LocusFocus Documentation

Release 1.4.9 alpha

Naim Panjwani

Apr 24, 2021

Table of Contents:

1 Quick Start **3**
- 1.1 Selecting the human coordinate system . . . 3
- 1.2 Primary dataset input . . . 4
  - 1.2.1 Simple Sum Colocalization . . . 4
  - 1.2.2 COLOC2 Colocalization . . . 5
  - 1.2.3 Formatting the primary dataset file input . . . 5
  - 1.2.4 Non-default column names . . . 5
- 1.3 Selecting an LD matrix . . . 5
  - 1.3.1 Selecting a publicly available 1000 Genomes population LD matrix . . . 7
  - 1.3.2 Computing the LD matrix from your GWAS population . . . 9
- 1.4 Secondary datasets . . . 9
  - 1.4.1 Selecting GTEx tissues as secondary datasets . . . 10
  - 1.4.2 Formatting custom secondary datasets . . . 10
- 1.5 Some important points to consider . . . 12

2 Retrieving a Previous Session **15**

3 Example Run, Usage and Interpretation **17**
- 3.1 Sample Data and Output . . . 17
- 3.2 Selecting LD (Linkage Disequilibrium) . . . 18
- 3.3 Selecting Secondary Datasets from GTEx . . . 19
- 3.4 Selecting Genes of Interest for Colocalization Testing . . . 19
- 3.5 Overriding the First-Stage Set-Based P-value Threshold . . . 19
- 3.6 Submit . . . 21
- 3.7 Saving and Retrieving Your Session . . . 21
- 3.8 Interpreting Data Output . . . 22
  - 3.8.1 Colocalization plot . . . 22
  - 3.8.2 Interpreting the Heatmap Plot . . . 23
  - 3.8.3 Simple Sum Table . . . 23
  - 3.8.4 COLOC2 Posterior Probability Results Table . . . 24

4 Versions **25**
- 4.1 Version History . . . 25
- 4.2 Datasets . . . 27
- 4.3 Programs . . . 27

5 Local installation of LocusFocus **29**

6 Contact 31

7 MIT License 33

8 Future Directions 35

LocusFocus is a web application to facilitate the exploration of a GWAS signal at a particular locus of the genome and its degree of colocalization with any SNP-level association data (e.g. expression quantitative trait loci for genes within +/- 1Mbp in the relevant GTEx tissues selected).

When paired with GTEx data, the aim is to annotate a GWAS (or region-based association) to the most probable gene(s) and tissue(s) that may be driving the observed GWAS signal.

In addition, users may upload other datasets to test colocalization with. For example, other phenotypic associations (i.e. PheWAS) may be uploaded for assessing pleiotropy, or eQTL data from other sources to obtain a formal colocalization test and visualization of the data.

The Simple Sum method is used for assessing the degree of colocalization of any two given datasets. When applied to GTEx, LocusFocus presents the degree of colocalization of genes nearby the GWAS association for all the tissues selected in an interactive heatmap plot.

COLOC2 colocalization testing is also available, and more colocalization methods may be made available in future version releases.

CHAPTER 1

Quick Start

This section provides a quick guide to preparing the required input files for colocalization analysis with LocusFocus.

Up to three files may be selected and uploaded, with a maximum combined file size of 100MB:

1. .txt or .tsv (required): tab-separated primary summary statistics (eg. GWAS)
2. .ld (optional): PLINK-generated LD matrix – must have the same number of SNPs as primary file
3. .html (optional): Secondary datasets to test colocalization with.
   - For uploaded secondary datasets, each data table must be preceded by an <h3> title tag describing the table
   - You may use the merge_and_convert_to_html.py script to ready your files for uploading
   - Press and hold the Ctrl/Cmd key to select multiple files

1.1 Selecting the human coordinate system

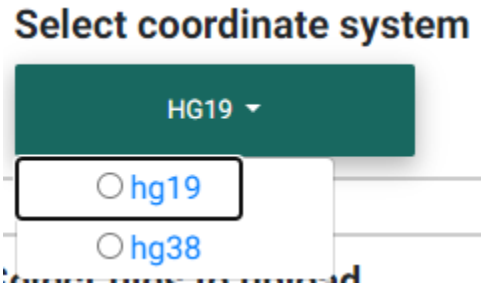


Fig. 1: Selecting the appropriate human genome coordinate system.

Before you begin, you must choose the appropriate human coordinate system for accurate visualization of the data in the colocalization plot, accurate matching with GTEx data (if selected), and accurate matching with the 1000 Genomes data for LD calculations (unless a custom matrix is uploaded). hg19 refers to the GRCh37 and hg38 to the GRCh38.

You may ignore this step if your purpose is to simply compute the colocalization between two datasets that you upload. In that case, simply ensure the secondary dataset(s) have been prepared using the merge_and_convert_to_html.py script or the merge_and_convert_to_html_coloc2.py script if also running COLOC2.

1.2 Primary dataset input

Please note that use of the web tool requires uploading your summary statistics to a public server.

The primary dataset will usually be a genome-wide association study (GWAS) file with summary statistics. You may download our sample dataset for meconium ileus around *SLC26A9* gene.

A tab-delimited text file input is preferred.

The first few lines of a sample input file are shown below.

Listing 1: *Example of a few lines of a tab-delimited input file*

```
> head MI_GWAS_2019_1_205500-206000kbp.tsv

#CHROM       BP       SNP      REF      ALT      BETA     SE       P        AF       N
1    205500022       rs114114637      C       A       -0.0739321001603956     0.
↪152350860711495       0.627481118153097       0.0260753323485968      6770
1    205500203       rs11240506       A       T       -0.028580509980184      0.
↪0720691836942088       0.691684232909143       0.123175036927622       6770
1    205500342       rs76069011       C       T       0.269353170260049       0.
↪246838504548712       0.275179551742949       0.0104350073855244      6770
1    205500450       rs45495498       C       G       -0.253245978881793      0.
↪252113050092583       0.31514069034813        0.0127828655834564      6770
1    205500543       rs11240507       T       C       0.0546597433178092      0.
↪0555106475800506       0.324785538604616       0.256992614475628       6770
1    205500586       rs146709978      C       G       0.184819472593451       0.
↪219156490990298       0.399048427185569       0.0155140324963072      6770
1    205500719       rs74142375       T       A       0.312872957911176       0.
↪227488579712488       0.169027675607645       0.013301329394387       6770
1    205500767       rs3838999        G       GA      -0.0717111663499279     0.
↪152168074740335       0.637453014437489       0.0259283604135894      6770
1    205500829       rs3795547        C       T       -0.00434473636812393    0.
↪12753158044846        0.972822985458176       0.0379985228951256      6770
```

1.2.1 Simple Sum Colocalization

For the Simple Sum method, two columns are absolutely necessary:

1. **ID** - The SNP rs id or chrom_pos_ref_alt_build format (e.g. 1_205860191_A_G_b37).
2. **P** - The association p-value

In this case, you must check the box "Use marker ID column to infer variant position and alleles" to map the provided rs ID's to the chromosomal position and alleles using dbSNP.

Alternatively, for more accurate matching of alleles, all four fields (chromosome, basepair position, reference and alternate alleles) may be provided in addition to the ID column field.

1. **ID** - The SNP rs id or chrom_pos_ref_alt_build format (e.g. 1_205860191_A_G_b37)
2. **#CHROM** - The chromosome column (either "X" or "23" is acceptable)
3. **POS** - The basepair position (in hg19 coordinates)

4. **REF** - The reference allele (on the plus strand, as defined in the 1000 Genomes)
5. **ALT** - The alternate allele (on the plus strand, as defined in the 1000 Genomes)
6. **P** - The association p-value

Use the web form to change the column names as found in your file if different from the default names. For the example above, you would change the position column default name (POS) to *BP* and MAF to *AF*.

1.2.2 COLOC2 Colocalization

To run COLOC2, check the box "Add required inputs for COLOC2". For this option, additional fields are required, and the corresponding required fields will be populated to allow for input of the column names if other than default:

1. **BETA** - The beta of the SNP in the association analysis
2. **SE** - The standard error of beta
3. **N** - The total sample size of the study
4. **MAF** - The minor allele frequency
5. **Study type** - Select either quantitative or case-control.

For the case-control case, you are required to provide the number of cases in the study.

If you have a csv or excel file, see below on *how to convert it to a tab-delimited file*.

We recommend having a dense set of SNPs for the region in order to obtain accurate results. If the number of SNPs in the region of interest is too small, it may not be possible to compute the Simple Sum p-value.

