Tissue heterogeneity is prevalent in gene expression studies (Supplementary Information)

Gregor Sturm, Markus List, Jitao David Zhang

2020-11-03

Contents

1 About this document

This document is supplementary information for

> Gregor Sturm, Markus List and Jitao David Zhang. Tissue heterogeneity is prevalent in gene expression studies

This document demonstrates step-by-step how we generated and validated tissue signatures, curated data and derive our results.

The source code and instructions how to reproduce this study are available from GitHub.

2 Validating Tissue Signatures

The authors of *BioQC* have taken three independent approaches to show that their signatures are valid and biologically meaningful (Zhang et al. 2017). (1) They checked the results for batch effects using surrogate variable analysis (SVR), (2) they ensured that the signatures are biologically meaningful by relating them to biological knowledge, and (3) used an independent method to derive the signatures, which yields comparable results.

However, they do not quantify the performance of the signatures using standardized performance measures, such as sensitivity and specificity. In principle, one can think of three methods to achieve such a validation (Gönen 2009): (1) internal validation, *i.e.* using the same data for generating and testing signatures, (2) split-sample validation, *i.e.* dividing the dataset in a test and training dataset and (3) independent-sample validation, *i.e.* using entirely unrelated samples for training and testing. While method (2) might be acceptable if sufficient data is unavailable, only method (3) can ensure that the signature does not reflect experimental biases.

To address this, we independently derived signatures on the GTEx dataset using *gini-index* as described previously (Zhang et al. 2017) and performed both a 10-fold cross validation on the same dataset and a cross-species, cross-platform validation on the mouseGNF GeneAtlas. To this end, we developed the python package pygenesig, a framework to create and validate signatures.

In this chapter, we

- perform a 10-fold cross-validation on the GTEx dataset, calculating the precision and recall for each signature.
- perform a cross-species, cross-platform validation of the signatures generated on the GTEx dataset
- identify a set of tissues, that can be reliably and unabmigously identified with the *BioQC* method.

2.1 Data

- The **Genotype Tissue Expression (GTEx)** project is a comprehensive resource of tissue-specific gene expression data. We use this dataset to derive tissue-specific signatures. The data is human only and was generated using Illumina sequencing.
- We use the **GNF Mouse GeneAtlas V3** as a control dataset to demonstrate that the gini-method is robust over multiple platforms and species. This dataset originates from mouse and was generated using the *Affymetrix Mouse Genome 430 2.0 Array (GPL1261)*.

2.2 Cross-Validation of signatures on the GTEx dataset

We use pygenesig to create and validate signatures on the GTEx v6 dataset. The data preparation steps are performed using these jupyter notebooks. The output of *pygenesig* can be viewed here. Below, we summarize the prodecues described in these documents.

We obtained the gene expression data and sample annotation from the GTEx portal. We collapsed gene expression data by HGNC symbol, aggregating by the sum. We aggregated samples of the same tissue by median. We generated signatures based on gini index as described in the BioQC paper. In brief, we calculated gini index for each gene across all tissues. Genes with a gini index >= 0.8 and expression of >= 5 TPM were added to the signatures of the 3 tissues with their highest expression.

Next, we performed a 10-fold cross-validation as follows: We split samples in 10 stratified folds, *i.e.* samples from all tissues are equally distributed across all folds. We use 9 folds to generate signatures using gini index. These signatures were applied to the remaining fold using BioQC. We iterated over the folds such that each fold has been used for training and testing.

The following heatmap shows the average BioQC score over all folds for each signature and each tissue.

As identifying contaminated/mislabled samples can be boiled down to a classification exercise, we are interested in the predictive performance of each signature. The following heatmap shows the confusion matrix of using the signatures for classification. A sample is considered as classified as a tissue, if the corresponding signature scores highest among all other signatures.

2.3 'Reference' Signatures

From the above matrices we learn that, while the vast majority of signatures yield a high score in the corresponding tissue, an unambigous classification of tissues is only viable for a subset of tissues. For instance, the different brain regions are hard to distinguish, and so are physiologically close tissues (e.g. large and small intestine).

Here, we reduce the dataset to a subset of tissues, which can be unambigously distinguished using the BioQC method (*i.e.* precision = recall = 1.0).

We manually map the tissues from GTEx to a reduced subset of tissue names. The results are available in this jupyter notebook and summarized below.

Again, the following heatmap shows the confusion matrix.

All tissues have been correctly identified at Precision = Recall = 1.0.

2.4 Cross-Platform Cross-Species Validation

Arguably, in the above experiment, we could have built signatures based on human-specific genes, genes that can only be detected by a certain experimental platform, or even experiment-specific batch effects instead of universally translatable marker genes.

To asses if the signatures translate across species and platforms, we tested the signatures generated above (human, Illumina sequencing) on the *mouseGNF tissue expression atlas* (mouse, Affymetrix microarray). The procedure is described in this notebook.

The following figure shows the score matrix of GTEx signatures against mouseGNF samples:

The signatures *Brain, Heart, Kidney, Liver, Skeletal Muscle, Pancreas, Skin* and *Testis* identify the respective tissue despite the species and platform differences at a high (>5) BioQC score.

As expected *Heart* and *Skeletal muscle* also identify each other, however *Heart* scores still higher on heart samples and *Skeletal muscle* scores higher on skeletal muscles samples, therefore we retain both signatures.

Figure 1: cross-validation of GTEx tissue signatures. Signatures are shown on the y-axis, the corresponding groups of samples on the x-axis. The tile shading indicates the average BioQC score of a signature on a group of samples. For better visibility of low scores, the colors are saturated at 30.

