GTEx expression data

-

Rainet project

Diogo Ribeiro, Lionel Spinelli, Andreas Zanzoni, Christine Brun

March 3, 2016

1 Introduction

We plan to use the GTEx 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.

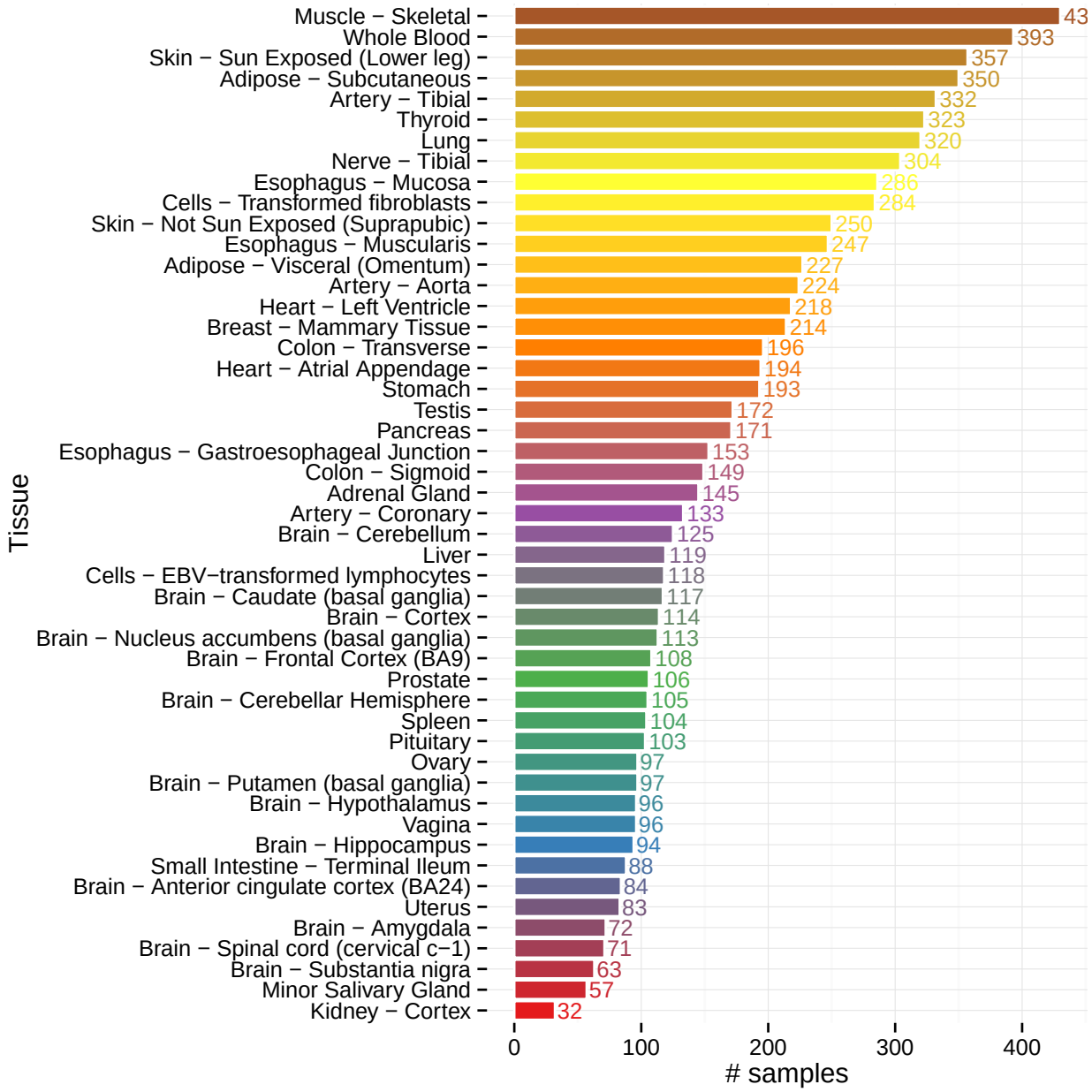
The GTEx dataset contains full RNA-seq experiments done on hundreds of individuals, and retrieved from several physical locations / tissues. The GTEx project provides downloadable data where the reads have been mapped to human genome and rpkm values were calculated for each GENCODE (v19) transcript (note 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: GTEx_Data_V6_Annotations_SampleAttributesDS.txt.

The file with the rpkm values contains values for each transcript for each sample: GTEx_Analysis_v6_RNA-seq_Flux1.6_transcript_rpkm.txt.

