Exploring regulation in tissues with eQTL networks: data, scripts and pipeline

Maud Fagny & John Platig

September 27, 2017

Introduction

This document describe all the data files and scripts necessary to replicates figures, supplementary figures, supplementary tables and datasets from the following paper:
Fagny M, Paulson JN, Kuijjer ML, Sonawane AR, Chen C.-Y., Lopes-Ramos CM, Glass K, Quackenbush J, Platig J. (2017) Exploring regulation in tissues with eQTL networks. *PNAS* **114(37)**:E7841-E7850. doi:10.1073/pnas.1707375114

General settings

List of necessary R packages and softwares

Running the following scripts requires to have the following softwares and packages installed:

Softwares

Software	Where to find it
plink2 v1.90 or higher	https://www.cog-genomics.org/plink2
R	https://cran.r-project.org/

R packages

Package	Where to find it
MatrixEQTL	https://cran.r-project.org/web/packages/MatrixEQTL/
condor	https://github.com/jplatig/condor
igraph	https://cran.r-project.org/web/packages/igraph/
data.table	https://cran.r-project.org/web/packages/data.table/
plyr	https://cran.r-project.org/web/packages/plyr/
gplots	https://cran.r-project.org/web/packages/gplots/
ggplot2	https://cran.r-project.org/web/packages/ggplot2/
RColorBrewer	https://cran.r-project.org/web/packages/RColorBrewer/
dendextend	https://cran.r-project.org/web/packages/dendextend/
survival	https://cran.r-project.org/web/packages/survival/
broom	https://cran.r-project.org/web/packages/broom/
metap	https://cran.r-project.org/web/packages/metap/
R Bioconductor	http://bioconductor.org

R Bioconductor packages

Package	Where to find it
Biobase	http://bioconductor.org/packages/release/bioc/html/Biobase.html
limma	http://bioconductor.org/packages/release/bioc/html/limma.html

List of required data files

Running these scripts requires to have the following data files:

List of files that are accessible via dbGaP

- A VCF file with genotyping data.
- An R object with filtered/normalized gene expression data (derived from read counts file using the R yarn package).
- A tab separated file containing a lookup table to match SNPs VCF ID with rsID.

List of files that are accessible via the GTEx portal or in the GTEx paper supplementary material

Where to find them?

- The GTEx portal: www.gtexportal.org
- The GTEx paper supplementary material: The GTEx Consortium (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans, *Science* **348**:648-660, doi: 10.1126/science.1262110

List of files

- A tab separated file with the first principal components of genotyping data.
- An R object derived from a file with the eQTL results obtained by the GTEx consortium

List of annotation files that are provided:

Description of file content	Path to data file
An R object with tissues names, shortname and description	data/tissue_names.RData
An R object with SNPs anotations	data/annotations_snps.RData
An R object with genes annotations	data/annotations_snps.RData
An R object with a lookup table for gene ENSEMBL ID/HGNC ID	data/genes_Ensembl_HGNC_corres.RData
An R object with the SNPs Epigenome roadmap annotations	data/epigenome_state_snps.RData
A curated version of the GWAS catalog (duplicated lines and entries without rsID have been removed)	data/gwas_catalog_curated.txt

Description of file content	Path to data file
A file with the list of GWAS traits and diseases corresponding to metabolic traits	data/gwas_metabolism_terms.txt
A file with the list of GWAS traits and diseases corresponding to autoimmune diseases	data/gwas_autoimmune_terms.txt
A tab separated file with the LD blocks	output/dosageMatrices/LDblockInfo.txt
A file with the genes from heart left ventricle community 86 annotated as "cellular respiration" by Gene Ontology	data/heart_left_ventricle_86_cellularrespiration_HGNC.txt

List of intermediate files that are provided and necessary to reproduce the results

Description of file content	Path to data file
An R object with the results of eQTL mapping	output/eqtls/all_tissues_eqtls_fdr0.20.2_1MB.Rdata
An R object with sample sizes for each tissue	output/eqtls/nb_samples.RData
A text file with the list of tissues that meet the criteria	code/tissues_gtex.txt

Scripts pipeline

To reproduce results and figures from our paper, run the following pipeline.

Data Normalization and eQTL mapping

All the scripts contained in this section require files from the GTEx project containing personal or identifying informations, and use data under retricted access. Consequenly, data files are not provided and these scripts can only be run by people that have requested and obtained access to the GTEx data and downloaded them. To gain access to the raw data used in this study, see the dbGap website. ### Genotype QC This script runs the genotype quality checks and generates the input files for matrix eQTL:

```
Rscript code/QCscript.R
Rscript code/create_matrix_eqtl_files.R
```

This script creates the file containing the list of tissues that match the filtering criteria:

```
code/create_list_tissues.sh data/ code/ tissues_gtex.txt
```

eQTL mapping

This script maps cis and trans eQTLs:

```
code/run_matrix_eqtl_gtex.R code/ tissues_gtex.txt
```