1.2.3 Formatting the primary dataset file input

The simplest way to convert either csv or excel file formats is to open the file in Excel, then choose to save the file as a tab-delimited file.

1.2.4 Non-default column names

Use the web form to change the column names as found in your file if different from the default names. For the example above, you would change the position column default name (POS) to *BP* and MAF to *AF*.

1.3 Selecting an LD matrix

You may either:

- *Select the appropriate 1000 Genomes population for your study from the dropdown*.
- *Compute the LD matrix from your population*.

For the most accurate colocalization statistics, we recommend uploading the LD matrix of your study. If this is unavailable, you may select the most appropriate 1000 Genomes population subset for your study.

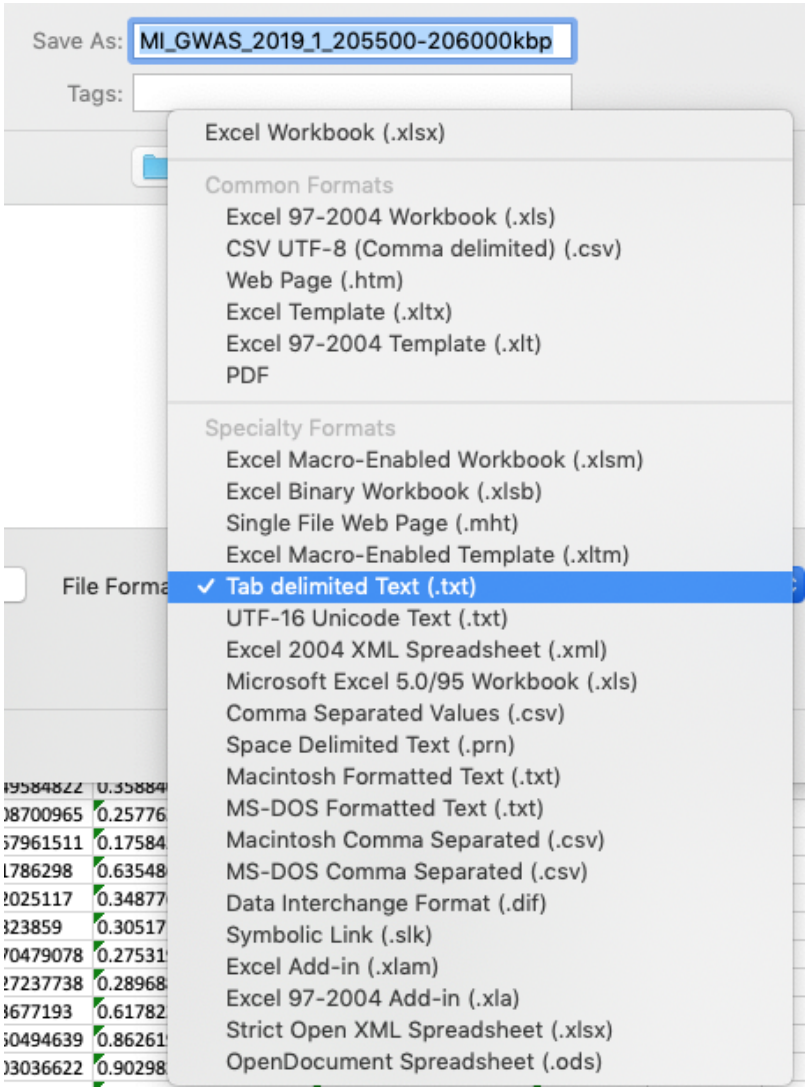


Fig. 2: Formatting your csv or Excel file as a tab-delimited file using Excel.

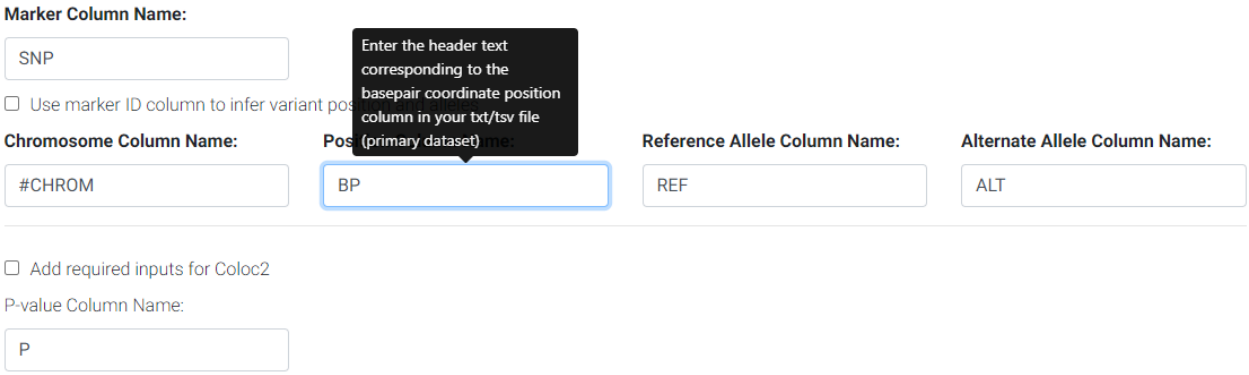


Fig. 3: Changing default column names to correspond to input file header names.

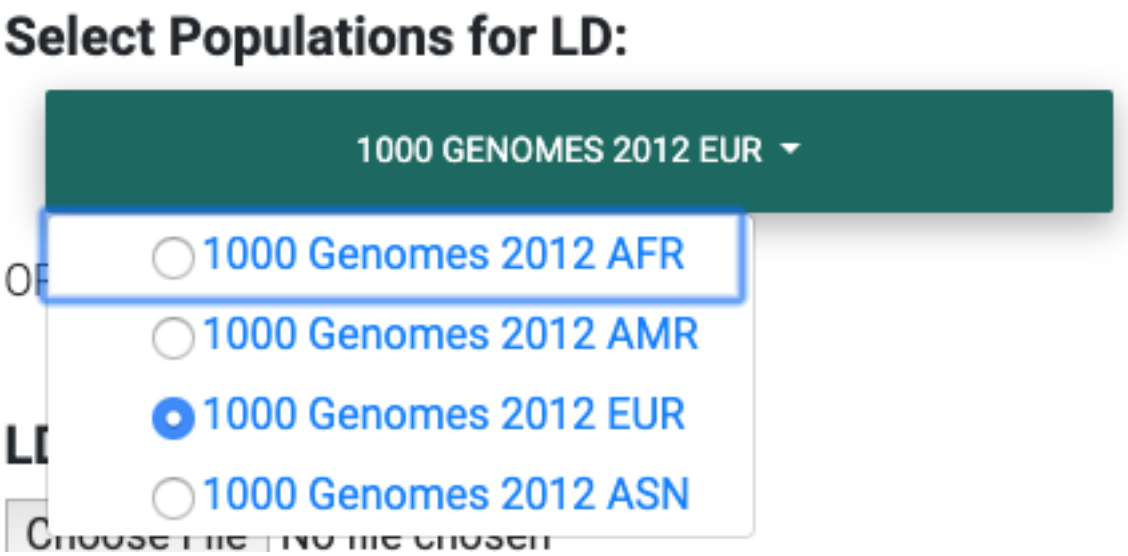


Fig. 4: Selecting the most appropriate 1000 Genomes population (hg19).

1.3.1 Selecting a publicly available 1000 Genomes population LD matrix

hg19

These datasets were obtained from LocusZoom. However, for more accurate results, we suggest *computing and uploading the LD matrix for your GWAS study*.

