

GTEEx expression data

-

Rainet project

Diogo Ribeiro, Lionel Spinelli, Andreas Zanzoni, Christine Brun

February 16, 2016

1 Introduction

We plan to use the GTEEx V6 dataset (<http://gtexportal.org>), Human genome-wide RNA-seq expression data, to add RNA expression information in our RAINET database. The purpose is to use the expression to filter out Protein-RNA interactions that unlikely to occur in vivo. We will apply a threshold of expression value to ascertain the potential presence of each protein and RNA in a certain tissue, and excluded interactions where one or both of the interaction partners are missing. We will extrapolate the protein presence or absence by the expression of their respective mRNA. It as been shown (REF) that prediction of protein presence is accurate when its RNA is expressed above ..X.. rpkm.

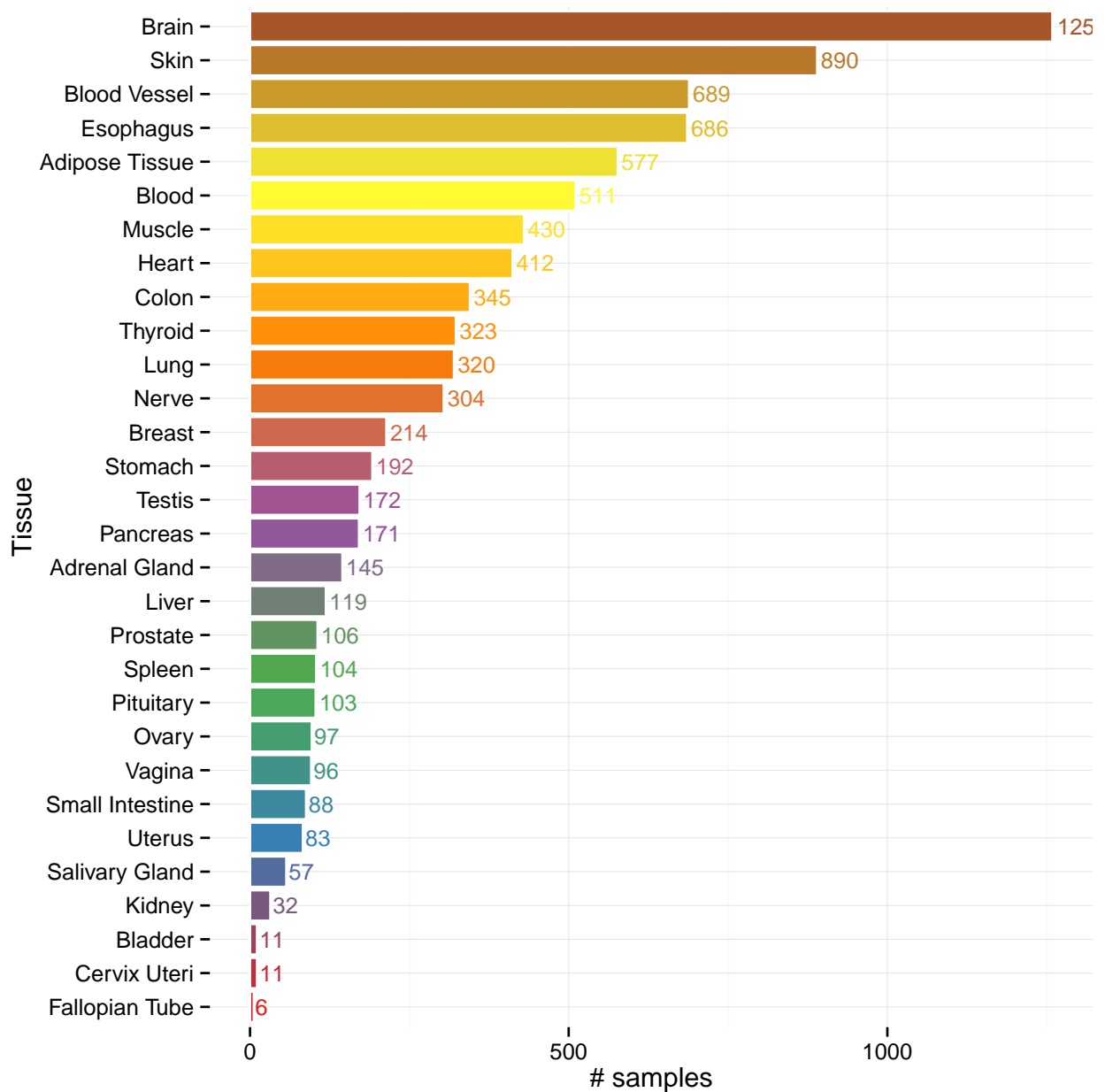
The GTEEx dataset contains full RNA-seq experiments done on hundreds of individuals, and retrieved from several physical locations / tissues. The GTEEx project provides downloadable data where the reads have been mapped to human genome and rpkm values were calculated for each GENCODE (v19) transcript (note that this includes all types of RNAs: mRNAs, lncRNAs, snoRNAs, etc). They provide several files, among them the description of the individual RNA-seq samples, including the tissue and body site: `GTEEx_Data_V6_Annotations_SampleAttributesDS.txt`.

