# GTEx analysis

June 19, 2024

## Contents



# 1 Introduction

The Adult Genotype Tissue Expression (GTEx) Project offers a large public resource of human RNA-seq data across tissues and individuals. This data offers an opportunity to identify tissue-shared and tissue-specific expression of genes, biological pathways/processes, and their transcriptional regulators.

We present a simple, baseline analysis to identify liver-specific pathways and transcription factors. Further analyses that should have been explored to make the analysis robust and exhaustive are presented in the discussion section.

The analysis is performed on a subset of tissues and protein-coding genes. We note that performing separate analysis for protein-coding genes can be motivated biologically and not solely by concerns of memory issues. For example, lncRNA have a different expression distribution than protein-coding genes and are sometimes treated as a separate modality (Malagoli et al. [2024]).

Performing analysis of highly heterogeneous GTEx data requires careful consideration. We start by a short literature review on possible concerns and analysis strategies for such data. We also quantitatively define terms such as "heterogeneous" and "specificity".

## 1.1 Quality Control and Normalization

The first step of our analysis is to perform quality control (QC) of the input data and normalization to eliminate technical variation. This includes:

1. Identifying mislabelled samples
2. Identifying contamination (e.g., in GTEx data, pancreas genes *PRSS1*, *CELA3A*, *PNLIP*, *CLPS* and esophagus genes *KRT4*, *KRT13* are expressed in other tissues (Nieuwenhuis et al. [2020]))
3. Filtering out samples of poor quality
    - In a tissue-aware manner
    - In a tissue-agnostic manner
4. Filtering out lowly expressed genes
    - In a tissue-aware manner
    - In a tissue-agnostic manner
5. Defining relevant groups of interest. In our case, this means choosing to group sub-tissues or keep them separate
6. Normalization
    - In a tissue-aware manner
    - In a tissue-agnostic manner

In our case, we won't address steps 1 and 2 in detail for the sake of time. We instead will rely on the way they have been addressed in prior literature analyzing GTEx data. For step 1, GTEX-11LO (not present in our samples of interest), underwent sex reassignment surgery and thus was assigned the wrong sex label (Paulson et al. [2017]). For step 2, a low-level contamination of some

genes should not significantly affect our analysis, under the hypothesis that it is at the same magnitude across tissues.

For step 3, we first rely on the GTEx-provided RNA Integrity Number to flag samples of poor quality. We also compute four quality metrics for each sample:

- Number of expressed genes
- Number of counts
- Percent of counts from mitochondrial genes
- Percent of counts coming from the 20 most expressed genes

Importantly, we can compute such metrics tissue by tissue rather than across tissues. This is important since distributions can vary between tissues, e.g., we find that heart samples have a much higher percent of counts from mitochondrial genes (10).

For step 4, we rely on the simple tissue-specific filtering procedure as defined and recommended by (Paulson et al. [2017]). "Tissue-agnostic" filtering removes genes with less than one count per million (CPM) in half of all samples. "Tissue-aware" filtering removes genes with less than one CPM in fewer than half of the number of samples of the smallest tissue.

For step 5, we rely once again on a simple grouping procedure by (Paulson et al. [2017]):

1. Group samples by tissue
2. Exclude the X, Y, and mitochondrial genes
3. Identify the 1000 most variable genes
4. Log2-transform raw counts
5. Check whether sub-tissues cluster separately on PCA or PCoA plots. If they don't form separate clusters, merge sub-tissues into a single tissue label

For step 6, the normalization strategy will depend on the downstream gene set identification approach.

- Using a "within sample" approach, clusters of genes and pathway activity are first defined separately for each sample, e.g., using WGCNA. For such within-sample analysis, TPM values are suitable. WGCNA was for example the approach used in the GTEx pilot study to construct coexpression networks and extract gene modules for each tissue. The modules are then used to identify active pathways across tissues (GTEx Consortium [2015]).

- Using a "between samples" approach, where samples' expression values are directly compared against one another using specificity metrics or differential expression. TPM values can't be used for such comparisons. Moreover, assumptions of usual normalization approaches might not hold (Paulson et al. [2017]). For example, that:
  - Most genes are not differentially expressed
  - Global differences in expression distribution are induced only by technical variation
  - The number and magnitude of up- and down-regulated genes are comparable

Rather than simple quantile normalization, smooth quantile normalization methods (e.g., qsmooth, SNAIL (Hicks et al. [2018], Hsieh et al. [2023])) have been developed to better compare different tissues from GTEx. They assume that the statistical distribution of each sample should be similar within a biological group, instead of globally across biological groups. The authors show this allows for better preservation of biologically known tissue-specific expression.

In our analysis, we choose to rely on smooth quantile normalization to enable direct comparisons across tissues and will not compare results with a within sample approache due to time constraints.

## 1.2 Definition of Tissue "Specificity" and Prior Expectations

After quality control and normalization, the next question is to quantitatively define tissue "specificity", which could be interpreted as:

- Strict presence or absence compared to other tissues (similar conceptually to the task of identifying "marker genes" (Pullin and McCarthy [2024])). Clearly, this second definition is not realistic to apply given expectations from prior literature (Aguet and Ardlie [2016]):

  > "Since most protein-coding genes appear to be expressed ubiquitously, tissue specificity is driven by the concerted, differential expression of multiple genes, or networks of genes, as part of specific cellular programs, and not the tissue-specific expression of a few select genes. [...] In general, gene expression varies far more across tissues than across individuals"

- A differential increase or decrease in activity compared to other tissues (testing with a one vs rest approach or pairwise), as measured by a statistical test or linear model (e.g., Wilcoxon, edgeR, limma, ...)

