Genetics, 2011, 88(3):283-293. PubMed Link
2. **For ECS and conditional gene-based association analysis** Miaoxin Li, Lin Jiang, T S H Mak, et al. A powerful conditional gene-based association approach implicated functionally important genes for schizophrenia. Bioinformatics, 2019, 35(4):628-635. PubMed Link
3. **For pathway or (gene-set) based association analysis:** Hongsheng Gui, Johnny S Kwan, Pak C Sham et al. Sharing of Genes and Pathways Across Complex Phenotypes: A Multilevel Genome-Wide Analysis. Genetics, 2017, 206(3):1601-1609. PubMed Link

Options

This analysis inputs the p-values of SNPs and outputs the p-values of genes. The tutorial command is:

```
java -Xmx4g -jar ../kggsum.jar \
   assoc \
   --sum-file ./scz_gwas_eur_chr1.tsv.gz \
              cp12Cols=CHR,BP \
              pbsCols=P \
              refG=hg19 \
   --ref-gty-file ./1kg_hg19_eur_chr1.vcf.gz \
                  refG=hg19 \
   --output ./t1
```

Flag	Description
`assoc`	Trigger the gene-based association analysis.
`--sum-file`	Specifies the file path of a GWAS summary file. The detailed description can be found here.
`--ref-gty-file`	Specifies the file path of a reference genotype file. The detailed description can be found here.
`--output`	Specifies the prefix of the output folder.
`--gene-model-database`	Specifies the gene boundaries for the variant-to-gene mapping. The default one is gencode. It can be refgene as well. The former contains more non-coding genes than the latter, while the latter may be more reliable. The exact boundaries are set by `--upstream-distance` and `--downstream-distance`. The default values are 1000 bp for both boundaries.
`--xqtl-file`	Specifies the target genes associated with each variant, typically using eQTL summary statistics. The detailed description can be found here. Note that the `--gene-model-database` and `--xqtl-file` options are mutually exclusive; specifying the latter will automatically disable the former. The `--xqtl-file` is usually used for causal gene inference via a Mendelian Randomization method (e.g., EMIC).

`--max-condi-gene`	Specifies the maximal number of significant genes for the conditional gene-based test. The default value is 1000. The value -1 disables this setting.

Output

The numeric results of gene-based association tests are saved in `GeneBasedAssociationTask\genes.hg38.assoc.txt`. These are the columns in the file:

Header	Description
RegionID	Region IDs or Gene symbols
Chromosome	Chromosome of the region or gene
StartPosition	The position of the first SNP in the region or gene
EndPosition	The position of the last SNP in the region or gene
#Var	Number of variants within the region or gene
GATES.P	p-value of ECS
ECS.P	p-value of GATES
CCT.P	the combined p-value of ECS and GATES by the Cauchy Combination test

The Q-Q plots for p-values of gene-based association tests by GATES or ECS are saved in `GeneBasedAssociationTask\genes.hg38.assoc.qq.pdf`.

In addition, the conditional gene-based association test is then carried out for significant genes to remove significant genes, mostly due to LD with the most significant gene in a local region. The results are stored in GeneBasedConditionalAssociationTask\genes.hg38.condi.assoc.txt. These are the columns in the file:

Header	Description
...	...
Condi.ECS.P	The p-value of the conditional gene-based test by ECS

In the above analysis, variants are mapped to genes according to their physical positions in the gene models (`--gene-model-database`). However, remote regulatory variants may not be included depending on the position. One can use xQTL to link distant variants to genes by option --eqtl-file.

Heritability

Gene/region-based heritability estimation by EHE

Heritability measures how well differences in people's genes account for differences in their phenotypes. This EHE analysis estimates each gene's heritability and performs gene-based association tests simultaneously (Miao et al. 2023).

Citations

Lin Miao, Lin Jiang, Bin Tang, Pak Chung Sham and Miaoxin Li. Dissecting the high-resolution genetic architecture of complex phenotypes by accurately estimating gene-based conditional heritability. The American Journal of Human Genetics (2023). 110(9):1534–1548. PubMed Link

Options

The tutorial command is:

```
java -Xmx4g -jar ../kggsum.jar \
   assoc \
   --calc-ehe \
   --sum-file ./scz_gwas_eur_chr1.tsv.gz \
              cp12Cols=CHR,BP \
              pbsCols=P \
              sampleSizeCols=Nca,Nco \
              refG=hg19 \
              prevalence=0.01 \
   --ref-gty-file ./1kg_hg19_eur_chr1.vcf.gz \
                  refG=hg19 \
   --output ./t2
```

Flag	Description
`assoc`	Trigger the gene-based association analysis.
--calc-ehe	Ask the program to estimate heritability when calculating association. It has two sub-options. calcCondi： also calculate the conditional heritability. Note it may be time-consuming. So it can be turned off by setting a value no. The default is yes. "topGeneNum=0
`--sum-file`	Specifies the file path of a GWAS summary file. For quantitative traits, a single column specifying the sample sizes is required. For binary traits, two columns indicating the case and control sample sizes are necessary. Additionally, for a disease phenotype, the disease prevalence must be specified. The detailed description can be found here.
`--ref-gty-file`	Specifies the file path of a reference genotype file. The detailed description can be found here.
`--out`	Specifies the prefix of the output folder.