This script merges all cis and trans eqtl outputs with the RsIDs provided by the gtex consortium. It Adds a column to track whether or not an eQTL assocation is cis or trans.

```
code/eqtl_annotation_merge_bytissue.sh 0.2 0.2 1 code/tissues_gtex.txt output/expression
output/eqtls/data/GTEx_Analysis_\2015-01-12_OMNI_2.5M_5M_450Indiv_chr1to22_genot_imput_
info04_maf01_HWEp1E6_VarID_Lookup_Table.txt
```

This script creates an R object with eQTL mapping results and an R object with sample counts by tissue.

```
Rscript code/summarize_eqtls.R
```

eQTL results analysis

Summary of results

This script plots the number of eQTL as a function of sample size (Figure S2)

```
Rscript code/plot_nbeqtl_samplesize.R
```

This script summarizes the proportion of SNPs that are cis- and trans-eQTLs (dataset S2)

```
Rscript code/count_eqtls.R
```

This script calculates the distances between SNPs and their associated genes

```
Rscript code/distance_to_TSS.R
```

Comparing eQTL results with the GTEx consortium results

This script compares our cis-eQTLs with the GTEx consortium cis-eQTLs (Table S2)

```
Rscript code/compare_eqtls_our_gtex.R
```

Study gene expression correlation within modules

This script plots the distribution of pairwise gene expression correlation for each tissue and each community (Figure S5) (Can only be run after code/extract_snp_genes_edges.R)

```
Rscript code/expression_correlation_within_modules.R
```

Network clustering and properties

Starting here, all scripts can be run using the provided intermediary files (scripts don't use any input files containing identifying or personal data).

Network clustering

This script clusters all eqtl networks using condor:

```
code/run_eqtl_network_clustering_fast.sh 0.2 0.2 1 code/tissues_gtex.txt output/eqtls FALSE
```

This script creates the R object containing clustering results and three R objects containing the lists of SNPs/Genes/Edges for each tissue and each community.

```
Rscript code/extract_snp_genes_edges.R
```

Analysis of network properties

This script plots modularity for each tissue (Figure 1B)

```
Rscript code/analyse_cluster_modularity.R
```

This script plots networks as matrix (Figure 1C and S4).

```
Rscript code/makeCondorMatrixPlot.R
```

This script plots the proportion of communities with SNPs and genes from more than 2 chromosomes (Figure S6).

```
Rscript code/plot_cluster_distrib_chr_summary.R
```

Biological characterization of communities

Gene Ontology analysis

This script performs over-representation of Gene Ontology categories among communities

```
Rscript code/analyze_clusters_GO.R
```

This script plots the heatmap and the bubble plots with Gene Ontology results (Figure 2A, 3B and S7, dataset S3). The html file that was used to generate the sankey plot is in the "Figures/" folder (sankey_th2.html, Figure 2B)

```
Rscript code/plot_GO_results.R
```

This script generates the gwas snp-traits/diseases lookup table.

```
Rscript code/generate_gwas_lookup.R
```

This script generates the input files to plot the circos plot (Figure 3A) and output annotation for genes and SNPs in the community (dataset S4). The code to plot the circos plot is given in the "code/circos_plot/" folder.

```
Rscript code/generate_circos.R
```

This scripts computes resampled p-values for GO terms in shared communities in each tissue, and plots their distribution for GO:0010468

```
code/run_resample_GO_results_shared.sh
Rscript plot_resampling_shared_GO_pvalues.R
```

Tissue-specificity of communities

This script plots the enrichment tissue-specific genes/SNPs/edges in tissue-specific communities compared to shared communities (Figure 4A).

```
Rscript code/tissue_specificity_network_gcc.R
```

This script plots the odds ratio for tissues-specific SNPs in tissue-specific activated regions among tissue-specific communities compared to shared communities (Figure 4B).

```
Rscript code/tissue_specificity_epigenome_roadmap_state.R
```

Genomic location of central SNPs

This script generates the table with number of genes within 1Mb of each SNP.

```
Rscript code/identify_snps_neighbors.R
```

This script generates plots of enrichement in chromatin states categories among high core-scores and high-degree SNPs (Figure 5A and 5B, and datasets S1 and S5)

```
Rscript code/plot_OR_epigenomic_states_across_tissues.R
```

This script plots examples of core-SNPs and hubs (Figure 5C and 5D)

```
Rscript code/draw_genome.R
```

GWAS SNPs properties analysis

This script calculatesand plots pvalues for observed/expected distribution of SNP degree among GWAS SNPs and plot figure

```
Rscript code/resample_degree_gwas.R
```

This scripts calculates and plot enrichment in GWAS SNPs among high core-score (Figures 6B and S10)

```
Rscript code/gwas_core_score_Qi_LRT.R
```

Relationship between Qi (core-score) and Qik (normalized core-score)

This script plots the Qi vs. Qik plot (Figure S11).

```
Rscript code/plot_qik_qi_distrib.R
```