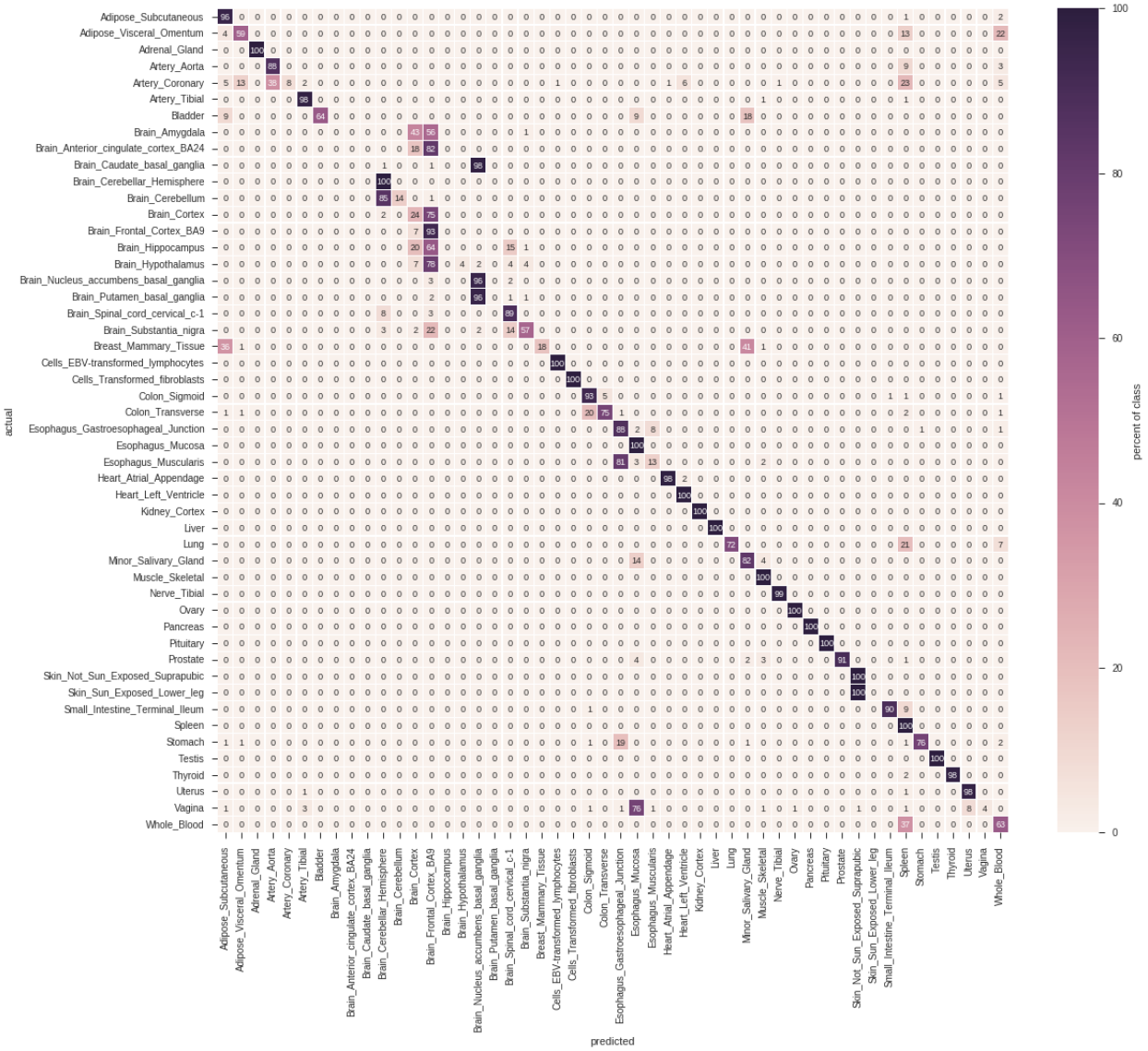

Figure 2: Confusion matrix of the cross-validation.

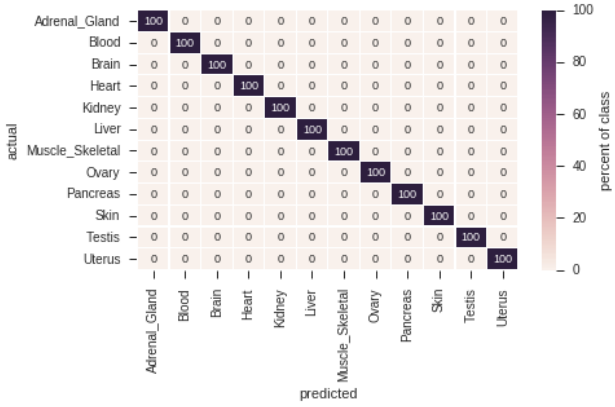

Figure 3: Confusion matrix of rosubst signatures.

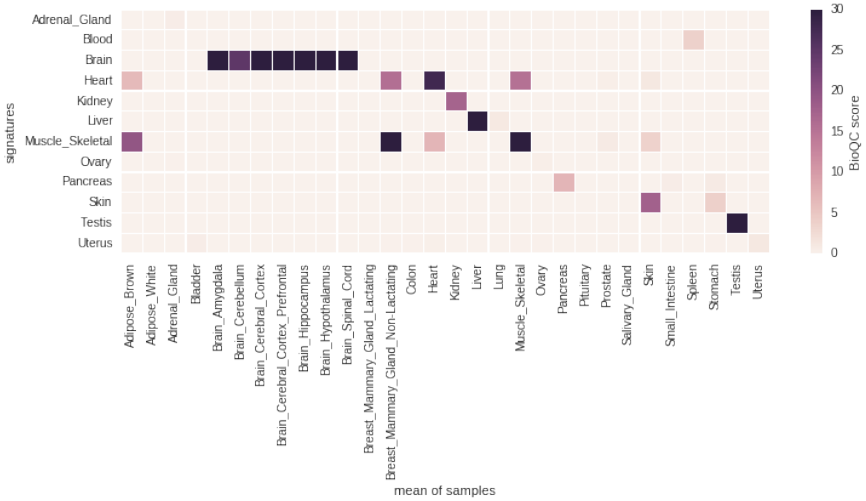

Figure 4: Cross-platform, cross-species validation of the robust signatures identified in the previous step.

Surprisingly, *Adrenal Gland*, *Ovary* and *Uterus* are not able to identify the respective samples, despite having a high score in the cross-validation. We therefore exclude these signatures from the reference signature set.

Unfortunately, *Blood* was not profiled in the mouseGNF dataset. We keep the signature nonetheless as it does not trigger any false positives.

3 Sample data and metadata

In this section, we describe how we obtained and curated gene expression data and sample metadata.

3.1 GEO

3.1.1 Downloading GEO data

We retrieved sample metadata for GEO using the GEOmetadb package. We download the studies with GEOquery and store them as R ExpressionSet using the R script geo_to_eset.R. We used the `annotGPL=TRUE` option of GEOquery's `getGEO` function to obtain gene symbols for the studies, where available. Since the tissue signatures use human gene symbols, we added human orthologs for all mouse and rat samples using the ribiosAnnotation package.