- A tissue specificity metric, such as those benchmarked in (Kryuchkova-Mostacci and Robinson-Rechavi [2017]). The authors divide methods into two categories: metrics that output a single number whether a gene is

tissue-specific or ubiquitous (Tau, Gini, TSI, Counts and Hg), and metrics that report for each tissue how specific the gene is to that tissue (z-score, SPM, EE and PEM). While we could use one metric of this second category, in our case, we are not looking for a full vector of tissue specificity but a single value for gene specificity to the liver. Moreover, these metrics do not report genes' statistical significance out of the box. Hence, we will opt for a simple approach based on differential expression.

### 1.3 Definition of Biological Pathway and Process

Next, we need to define the meaning of "pathway" or "biological process" given our goal to compare different tissues. Here we will not explore in detail which resource is most relevant but use one out of the box: the HALLMARK collection from the Molecular Signatures Database (MSigDB) (Liberzon et al. [2011]).

### 1.4 Definition of Transcriptional Regulators

Finally, we define transcriptional regulators based on CollecTRI (Müller-Dott et al. [2023]), a meta-resource containing signed TF-gene interactions for 1186 TFs extracted from multiple databases.

## 2 Results

Starting with raw counts from the GTEx website, we retained protein-coding genes and heart, kidney, liver, lung, muscle, pancreas, spleen, stomach, pituitary gland, and thyroid samples.

Subtissues were grouped for kidney and considered separately for heart based on scatterplot visualization as described in (Paulson et al. [2017]) (12, 13). Kidney Medulla sub-tissue was excluded by RNA integrity number quality control anyway, thus the decision to merge ended up having no impact.

Samples that passed both RNA integrity quality control and tissue-specific quality control were smooth quantile normalized (9).

Differentially expressed genes were defined based on a Wilcoxon and the default scanpy procedure, which uses a one vs rest approach (i.e., test liver vs all other tissues rather than pairwise testing of tissue) (1).

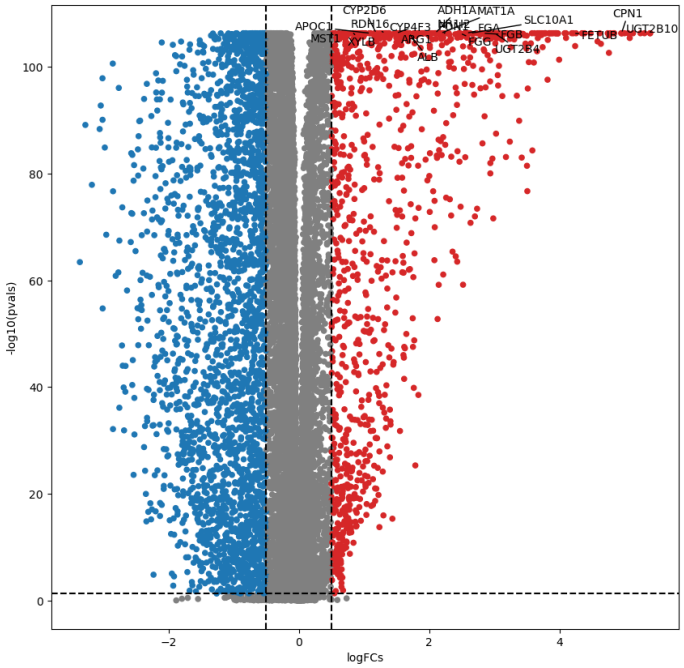


Figure 1: Volcano plot of DE genes found by Wilcoxon test of liver vs all other tissues considered

Wilcoxon test scores were given as input to decoupleR's univariate linear model (Badia-I-Mompel et al. [2022]) to predict transcription factor activities and p-values based on prior knowledge of the transcription factor (TF) to target network from collecTRI (Müller-Dott et al. [2023]) (3,3).