The file with the rpkm values contains values for each transcript for each sample: `GTEEx_Analysis_v6_RNA-seq_Flux1.6_transcript_rpkm.txt`.

For insertion of expression data into our RAINET database we want to transform the GTEEx data to have a single expression value for each RNA-tissue pair, therefore, exclude the sample/individual dimension from the data. We will first perform a rough analysis of expression values distributions across samples to reach the best solution for transforming this data.

2 Samples variability per tissue

Number of samples per tissue using the whole `GTEEx_Data_V6_Annotations_SampleAttributesDS.txt` file with the attribute SMTS (the less-specific tissue terms).



```
## [1] "Total # of tissues : 30"
## [1] "Total # of samples : 8553"
```

We can see that number of samples per tissue is highly variable. Perhaps this should be considered in the further analysis.

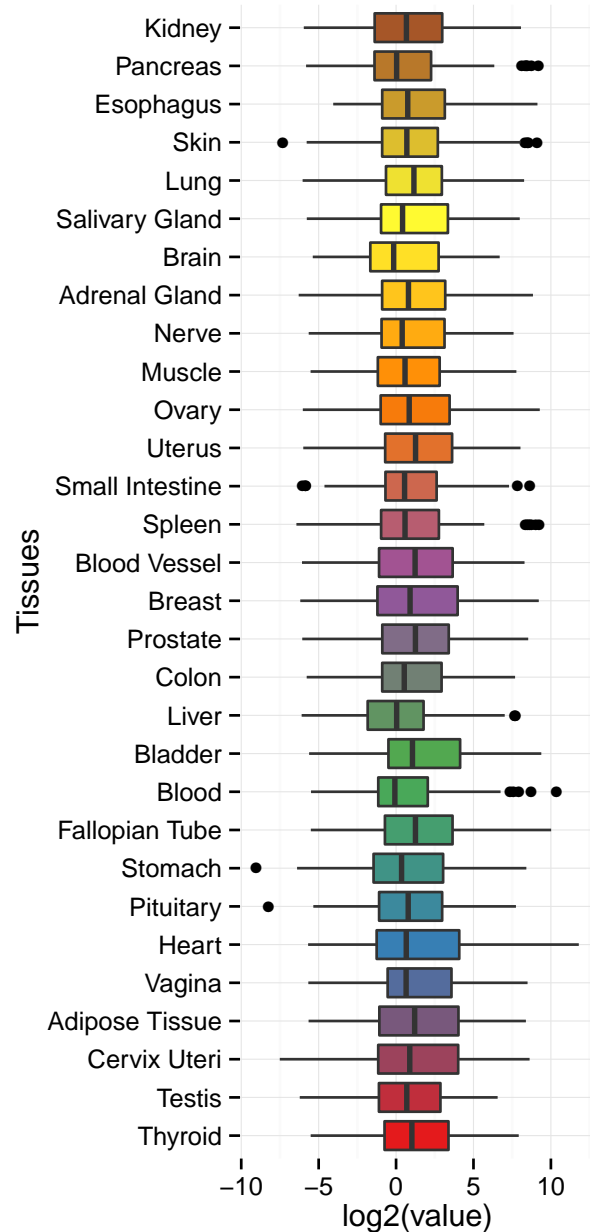
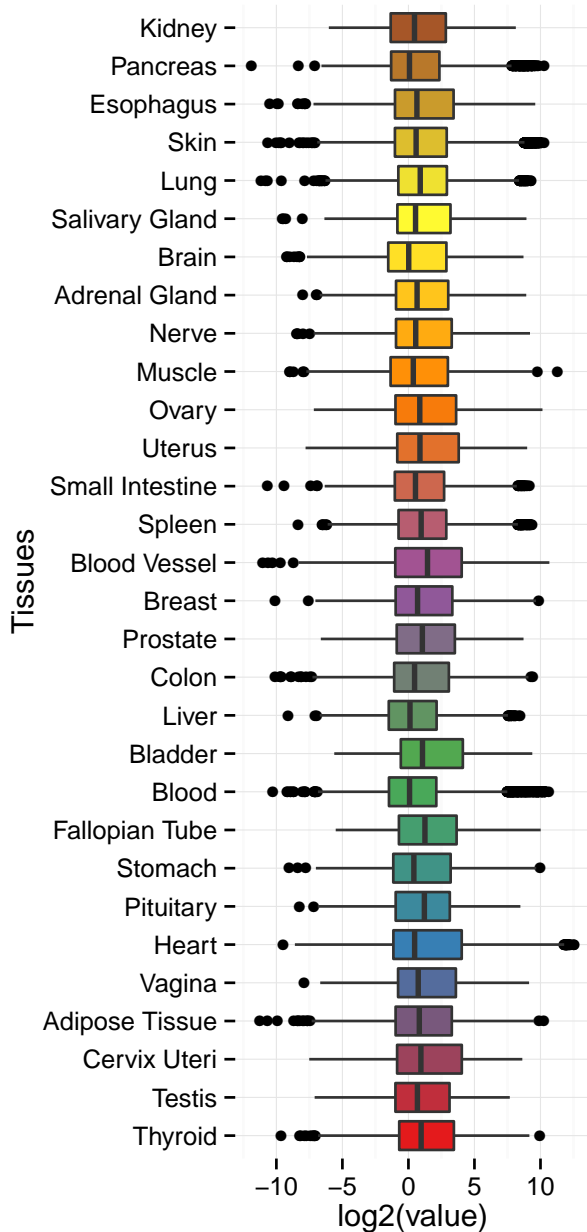
3 Expression variability per tissue

Distribution of all the expression values per tissue of 50 randomly sampled transcripts (out of 190 thousand), sampled from the whole GTEx rpkm file. Keep in mind that different tissues have different sample sizes, and thus a different amount of values are plotted on each tissue.

The purpose here is not to compare which tissue has higher or lower expression, or higher or lower variability in expression, but to understand the variability of expression between each tissue, coming from both biological variability and sampling variability.