The 1000 Genomes population dataset (phase 1, release 3) consists of:

- EUR: European population of 379 individuals
  - 85 CEU - Utah Residents (CEPH) with Northern and Western European Ancestry
  - 93 FIN - Finnish in Finland
  - 89 GBR - British in England and Scotland
  - 14 IBS - Iberian Population in Spain
  - 98 TSI - Toscani in Italia
- AFR: African population of 246 individuals
  - 61 ASW - Americans of African Ancestry in SW USA
  - 97 LWK - Luhya in Webuye, Kenya
  - 88 YRI - Yoruba in Ibadan, Nigeria
- ASN: Asian population of 286 individuals
  - 97 CHB - Han Chinese in Beijing, China
  - 100 CHS - Southern Han Chinese
  - 89 JPT - Japanese in Tokyo, Japan
- AMR: Ad Mixed American population of 181 individuals
  - 60 CLM - Colombians from Medellin, Colombia
  - 66 MXL - Mexican Ancestry from Los Angeles USA

  - 55 PUR - Puerto Ricans from Puerto Rico

Descriptions of the population codes can be found in the IGSR: The International Genome Sample Resource. Also, please note that the known cryptic relationships in the 1000 Genomes were not removed for the non-European populations.

hg38

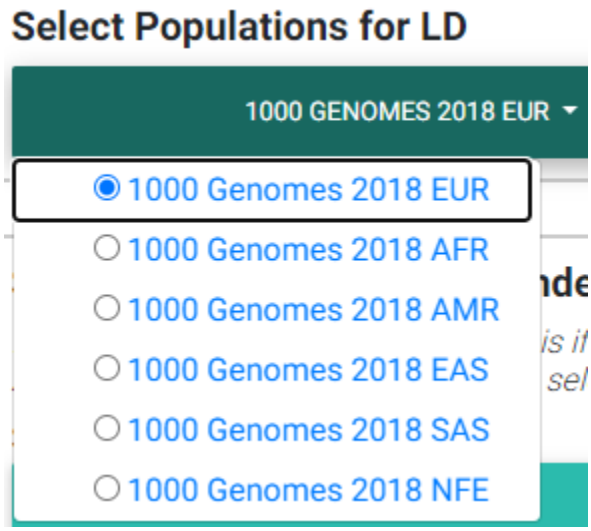


Fig. 5: Selecting the most appropriate 1000 Genomes population (hg38).

The hg38 version of the 1000 Genomes population is computed from a fully realigned call set against GRCh38. The biallelic SNV call set is used and more details are available on the 1000 Genomes Project FTP site.

A total of 2,548 individuals are available on the 1000 Genomes, and 2,507 have been assigned a population and super-population code. The distribution of populations available for selection on the app are as follows:

- EUR: European population of 502 individuals
  - 99 CEU - Utah Residents (CEPH) with Northern and Western European Ancestry
  - 89 GBR - British in England and Scotland
  - 107 IBS - Iberian Population in Spain
  - 111 TSI - Toscani in Italia
  - 96 FIN - Finnish in Finland
- NFE: Non-Finnish European (406 individuals)
- AFR: African population of 671 individuals
  - 97 ACB - African Caribbeans in Barbados
  - 61 ASW - Americans of African Ancestry in SW USA
  - 100 ESN - Esan in Nigeria
  - 113 GWD - Gambian in Western Divisions in the Gambia
  - 103 LWK - Luhya in Webuye, Kenya
  - 90 MSL - Mende in Sierra Leone

  - 107 YRI - Yoruba in Ibadan, Nigeria
- AMR: Ad Mixed American population of 333 individuals
  - 92 CLM - Colombians from Medellin, Colombia
  - 64 MXL - Mexican Ancestry from Los Angeles USA
  - 85 PEL - Peruvians from Lima, Peru
  - 92 PUR - Puerto Ricans from Puerto Rico
- EAS: East Asian population of 509 individuals
  - 100 CDX - Chinese Dai in Xishuangbanna, China
  - 106 CHB - Han Chinese in Beijing, China
  - 99 CHS - Southern Han Chinese
  - 105 JPT - Japanese in Tokyo, Japan
  - 99 KHV - Kinh in Ho Chi Minh City, Vietnam
- SAS: South Asian population of 492 individuals
  - 106 GIH - Gujarati Indian from Houston, Texas
  - 102 ITU - Indian Telugu from the UK
  - 96 PJL - Punjabi from Lahore, Pakistan
  - 102 STU - Sri Lankan Tamil from the UK

1.3.2 Computing the LD matrix from your GWAS population

Before you compute the LD matrix, please ensure that the number and order of SNPs matches that of the original uploaded GWAS or primary dataset file.

The easiest way to compute the LD from your own population is using PLINK.

Assuming your GWAS dataset is in binary PLINK format (ie. bed/bim/fam fileset), and you have subset the region, an example run would be:

Listing 2: *Example PLINK command for calculating the LD matrix*

```
plink --bfile <plink_filename> --r2 square --make-bed --out <output_filename>
```

In the above command, please replace plink_filename and output_filename with appropriate substitutes.

1.4 Secondary datasets

All 48 *GTEx (v7)* (hg19) and 49 *GTEx (v8)* tissues are provided for selection as secondary datasets to test colocalization with. The genes found in the coordinates entered (*GENCODE v19 (hg19) or GENCODE v26 (hg38)*) can be chosen for colocalization testing (thus, the number of secondary datasets and colocalization tests performed is the number of tissues selected times the number of genes selected in the region).

In addition to GTEx tissues, several user-specified datasets may be uploaded as a merged HTML file. For further instructions on how to *create a merged HTML file*, see the section below.

Alternatively, you may skip selection of GTEx tissues altogether and only focus on the colocalization tests for your uploaded secondary dataset(s). Please note that you *can* have both custom and GTEx datasets analyzed as secondary datasets.

1.4.1 Selecting GTEx tissues as secondary datasets

You may select all the necessary tissues to test colocalization with your primary dataset. Computation times, however, increase the more tissues and genes you select, so please be selective here if possible. Most colocalization analyses will finish within 10-15 minutes, but a gene-rich region may take 30 minutes or longer to compute.

Also note that computing a large number of Simple Sum p-values is computationally demanding for a web server, and doing so may delay or prevent others from accessing the website. In a later release, we plan to add a queue system for a better experience.

1.4.2 Formatting custom secondary datasets

In order for LocusFocus to recognize a dataset as secondary and perform colocalization testing, you must format your dataset in HTML format. The HTML format allows several datasets to be merged in a single HTML file. We suggest each dataset be preceded by an <h3> tag with the description title of the dataset.

We provide a python script to simplify the generation of the merged HTML dataset. To run both Simple Sum and COLOC2, please use the merge_and_convert_to_html_coloc2.py script.

The first step in creating the HTML file is to create a tab-separated descriptor file containing the list of files to be merged together (first column). The second column (tab-delimited) may contain descriptions of the datasets. The remaining columns specify the column names for chromosome, basepair position, SNP name, P-value (in that order).

For example, suppose we had three genomewide association analyses (from Ben Neale) from the UK Biobank:

1. Forced vital capacity (FVC) - 3062
2. Forced expiratory volume in 1-second (FEV1) - 3063
3. Peak expiratory flow (PEF) - 3064

The first few lines for FVC look as follows:

Listing 3: *Few lines of a tab-delimited FVC association analysis file from Ben Neale's analysis of the UK Biobank*

```
> head 3062.assoc.mod.ROI.slc26a9.tsv

variant rsid    nCompleteSamples        AC      ytx     beta    se      tstat   pval
↪  chr     pos     variant ref     alt
1:205860191:A:G rs149104610     307638  7.58684e+03     2.89852e+02     -1.06502e-02
↪  9.01042e-03     -1.18199e+00    2.37211e-01     1       205860191       A       G
1:205860462:G:A rs183927606     307638  7.58336e+03     2.94623e+02     -1.03192e-02
↪  9.01499e-03     -1.14467e+00    2.52347e-01     1       205860462       G       A
1:205860763:T:C rs182878528     307638  7.57953e+03     2.92349e+02     -1.04035e-02
↪  9.01627e-03     -1.15386e+00    2.48558e-01     1       205860763       T       C
1:205860874:A:G rs573870089     307638  1.06676e+03     6.36134e+01     2.08202e-02
↪  2.49968e-02     8.32914e-01     4.04894e-01     1       205860874       A       G
1:205861028:G:A rs36039729      307638  4.44225e+04     2.35233e+03     9.40215e-04
↪  3.81097e-03     2.46713e-01     8.05131e-01     1       205861028       G       A
1:205861107:C:T rs9438396       307638  2.85296e+04     1.97301e+03     9.23118e-03
↪  4.68786e-03     1.96917e+00     4.89347e-02     1       205861107       C       T
1:205861225:G:A rs115170053     307638  5.08637e+03     3.62211e+02     6.28049e-03
↪  1.11434e-02     5.63607e-01     5.73022e-01     1       205861225       G       A
```

(continues on next page)

(continued from previous page)

```
1:205861433:G:C rs6670490       307638  2.88292e+04     1.99373e+03     8.56608e-03
↪  4.66254e-03     1.83721e+00     6.61794e-02     1       205861433       G       C
1:205862075:T:C rs6665183       307638  2.18993e+04     1.26147e+03     1.50666e-04
↪  5.70059e-03     2.64299e-02     9.78914e-01     1       205862075       T       C
```

Note that we modified the original file by adding the chromosome and position columns

The FEV1 and PEF phenotype association files look similar.