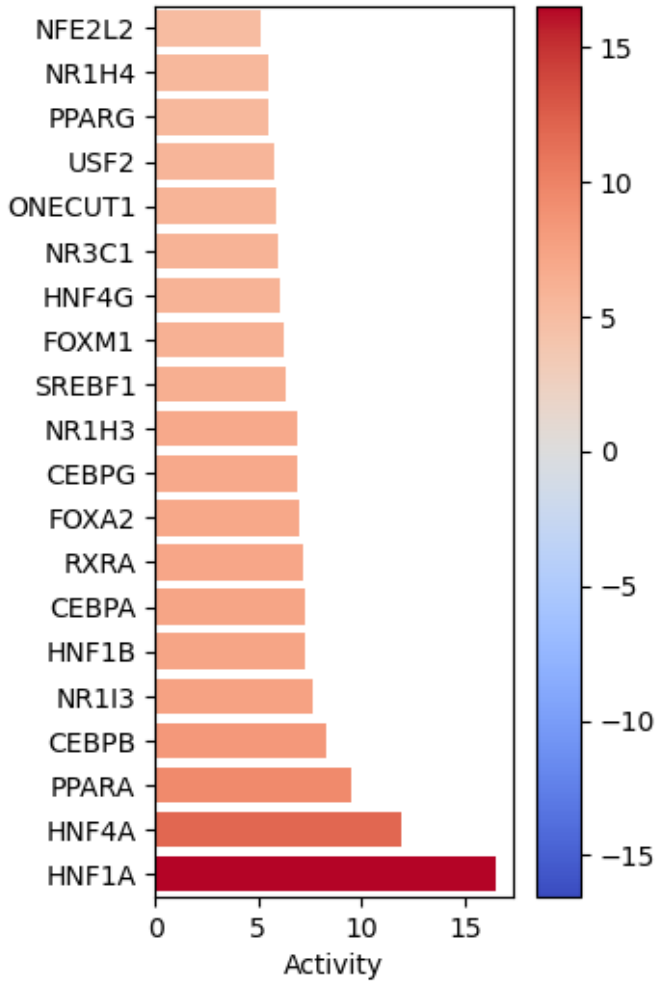


Figure 2: Top 20 significant TF activities found using decoupleR's univariate linear model and collecTRI TF-target database, fitted on Wilcoxon scores of liver vs all other tissues

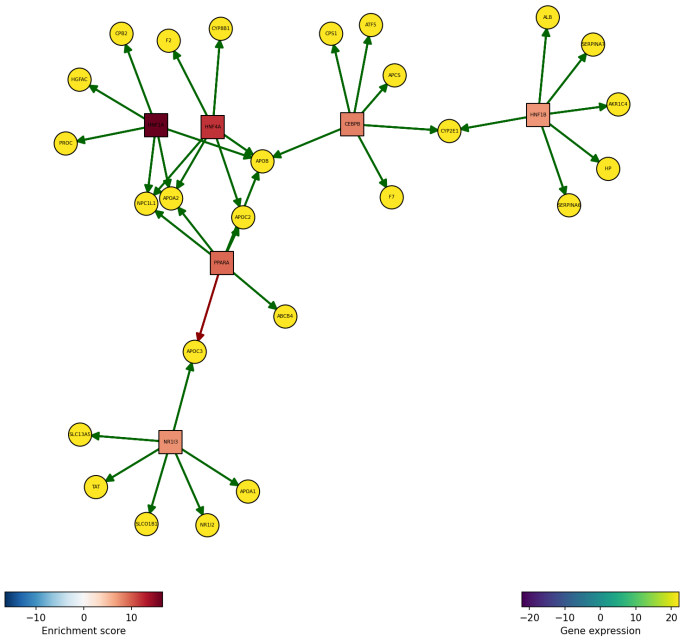


Figure 3: Network of top significant TFs and a few of their target genes. Colored by activities found using decoupleR's univariate linear model and collecTRI TF-target database, fitted on Wilcoxon scores of liver vs all other tissues

Wilcoxon test scores were also given as input for gene set enrichment analysis (GSEA) of terms in the MSigDB HALLMARK collection (4). Finally, to obtain TFs specific to a pathway, we simply use previously computed significant TF activities, subsetting to significant TFs connected to leading genes reported by GSEA for that pathway. For Xenobiotic metabolism (top GSEA pathway), this yielded largely unchanged results compared to previously identified non-pathway-specific significant TFs (5,6).

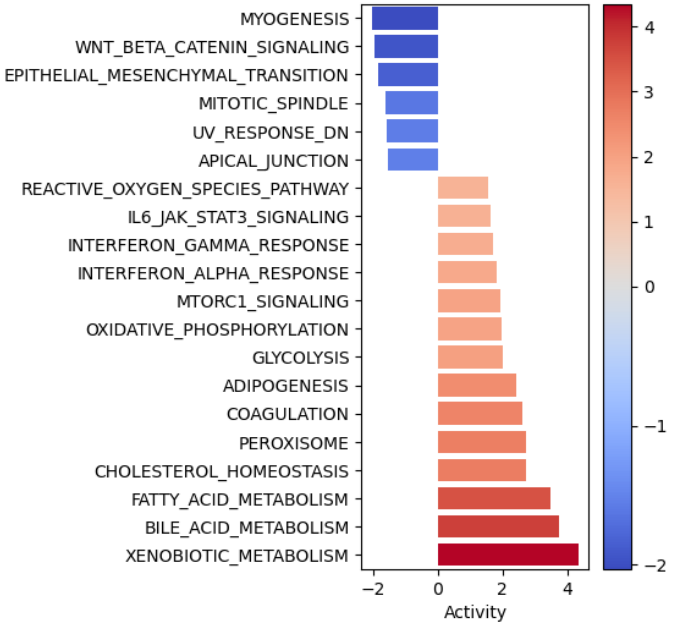


Figure 4: Top 20 pathways' normalized enrichment scores found using decoupleR's GSEA and MSigDB HALLMARK gene sets, with Wilcoxon scores of liver vs all other tissues as input

