Joint Likelihood Mapping (JLIM) 2.5

JLIM is a cross-trait test for shared causal effect described in Chun et al. Nature Genetics 2017. JLIM tests whether two traits – primary and secondary – are driven by shared causal effect or not. Typically, the primary trait is a large GWAS study, and the secondary trait is an expression Quantitative Trait Loci (eQTL) association study. JLIM needs only summary-level association statistics for both primary and secondary traits but can also run on permutation data generated from individual-level genotype data. The latest version is 2.5.0.

Download

Source code 2.5.0 (tar.gz) released May 11, 2021.

Installation

1. JLIM core module is implemented as an R extension (**jlimR**). **R** (version 3.6 or newer) and **getopt** package are required to run JLIM. The **getopt** package can be installed by running:

```
Rscript -e 'install.packages("getopt", repos="http://cran.r-project.org")'
```

2. [**Optional**] JLIM provides a built-in support for the eQTL Catalogue data format. To use this feature, **seqminer** and **dplyr** packages are needed. Install these packages as follows:

```
Rscript -e 'install.packages(c("seqminer", "dplyr"),
repos="http://cran.r-project.org")'
```

3. [**Optional**] To read GTEx parquet files, JLIM uses **arrow** package with snappy codec. This package is necessary only if GTEx parquet files are used. The **arrow** package can be installed as below (See here for the full installation instruction):

```
Rscript -e 'Sys.setenv(LIBARROW_BINARY="false"); Sys.setenv(NOT_CRAN="true");
install.packages("arrow", repos="http://cran.r-project.org")'
```

4. After this, **jlimR** (included in the distribution file) can be installed by:

```
R CMD INSTALL jlimR_2.5.0.tar.gz
```

5. [**Optional**] We provide an optional Python module to generate permutation data for meta-analyzed secondary trait cohorts. This feature is not needed for typical eQTL analyses. To use this functionality, the following packages are additionally required:

- **Python** (2.7.9 or newer)
- **numpy** (1.14.3 or newer)
- **scipy** (1.0.0 or newer)

6. We provide pre-processed reference genotype panels to go with JLIM. Reference genotype panels should be downloaded separately and unpacked before running JLIM:

- 1000GP Non-Finnish Europeans (NFE) b37 b38
- 1000GP Finnish (FIN) b37 b38
- 1000GP East Asians (EAS) b37 b38
- 1000GP Admixed Americans (AMR) b37 b38
- 1000GP Africans (AFR) b37 b38

Change Log

Changes to 2.5

- Support downloadable reference genotype panels
- Estimate the p-value by resampling of reference genotypes (no individual-level genotype data required)
- Built-in eQTL Catalogue support
- Built-in GTEx parquet file support
- Usability enhancement
- Discontinued support for the batch run mode (.cfg files)

Changes to 2.0

- Support permutation of a meta-analyzed secondary trait cohort. This allows to directly account for the subcohort-level variation in data processing and QC in meta-analyzed secondary trait data.

How to Run JLIM

Example data for eQTL analysis

We provide Multiple Sclerosis (MS) GWAS data and Lymphoblastoid Cell Lines (LCL) eQTL data for two loci (chr1:160,697,074-160,933,065 and chr12:57,626,582-58,488,667) in the `examples/` directory. The MS GWAS data are from IMSGC. Nature Genetics 2013, and the LCL eQTL data are from Lappalainen et al. Nature 2013 (Geuvadis; non-Finnish Caucasian samples only).

To test if the MS GWAS peak colocalizes with SLAMF7 eQTL in LCL, run JLIM as follows:

```
run_jlim.sh --maintr-file examples/MS/MS.1.160697074.160933065.txt \
--sectr-file
examples/LCL/locus.1.160697074.160933065/LCL.SLAMF7_ENSG00000026751.11.assoc.l
inear.gz \
--ref-ld refld.1kg.nfe.b37/ \
--index-snp 1:160711804 \
--output-file examples/MS-Geuvadis_LCL_SLAMF7.out \
--sectr-sample-size 278
```

The command line arguments are as below:

