A tool for rapid aggregation of eQTLs in human cells

Sebastian Hofer, Matthias Samwald

Section for Medical Expert and Knowledge-Based Systems
Center for Medical Statistics, Informatics, and Intelligent Systems
Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

Introduction

Expression quantitative trait loci (eQTLs) are genetic variants that explain gene expression levels. They are usually identified in studies combining genome-wide association studies with global gene expression measurements, by using gene expression levels as quantitative traits. eQTLs can be characterized by:

- Mode of action: cis- or trans-eQTL

  *eQTLs can modify gene expression locally, such that the genetic variant is located in close proximity of the affected gene, or distantly, possibly even by acting on a different chromosome. They are thus referred to as **cis- (local)** or **trans-eQTLs (distant)**.*

- Genetic variant type: SNP, CNV, indel

  Most eQTLs are single nucleotide polymorphisms, sometimes referred to as eSNPs. However, eQTLs can also be short (< 100 basepairs) insertions or deletions, or indels. The third, and rarest, type of eQTL is a structural variant or copy number variation.

- Tissue: blood, liver, ...

  eQTLs can be tissue-specific, or act in multiple tissues. While a considerable proportion of cis-eQTLs affect multiple tissues, trans-eQTLs are generally thought to be more tissue specific.

Objective

eQTL research has come a long way, but some practical problems when trying to work with these data still exist:

- Reproducibility

  Previously published eQTLs can often not be reproduced in similar studies. For example, a paper from 2012 comparing 4 liver eQTL studies[1] concluded that cis-eQTLs are relatively reproducible (67% overlap between two studies), while trans-eQTLs are not (only 6% overlap). Furthermore, the four studies that were compared are all eQTL studies available for human liver tissue back then.

- Data availability

  While authors try to make their data public, there is no consensus about the best way to do so for eQTLs. This has led to a number of databases (SCAN, seeQTL, Genevar, dbGaP & GETx) with a common limitation: all of them are restricted to their respective data sets. Researchers working in a specific field may find data that is relevant to them either scattered across multiple of these databases or not included at all.

The combination of these two problems makes it hard for researchers to come up with a comprehensive list of eQTLs that could be used for knowledge discovery, e.g., using machine learning techniques. We thus propose a workflow that is more flexible than relying on existing databases: The direct use of supplementary material published by the authors. This simplifies the combination of results from previous studies, as well as giving the researcher full control over which studies are included. We thus developed software demonstrating this process by processing data from 4 previously published eQTL studies.

	Westra et al.	Innocenti et al.	Schröder et al.	Kabakchiev et al.
Tissue	Blood	Liver	Liver	Intestine
Sample size	5311	206 + 60 + 266*	149	173
eQTLs considered**	923,021	12,480	1,201	15,091
VIP eQTLs***	668	5	14	27

Table 1. Data about eQTL studies that were included in this analysis.
* Innocenti et al. reported eQTLs from three different studies.
** the count of eQTLs in the considered supplementary material.
*** eQTLs that are associated with a VIP gene and pass a p-value threshold (10^-4 for Schröder et al., 10^-8 for others).

Methods

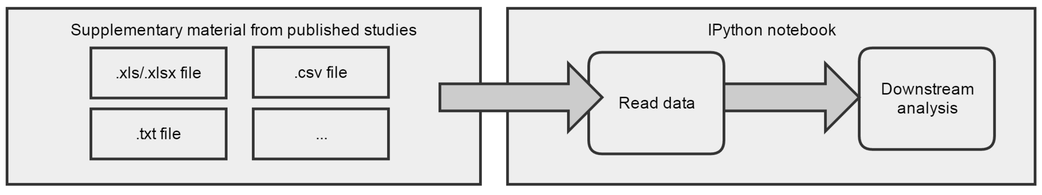

Figure 1. A possible workflow using the IPython notebook.

The software was written in Python and is available as an IPython notebook, which can be extended to include different data sources.

We downloaded the supplementary materials for the included studies and identified files with data relevant for our purposes. While different kinds of eQTLs exist, we have only included SNPs that act locally (cis-eSNPs), as these were available for all 4 studies. The most important values in these files are: the eQTL identifier, the gene symbol and the p-value.

These files are then parsed by the IPython notebook via a single function call that returns a pandas dataframe, which is suitable for downstream analysis in Python. That function and/or its parameters are the only thing that have to be adapted if one wishes to integrate additional data sources, for example from similar studies. We also added capabilities for unified data access and to filter the input lines, e.g. by only including eQTLs that pass a specified p-value cut-off or act on specific genes.

We analyzed the data from a Pharmacogenomics point of view, by incorporating data about PharmGKBs Very Important Pharmacogenes (VIP, a small set of genes that are important for drug metabolism) and analyzing only eQTLs that act on these genes, here referred to as VIP eQTLs.

Results

We created an IPython notebook to compare supplementary material from 4 different studies. As expected from different tissues and diverse sample sizes, the count of eSNPs identified varied between the studies. While none of the VIP eQTLs in the liver studies replicated in the other, after combining them to a single set, we found that some of them are shared by other tissues:

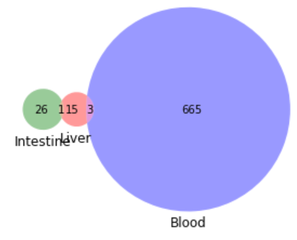

Figure 2. A venn diagram of eSNPs in different tissues.

Specifically, the SNPs in shared by blood and liver tissue are rs10489185 (in gene F5), rs407257 (GSTT1) and rs6591256 (GSTP1), whereas the SNP shared by intestine and liver tissue is rs2303222 (VKORC1). We also wanted to better understand how eQTLs are currently integrated into PharmGKB. We found that 191 out of the VIP eQTLs in blood, 11 in liver and 13 in intestine tissue are currently annotated in PharmGKB.

Finally, we are working on a new version of the IPython notebook, which includes improvements such as:

- A SnpDb class that stores and exposes the information
- Lazy access to the SnpDb objects
- A method-chaining approach that allows you to write code like this:

```
intestine_db = SnpDb(partial(read_xls, os.path.join(DBDIR, 'eqtl-intestine', 'mmc1.xls'))) \
        .rsid_column('SNP') \
        .gene_column('Gene') \
        .filter(partial(pass_below, filter=('P-value', 1e-8)))
```

This work was supported in part by the Austrian Science Fund (FWF): [P 25608-N15].

The analysis:
http://nbviewer.ipython.org/urls/gitlab.com/medication-safety/ms-ipython/raw/master/pharmacogenomic_eQTLs/pharmacogenomic_eQTLs.ipynb

The repository:
https://gitlab.com/medication-safety/ms-ipython/tree/master/pharmacogenomic_eQTLs