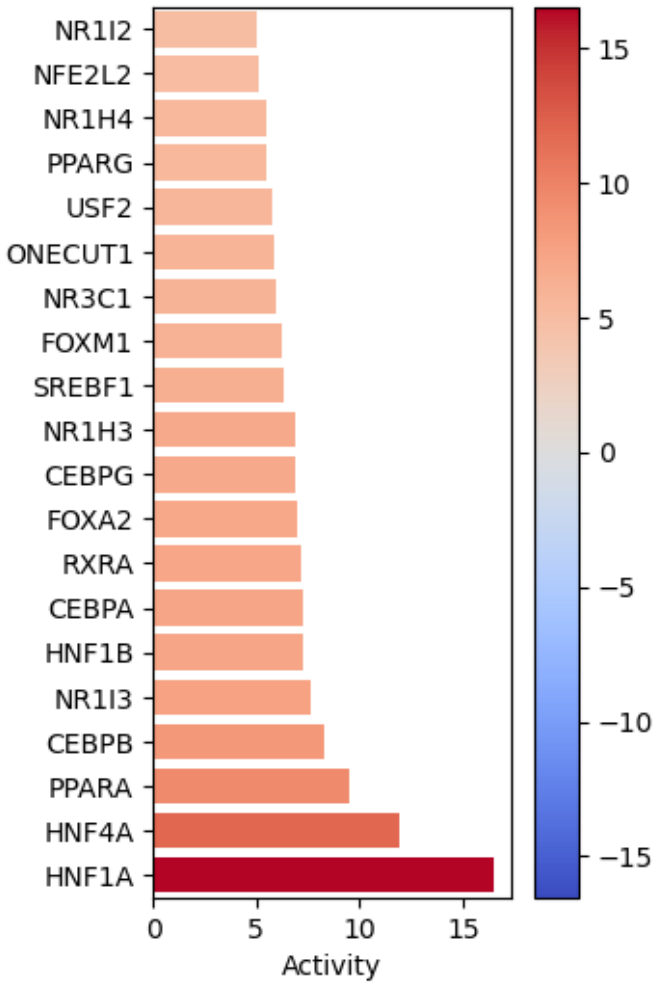


Figure 5: Xenobiotic metabolism (top GSEA pathway) top 20 TF activities found using decoupleR's univariate linear model and collecTRI TF-target database, fitted on Wilcoxon scores of liver vs all other tissues

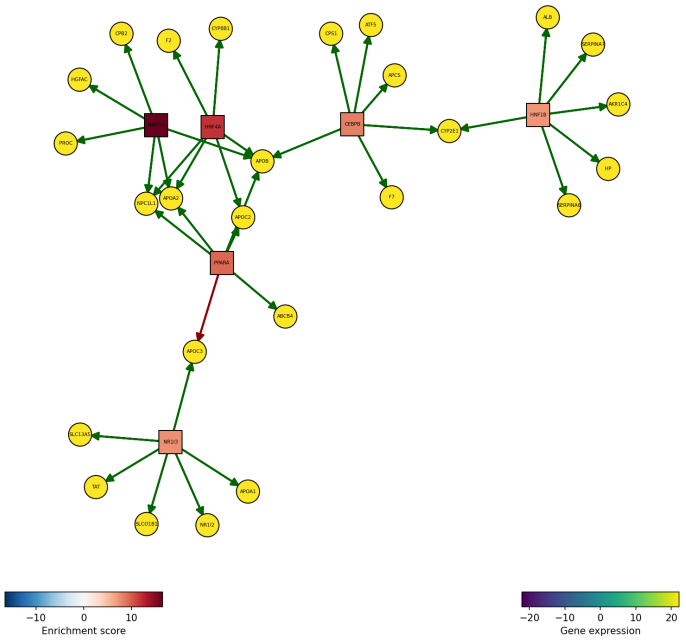


Figure 6: Xenobiotic metabolism (top GSEA pathway) network of top TF and a few of their target genes. Colored by activities found using decoupleR's univariate linear model and collecTRI TF-target database, fitted on Wilcoxon scores of liver vs all other tissues

# 3 Discussion

Our analysis presented a simple pipeline to identify liver-specific pathways and associated regulators. Although simple, top hits identified by this pipeline are TFs and pathways that make biological sense given functions of the liver.

Our analysis could have been further improved by:

- Being more comprehensive in testing combinations of normalization methods and metrics to define specificity or pathways. For example, we could also use an ensemble of DE methods with possible use of covariates, e.g. for patient or sex
- Testing single sample approaches mentioned in the introduction rather than directly comparing normalized samples.
- Testing for changes in gene co-expression rather than gene expression. This has been performed and reported to yield better tissue specificity in prior literature (Sonawane et al. [2017]).

# 4 Methods

## 4.1 Quality control and preprocessing

Data were processed using scanpy (Wolf et al. [2018]) and yarn (Paulson et al. [2017]). All functions were used with default parameters. Package versions necessary for reproducibility are provided on github in the .yml file.

We removed, in accordance with GTEx guidelines, samples with RNA Integrity Number below and samples with median absolute deviation 6 times above or below the median (only above for the mitochondrial filter) for any of the four quality metrics, computed separately for each tissue:

- number of expressed genes
- number of counts
- percent of counts from mitochondrial genes
- percent of counts coming from the 20 most expressed genes.

We performed "tissue-aware" filtering using yarn (Paulson et al. [2017]), removing genes with less than one CPM in fewer than half of the number of samples of the smallest tissue.