- `–maintr-file`: primary trait association file (summary statistics)
- `–sectr-file`: secondary trait association file (summary statistics)
- `–ref-ld`: reference genotype panel (separately downloaded and unpacked)
- `–index-snp`: coordinate of the lead SNP for primary trait ([CHR]:[BP])
- `–output-file`: output file path
- `–sectr-sample-size`: sample size of secondary trait cohort

By default, a 200-kb analysis window is set up centered at the specified index SNP. In the analysis window, JLIM uses only the SNPs that are shared across all data: namely, primary and secondary trait association files and reference genotype panel. InDels, rare variants (MAF < 5%), tri-allelic SNPs and duplicated SNPs are excluded from the analysis. All input data files should be mapped to the same genome assembly.

Output:

JLIM generates an output file as below (`examples/MS-Geuvadis.LCL.SLAMF7.out`). The full details on the output columns can be found here. Note that JLIM reports an empirical **pvalue** of **0** when a sampled null distribution does not produce a statistic as or more extreme than the actual statistic. In such a case, the expectation of true p-value is less than 1 over **executedPerm** (the total count of permutation/sampling), namely, 1/10,000 in this example. The total number of sampling/permutation iterations can be controlled by the ''--max-perm''.

userIdxBP	actualIdxBP	STAT	pvalue	usedSNPsNo	startBP	endBP	sectrSampleSize	sectrGeneName	executedPerm	desc
160711804	160711804	3.75229222129719	0	106	160698276	160807409	278	NoGeneName	10000	executed

eQTL Catalogue support:

A user can run JLIM against remote datasets available in eQTL Catalogue using RESTful API. For this, we provide the option `–sectr-ref-db eQTLCatalogue:remote`. The secondary trait association file (`–sectr-file`) should be an eQTL Catalogue dataset name: e.g. **GENCORD_ge_LCL** for **GENCORD_ge_LCL.all.tsv.gz**. The full list of available datasets in eQTL Catalogue can be found here (See the **ftp_path** column). JLIM will test for the colocalization against all genes for which cis-eQTL association data are available. With the `–sectr-ref-db` option, the sample size of eQTL dataset will be automatically loaded into the test. We also support running JLIM on a local download of eQTL Catalogue

for heavy usage (See ''--sectr-db eQTLCatalogue'').

The eQTL Catalogue is based on the reference genome b38, thus b38 reference genotype panel should be used (e.g. `–ref-ld refld.1kg.nfe.b38/`). As an example, we provide an MS GWAS file lifted over from b37 to b38 (`examples/MS/MS.1.160697074.160933065_38.txt`). Note that some datasets in eQTL Catalogue are not from the non-Finnish European population. A user is responsible for assigning an appropriate reference genotype panel for such studies.

```
run_jlim.sh --maintr-file examples/MS/MS.1.160697074.160933065_38.txt \
--sectr-file GENCORD_ge_LCL \
--ref-ld refld.1kg.nfe.b38/ \
--index-snp 1:160742014 \
--output-file examples/MS-GENCORD_ge_LCL.out \
--sectr-ref-db eQTLCatalogue:remote
```

This will produce the following results (See `examples/MS-GENCORD_ge_LCL.out` for the full output). The **sectrGeneName** of ENSG00000026751 corresponds to SLAMF7. By default, JLIM does not run a colocalization test unless a gene has at least one SNP with the association p-value < 0.05 in the analysis window. This is because there is not enough evidence for any genetic association to a secondary trait (e.g. ENSG00000066294 below), thus there is no need to test for colocalization of causative genetic effects between traits. This minimum p-value threshold can be lowered by using the option ''--min-pvalue''.

userIdxBP	actualIdxBP	STAT	pvalue	usedSNPsNo	startBP	endBP	sectrSampleSize	sectrGeneName	executedPerm	desc
160742014	160742014	4.01572683513747	0	115	160728486	160840829	190	ENSG00000026751	10000	executed
160742014	160742014	NA	NA	91	160728486	160830769	190	ENSG00000066294	0	Second trait min pvalue is higher than threshold: 0.05
160742014	160742014	-1.92857440859272	0.618	92	160728486	160840829	190	ENSG00000117090	10000	executed
160742014	160742014	-1.22821917489519	0.3512	89	160728486	160830769	190	ENSG00000117091	10000	executed
…	…	…	…	…	…	…	…	…	…	…
160742014	NA	NA	NA	85	160766482	160840829	190	ENSG00000118217	0	index SNP has been filtered out.
160742014	NA	NA	NA	85	160766482	160840829	190	ENSG00000226889	0	index SNP has been filtered out.