3.1.2 Filtering GEO data

We filtered GEO samples by the following criteria:

1. the tissue or origin is annotated,
2. gene symbols are annotated,
3. the readout was performed on a single-channel microarray, and
4. the tissue could be mapped to our controlled vocabulary (CV).
5. We only retained samples from the three major organisms: human, rat and mouse.
6. We removed studies which have been normalized per-gene and where ubiquitous house-keeping genes were not expressed.
7. Finally, we only retained samples originating from tissues for which a *reference signature* is available.

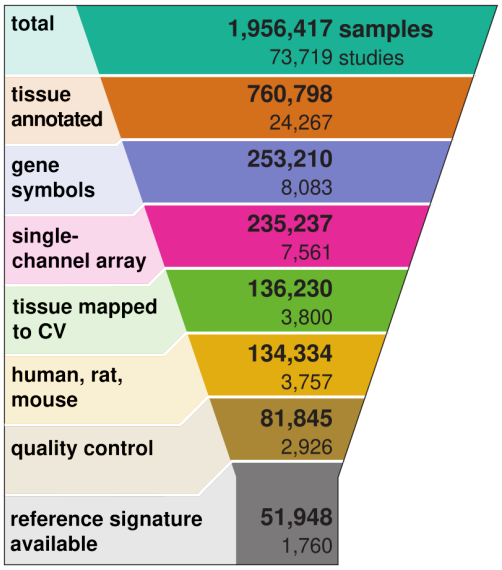

Figure 5: Summary of filtering steps on GEO samples

3.2 ARCHS4

In addition to GEO, we used data from ARCHS4, a publicly available data collection of annotaed, consistently processed gene expression datasets based on RNA-sequencing. We downloaded gene expression and metadata as `RData` objects from the ARCHS4 website (version 8.0).

We filtered samples by the following criteria:

1. The library is a transcriptomic cDNA library, the library strategy is RNA-seq, and either polyA or total RNA were extracted.
2. We excluded single-cell RNA-seq samples (none of the annotation fields may contain the keywords "single-cell", "single cell" or "smartseq")
3. At least 500,000 reads could be mapped to genes.
4. The tissue could be mapped to our controlled vocabulary (CV).
5. Finally, we only retained samples originating from tissues for which a *reference signature* is available.

Gene counts were normalized into TPM values before analysing them with BioQC.

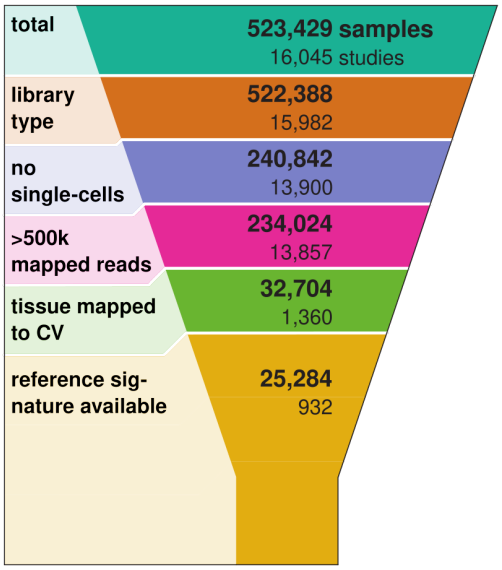

Figure 6: Summary of filtering steps on ARCHS4 samples

3.3 Normalize Tissue Names

The annotation of tissues is inconsistent within GEO. A "liver" sample can be termed *e.g.* "liver", "liver biopsy" or "primary liver". We, therefore, need a way to *normalize* the tissue name. We manually mapped the most abundant tissues to a controlled vocabulary in this Excel sheet.

In order to find out which samples show *tissue heterogeniety*, we first need to define which signatures we would expect in a certain tissue. We mapped signatures to the respective tissue type in this Excel sheet. For example, we map the signatures `Intestine_Colon_cecum_NR_0.7_3` and `Intestine_Colon_NR_0.7_3` to *colon*.

Since the "reference signatures" are not as specific as the annotation in the GEO, we created *tissue sets* to combine them into groups. For instance, it is hard to distinguish *jejunum* from *colon*, but easy to distinguish the two from other tissues. We therefore created a tissue set *intestine*, which contains both *jejunum* and *colon* and references all signatures associated with the two tissues. This information is part of the same Excel sheet.

Table 1: reference signatures

Reference Signature	Tissue
GTEX_Blood	blood
GTEX_Brain	brain
GTEX_Heart	heart
GTEX_Kidney	kidney
GTEX_Liver	liver
GTEX_Muscle_Skeletal	skeletal muscle
GTEX_Pancreas	pancreas
GTEX_Skin	skin
GTEX_Testis	testis

4 Testing for tissue heterogeneity

4.1 Tissue signatures

In section 2, we identified a set of 9 *reference signatures* (table 1) which unambiguously identify their corresponding tissue across platforms and species. In addition to that, we use 120 tissue signatures from the BioQC publication, which we refer to as *query signatures*.

4.2 Testing samples for heterogeneity

We tested for enrichment of 120 selected signatures provided by BioQC, the 9 reference signatures generated by us and one random control signature of 100 randomly drawn genes on all 76576 selected samples resulting in a list of 10031456 (sample, signature, pvalue) pairs.

Our intention is to identify samples that show *tissue heterogeneity*, *i.e.* unintentional profiling of cells of other origin than the target tissue of profiling. We classify samples into *heterogenous* and *not heterogenous*. We call a classification *true-positive* if the given sample is classified as *heterogenous* and the sample indeed contains cells different from the annotated tissues. Analogous, we call a classification *false-positive* if the given sample is classified as *heterogenous* but in reality only contains cells from the annotated tissue.

Naively, we would label a sample as heterogenous, if a signature unrelated to the annotated tissue exceeds a certain score. The problem with this approach is, that some signatures overlap; the resulting scores are therefore correlated and will lead to false-positives. One cannot simply solve this problem by excuding genes that are members of multiple signatures, as it is easily possible to build two (in fact many) distinct, non-overlapping signatures matching the same tissue.