Output

The gene-based association p-values and heritability estimates are saved in `GeneBasedAssociationTask\genes.hg38.assoc.txt`. These are the columns in the file:

Header	Description
...	...
eH2	The estimated heritability of the region or gene by EHE.
eH2.SE	The standard error of the estimated heritability.

In addition, a conditional gene-based estimation is then carried out for significant genes to remove genes that have heritability merely due to LD with the most significant gene in a local region. The results are stored in GeneBasedConditionalAssociationTask\genes.hg38.condi.assoc.txt. These are the columns in the file:

Header	Description
...	...
Condi.eH2	The estimated conditional heritability of the region or gene by EHE.
Condi.eH2.SE	The standard error of the estimated conditional heritability.

CellTypes

DESE

DESE (**D**river-tissue/cell **E**stimation by **S**elective **E**xpression; Jiang et al.. 2019) 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 develop primarily, which are called driver tissues. Therefore, a driver tissue is likely to be enriched with the selective expression of susceptibility genes of a phenotype.

DESE initially analyzed the association 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. KGGSum implements DESE and eDESE with an improved effective chi-squared statistic to control type I error rates and remove redundant associations (Li et al. 2022).

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 and outputs p-values of phenotype-tissue associations and conditional p-values of genes.

Citations

1. **For phenotype-associated tissue estimation by DESE:** Lin Jiang, Chao Xue, Sheng Dai, et al. DESE: estimating driver tissues by selective expression of genes associated with complex diseases or traits. Genome biology, 2019, 20(1):1-19. PubMed Link
2. **For phenotype-associated tissues' susceptibility genes and isoforms estimation:** Xiangyi Li, Lin Jiang, Chao Xue, et al. A conditional gene-based association framework integrating isoform-level eQTL data reveals new susceptibility genes for schizophrenia. Elife. 2022 Apr 12;11:e70779. PubMed Link
3. **For phenotype-associated cell-type estimation by DESE**: Xue C#, Jiang L#, Zhou M, Long Q, Chen Y, Li X, Peng W, Yang Q, Li M. PCGA: a comprehensive web server for phenotype-cell-gene association analysis. Nucleic Acids Res. 2022 May 26;50(W1):W568-76.

Options

The tutorial command is:

```
java -Xmx4g -jar ../kggsum.jar \
   assoc \
   --sum-file ./scz_gwas_eur_chr1.tsv.gz \
              cp12Cols=CHR,BP \
              pbsCols=P \
              refG=hg19 \
   --ref-gty-file ./1kg_hg19_eur_chr1.vcf.gz \
                  refG=hg19 \
   --gene-score-file ../resources/GTEx_v8_TMM_all.gene.meanSE.txt.gz \
   --output ./t3
```

Flag	Description
`assoc`	`--sum-file` , `--ref-gty-file` , and `--out` have the same functions as previously described.
`--gene-score-file`	Specifies a gene score file. The scores can represent various attributes, such as RNA expression, protein expression, epigenetic markers, or perturbation profiles at genes. See more at the Input Data Description
`--gene-p-cut`	Set p-value threshold to select significant genes for the conditional gene-based test. The default value is 0.05.
`--gene-multiple-testing`	Specifies the method for multiple testing correction with a given p-value threshold to select significant genes for the conditional gene-based test. It has three alternative method labels: bonf: Bonferroni correction with family-wise threshold specified by `--gene-p-cut` benfdr: Filter by the false discovery rate (FDR) calculated by Benjamini-Hochberg procedure. The threshold is also defined by `--gene-p-cut` fixed: Filter by the p-value threshold specified by `--gene-p-cut` without any multiple testing correction. The default value is bonf.
`--max-condi-gene`	Set the maximal number of significant genes for the conditional gene-based test. The default value is 1000. The value -1 disables this setting.
`--permutation-`	Set the number of permutations to adjust the p-value for driver-tissue or -celltypes inference due to selection

`num`	bias and multiple testing. The default value is 100. A larger number will take more running time.

Output files

This function produces three sets of results: the gene-based association summary statistics saved in `GeneBasedAssociationTask\genes.hg38.assoc.txt`, the gene-based conditional association summary statistics saved in `GeneBasedAssociationTask\genes.hg38.condi.assoc.txt,` and the integrative enrichment summary statistics saved in `GeneBasedConditionalAssociationTask\$scoreFileName.enrichment.txt`. Basically, this is the result of the Wilcoxon rank-sum test, which tests whether the selective expression median of the phenotype-associated genes is significantly higher than that of the other genes in the interrogated tissue. The file contains four columns:

Header	Description
Condition	Name of the tissue being tested
Unadjusted(p)	Unadjusted p-values for the tissue-phenotype associations
Adjusted(p)	Adjusted p-values were calculated by adjusting both selection bias and multiple testing by permutation of gene scores within each condition.
Median(IQR)SigVsAll	Median (interquartile range) expression of the conditionally significant genes and all the background genes Heritability

Drugs

DESE

Infer effective drugs for a GWAS disease with selective perturbation gene expression profile by DESE. The assumption is that effective drugs may treat disease by specifically disturbing the expression of disease-susceptible genes. A detailed explanation can be found in this paper. The options and input format are the same as those of the above analyses for associated cell types. The difference is just what expression profiles are input. Instead of the gene expression profiles of various cell types or tissues, the perturbed gene expression profile by various drugs are specified by `--gene-score-file`.