To merge the summary statistics from these three files into a merged HTML file, we would first create a descriptor file of all the files we would like to merge. See below for the case of combining these three files:

Listing 4: *Description file of all the secondary dataset files we would like to merge into an HTML file*

```
> cat slc26a9_uk_biobank_spirometry_files_to_merge.txt

3062.assoc.mod.ROI.slc26a9.tsv       Forced vital capacity (FVC) - 3062      chr
↪pos     rsid pval
3063.assoc.mod.ROI.slc26a9.tsv       Forced expiratory volume in 1-second (FEV1) -
↪3063      chr     pos     rsid    pval
3064.assoc.mod.ROI3.slc26a9.tsv      Peak expiratory flow (PEF) - 3064       chr
↪pos     rsid    pval
```

Each column (tab-separated) defines:

1. Filename
2. Description title
3. Chromosome column name
4. Basepair coordinate position column name
5. rs ID column name (alternatively, you may specify a column with variant ID formatted as chrom_pos_ref_alt_b37; e.g. 1_205860191_A_G_b37)
6. P-value

Note that COLOC2 assumes an eQTL dataset as secondary input, and the merge_and_convert_to_html_coloc2.py script must be used, which requires the same inputs as above, plus the following:

7. Beta
8. Standard error
9. Number of samples
10. A1 (minor) or alternate allele
11. A2 (major) or reference allele
12. Minor allele frequency (MAF)
13. Probe ID

You may refer to the github page for more examples of datasets, where a sample descriptor file and sample command are provided for guidance to build a secondary dataset to also run COLOC2 colocalization.

While running COLOC2 is possible, we proceed below with the simpler example without COLOC2. The steps to also include COLOC2, however, are similar.

Then, to generate the merged html file while subsetting the region we may issue the command as follows:

Listing 5: *Example command to merge three summary statistic datasets into an HTML file using merge_and_convert_to_html.py*

```
> python3 merge_and_convert_to_html.py slc26a9_uk_biobank_spirometry_files_to_merge.
↪txt 1:205860000-205923000 slc26a9_uk_biobank_spirometry_merged.html
```

A description of the positional arguments may be issued with the -h or –help arguments:

Listing 6: *Description of positional arguments for merge_and_convert_to_html.py script*

```
> python3 merge_and_convert_to_html.py -h

usage: merge_and_convert_to_html.py [-h]
   filelist_filename coordinates outfilename

Merge several datasets together into HTML tables separated by <h3> title tags

positional arguments:
   filelist_filename  Filename containing the list of files to be merged
                  together. The second column (tab-delimited) may contain
                  descriptions of the datasets. The remaining columns
                  specify the column names for chromosome, basepair
                  position, SNP name, P-value (in that order).
   coordinates        The region coordinates to subset from each file (e.g.
                  1:500,000-600,000
   outfilename        Desired output filename for the merged file

optional arguments:
   -h, --help         show this help message and exit
```

The above command will generate the merged slc26a9_uk_biobank_spirometry_merged.html, file which can be used with LocusFocus.

1.5 Some important points to consider

- Please note that your GWAS and secondary datasets must be subset in order to reduce the file size for uploading purposes (current combined limit is 100 MB).
- You must also enter the genomic location you are interested in the *HG19 Coordinates* field. The format must be *chromosome:start-end*, where *start* is the starting basepair position and *end* is the ending basepair position.
- The region size entered in *Coordinates* field cannot be larger than 2 MBbp.
- If the SNP column has multiple rsid's separated by semicolon, the first rsid will be used.
- The SNP rs id is not used for determining the presence of the SNP in the 1000 Genomes population for the LD calculation; the chromosome and position columns determine this. If the chromosome:position combination is found in the 1000 Genomes, then the pairwise LD will be calculated for that particular SNP.
- The LD matrix is calculated for chromosome:position SNPs available in both GWAS input and the selected 1000 Genomes population datasets.
- Only overlapping SNPs are used for the Simple Sum calculations, and only overlapping SNPs with the primary dataset are plotted.

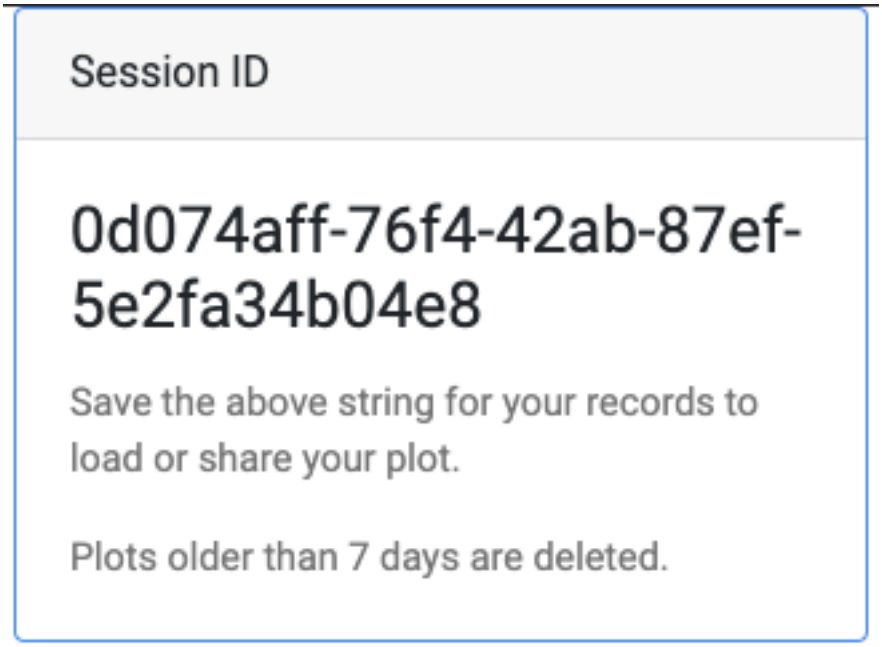
- A region within +/- 0.1 Mbp is selected around the top SNP to compute the Simple Sum p-value. This area is shaded gray in the first plot. This is done for each gene found within +/- 1 Mbp of the top SNP for all the tissues selected.
- It is important to have a dense set of genotyped SNPs to get an accurate assessment of the Simple Sum p-value calculation.

CHAPTER 2

Retrieving a Previous Session

Every session that is submitted and run can be quickly retrieved without the need to re-run the entire analysis again. These sessions are saved for 7 days or longer.

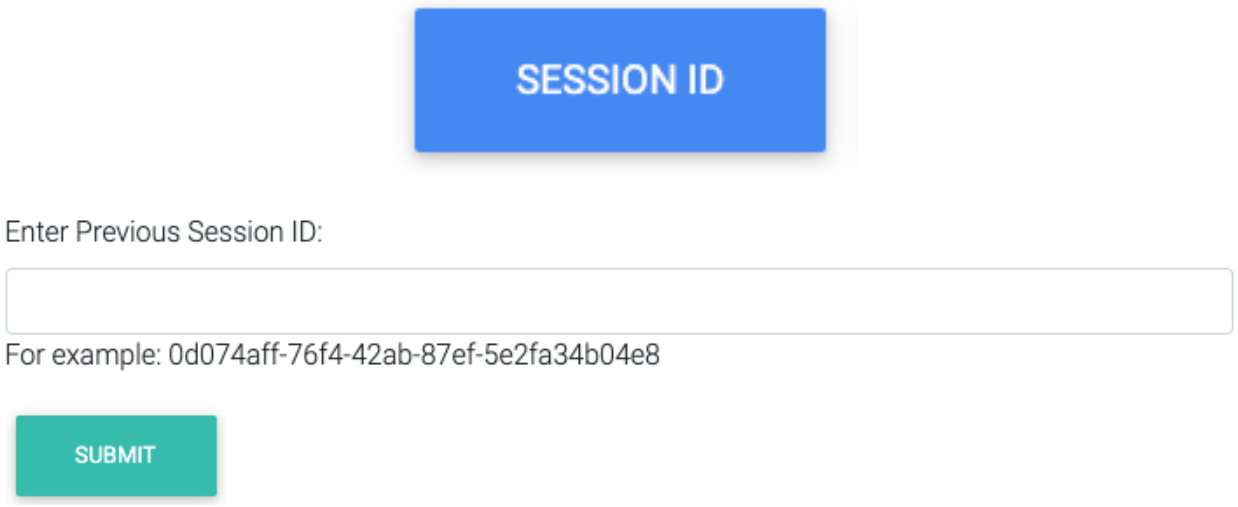
To do so, you must copy and store the session ID string that is given at the top right corner of the output page. An example is shown below.