In section 2 we haven created toroughly validated *reference signatures* for 9 tissues. Even though we have demonstrated that each signature unambigously identifies its corresponding tissue (*i.e.* scores highest), the signatures could still be correlated. Some of them in fact are, e.g. cardiac muscle and skeletal muscle (see figure 7). Moreover, we lack sufficient data to perform an independent-sample validation on the signatures provided by BioQC. Accounting for correlation with the reference signatures, we can still avoid false-positives, even if a signature is characteristic for a tissue histologically close to the annotated tissue.

We therefore take the following approach to account for the correlation of signatures: A given sample s annotated as tissue t is tested for enrichment with signature k_{test} resulting in a p-value p_{test}. Let k_{ref} be the reference signature associated with tissue t and p_{ref} the p-value of testing s for enrichment of k_{ref}. Let τ be a certain false discovery rate (FDR)-threshold.

(1) If the Benjamini-Hochberg (BH)-adjusted $p_{\text{test}} \geq \tau$, we assume that s is not heterogenous; else continue.

(2) We fit a linear model using `rlm` from the `R MASS` package of $|log10(p_{\text{test}})|$ against $|log10(p_{\text{ref}})|$ for all samples annotated as t. We assume that the residuals R of the linear model follow a normal distribution $R \sim \mathcal{N}(0, \sigma^2)$, where σ is the standard deviation of the residuals.
(3) We extract the residual r corresponding to sample s. We calculate the p-value $p_{\text{corr}} = 1 - \text{CDF}_R(r)$ where CDF_R is the cumulative density function of $\mathcal{N}(0, \sigma^2)$.
(4) If the BH-adjusted $p_{\text{corr}} < \tau$, we reject the hypothesis that k_{test} is enriched only due to correlation and label the sample as heterogenous.

4.2.1 Determining τ.

We desire an overall FDR of 0.01. Since we test in a two-step procedure, we choose a threshold τ, such that the overall FDR equals 0.01.

$$\text{FDR} = \tau + (1 - \tau)\tau$$

Explanation: Of all positives in step 1, there are, per definition, a fraction of τ false-positives and $1 - \tau$ true-positives. To all positives, the step 2 test is applied. In addition to the false-positives from step 1, there is a fraction of $\tau(1 - \tau)$ false-positives among the true-positives from step 1.

Solving the equation yields:

$$0.01 = \tau + (1 - \tau)\tau \Leftrightarrow 0 = \tau^2 - 2\tau + 0.01$$

```
tau = polyroot(c(FDR_THRES, -2, 1))[1]
tau
```

```
## [1] 0.005012563-0i
```