Citations

1. Li X, Xue C, Zhu Z, Yu X, Yang Q, Cui L, Li M. Application of GWAS summary data and drug-induced gene expression profiles of neural progenitor cells in psychiatric drug prioritization analysis. Mol Psychiatry. 2025 Jan;30(1):111-121.

Options

The tutorial command is:

```
java -Xmx4g -jar ../kggsum.jar \
   assoc \
   --sum-file ./scz_gwas_eur_chr1.tsv.gz \
              cp12Cols=CHR,BP \
              pbsCols=P \
              refG=hg19 \
   --ref-gty-file ./1kg_hg19_eur_chr1.vcf.gz \
                  refG=hg19 \
   --gene-score-file https://idc.biosino.org/pmglab/resource/kgg/kggsum/datasets/drugs/GEO_expression_profiles/hipsc_
ctrl_with_se_drug_induced_foldchange.txt.gz \
   --threads 20 \
   --output ./t4
```

The options are identical to those for the associated Cell-type inference.

Output

The output files are the same as those of the CellTypes association analyses. The prioritized drugs are saved in `GeneBasedConditionalAssociationTask\$scoreFileName.enrichment.txt`, in which the Wilcoxon rank-sum test produces the enrichment scores, and the permutation approach is used to make valid statistical p-values with the consideration of multiple testing, selection bias, and internal correlation of gene perturbation scores.

Spatiality

Infer the spatial heterogeneity of cell types associated with complex diseases using DESE. We pre-integrated large-scale single-cell transcriptomics and spatial transcriptomics data to generate high-quality gene expression profiles of spatially specific cell types. The gene expression profiles can be input into DESE to estimate disease-associated spatially specific cell types and genes. Given the need for complex interactive visualizations, this functionality is implemented on the PSC web server (https://pmglab.top/psc), enabling convenient and rapid analysis and visualization of the results. Incidentally, the underlying program of PSC is still powered by the KGGSum platform.

Citations

1. Xue C., Liu M., Zhou M., Li M., PSC: a comprehensive web server for resolving spatial heterogeneity of cell types associated with complex phenotypes, 2024. Unpublished manuscript.

Output

The output results can be obtained and visualized on the website (https://pmglab.top/psc).

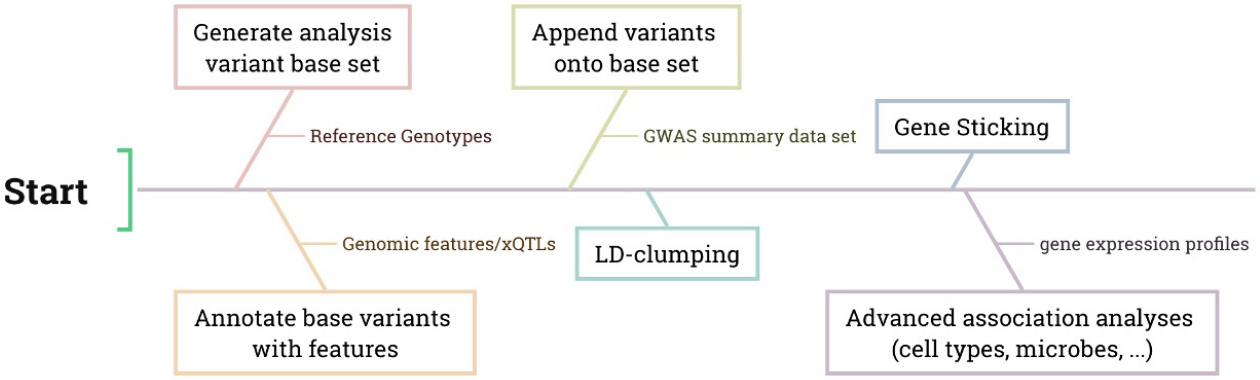
Causation inference

About

The `causation` module provides advanced Mendelian randomization methods for inferring causation from genes, lifestyle, or microbes (as exposures) to phenotypes (as outcomes). It enables rapid inference screening of tens of exposures and outcomes simultaneously.

Main Workflow of the Causation Module

1. **Generation**: Extract variant coordinates and frequencies from the VCF or GBC file to create a root variant set for further analysis.
2. **Annotation**: Annotate the root variant with gene features or xQTLs.
3. **Append**: Integrate GWAS variants and their summary statistics into the annotated root variant set.
4. **LD Clumping**: Select the significant variants of exposures as IVs and remove redundant IVs according to LD.
5. **Gene Sticking**: Link IV to the potential target gene according to LD. Note that this is a rough linking, and some target genes may not be true.
6. **MR analysis**: Infer the causality by suitable MR methods (e.g., EMIC or PCMR)
7. **…**: *(Additional analysis as specified).*



Basic Usage

```
java -jar kggsum.jar causal --sum-file <input1> --ref-gty-file <input2> --output <output> [options]
```

Genes

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 (Jiang et al. 2022).