The left-side of the following boxplot contains the distribution of expression values for the sample transcripts using all available samples (e.g. thousands in brain, a handful in Fallopian tube), whereas the right-side boxplot displays the data for the same transcript when only considering 6 (random) samples for each tissue.

```
## [1] "ALL SAMPLES: summary(expression_df$value[expression_df$variable == Brain]"
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.0000  0.0000  0.0000  3.8380  0.6706 419.8000
## [1] "ALL SAMPLES: summary(expression_df$value[expression_df$variable == Fallopian Tube]"
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00  0.00  0.06  13.61  2.49 1035.00  63903
## [1] "SAMPLED: summary(expression_df$value[expression_df$variable == Brain]"
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00  0.00  0.00  3.75  0.64  103.30  63903
## [1] "SAMPLED: summary(expression_df$value[expression_df$variable == Fallopian Tube]"
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00  0.00  0.06  13.61  2.49 1035.00  63903
```



We can see that the number of samples does not drastically change the overall distribution of expression per tissue, only that the number of extreme values (i.e. the black dots) are decreased when reducing number of samples. Thus, as we plan to use some averaging metric between samples of the same tissue for each transcript, we use a different number of samples per tissue, without impairing the credibility or comparability of the average expression value.

4 Expression distribution per transcript

Selected 5 tissues with different sample numbers, randomly selected 12 transcripts