We merged or not sub-tissues based on the procedure by (Paulson et al. [2017]):

1. group samples by tissue.
2. exclude the X, Y, and mitochondrial genes
3. identify the 1000 most variable genes

4. log2-transform raw counts

5. check whether sub-tissues cluster separately on PCA or PCoA plots. If they don't form separate cluster, merge sub-tissues into a single tissue label

### 4.2 Differential expression test

We applied scanpy's Wilcoxon test to log1p transformed, smooth quantile normalized data to find DE genes for liver samples against samples from all other tissues grouped together.

### 4.3 GSEA and TF activity inference

We used decoupleR's implementation of GSEA and the univariate linear model for TF activity inference. All parameters were used in their default settings. Significant TFs were defined as those having p-value below 0.05.

### 4.4 Data availability statement

Data can be obtained by download of files from the provided GTEx website URL as well as GENCODE v26 for protein coding genes metadata. We also provide the data folder at [drive url] for reproducibility.

## 5 Supplementary figures

### 5.1 Distribution of samples retained

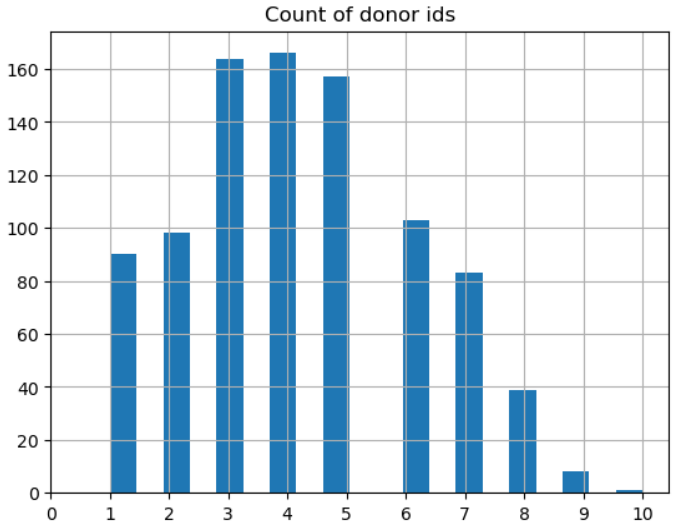


Figure 7: Count of donor IDs

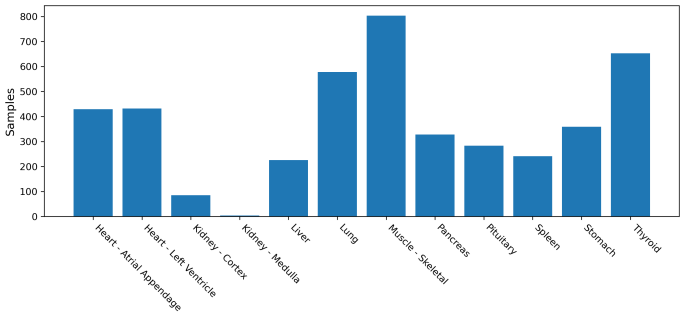


Figure 8: Number of samples per tissu before QC

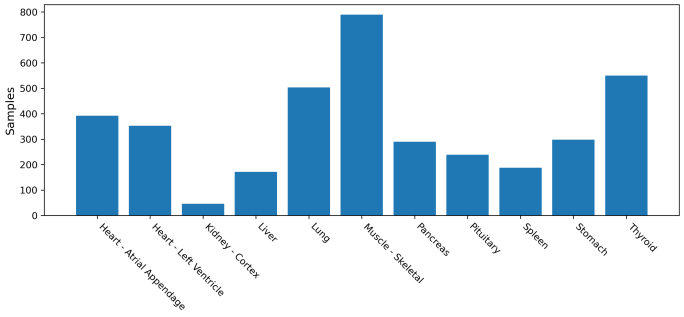


Figure 9: Number of samples per tissue after QC filtering

## 5.2 Global QC metrics

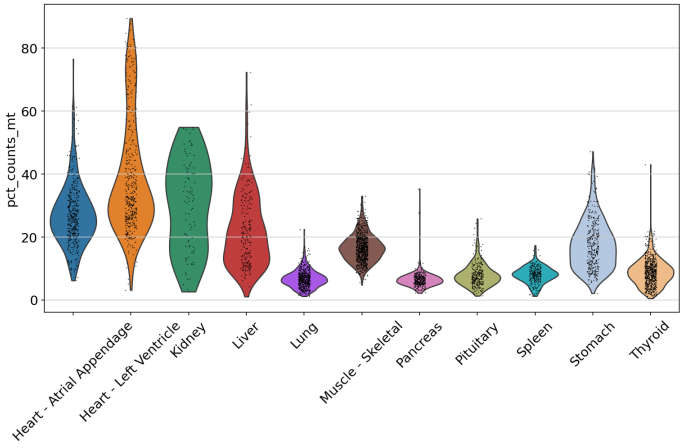


Figure 10: Example of one QC metric: violin plots of percent of mitochondrial genes for each tissue reveals tissue specificity