Citations

[1] Lin Jiang, Lin Miao, Guorong Yi, Xiangyi Li, Chao Xue, Mulin Jun Li, Hailiang Huang and Miaoxin Li. Powerful and robust inference of causal genes of complex phenotypes with dependent expression quantitative loci by a novel median-based Mendelian randomization. Am J Hum Genet. 2022 May 5;109(5):838-856. PubMed

Options

The tutorial command is:

```
java -Xmx4g -jar ../kggsum.jar \
   causal \
   --xqtl-file https://idc.biosino.org/pmglab/resource/kgg/kggsum/datasets/gtex/eqtl/hg38/Brain_Frontal_Cortex_BA9_eu
r_v8_tmm_p01.gene.hg38.cov.eqtl.tsv.gz \
               refG=hg38 \
   --sum-file ./scz_gwas_eur_chr1.tsv.gz \
               cp12Cols=CHR,BP,A1,A2 \
               pbsCols=P,OR,SE \
               betaType=2 \
               prevalence=0.01 \
   --ref-gty-file ./1kg_hg19_eur_chr1.vcf.gz \
               refG=hg19 \
   --threads 18 \
   --output ./test/ba9_scz_causal
```

Flag	Description	Default
`causal`	Trigger the causality inference. The default analysis is EMIC to infer causal genes of phenotypes.	-
`--xqtl-file`	Specifies a file containing SNP effects on gene or transcript expression. The file should be a text table, where each row represents a single SNP, and columns are delimited by tabs or spaces. This is a combination parameter; further details can be found in the description of --xqtl-file.	-
...	...	...

Output

The numeric results of EMIC are saved in `GeneBasedCausationTask/genes.hg38.emic.tsv`. There are nine columns in the file:

Header	Description
SymbolID	The gene symbol
ChromosomeID	Chromosome of a gene
Start	The coordinate of the first SNP.
End	The coordinate of the last SNP.
ExpressionID	The gene ID.
::IVNum	Number of IVs within the gene.

::Effect	The estimated causality effect.
::SE	The standard error of the estimated causality effect.
::P	p-value of EMIC for statistical causality test.

Phenotypes

PCMR

PCMR (Pleiotropic Clustering model for MR analysis) is a tool for analyzing GWAS summary statistics (provided via --sum-file) to **infer causal relationships between phenotypes**. It is designed to tackle correlated horizontal pleiotropy, a common challenge in Mendelian Randomization (MR) studies.

By extending the zero modal pleiotropy assumption (ZEMPA), PCMR improves causal inference even in the presence of a high proportion of pleiotropic variants. It tackles the difficulty of distinguishing between correlated pleiotropic effects and true causal effects by combining them into a single “correlated HVP effect,” modeled using a Gaussian Mixture Model. This allows PCMR to categorize instrumental variables (IVs) effectively, including identifying those with causal effects.

PCMR also includes a pleiotropy test to detect correlated horizontal pleiotropy and enhances causal inference in these scenarios. This makes it a powerful tool for evaluating the causal effects of gene expression on complex phenotypes. (Tang et al., 2024).

Citations

[1] Bin Tang, Nan Lin, Junhao Liang, Guorong Yi, Liubin Zhang, Wenjie Peng, Chao Xue, Hui Jiang, Miaoxin Li. Leveraging Pleiotropic Clustering to Address High Proportion Correlated Horizontal Pleiotropy in Mendelian Randomization Studies. Nat Commun. 2025 Mar 21;16(1):2817 PubMed

Options

This main analysis inputs GWAS summary of SNPs and outputs p-values of genes. The following are options for an example:

```
java -Xmx4g -jar ../kggsum.jar \
  causal \
   --pcmr 1T2,2T1 \
   --sum-file ./smoking_chr1.tsv.gz \
              cp12Cols=CHR,POS,A1,A2 \
              pbsCols=Pval,Beta,SE \
              prevalence=0.05   \
              betaType=1   \
   --sum-file ./scz_gwas_eur_chr1.tsv.gz \
              cp12Cols=CHR,BP,A1,A2 \
              pbsCols=P,OR,SE \
              betaType=2 \
              prevalence=0.01 \
   --ref-gty-file ./1kg_hg19_eur_chr1.vcf.gz \
   --threads 10 \
   --output ./test/smk_scz_pcmr \
   --exclude-complementary-allele
```