To retrieve the session, simply click on the Session ID button on the main page.

And then paste the session ID string in the input form and click submit.

As an example, you can try submitting session ID 0d074aff-76f4-42ab-87ef-5e2fa34b04e8



The example above will bring the plots associated with the GWAS association results for meconium ileus, a complication in cystic fibrosis patients where newborns experience bowel obstruction at birth due to thick meconium, and assesses the *SLC26A9* gene eQTLs for each relevant digestive tract tissue from GTEx (version 7). The heatmap provides an overview of the strength of colocalization using the Simple Sum p-value across the digestive tissues and all the genes found within 1 Mbp of the top SNP in the region.

CHAPTER 3

Example Run, Usage and Interpretation

Here we show an example run using the GWAS summary statistics for meconium ileus around the *SLC26A9* association locus as the primary dataset, and its colocalization with nearby genes in GTEx tissues.

3.1 Sample Data and Output

You may follow along by downloading the sample data for meconium ileus. The output after a run using this dataset can be accessed by entering the session ID 0d074aff-76f4-42ab-87ef-5e2fa34b04e8

1. Click on "Choose File" to upload the GWAS summary statistics. As stated, the file must be tab-delimited and not exceed 100MB.

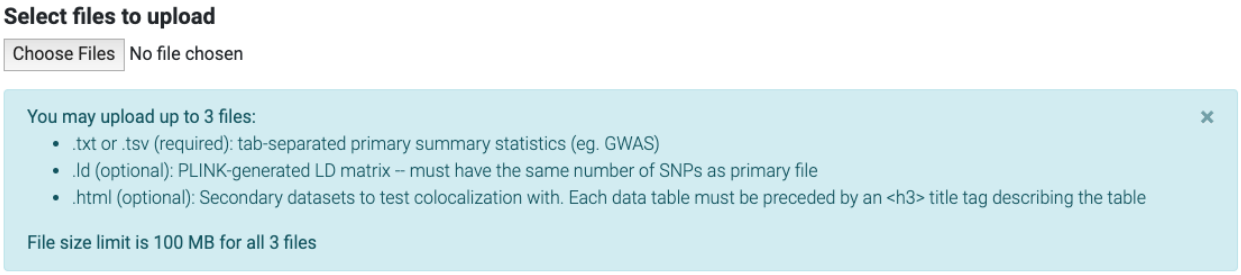


Fig. 1: Upload button allows for uploading multiple files

- This step allows uploading multiple files. Two additional files may be uploaded here (optional):
  - *A PLINK-generated .ld matrix file for your sample population*
  - *An HTML file of secondary datasets*

2. Next, if your GWAS summary statistics have column names that differ from the defaults, you may enter them in the respective columns.

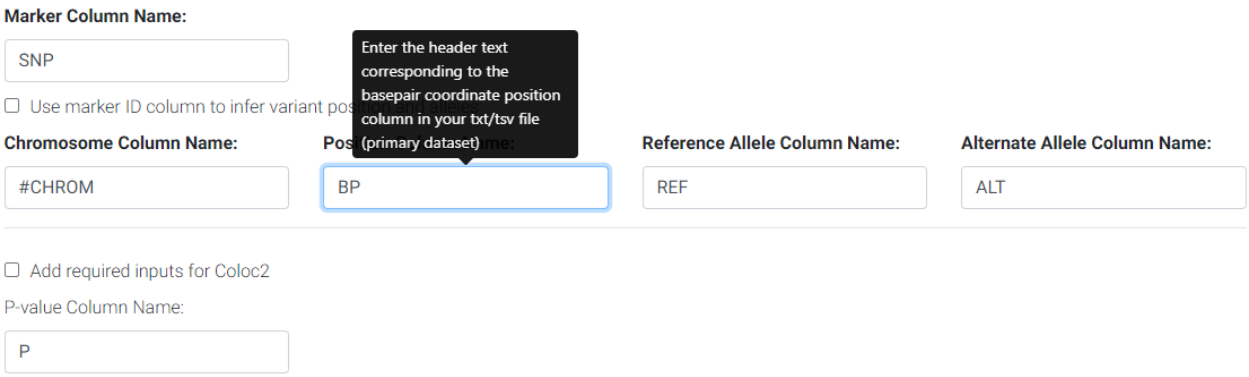
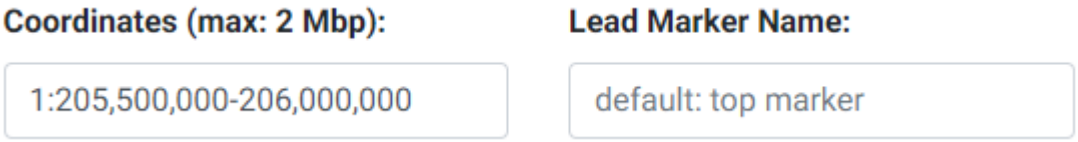


Fig. 2: Enter the column names for your primary/GWAS dataset

3. Enter the region to subset and, optionally, enter a lead marker name. The format of the region must be chr:start-end. The region cannot be more than 2 Mbp wide. This input field is checked for errors and a message will appear below it if an inappropriate entry is made. The coordinates specified here are used to look up the genes available (GENCODE v19 for hg19 and GENCODE v26 for hg38) for colocalization testing using GTEx. The marker string entered must match the string in the SNP column of the uploaded GWAS file. If no marker is chosen, then the top SNP (with lowest P-value) will be automatically used. If the top SNP does not match an entry in the 1000 Genomes, the next best SNP will be automatically chosen.



3.2 Selecting LD (Linkage Disequilibrium)

Please see the section on *selecting an LD matrix*. In brief, you can select one of the 1000 Genomes populations LD structure, or upload the LD structure for your population. We recommend users calculate the LD matrix for the uploaded region for their population for more accurate results. You may download the matrix for this example here. For this example, we selected the 1000 Genomes European population instead.

In the image above,

- EUR: European population
- AFR: African population
- AMR: Ad Mixed American population
- EAS: East Asian population
- SAS: South Asian population
- NFE: Non-Finnish European

The LD matrix file must have the same number and order of SNPs as the GWAS summary statistics file.

For details on the populations available for LD calculations in either hg19 or hg38 builds, see the section on *Selecting a publicly available 1000 Genomes population*

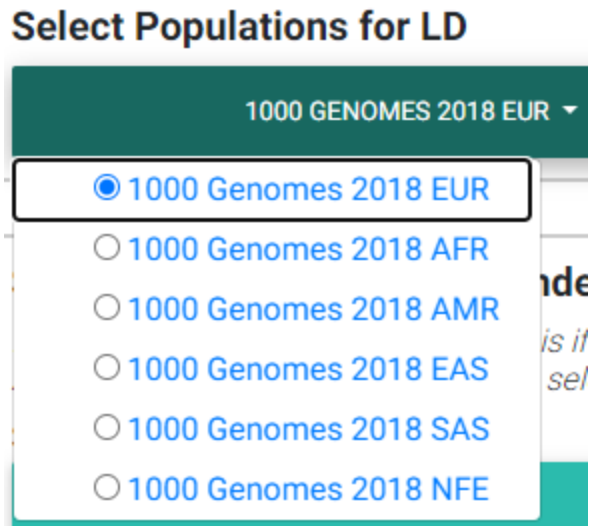


Fig. 3: Selecting most appropriate 1000 Genomes population (ignored if uploading .ld file)

3.3 Selecting Secondary Datasets from GTEx

In the dropdown, we list all 48 tissues analyzed by GTEx v7 (hg19) or 49 tissues for GTEx v8 (hg38).

Please note that while we allow the ability to select all the tissues, this increases the amount of time for computing all Simple Sum p-values due to the number of gene-tissue pairs at a particular genomic region.

You may search through these tissues in the search box provided in the dropdown. Please note that the search is case-sensitive and that all tissue names start with a capital letter.

3.4 Selecting Genes of Interest for Colocalization Testing

The next field requests for selecting genes found in the region that was entered previously in the *Coordinates* field. Colocalization testing is performed on the eQTL data for the gene/tissue pairs selected.

Please note that colocalization testing is computationally demanding and may take some time to complete if selecting many gene/tissue pairs.

3.5 Overriding the First-Stage Set-Based P-value Threshold

The Simple Sum method first assesses which secondary/eQTL datasets pass a first-stage significance test. The threshold for this passing this test is set at the Bonferroni level (0.05 divided by the number of secondary datasets). If you would like to override this p-value threshold, you may enter your desired threshold in the input field.