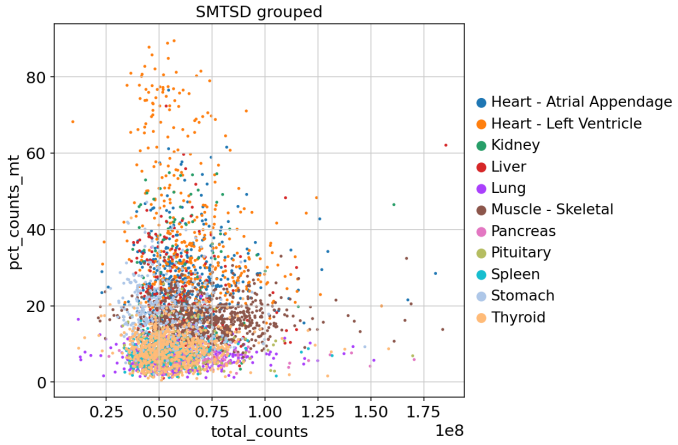


Figure 11: Heart Atrial Appendage QC

## 5.3 Scatter plots used to merge sub-tissues or not

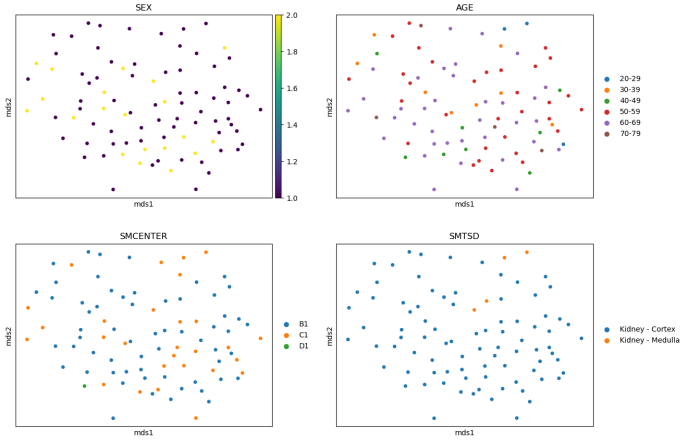


Figure 12: Kidney sub-tissue QC: they are not distinct so merging was chosen


Figure 13: Heart sub-tissue QC: they are distinct so merging was not chosen

## 5.4 Samples excluded by tissue-specific QC metrics

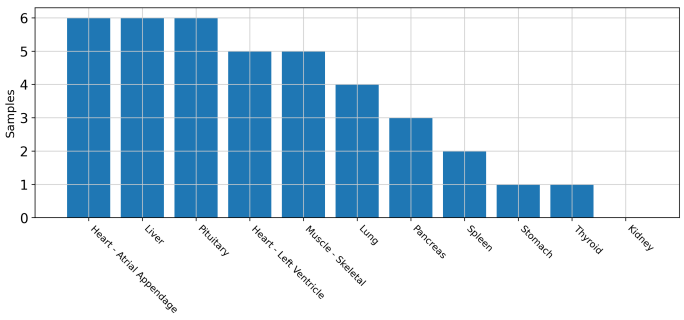


Figure 14: Samples excluded by tissue-specific QC metrics

## 5.5 Examples of tissue-specific QC metrics (for all organs go in figures folder)

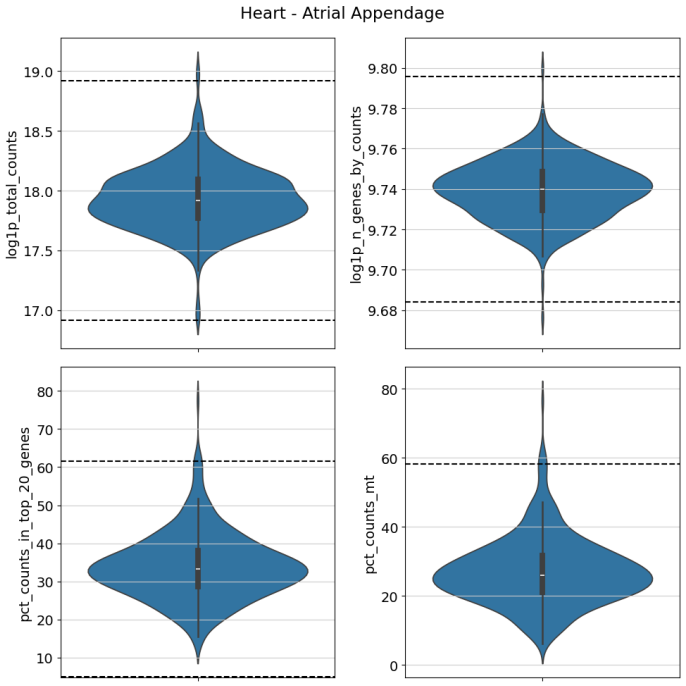


Figure 15: Heart Atrial Appendage QC

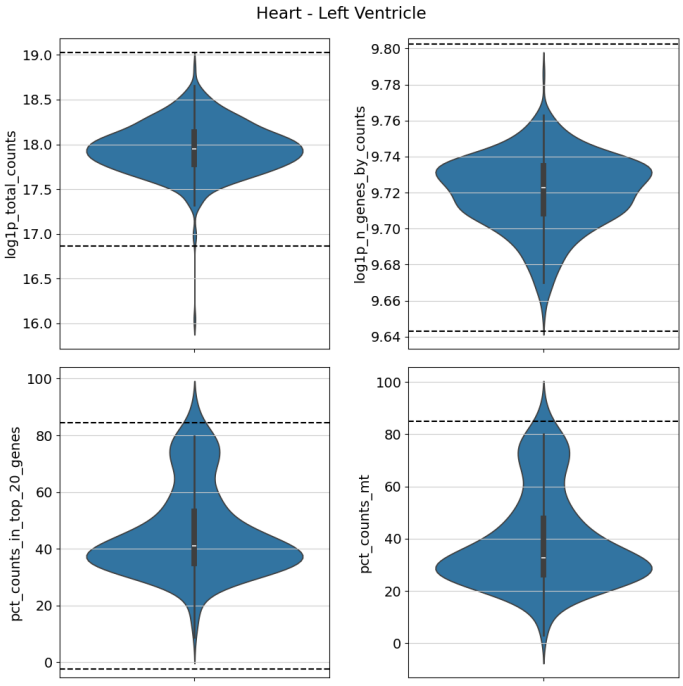


Figure 16: Heart Left Ventricle QC

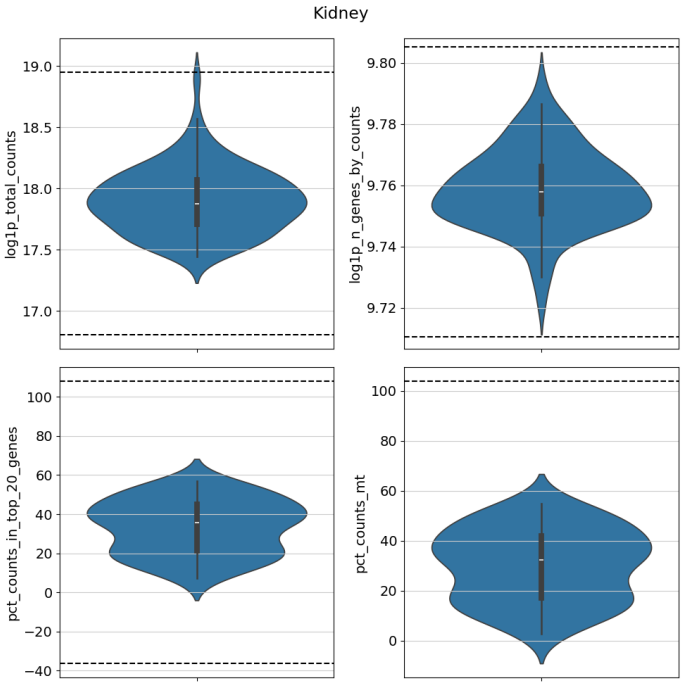


Figure 17: Kidney QC

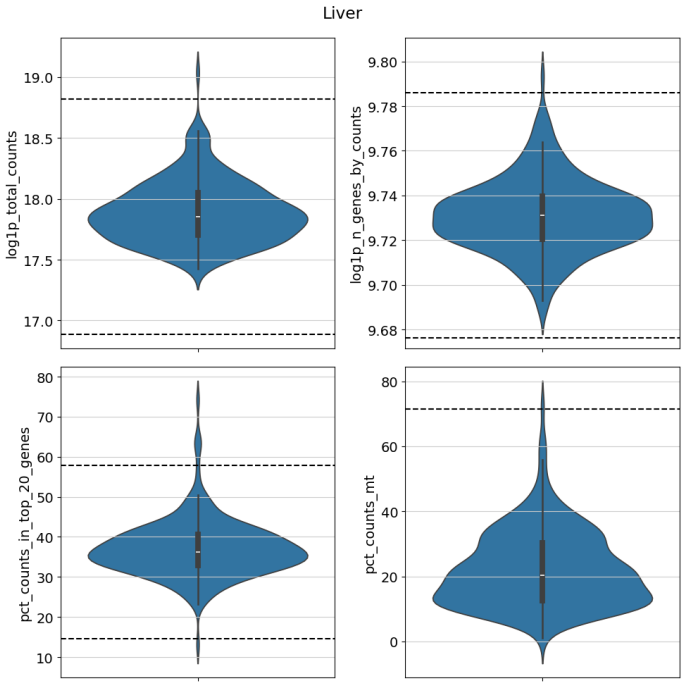


Figure 18: Liver QC

## 5.6 Examples of contaminating genes

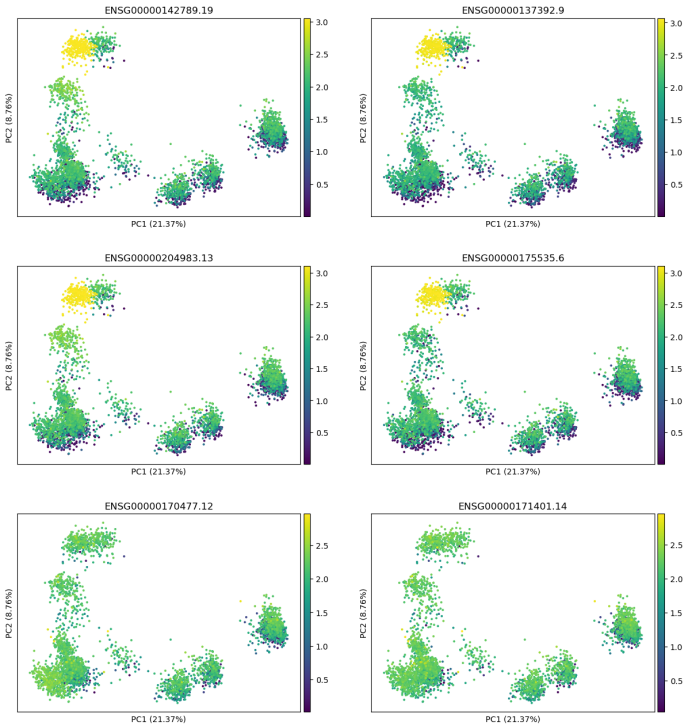


Figure 19: Examples of contaminating genes

## 5.7 PCA of smooth qunatile normalized data

# References

François Aguet and Kristin G Ardlie. Tissue specificity of gene expression. *Curr. Genet. Med. Rep.*, 4(4):163–169, December 2016.