Format	Description	Default
	Triggers the PCMR analysis. This is a combination parameter with the following options: • `causalPair` : Defines the direction of causal inference, where traits (indicated by their order number) are specified in the `--sum-file` . For example, a value of 1T2 indicates an inference of causation from the phenotype(s) specified by the first `--sum-file` to the phenotype(s) listed in the second `--sum-file` . • `effIVPCut` : Sets the p-value threshold for selecting instrumental variables. • `effIVPCorrect` : Set a method for multiple testing of p-values for selecting instrumental variables. There are three candidates methods: fixed (no correction), bonf (Bonferroni correction), and bhdfr (Benjamini and Hochberg FDR).	causalPair=1T2,2T1 effIVPCut=5E-8 effIVPCorrect=fixed

	• `ldPruneCut` : Sets the r² threshold for LD clumping.
	• `initIVPCut` : Sets the p-value threshold for selecting instrumental variables to model uncorrected pleiotropic effects.
	• `ldStickCut` : Sets the LD r² threshold for clustering genes whose SNPs are in LD with an instrumental variable.
	Format:
	`--pcmr causalPair= effIVPCut=[p-value] effIVPCorrect=[fixed] ldPruneCut=[r²] initIVPCut=[p-value] ldStickCut= [r²]`
	
	Example:
	`--pcmr 1T2,2T1 5E-8 fixed 0.1 0.5 0.8`
		ldPruneCut=0.1
		initIVPCut=0.5
		ldStickCut=0.8

`--exclude-complementary-allele`	If specified, variants with complementary alleles (e.g., A/T and C/G) are excluded from the analysis.	-
	The description of other options is the same as that for association analyses.	

Output

At the end of the PCMR analysis, the results are summarized and stored in a file named MendelianRandomization.summary.tsv. Meanwhile, the main causal inference results are detailed on the screen. Here is a case example:

```
2024-11-19 21:15:04 Clustering (2 categories) phi: [0.29485845139124117,0.3508103656075827]
2024-11-19 21:15:04 Heterogeneity test by P_plei-test for correlated horizontal pleiotropy: 0.86885
2024-11-19 21:16:17 Correlated horizontal pleiotropy may be absent (P_plei-tes >= 0.20), and the estimate causal effe
ct is:

- By the one-category model of PCMR:
  - The causal effect(SE): 0.323(0.0523); OR: 1.38(1.25-1.53)
   - PCMR's causality evaluation p-value: 6.56e-10

- By Inverse-Variance Weighted MedianMR:
  - The causal effect(SE): 0.321(0.0634); OR: 1.38(1.22-1.56)
  - Median-based causality evaluation p-value: 4.03e-07
```

According to PCMR's heterogeneity test, it indicates the failure to reject the null hypothesis of no correlated pleiotropy ($P_{plei-tes} >= 0.20$). The causality may be more appropriately inferred by the one-category model of PCMR and conventional inverse-variance weighted median MR.

In addition, the summary statistics of IVs for MR are saved in the file named variants.hg38.tsv.gz under the subdirectory of PCMRTask. There are thirteen columns in the file:

Header	Description
CHROM	Chromosome of the gene
POS	The coordinate of the IV with the lowest GWAS p-value
REF	The reference sequence base
ALT	The alternative sequence base
MarkFeatureGene	The Gene annotated with the SNP
MarkGeneFeature	The feature of the gene annotated with the SNP
[exposure]::P	The P value of this SNP on exposure
[exposure]::Beta	The effect of this SNP on exposure
[exposure]::SE	The effect's standard error of this SNP on exposure
[outcome]::P	The P value of this SNP on the outcome
[outcome]::Beta	The effect of this SNP on the outcome
[outcome]::SE	The effect's standard error of this SNP on the outcome
Class	The category given by PCMR, may be 1, 2, 3, etc.

A graphical result file is also presented as IVScatterPlots.pdf in the same directory. The horizontal axis of the graph represents the effect of SNPs on the exposure variable, while the vertical axis represents the effect of SNPs on the outcome variable. Each point signifies an SNP selected by PCMR, along with the confidence interval of its corresponding effect size. Different colors are used to distinguish between different types of points identified by PCMR. The slope of the diagonal line with the same color represents the effect of the exposure on the outcome by the corresponding category of SNPs.

There are also some intermediate output files in the directory Actinobacteria_AN_pcmr. A brief introduction is provided in the following table.

File	Description
ConvertVCF2GTBTask\EUR.hg19.gtb	the gtb format file of input VCF file
GenerateRootVariantSetTask\variants.annot.hg38.gtb	An hg38 file in gtb format with genotype data removed and annotation information added
IncorporateVariants2RootVariantSetTask\{exposure_file}.gtb	the gtb format file of the exposure sum file. By default, all coordinates are converted to hg38.
IncorporateVariants2RootVariantSetTask\{outcome_file}.gtb	the gtb format file of the outcome sum file. By default, all coordinates are converted to hg38.
IncorporateVariants2RootVariantSetTask\variants.annot.hg38.gtb	The gtb format file of the SNPS selected by P-value in the exposure and outcome files
GeneFeatureAnnotationTask\variants.annot.hg38.gtb	variants.annot.hg38.gtb in IncorporateVariants2RootVariantSetTask with annotation
LDGeneStickingTask\variants.annot.hg38.gtb	The loci annotated to the nearest gene region
LDPruningTask\variants.annot.hg38.5.0E-8.gtb variants.annot.hg38.0.5.gtb	The files containing the retained SNPs after LD clumping.