For example, if you selected 3 tissues and 4 genes for testing, and uploaded 3 other secondary datasets, you have a total of $3 \times 4 + 3 = 15$ secondary datasets or tests for the first stage. The default Bonferroni-corrected p-value threshold of $0.05 / 15 = 0.0033$ will be used for first stage significance testing. Secondary datasets that pass the first stage threshold undergo colocalization testing in the next stage.

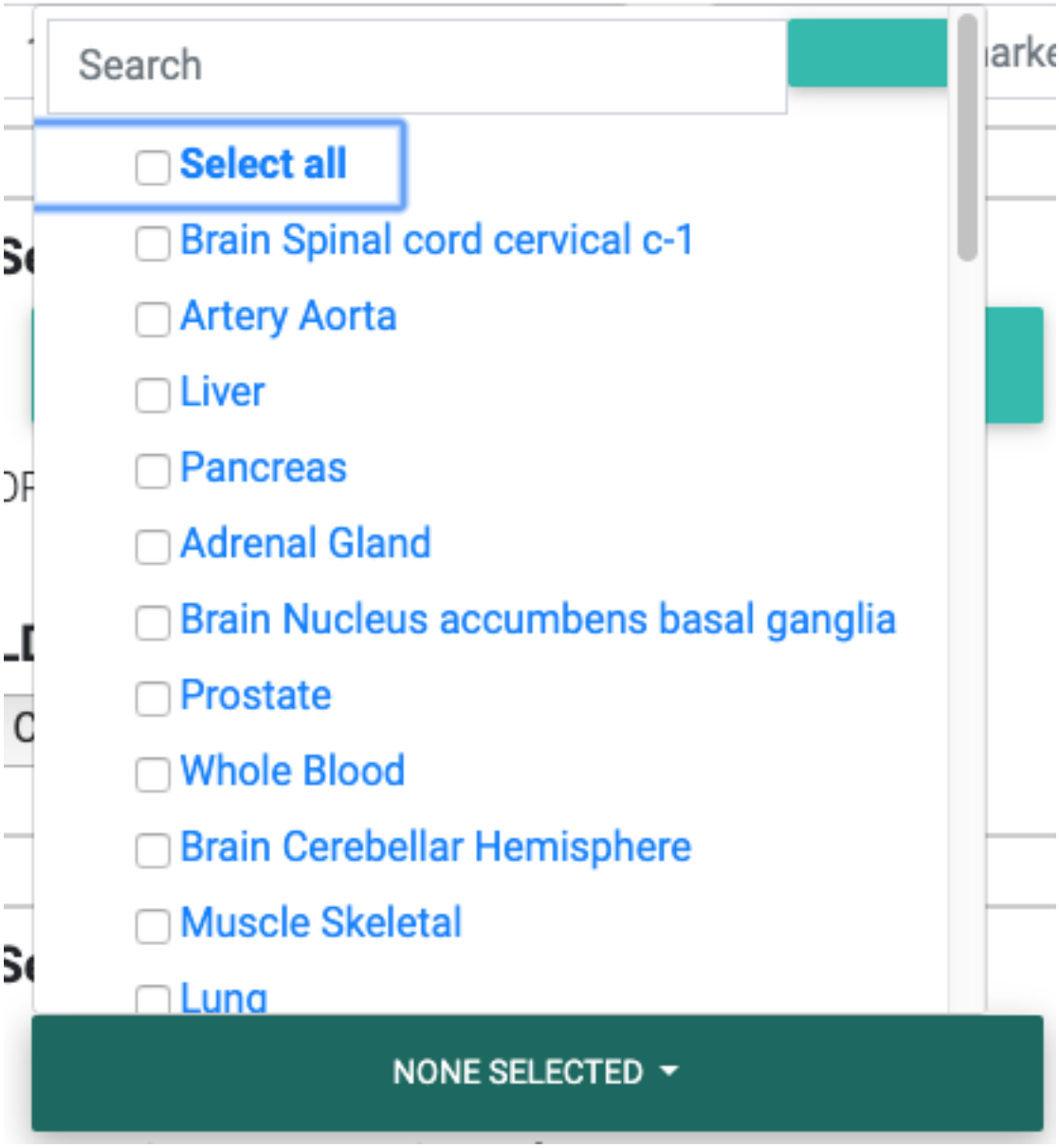


Fig. 4: Selecting appropriate tissues for colocalization testing from the GTEx Project.

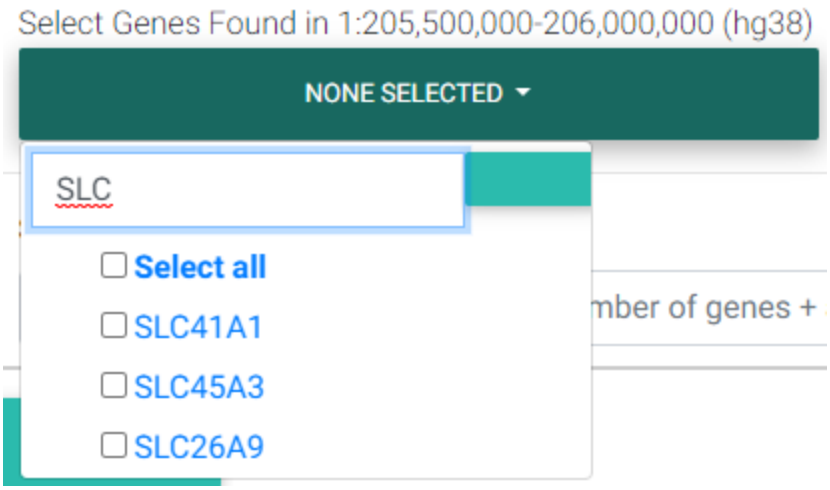


Fig. 5: Selecting genes found in the region of interest

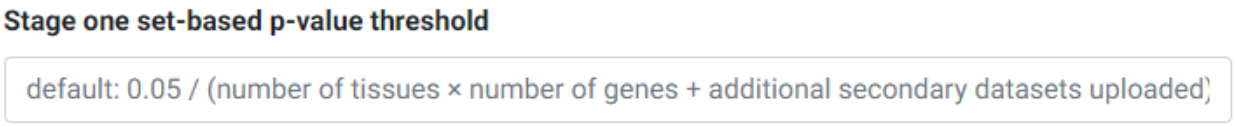
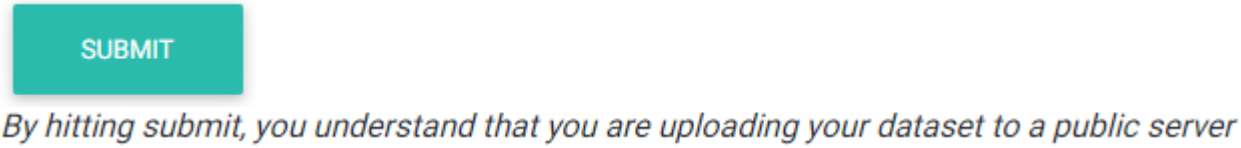


Fig. 6: Overriding the first-stage set-based Bonferroni p-value threshold

3.6 Submit

Good job! You are now ready to hit Submit!



Please click the submit button just once. Depending on how many tissues and tissues you have selected, the process may take anywhere from a few minutes up to 30-45 minutes for a gene-rich region with all tissues selected.

3.7 Saving and Retrieving Your Session

After the program has computed the colocalization tests, the page will refresh to show the plots and a session ID on top of the page.

Please save this session ID string for your records in order to retrieve the page without running the full computation again. See session retrieval for help on this.

3.8 Interpreting Data Output

3.8.1 Colocalization plot

Plots are generated using Plotly.