Pau Badia-I-Mompel, Jesús Vélez Santiago, Jana Braunger, Celina Geiss, Daniel Dimitrov, Sophia Müller-Dott, Petr Taus, Aurelien Dugourd, Christian H Holland, Ricardo O Ramirez Flores, and Julio Saez-Rodriguez. decoupler: ensemble of computational methods to infer biological activities from omics data. *Bioinform Adv*, 2(1):vbac016, March 2022.

GTEx Consortium. Human genomics. the Genotype-Tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235): 648–660, May 2015.

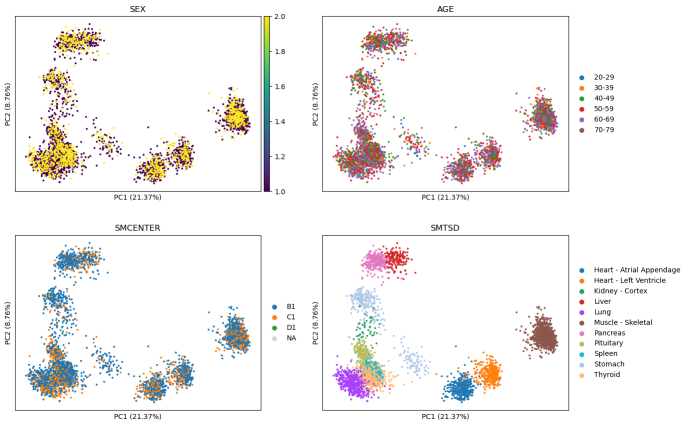


Figure 20: PCA of smooth quantile normalized data, colored by covariates

Stephanie C Hicks, Kwame Okrah, Joseph N Paulson, John Quackenbush, Rafael A Irizarry, and Héctor Corrada Bravo. Smooth quantile normalization. *Biostatistics*, 19(2):185–198, April 2018.

Ping-Han Hsieh, Camila Miranda Lopes-Ramos, Manuela Zucknick, Geir Kjetil Sandve, Kimberly Glass, and Marieke Lydia Kuijjer. Adjustment of spurious correlations in co-expression measurements from RNA-Sequencing data. *Bioinformatics*, 39(10), October 2023.

Nadezda Kryuchkova-Mostacci and Marc Robinson-Rechavi. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.*, 18(2):205–214, March 2017.

Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, June 2011.

Gabriele Malagoli, Filippo Valle, Emmanuel Barillot, Michele Caselle, and Loredana Martignetti. Identification of interpretable clusters and associated signatures in breast cancer single-cell data: A topic modeling approach. *Cancers (Basel)*, 16(7):1350, March 2024.