Microbes

Infer causality from microbes to phenotypes by MR methods

We provide GWAS summary statistics of microbes to enable users to identify causal microbes associated with phenotypes using Mendelian Randomization (MR) methods. The advanced MR method, PCMR, which is more robust to correlated horizontal pleiotropy (see details above), is applied as the primary analysis. Simultaneously, the presence of correlated horizontal pleiotropy is assessed. If no significant correlated horizontal pleiotropy is detected, a conventional IVW-based MR method is subsequently performed to evaluate the significance of the estimated causal effects.

Citations

[1] Bin Tang, Nan Lin, Junhao Liang, Guorong Yi, Liubin Zhang, Wenjie Peng, Chao Xue, Hui Jiang, Miaoxin Li. Leveraging Pleiotropic Clustering to Address High Proportion Correlated Horizontal Pleiotropy in Mendelian Randomization Studies. Nat Commun. 2025 Mar 21;16(1):2817 PubMed

Options

This main analysis inputs a GWAS summary of SNPs and outputs p-values of genes. The following are options for an example:

```
java -Xmx4g -jar ../kggsum.jar \
  causal \
   --pcmr 1T2 \
          effIVPCut=1E-3 \
   --sum-file './microbiome/mibiogen/k__*.tsv.gz' \
              cp12Cols=CHR,BP,A1,A2 \
              pbsCols=P,BETA,SE \
              sep=TAB \
```

```
              betaType=0 \
    --sum-file ./scz_gwas_eur_chr1.tsv.gz \
              cp12Cols=CHR,BP,A1,A2 \
              pbsCols=P,OR,SE \
              betaType=2 \
              prevalence=0.01 \
    --ref-gty-file ./1kg_hg19_eur_chr1.vcf.gz \
    --threads 10 \
    --output ./test/microb_scz_casual \
    --exclude-complementary-allele
```

Format	Description	Default
`--pcmr`	Triggers the PCMR analysis. This is a parameter set with multiple sub-options, as described above. **Example**: `--pcmr 1T2,2T1 5E-5 fixed 0.1 0.5 0.8`	causalPair=1T2,2T1 effIVPCut=5E-8 effIVPCorrect=fixed ldPruneCut=0.1 initIVPCut=0.5 ldStickCut=0.8
..	...	-

Output

The output results of analyses are also the same as that of the phenotype causality.

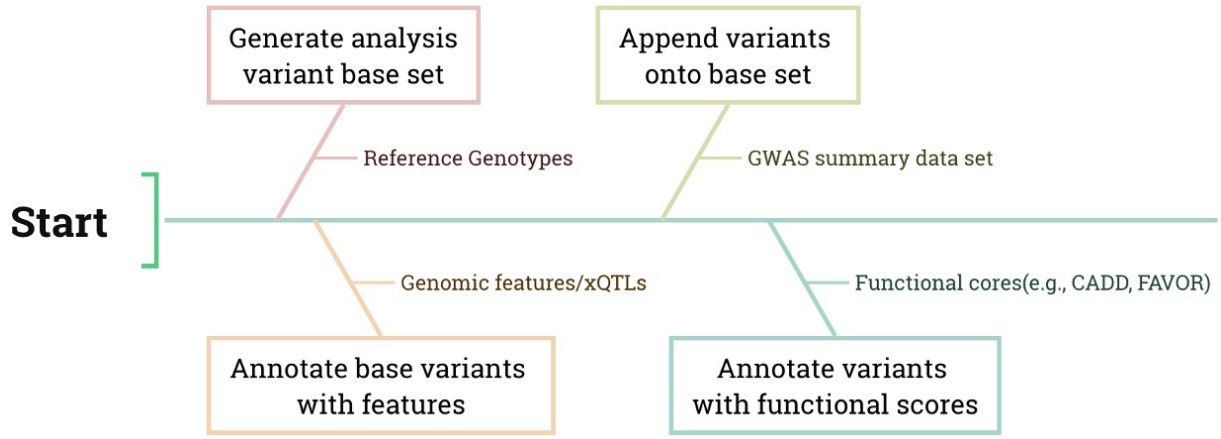
Variant Annotation & Filtration

About

The `annotate` module enables rapid annotation of millions of variants with genomic features using one or multiple different databases, leveraging the full or partial fields of these databases. Additionally, it offers a variety of filtering functions based on annotation results, such as gene feature filtering, population frequency, conservation and epigenetic modification, to assist in interpreting and deciphering the significant association signals.

Workflow of the Annotate Module

1. **Generation**: Extract variant coordinates and frequencies from the VCF or GBC file to create a root variant set for further analysis.
2. **Gene Annotation**: Annotate the root variant with gene features or xQTLs.
3. **Append**: Integrate GWAS variants and their summary statistics into the annotated root variant set.
4. **Variant Annotation**： Annotating variants with databases (e.g., CADD - Combined Annotation Dependent Depletion, and gnomAD) stored in GTB format to gain comprehensive insights into your genetic variations. KGGSum allows for rapid, one-stop annotation of hundreds of fields from one or multiple databases.