The first plot that is generated consists of:

- The GWAS p-values uploaded with the lead marker used as reference to show the degree of pairwise LD with the lead marker. These are shown as circles. The color pattern is similar to that followed by LocusZoom, where the strength of r^2 with the lead marker is broken down by the following color-coding scheme:
  - dark blue circles - low LD (< 0.2)
  - light blue circles - LD between 0.2-0.4
  - green circles - LD between 0.4-0.6
  - orange circles - LD between 0.6-0.8
  - red circles - high LD greater than 0.8
  - the purple circle (slightly larger than the rest) is the lead marker
  - gray circles are markers that could not be found in the 1000 Genomes (phase 1, release 3)
- Lines showing the (rough) eQTL p-value patterns followed for the particular gene and tissues selected.
  - These lines are connected by taking the lowest p-value in a moving window.
  - The size of these windows varies according to the size of the region entered as follows:
    - Region size (in basepairs) divided by 100,000 then times 15 (i.e. (regionsize/100,000) * 15)
- Circles (hidden by default) to show the eQTL data for the user-entered gene and tissues. This is the underlying raw data used to draw the (rough) line patterns.
  - To show these circles, simply click on the corresponding tissue name in the legend for which you would like to observe the eQTL data for.
- A gray-shaded region that spans approximately 100 Kbp on each side of the lead marker. Markers that fall in this shaded region are used for calculating the Simple Sum p-values. Note that while only markers in this shaded region are used for the Simple Sum p-value calculation, all genes that fall in the region entered get a Simple Sum value computed for them using the markers in this shaded region (i.e. while the gene may be far away from the shaded region, markers in the shaded region may fall in a *cis*-regulatory element that influences the expression of that gene).
- If there are genes in the region, the collapsed gene transcript model is shown under the plot. An attempt is made to display the gene name under or above the gene. However, if there are many genes in the region, some text is hidden to avoid crowding. If that's the case, one can always hover over the start, end or middle of the gene to display the gene name in a tooltip.

Plotly has several functionalities to permit the interactive exploration of the plot. On top of the plot, you will notice a toolbar to allow for several functions.

Some of the functions of this toolbar include saving the plot, zooming, panning, selection tools, and data exploratory tools such as spike lines and vertical data point comparisons (e.g. if you have the GWAS and eQTL circles shown, you may select the "Compare data on hover" and compare the same association p-values for the GWAS and eQTL SNPs simultaneously - see example figure below).

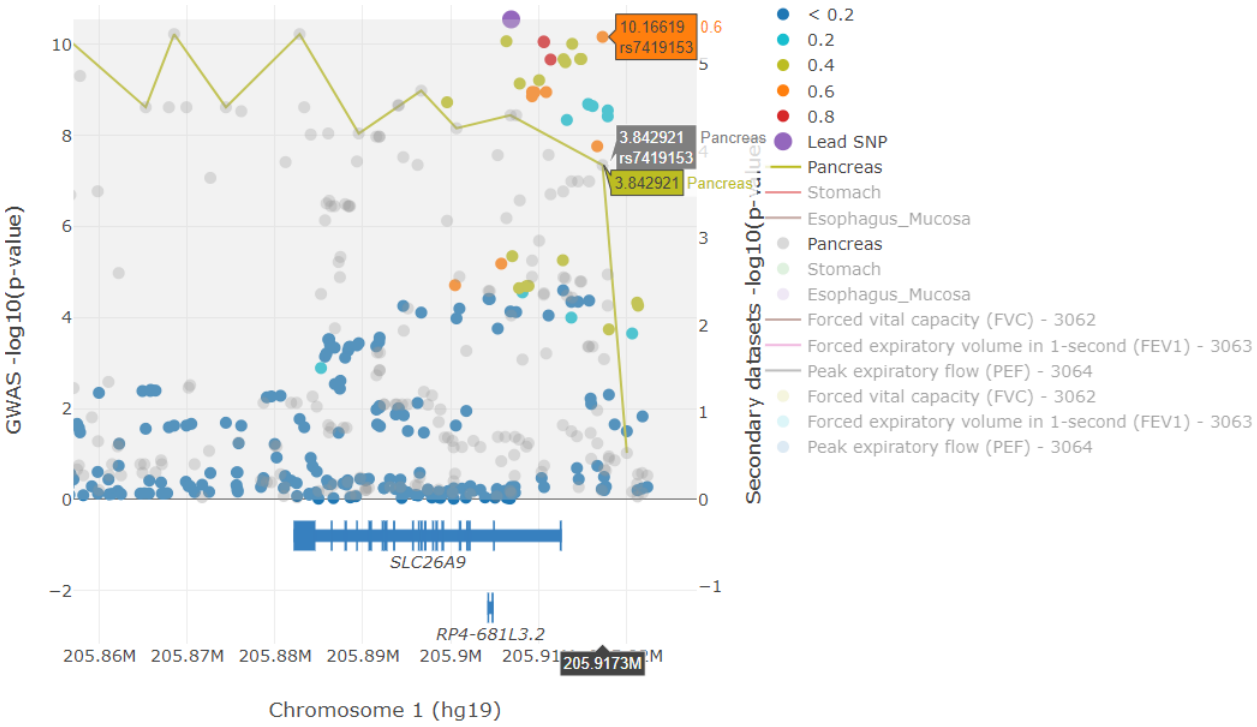


Fig. 7: Example colocalization plot illustrating the "compare data on hover" feature of plotly.

In the example image above, we find a particular top GWAS SNP (rs7419153) that also has a high -$\log_{10}$ eQTL P-value in the Pancreas. To get this result, simply zoom into the example session, click on the "Compare data on hover" tool, and hover over the SNPs (if the SNP data is dense, it is easier to first zoom in and show only the top GWAS hits - you could deselect the SNPs with low LD by clicking on the legend). The y-axes can be rescaled by clicking and dragging at the corners; clicking and dragging the y-axes from the middle repositions the zero line.

3.8.2 Interpreting the Heatmap Plot

The heatmap plot shows the -$\log_{10}$ Simple Sum P-values and their *relative* strength compared to all the other GTEx gene-tissue pairs for the session. If the Simple Sum p-value could not be calculated for a particular gene-tissue pair, it will show as a negative number.

Reasons for reporting a negative number are further broken down in three cases and an *interactive table* output below the heatmap describes the exact reason.

3.8.3 Simple Sum Table

The -$\log_{10}$ Simple Sum colocalization p-values are reported for the gene-tissue pairs that passed the first stage set-based test for significance (after Bonferroni correction by default, unless overriden by the user).

There are three cases in which colocalization p-values may not be calculated, and each of those particular cases is given a negative numeric value as described below:

- -1 value is given to gene-tissue pairs with no eQTL data (usually due to little or no expression)
- -2 value is given to gene-tissue pairs that did not pass the Bonferroni-corrected first stage testing for signficance among the secondary datasets chosen

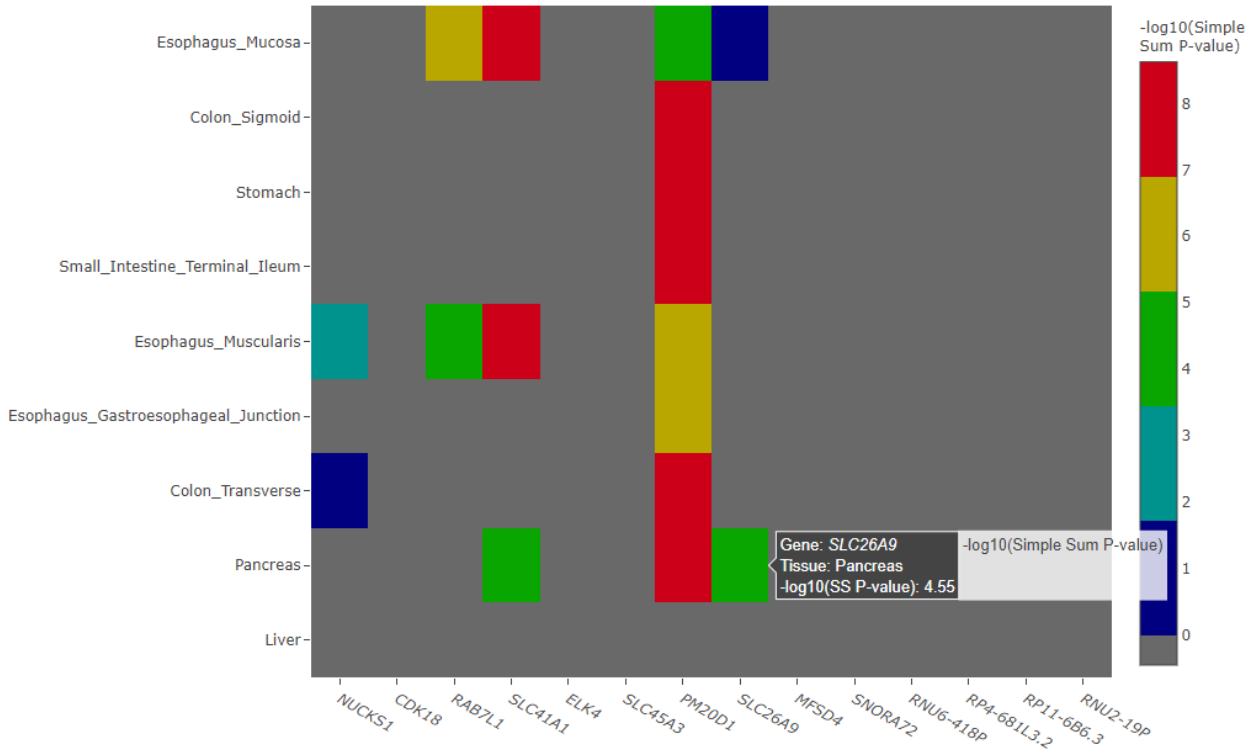


Fig. 8: Example heatmap plot of -$\log_{10}$ Simple Sum p-values

- -3 value is given to gene-tissue pairs where the Simple Sum P-value computation failed, likely due to insufficient SNPs