In the above example, for some genes (e.g. ENSG00000118217 and ENSG00000226889), JLIM reports that "**index SNP has been filtered out.**" This is because, in the eQTL Catalogue, the associations were tested only for SNPs within +/- 1 Mb from the transcription start site of each gene. Therefore, for genes that are ~1 Mb away from the index SNP, eQTL association data are available only for a partial segment of the analysis window (160,766,482-160,840,829 in this case). JLIM automatically skips if the index SNP is not included in the analysis window. *Users are responsible to additionally check if enough SNPs are included in the analysis window on both sides of index SNP by examining the **startBP**-**endBP** in the output file.*

GTEx v8 support:

JLIM supports the GTEx v8 parquet data file format. The GTEx v8 summary statistics are available for download at GTEx Google Bucket. Specifically, we recommend using the European-American summary statistics (available in the `GTEx_Analysis_v8_EUR_eQTL_all_associations` subdirectory in the Google Bucket) instead of the full trans-ethnic association statistics. For the European-American GTEx v8 data, the sample size of eQTL data in each tissue can be automatically loaded with the option `–sectr-`

`ref-db GTEx.v8.EUR`, and `refld.1kg.nfe.b38` is recommended as a reference genotype panel. For other GTEx data, `–sectr-ref-db` option is currently not available. A user is responsible to provide the correct sample size of each tissue and matching reference LD panel.

```
run_jlim.sh --maintr-file examples/MS/MS.1.160697074.160933065_38.txt \
--sectr-file
your_download_path/GTEx_Analysis_v8_QTLs_GTEx_Analysis_v8_EUR_eQTL_all_associa
tions_Cells_EBV-transformed_lymphocytes.v8.EUR.allpairs.chr1.parquet \
--ref-ld refld.1kg.nfe.b38/ \
--index-snp 1:160742014 \
--output-file examples/MS-GTEx_v8_EUR_LCL.out \
--sectr-ref-db GTEx.v8.EUR
```

Backward compatibility to JLIM 1.0 and 2.0:

Previous versions of JLIM required a pre-computed permutation file to run JLIM. We still support this mode to allow users to have full control of permutation with the `–perm-file` option. The `–ref-ld` option supports the downloadable reference panels new to v2.5 as well as reference LD files prepared by `fetch.refld0.EUR.pl` script in previous JLIM versions. If the `–maintr-ld` or `–sectr-ld` option is omitted, the reference LD will be assumed. See the backward compatibility section for further details on how to prepare input data. The following set of command line options will replicate the behaviors of previous versions of JLIM:

```
run_jlim.sh --maintr-file examples/MS/MS.1.160697074.160933065.txt \
--sectr-file
examples/LCL/locus.1.160697074.160933065/LCL.SLAMF7_ENSG00000026751.11.assoc.l
inear.gz \
--ref-ld refld.1kg.nfe.b37/ \
--sectr-ld examples/LCL/locus.1.160697074.160933065/LCL.ped.gz \
--perm-file
examples/LCL/locus.1.160697074.160933065/LCL.SLAMF7_ENSG00000026751.11.mperm.d
ump.all.gz \
--index-snp 1:160711804 --output-file examples/MS-Geuvadis_LCL_SLAMF7-perm.out
```

Example data for meta-analyzed secondary trait cohort (JLIM 2.0)

JLIM 2.0 provides python scripts to generate permutation data for meta-analyzed secondary trait cohort. This allows to directly account for the subcohort-level variation in data processing and QC in meta-analyzed secondary trait data. For the details, see JLIM 2.0 examples.