Basic Usage

```
java -jar kggsum.jar annot --sum-file <input1> --ref-gty-file <input2> --output <output> [options]
```

Variant Annotation and Filtration with External Databases enhance the analysis of genetic variations by leveraging external databases to provide additional information about variants. This process aids in filtering and prioritizing variants based on various annotations.

Gene

Gene feature Annotation & Filtration

Gene feature annotation is used to identify which gene or transcript is affected by a variant and what functional role it has on known genes. We now support two gene definition systems: RefSeq genes (refgene) and GENCODE genes (gencode).

Options

The tutorial command is:

```
java -Xmx8g -jar ../kggsum.jar \
   annot \
   --sum-file ./scz_gwas_eur_chr1.tsv.gz \
              cp12Cols=CHR,BP,A1,A2 \
              pbsCols=P,OR,SE \
              betaType=2 \
              prevalence=0.01 \
   --ref-gty-file ./1kg_hg19_eur_chr1.vcf.gz \
              refG=hg19 \
   --gene-model-database refgene \
   --gene-model-database gencode \
   --threads 18 \
   --output ./ta1
```

Flag	Description	Default
`annot`	Trigger the annotation procedure.	-
`--gene-model-database`	et the reference gene model database(s) used for gene feature annotation. It is a combination of parameters. The `<name>` is identified as the database name (such as refgene or gencode). The `<path>` identifies the path to the database, which can be a local file path or an FTP/HTTP file path. If the path of the resources folder has been specified, the `<path>` parameter can be bypassed. Note This option conflicts with `--xqtl-file` . Format: `--gene-model-database <name> path=<path>` Example: `--gene-model-database refgene`	-
`--upstream-distance`	Set the region length (bp) of upstream from the transcription start site. Format: `--upstream-distance <int>` Example: `--upstream-distance 1000` Valid setting: [int] >=1	1000
`--downstream-distance`	Set the region length (bp) of downstream from the transcription start site. Format: `--downstream-distance <int>` Example: `--downstream-distance 1000` Valid setting: [int] >=1	1000
`--disable-gene-feature`	Disable gene feature annotation. Format: `--disable-gene-feature`	
`--gene-feature-in`	Retain variants with specified annotated gene features . Format: `--gene-feature-in <int~int>,<int>,...` Example: `--gene-feature-in 0~6,9,10` Valid setting: [int] 0 ~ 17	0 ~ 17
...	...	...

KGGA has 18 number codes for the gene features after annotation.

Feature	Code	Explanation
Frameshift	0	Short insertion or deletion results in a completely different translation from the original.

Nonframeshift	1	Short insertion or deletion results in loss of amino acids in the translated proteins.
Start-loss	2	Indels or nucleotide substitution results in the loss of the start codon (ATG) (mutated into a non-start codon).
Stop-loss	3	Indels or nucleotide substitution results in the loss of stop codons (TAG, TAA, TGA).
Stop-gain	4	Indels or nucleotide substitution result in the new stop codons (TAG, TAA, TGA), which may truncate the protein.
Splicing	5	Variant is within 3-bp of a splicing junction (use `--splicing-distance x` to change this; the unit of x is base-pair).
Missense	6	Nucleotide substitution results in a codon coding for a different amino acid.
Synonymous	7	Nucleotide substitution does not change amino acids.
Exonic	8	Due to the loss of sequences in the reference database, this variant can only be mapped into the exonic region without more precise annotation.
UTR5	9	Within a 5' untranslated region.
UTR3	10	Within a 3' untranslated region.
Intronic	11	Within an intron.
Upstream	12	Within 1-kb region upstream of transcription start site (use `--upstream-distance x` to change this, the unit of x is base-pair).
Downstream	13	Within 1-kb region downstream of the transcription end site (use `--downstream-distance x to change this; the unit of x is base-pair).
ncRNA	14	Within a transcript without protein-coding annotation in the gene definition.
Intergenic	15	Variant is in intergenic region.
Monomorphic	16	It is not a sequence variation, which may result from bugs in the reference genome in variant calling.
Unknown	17	Variants has no annotation.

Output

The gene feature annotation results are saved in `OutputVariants2TSVTask/variants.hg38.tsv.gz`. There are two relevant columns in the file:

Header	Description
...	...
MarkFeatureGene	The gene where a variant is located. When a variant is mapped onto multiple genes, the genes led to the smallest code is called the mark gene.
MarkGeneFeature	The coordinate of the first SNP.
...	..

Function

Functional score annotation at variants

In KGGSum, the GWAS variants can also be annotated with multiple genomic features. Three databases are available: gnomAD for allele frequency annotation, CADD for variant function annotation, and ClinVar for disease linkage annotation. Note that the annotation datasets should be downloaded from an independent resource domain of KGGA.