At this stage, it is up to each study to determine a reasonable p-value threshold to determine if a particular Simple Sum p-value should be considered significant. A conservative approach would be to take a Bonferroni-corrected threshold where the alpha level is divided by the number of tests performed (i.e. the number of gene-tissue pairs and other uploaded datasets that passed the first-stage test of significance). For example, if a user selected 3 tissues and 4 genes for testing, and 3 other secondary datasets (a total of $3 \times 4 + 3 = 15$ tests) and among these, 6 datasets passed the first-stage test and were tested for colocalization, then one would conservatively choose to consider a Bonferroni-corrected p-value threshold of $0.05 \div 6 = 8.3 \times 10^{-3}$ for a 0.05 alpha level.

If you have uploaded custom secondary datasets, a separate interactive table is output below the GTEx's Simple Sum interactive table.

3.8.4 COLOC2 Posterior Probability Results Table

If you opted to run COLOC2, the posterior probabilities for H4 (the most directly comparable to the Simple Sum - see bioRxiv paper) are output in an interactive table.

CHAPTER 4

Versions

4.1 Version History

- v0.0.1 (released Sept. 6, 2019)
- **v0.0.2 (released Nov. 8, 2019)**
  - Fixed file upload issue; up to 3 files may be uploaded at once and file types are auto-detected
  - Increased file size limit to 100 MB
  - Added ability to change the eQTL gene within the plot output page
  - Added interactive table of Simple Sum -log10 P-value results, with ability to download the table in various formats
  - Added t^2 test for testing whether the given secondary datasets have a significant signal above a Bonferroni adjustment
  - Added ability to upload SNP names in chr_pos_ref_alt_b37 format for cases when the rs ID is not available
- **v1.0 alpha (released Dec. 5, 2019)**
  - Enabled ability to upload secondary datasets as a merged HTML file in addition to any selected GTEx tissues
- **v1.0.1 alpha (released Dec. 18, 2019)**
  - Added ability to transpose table and adjust plot figure drawing parameters
- **v1.1.0 alpha (released Apr. 1, 2020)**
  - Internal change to calculate the Simple Sum using an R script instead of Python.
  - This change enables the use of the app in Windows as the rpy2 package is no longer a requirement
- **v1.3.0 alpha (released Jul. 31, 2020)**
  - Addition of hg38 coordinate support

  - Added the latest GTEx version 8 (hg38) eQTL analyses for use as secondary datasets for colocalization testing
  - Added GRCh38 re-aligned 1000 Genomes (phase 3) as option for LD matrix
  - Using GENCODE v26 for hg38 gene track
  - Added support for COLOC2 colocalization testing

- **v1.4.0 alpha (released Aug. 6, 2020)**
  - Added ability to export images as svg vector format
  - Bug fix for the merge_and_convert_to_html_coloc2.py script
  - More sample datasets added that are compatible for COLOC2 runs
- **v1.4.1 alpha (Oct. 6, 2020)**
  - Bug fix in identifying lead SNP
- **v1.4.2 alpha (Oct. 9, 2020)**
  - Fixed issue where SS was not being computed due to non-singular matrix error
  - Improved initial plotting time of colocalization plot by filtering out GWAS p-values less than 0.1 by default
- **v1.4.3 alpha (Nov. 10, 2020)**
  - Fixed issue where the Simple Sum calculation was not being performed in the case where no missing data was present
- **v1.4.4 alpha (Nov. 16, 2020)**
  - Fixed rsid mapping bug
  - Fixed bug when matching the top SNP with secondary datasets
- **v1.4.5 alpha (Feb. 04, 2020)**
  - Fixed issue where the SS window did not follow the user-specified lead SNP
  - Added a function to clean up uploaded dataset of common file input reading issues
- **v1.4.6 alpha (Feb. 16, 2020)**
  - Improved SNP rs ID matching with GTEx rs ID's
  - Warning text is now displayed if a large number of SNPs do not match variants in GTEx
- **v1.4.7 alpha (Feb. 17, 2020)**
  - Added output table for parameters used in a particular app run
  - Added output table to help guide and interpret SS colocalization results more easily
- **v1.4.8 alpha (Mar. 24, 2020)**
  - Fixed SS computation issue when LD matrix has missing values
  - Fixed wrong output for the total number of GTEx datasets in the SS guidance table
  - Note: this version had an issue with uploaded secondary datasets with missing tables where the SS statistics were not assigned to the correct dataset
- **v1.4.9 alpha (Apr. 23, 2020)**
  - Fixed issue where uploaded secondary datasets with missing tables are properly dealt with

4.2 Datasets

- Human reference: hg19 (GRCh37.p13) and hg38 (GRCh38.p7)
- GTEx: versions 7 (hg19) and 8 (hg38)
- GENCODE: version 19 (hg19) (the transcript models were collapsed into a single gene model)
- GENCODE: version 26 (hg38) (the transcript models were collapsed into a single gene model)
- 1000 Genomes (phase 3) aligned to GRCh37 biallelic SNV call set
- 1000 Genomes (phase 3) biallelic SNV call set re-aligned to GRCh38
- dbSNP151 GRCh37.p13
- dbSNP151 GRCh38.p7

4.3 Programs

All required programs and versions are specified in the yml file or conda spec file.

CHAPTER 5

Local installation of LocusFocus

LocusFocus may be cloned from the GitHub repository and run as a local webserver in Mac or Linux distributions.

The GTEx database, and 1000 Genomes PLINK files are optional. GTEx may be built using the initdb_GTExV7.py and initdb_GTExV8.py scripts. The script pushes GTEx eQTL summary statistics into a NoSQL MongoDB database.

The app has been tested in Linux and Mac environments. For Windows, you may use the Windows Subsystem for Linux (WSL) to emulate a Linux environment. The app has been tested using the Ubuntu WSL system.

All required programs and packages are listed in a yml file and a conda spec file are also available for direct explicit installation of the exact conda virtual environment the app requires for a successful run.

To clone and install the required packages, simply run:

Listing 1: *Clone and install required conda environment*

```
> git clone https://github.com/naim-panjwani/LocusFocus.git
> conda create --name <env> --file spec-file.txt
```

Next, to run the application locally, simply run the app.py script:

Listing 2: *Run the app*

```
> python app.py
```

The above command will issue a web address to access the application locally. Simply type the address given in your browser.

CHAPTER 6

Contact

For any questions, you may contact Naim Panjwani via email:

naim.panjwani at sickkids.ca

On the subject line, please type LocusFocus. If applicable, please also paste your session ID and the parameters used during your run.

CHAPTER 7

MIT License

CHAPTER 8

Future Directions

LocusFocus is open source and continuously undergoing development and improvements.

Below is a list of improvements and features currently under development:

- Enable uploading of compressed files (then convert to bgzip/tabix)
- Implement a queue system to enable job submission and later retrieval
- Enable ability to upload a config file to replicate results
- Implement SMR colocalization method
- Enable immediate deletion of the session data
- Enable option to match variant names by just chrom:pos information (currently should only be either all rsid's or all variant_id format)
- Generate updated 1KG Phase 3 binary PLINK files for hg19 LD calculations (currently only have 2012 1KG file)
- Enable user to set the window sizes for eQTL lines; and make calculation clearer
- Make a table of GWAS and eQTL merged results
- Plot of beta correlations
- Make P-P plot

A list of known bugs currently being addressed:

- chrX plots do not show correctly
- SS coordinate input field checking not working - should also check if it's a subregion of full coordinate/locus region
- Remove related individuals prior to calculating LD from the 1000 Genomes
- Include option for NFE LD (Non-Finnish European subset) for hg19 1000 Genomes
- File too large error handler not working
- Fix COLOC2 chromosome X issue

- Add requirement for having REF and ALT columns for secondary datasets for better SNP matching