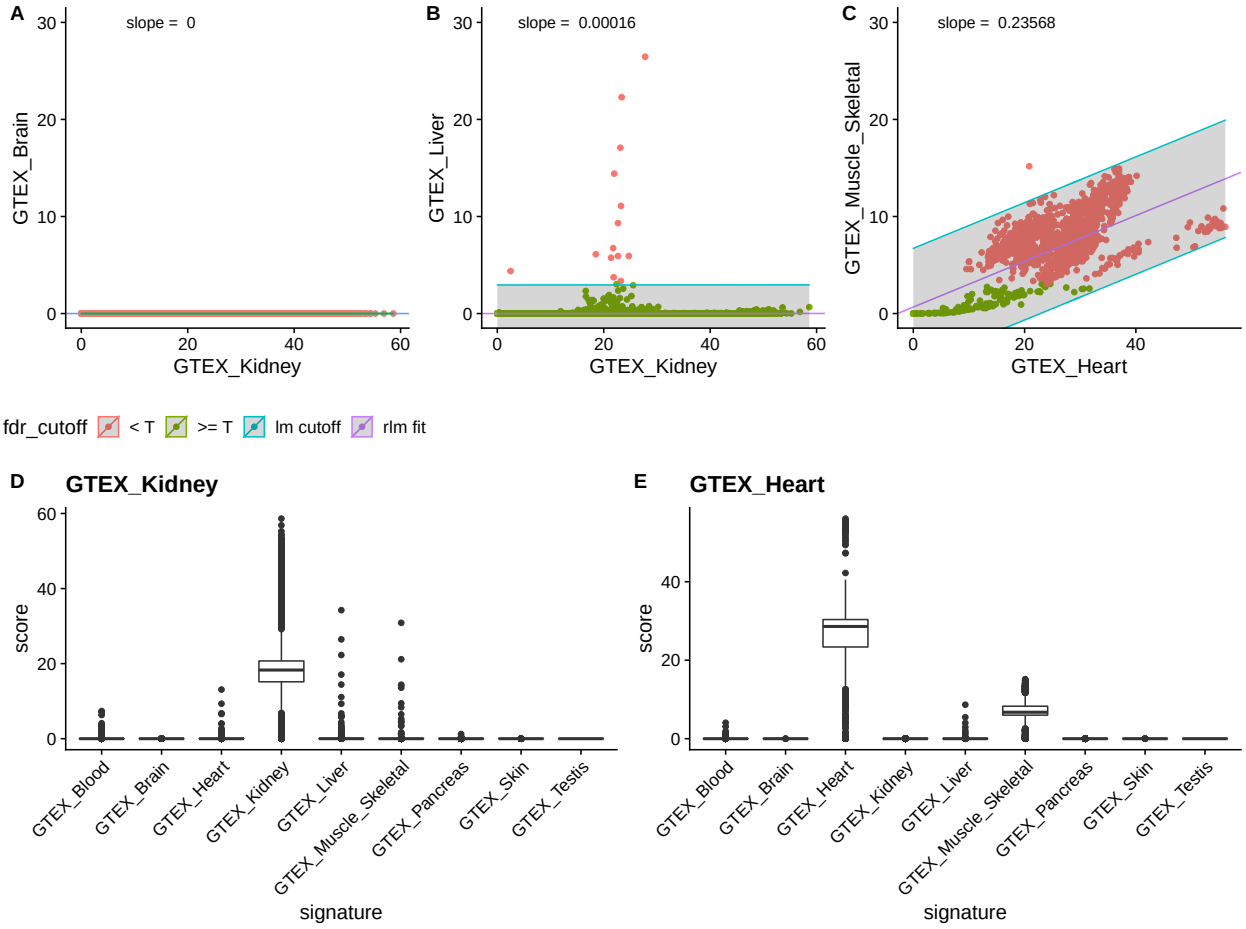

Figure 7: Examples of signature correlation. Panels A-C: correlation of the signature scores (y-axis) against the scores of a reference signature (x-axis). The purple line indicates the model fitted to the data. Points falling into the shaded area can be explained by the linear model. Points are colored according to whether they would be discarded already in the FDR-filtering step. (A) Brain scores of kidney samples against the scores of the kidney signature. There is virtually no correlation, nor are there outliers. Appraently, no kidney samples are contaminated with neuronal cells. (B) Liver scores of kidney samples against scores of the kidney signature. The samples are not correlated, however some outliers are detected which are samples potentially containing liver cells. (C) Cardiac muscle scores of skeletal muscle samples against skeletal muscle scores. The scores are highly correlated. While most of the poinst exceed the FDR threshold, they will not be classified as outliers because the elevated score can be explained with the correlation. Panels D and E show the boxplots of the scores of various signatures on kidney and heart samples respectively.

5 Results

5.1 Heterogeneity using GTEx signatures only

See figure **??**

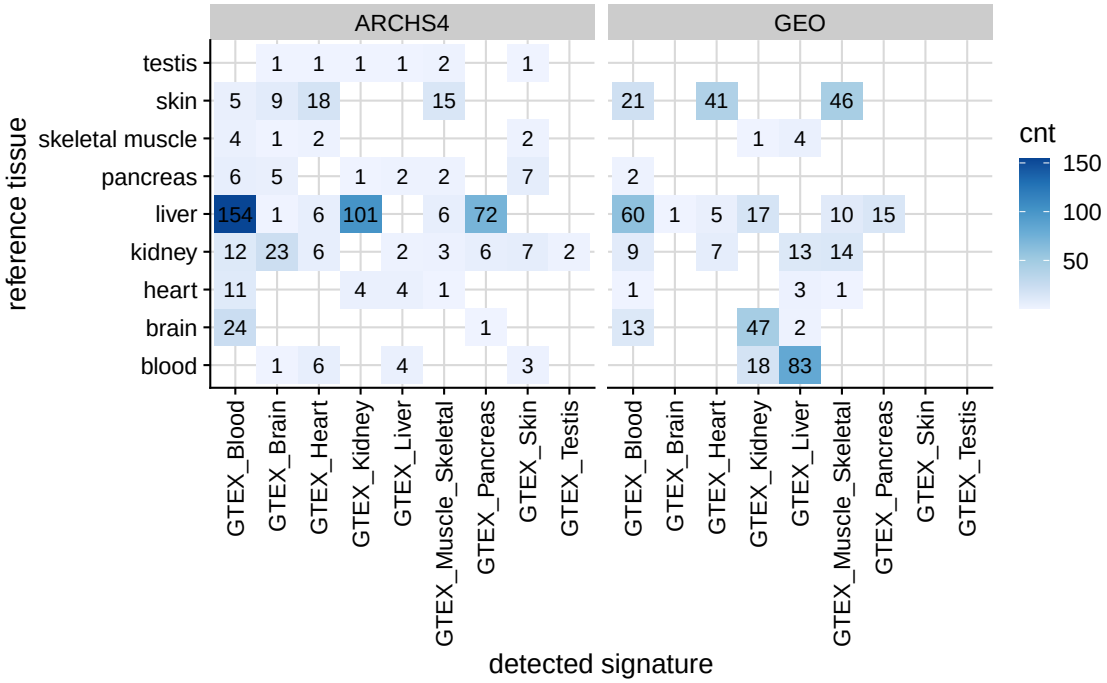

Figure 8: Tissue heterogeneity assessed with the reference signatures. The annotated tissues are listed in rows, the significantly enriched signatures in columns. If a signature has been found to be significantly enriched in a sample, the sample will count towards the number indicated in the matrix. All contaminations per sample are included, i.e. a sample can appear multiple times in a row.

5.2 Heterogeneity using all signatures

See figures 9, 10, 11, and 12.