Database Name	Short Description	Tag

CADD	Combined Annotation Dependent Depletion (CADD) is a widely used matrix for mutation deleteriousness and integrates more than 100 annotations for all possible single-nucleotide variants (SNVs) of the GRCh38/hg38 human reference genome.	cadd
Favor	Functional Annotation of Variants - Online Resource (FAVOR) provides comprehensive multi-faceted variant functional annotations that summarize findings of all possible nine billion SNVs across the genome (build GRCh38).	favor
ClinVar	ClinVar is a public database managed by the National Center for Biotechnology Information (NCBI) that provides information about the relationship between genetic variation and human health.	clinvar

Options

The tutorial command is:

```
java -Xmx8g -jar ../kggsum.jar \
   annot \
   --sum-file ./scz_gwas_eur_chr1.tsv.gz \
              cp12Cols=CHR,BP,A1,A2 \
              pbsCols=P,OR,SE \
              betaType=2 \
              prevalence=0.01 \
   --ref-gty-file ./1kg_hg19_eur_chr1.vcf.gz \
              refG=hg19 \
   --variant-annotation-database cadd \
              field=Epigenetics::EncodeDNase-max,Epigenetics::EncodeDNase-sum,ProteinFunction::CADD_PHRED \
   --threads 18 \
   --output ./ta2
```

Annotation option	Description	Default
`--variant-annotation-database`	Set the reference databases used for annotation at variants. `--variant-annotation-database` is a combination of parameters, and the usage rules are the same as those of `--freq-database`. Format: `--variant-annotation-database <name> path=<path> field=[field1,field2,...]` Example: `--variant-annotation-database cadd field=ProteinFunction::CADD_PHRED`	[OFF]

Additional epigenetic resources from third-party databases

To facilitate the convenient use of more resources, KGGSum provides an interactive approach that allows users to specify customized third-party resources for annotation. For example, by setting the file name documented in EpiMap, KGGSum can directly download epigenetic marker resources from the EpiGenome public domain, specifically from the EpiMap Repository.

Database Name	Short Description
EpiMap	EpiMap is one of the most comprehensive maps of the human epigenome, provides approximately 15,000 datasets across 833 bio-samples and 18 epigenomic marks, delivers a rich of gene-regulatory annotations encompassing chromatin states, high-resolution enhancers, activity patterns, enhancer modules, upstream regulators, and downstream target genes.

Annotation Option	Description	Default
--region-annotation-database	Specifies the interval/regional databases for annotation. **Format**: --region-annotation-database subID=[] marker=[] path= field=[field1,field2,...] **Example**: --region-annotation-database EpiMap subID=BSS00001 marker=H3K4me3 - name: Database name (e.g., EpiMap). - subID: Subject ID from EpiMap. - marker: Epigenetic marker (e.g., H3K4me3). - path: Path to the database file (optional for EpiMap). - field: Specific fields to include. **Note**: For name=EpiMap, KGGA automatically downloads data from the EpiMap Repository if not locally available.	

Frequency

Allele Frequency Annotation & Filtration

Allele frequency annotation allows users to incorporate population-level allele frequency information into the analysis of genetic variants. Click to view the provided allele frequency annotation databases.

Database Name	Short Description	Tag
gnomAD	Allele frequency data in the Genome Aggregation Database (gnomAD) v4 dataset (GRCh38) is derived from 730,947 exomes and 76,215 genomes from unrelated individuals of diverse ancestries.	gnomad

Options

The tutorial command is:

```
java -Xmx4g -jar ../kggsum.jar \
   annot \
   --sum-file ./scz_gwas_eur_chr1.tsv.gz \
              cp12Cols=CHR,BP,A1,A2 \
              pbsCols=P,OR,SE \
              betaType=2 \
              prevalence=0.01 \
   --ref-gty-file ./1kg_hg19_eur_chr1.vcf.gz \
              refG=hg19 \
   --freq-database gnomad \
              field=gnomAD_joint::ALL,gnomAD_joint::NFE \
   --threads 18 \
   --output ./ta3
```

Annotation option	Description	Default
`--freq-database`	Set the reference databases used for allele frequency annotation. `--freq-database` is a combination of parameters. The `<name>` is identified as the database name (such as gnomad). The `<path>` identifies the path to the database, which can be a local file path or an FTP/HTTP file path. If the path of the resources folder has been specified, the `<path>` parameter can be bypassed. The `<field>` is identified as the specified field filtered under this database. If no value is set, all fields of the specified database are selected by default. Format: `--freq-database <name> path=<path> field=[field1,field2,...]` Example: `--freq-database gnomad field=gnomAD::EAS`	[OFF]

Once the reference databases for allele frequency annotation have been properly configured, you can effectively filter variants by examining their allele frequencies within the reference population.

Filtration option	Description	Default
`--db-af`	Exclude variants with alternative allele frequency (AF) outside the range [`min` , `max`] in allele frequency databases. Format: `--db-af <min>~<max>` Example: `--db-af 0.05~1.0` Valid setting: [float] 0.0 ~ 1.0	[OFF]
`--db-maf`	Exclude variants with minor allele frequency (MAF) outside the range [`min` , `max`] in allele frequency databases. Format: `--db-maf <min>~<max>` Example: `--db-maf 0.05~0.5` Valid setting: [float] 0.0 ~ 0.5	[OFF]