Input File Format

Association statistics file

The association statistics files for primary and secondary traits should be either a Plink assoc file or space/tab-delimited table. For the latter, JLIM expects the columns for **CHR** (chromosome), **BP** (base pair position), and association statistics. The association statistics can be **STAT**, **T**, or **P**. If the association statistics files use column names which are different from the above default, the column names can be specified using the command line parameters without changing the original files.

Output File Format

By default, JLIM tests for the colocalization against all genes found in the secondary trait file. Each row in the output file corresponds to a test against each gene. The name of output file is specified by `–output-file` option. The columns in the JLIM output files have the following information:

- **userIdxBP**: base pair position of the index SNP provided by user
- **actualIdxBP**: base pair position of the SNP that is most associated to the primary trait; automatically selected in the analysis window by JLIM
- **STAT**: JLIM statistic
- **pvalue**: estimated JLIM p-value
- **usedSNPsNo**: total number of SNPs included in the analysis after all filtering
- **startBP**: position of first SNP in the analysis window; by default, up to -100 kb from **userIdxBP**
- **endBP**: position of last SNP in analysis window; by default, up to +100 kb from **userIdxBP**
- **sectrSampleSize**: sample size of secondary trait association study
- **sectrGeneName**: gene name if the secondary trait association file contains multiple genes
- **executedPerm**: actual number of permutations/sampling iterations performed to calculate p-values
- **desc**: JLIM status
    - *executed* – completed without an error;
    - *too few common SNPs to run JLIM* – not tested because there are too few SNPs in the analysis window;
    - *index SNP has been filtered out* – not tested because the index SNP has been filtered out (e.g. InDels, tri-allelic, or low MAF);
    - *second trait min pvalue is higher than threshold* – not tested since the association to secondary trait is too weak

Full List of Command Line Options

Primary trait association statistics file:

- `–maintr-file <FILE>`: primary trait association statistics file (required)

- `–maintr-colname-chr <COLUMN NAME>`: column name for chromosomes (default: **CHR**)
- `–maintr-colname-bp <COLUMN NAME>`: column name for base pair positions (default: **BP**)
- `–maintr-colname-p <COLUMN NAME>`: column name for association p-values (default: **P**)

Secondary trait association statistics file:

- `–sectr-file <FILE>`: second trait association statistics file (required)
- `–sectr-colname-chr <COLUMN NAME>`: column name for chromosomes (default: **CHR**)
- `–sectr-colname-gene <COLUMN NAME>`: column name for gene IDs if the secondary trait file contains multiple genes or traits (default: assume that the secondary trait file contains a single gene/trait)
- `–sectr-colname-bp <COLUMN NAME>`: column name for base pair positions (default: **BP**)
- `–sectr-colname-p <COLUMN NAME>`: column name for association p-values (default: **P**)
- `–sectr-colname-variant-id <COLUMN NAME>`: column name for variant ids in the format of chr4_87424296_G_C_b38; if specified, the chromsome and base pair position will be extracted from this column (default: none)
- `–sectr-gene-filter <GENE ID>`: test only against the specified gene (default: test all genes)

LD and reference genotype panels:

- `–ref-ld <DIR/FILE>`: directory to the unpacked reference genotype panel; alternatively, a gzipped or plain-text reference genotype file prepared with `fetch.refld0.EUR.pl` for backward compatibility (required unless both `–maintr-ld-file` and `–sectr-ld-file` are specified)
- `–maintr-ld-file <FILE>`: LD structure of primary trait cohort if it is different from the reference panel; FILE should be a gzipped or plain-text genotype file in the Plink .ped format and with the identical set and order of SNPs as in the primary trait association file (default: if unspecified, the reference LD will be assumed)
- `–sectr-ld-file <FILE>`: LD structure of secondary trait cohort if it is different from the reference panel; FILE should be a gzipped or plain-text genotype file in the Plink .ped format and with the identical set and order of SNPs as in the secondary trait association file (default: if unspecified, the reference LD will be assumed)

JLIM analysis window:

- `–index-snp <CHR>:<POS>`: chromosome and base pair position of index SNP (required)
- `–window-size <NUMBER>`: size of analysis window, centered at the user-specified index SNP (default: 200,000 bp)
- `–manual-window-boundary <STARTPOS>-<ENDPOS>`: Override the default placement of analysis window with STARTPOS to ENDPOS (default: +/- 100kb from the index SNP)