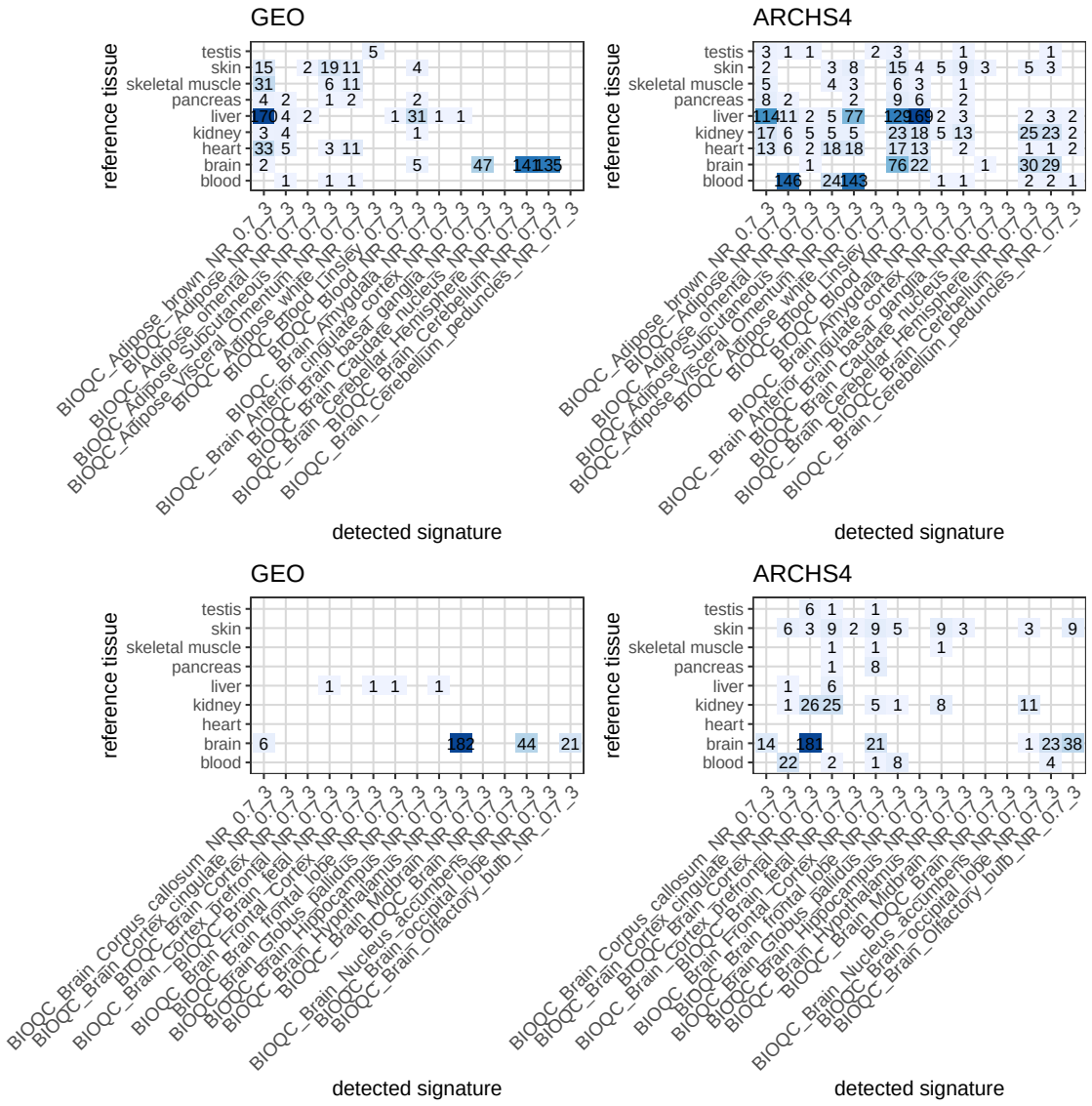

Figure 9: Tissue heterogeneity assessed with the BioQC signatures. The annotated tissues are listed in rows, the significantly enriched signatures in columns. If a signature has been found to be significantly enriched in a sample, the sample will count towards the number indicated in the matrix. All contaminations per sample are included, i.e. a sample can appear multiple times in a row.

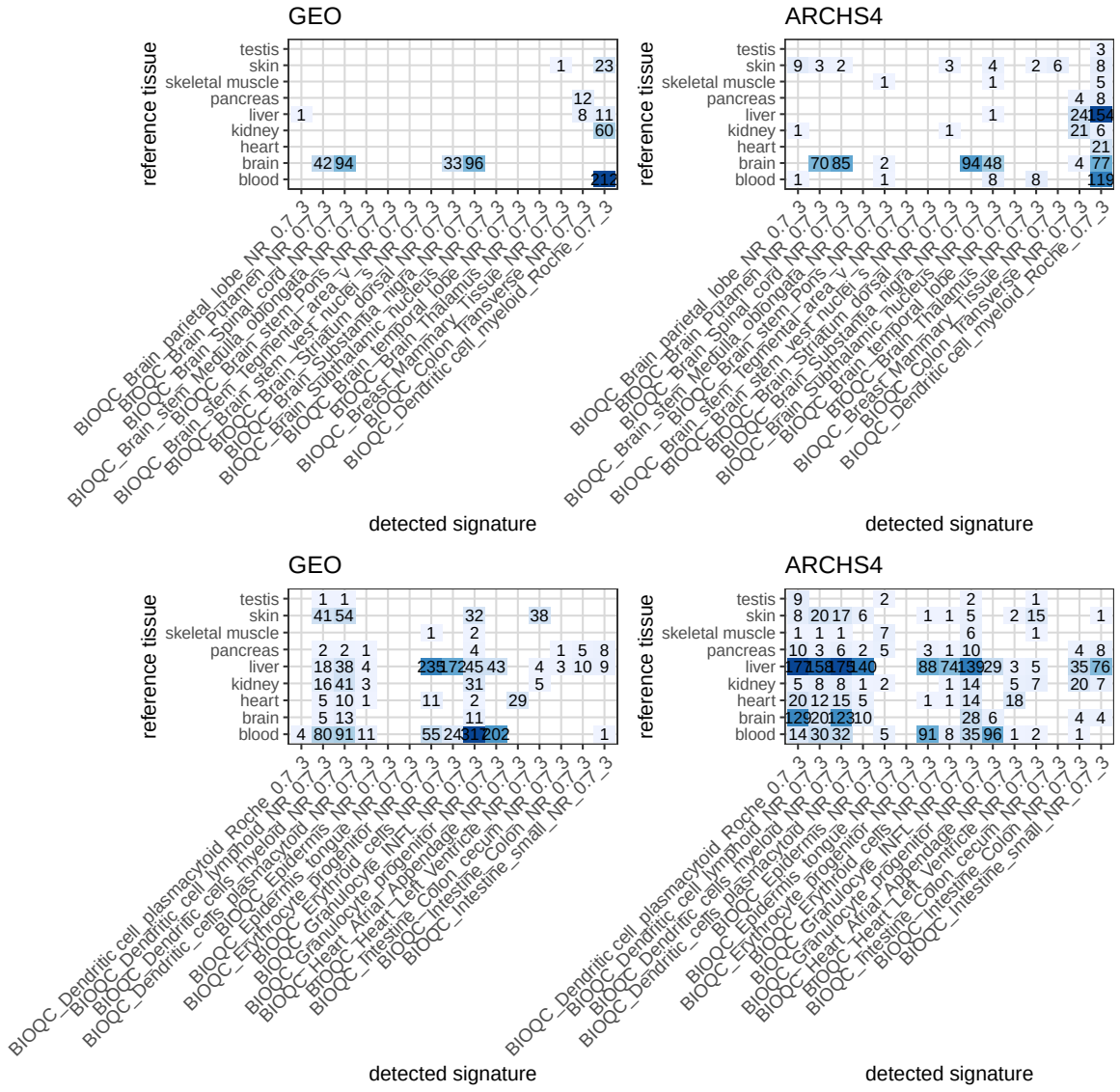

Figure 10: Tissue heterogeneity assessed with the BioQC signatures (continued)