Sophia Müller-Dott, Eirini Tsirvouli, Miguel Vazquez, Ricardo O Ramirez Flores, Pau Badia-I-Mompel, Robin Fallegger, Dénes Türei, Astrid Lægreid, and Julio Saez-Rodriguez. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Res.*, 51(20):10934–10949, November 2023.

Tim O Nieuwenhuis, Stephanie Y Yang, Rohan X Verma, Vamsee Pillalamarri, Dan E Arking, Avi Z Rosenberg, Matthew N McCall, and Marc K Halushka.

Consistent RNA sequencing contamination in GTEx and other data sets. *Nat. Commun.*, 11(1):1933, April 2020.

Joseph N Paulson, Cho-Yi Chen, Camila M Lopes-Ramos, Marieke L Kuijjer, John Platig, Abhijeet R Sonawane, Maud Fagny, Kimberly Glass, and John Quackenbush. Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. *BMC Bioinformatics*, 18(1):437, October 2017.

Jeffrey M Pullin and Davis J McCarthy. A comparison of marker gene selection methods for single-cell RNA sequencing data. *Genome Biol.*, 25(1):56, February 2024.

Abhijeet Rajendra Sonawane, John Platig, Maud Fagny, Cho-Yi Chen, Joseph Nathaniel Paulson, Camila Miranda Lopes-Ramos, Dawn Lisa DeMeo, John Quackenbush, Kimberly Glass, and Marieke Lydia Kuijjer. Understanding Tissue-Specific gene regulation. *Cell Rep.*, 21(4):1077–1088, October 2017.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, February 2018.