P-value estimation:

- `--sectr-sample-size <NUMBER>`: sample size of secondary trait cohort (required unless `--sectr-ref-db` is used)
- `--max-perm <NUMBER>`: maximum number of adaptive sampling/permutation runs (default: 10,000 or the size of permutation file)
- `--perm-file <FILE>`: calculate p-values using pre-permuted secondary trait association data as in JLIM 1.0 and 2.0; FILE should be a gzipped Plink `.mperm.dump.all` file (default: p-value is calculated indirectly by sampling from a reference genotype panel)

GTEx and eQTL Catalogue support:

- `--sectr-ref-db <PRESET CONFIGURATION>`: preset configuration for GTEx and eQTL Catalogue. If specified, the data format of the secondary trait association file is directly recognized, and the sample size of the secondary trait cohort will be automatically loaded. Currently, we support the following three CONFIGURATIONs:
  - `--sectr-ref-db GTEx.v8.EUR` This is for GTEX v8 summary statistics calculated using only European American individuals (genome assembly b38). The data files (.parquet) have to be downloaded from GTEx Google Bucket and specified by `--sectr-file` argument. JLIM will automatically run with `--sectr-colname-variant-id variant_id --sectr-colname-gene phenotype_id --sectr-colname-p pval_nominal`, and the `--sectr-sample-size` option will be set to the sample size of corresponding tissue.
  - `--sectr-ref-db eQTLCatalogue` This is for downloaded summary statistic files from eQTL Catalogue. The file path to tab-delimited flat files (.tsv) should be passed by `--sectr-file`. JLIM expects the corresponding tabix index files (.tbi) to be present in the same directory with the .tsv file. JLIM will automatically run with `--sectr-colname-chr chromosome --sectr-colname-bp position --sectr-colname-gene molecular_trait_id --sectr-colname-p pvalue`, and the `--sectr-sample-size` option will be set to the sample size of corresponding study. Note that some datasets in eQTL Catalogue are multi-ethnic association data, for which we do not provide a downloadable reference genotype panel to match the LD structure. The eQTL Catalogue data are based on the genome assembly b38.
  - `--sectr-ref-db eQTLCatalogue:remote` This is similar to `--sectr-ref-db eQTLCatalogue`, but summary statistics do not have to be locally downloaded. The slice of eQTL data corresponding to the analysis window will be retrieved from the eQTL Catalogue server using the RESTful API. For the file name in the `--sectr-file` option, please use the dataset name in the ftp_path without the file extention "**.all.tsv.gz**" (e.g. "**Alasoo_2018_ge_macrophage_naive**").

SNP filtering:

- `--min-MAF <NUMBER>`: minimum minor allele frequency for a variant; rare variants below this threshold will be excluded (default: 0.05)

- `–min-pvalue <NUMBER>`: run JLIM only if in the secondary trait file, there exists a SNP with the p-value of association lower than this limit (default: 0.05)

Other options:

- `–output-file <FILE>`: output file name (required)
- `–min-SNPs-count <NUMBER>`: run JLIM only if there exist a sufficient number of SNPs available in the analysis window (default: 50 SNPs)
- `–r2-resolution <NUMBER>`: genetic resolution of JLIM test in r^2 (default: r^2=0.8)

Citations

JLIM 1.0

> Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types.
> Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, Sunyaev SR, Cotsapas C.
> Nat Genet. 2017 Apr;49(4):600-605. doi: 10.1038/ng.3795.

JLIM 2.0

> Leveraging pleiotropy to discover and interpret GWAS results for sleep-associated traits.
> Chun S*, Akle S*, Teodosiadis A, Cade BE, Wang H, Sofer T, Evans DS, Stone KL, Gharib SA, Mukherjee S, Palmer LJ, Hillman D, Rotter JI, Hanis CL, Stamatoyannopoulos JA, Redline S, Cotsapas C*, Sunyaev SR*.
> BioRxiv. 2022. doi: 10.1101/832162 (preprint).
> * equal contribution.