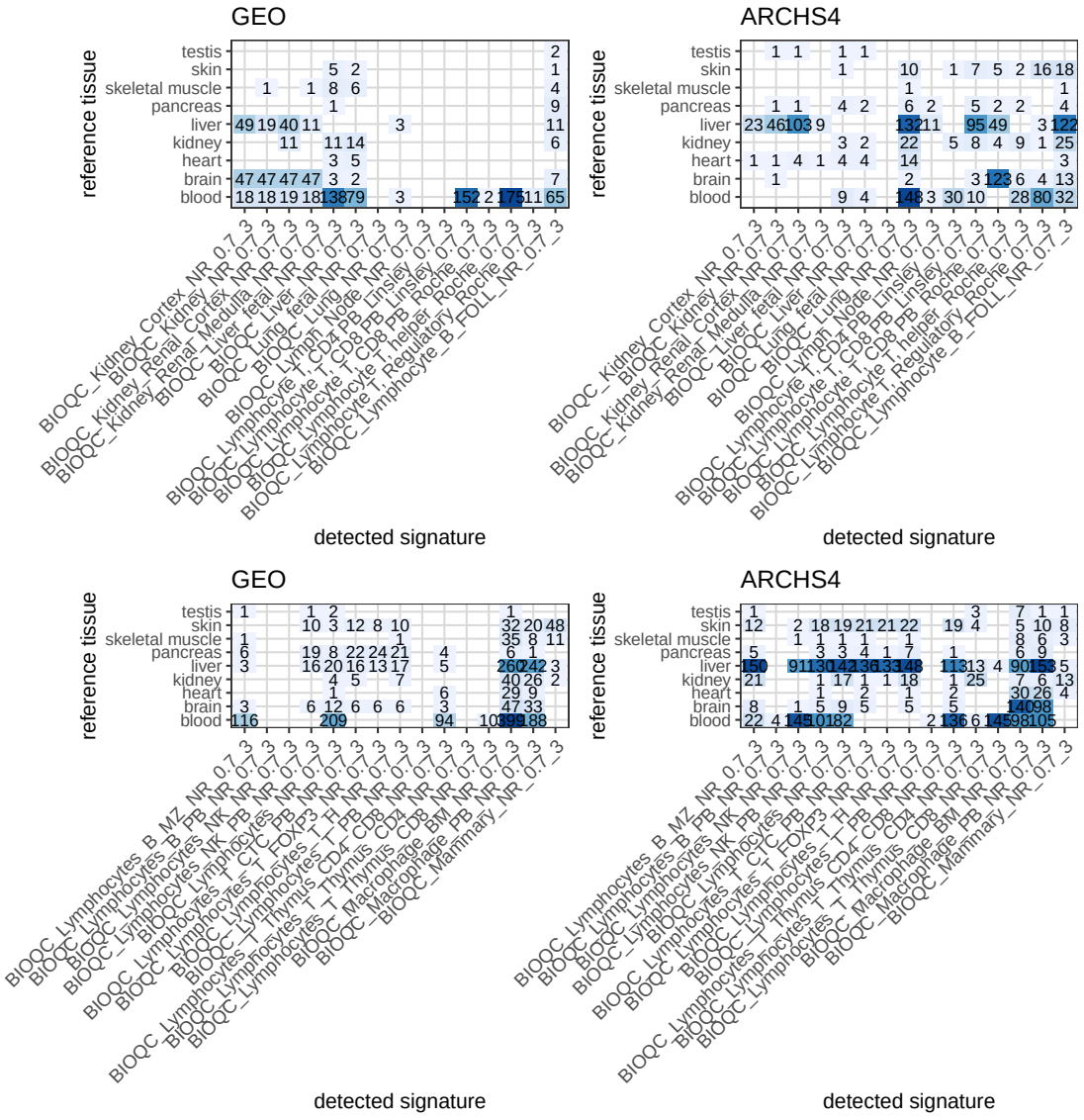

Figure 11: Tissue heterogeneity assessed with the BioQC signatures (continued)

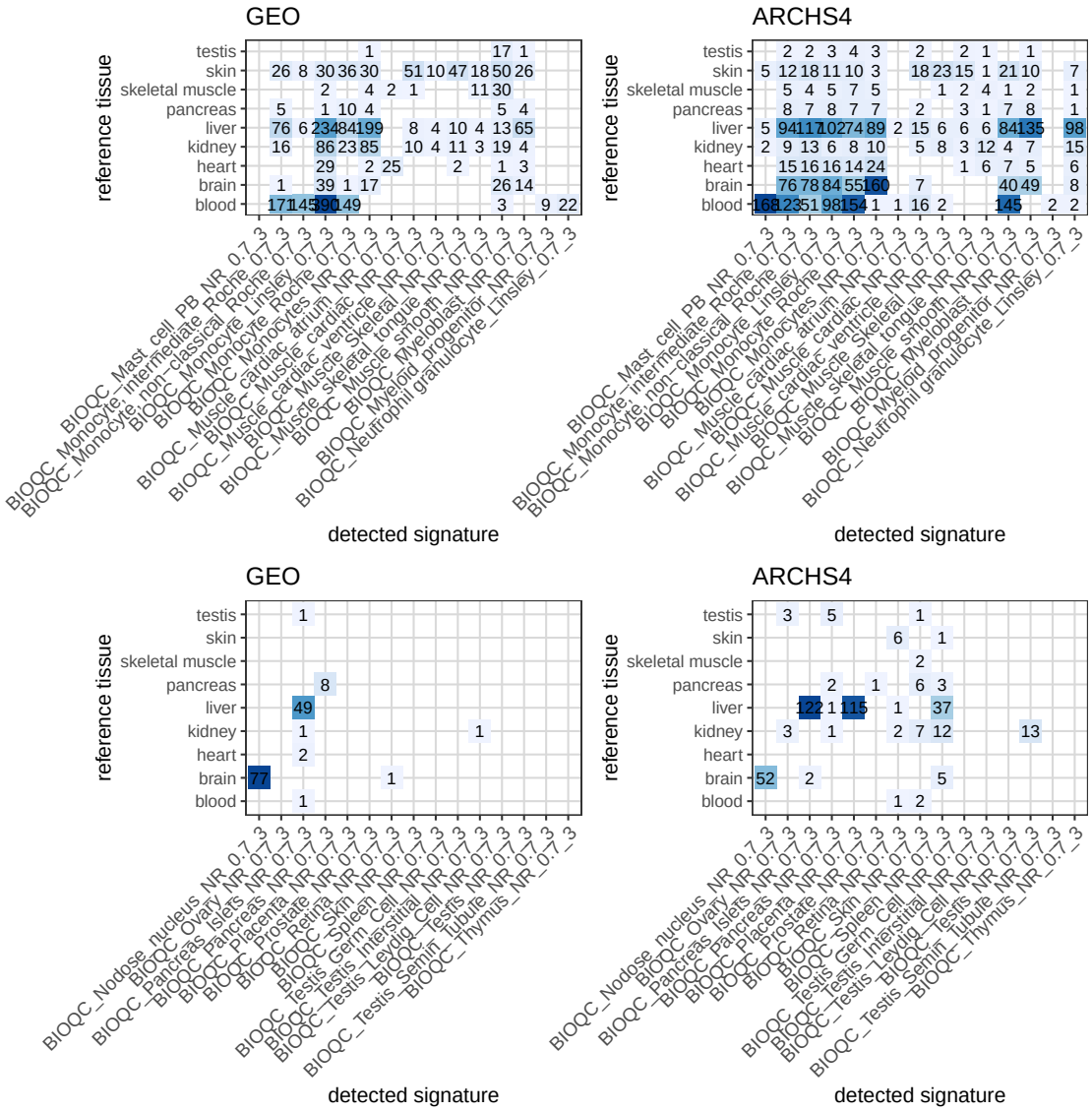

Figure 12: Tissue heterogeneity assessed with the BioQC signatures (continued)

6 Figures for publication

Same as in the section before, but all bioqc signatures aggregated by tissue groups. Additional, we define samples as being "severely heterogeneous", if the reference signature, i.e. the signature that should be present according to the annotation, is not enriched at an unadjusted p-value < 0.05.

We use bootstrapping (R package `boot`) to derive confidence intervals.

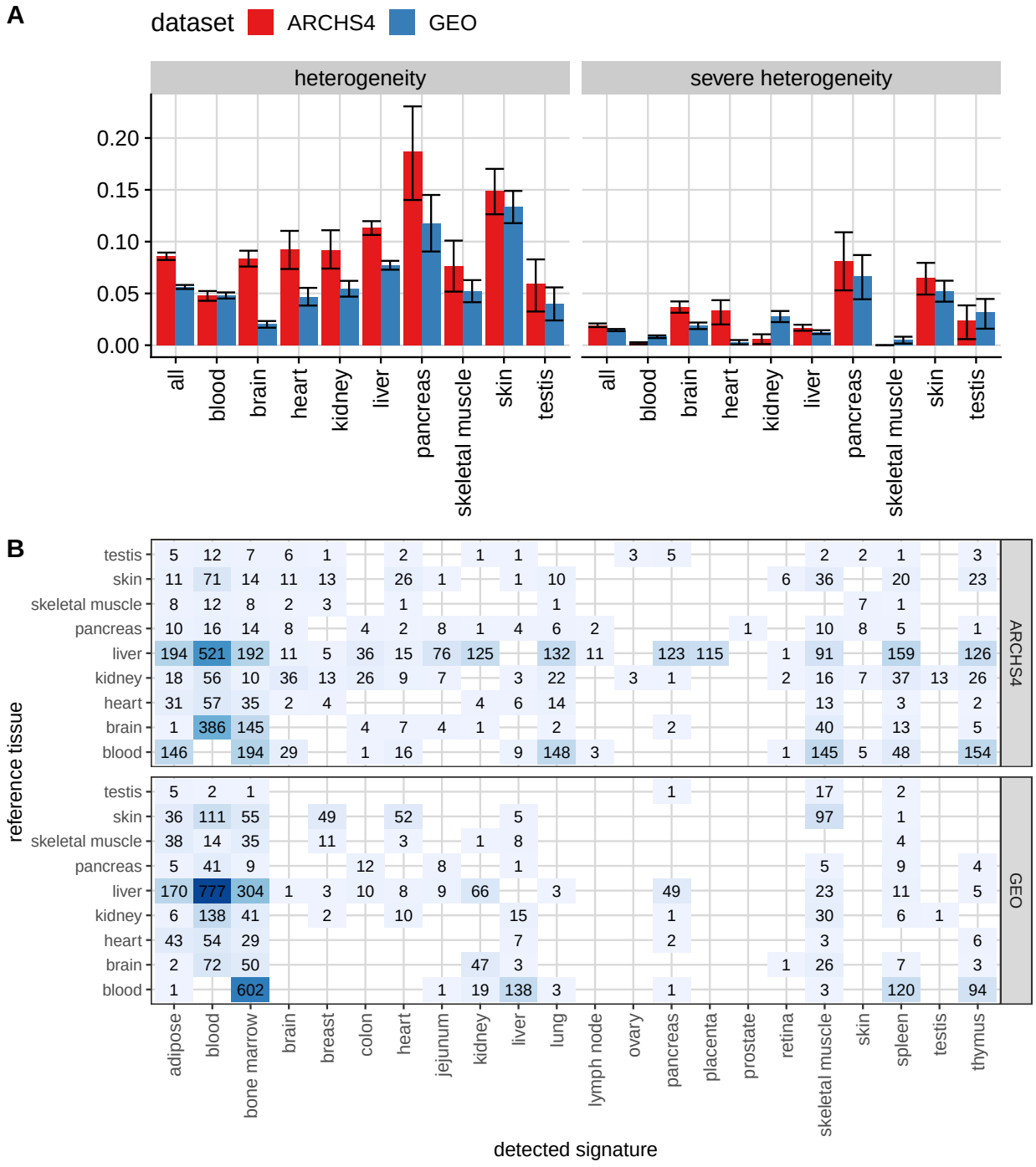

Figure 13: Main figure for paper. (A) Fractions of heterogeneous samples per tissue. (B) Sample confusion matrix.

7 References

Gönen, Mithat. 2009. "Statistical aspects of gene signatures and molecular targets." *Gastrointest Cancer Res* 3 (2 Suppl): 19–21. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2684735%7B/&%7Dtool=pmcentrez%7B/&%7Drendertype=abstract.

Zhang, Jitao David, Klas Hatje, Gregor Sturm, Clemens Broger, Martin Ebeling, Martine Burtin, Fabiola Terzi, Silvia Ines Pomposiello, and Laura Badi. 2017. "Detect tissue heterogeneity in gene expression data with BioQC." *BMC Genomics* 18 (1): 277. https://doi.org/10.1186/s12864-017-3661-2.