For insertion of expression data into our RAINET database we want to transform the GTEx 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 GTEx_Data_V6_Annotations_SampleAttributesDS.txt file with the attribute SMTSD (the more-specific tissue terms).




```
## [1] "Total # of tissues :  49"
## [1] "Total # of samples :  8527"
```

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

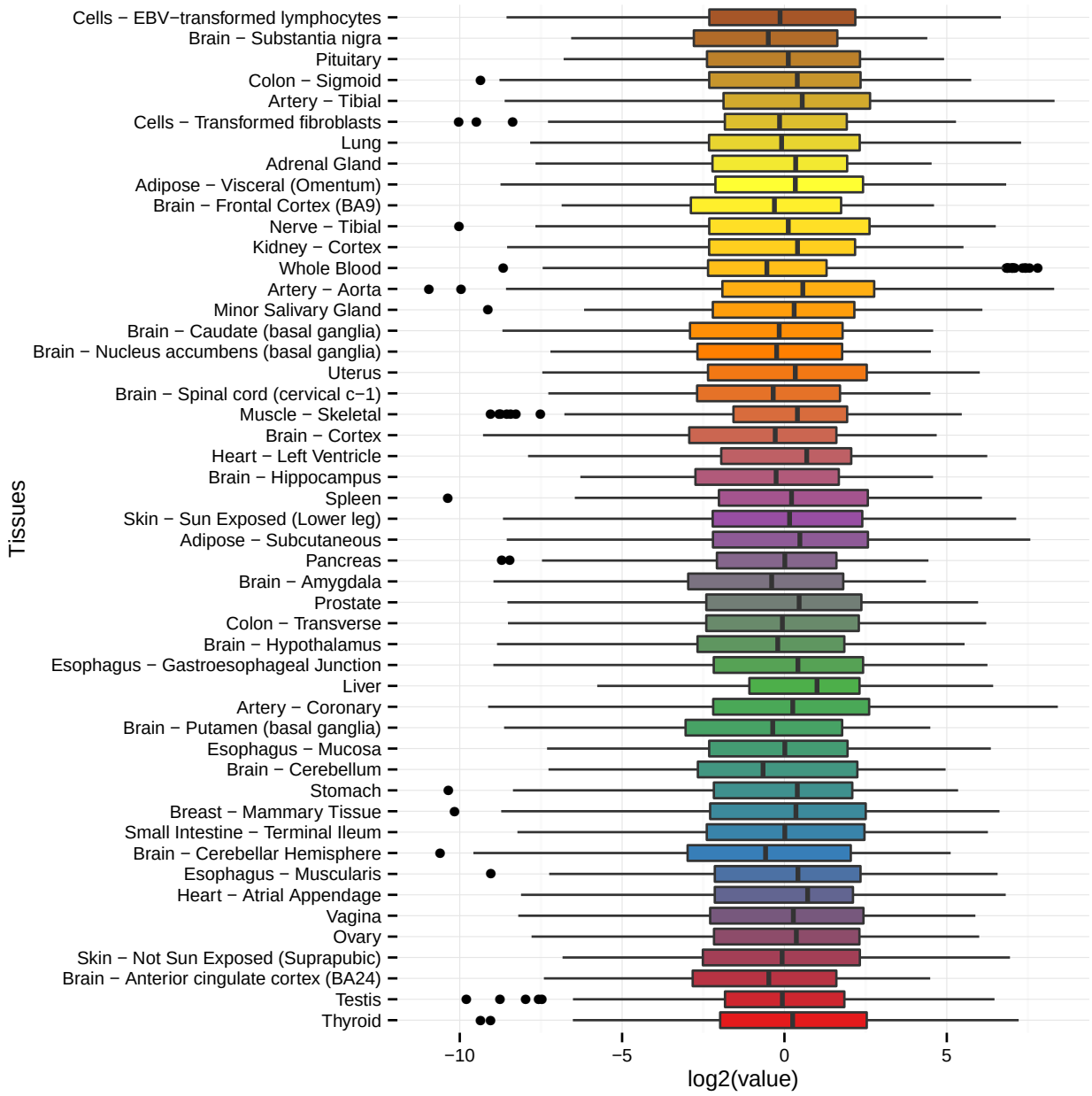
3 Expression variability per tissue

Distribution of all the expression values per tissue of 50 randomly sampled transcripts (out of 195.747 thousand from GENCODE), 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.

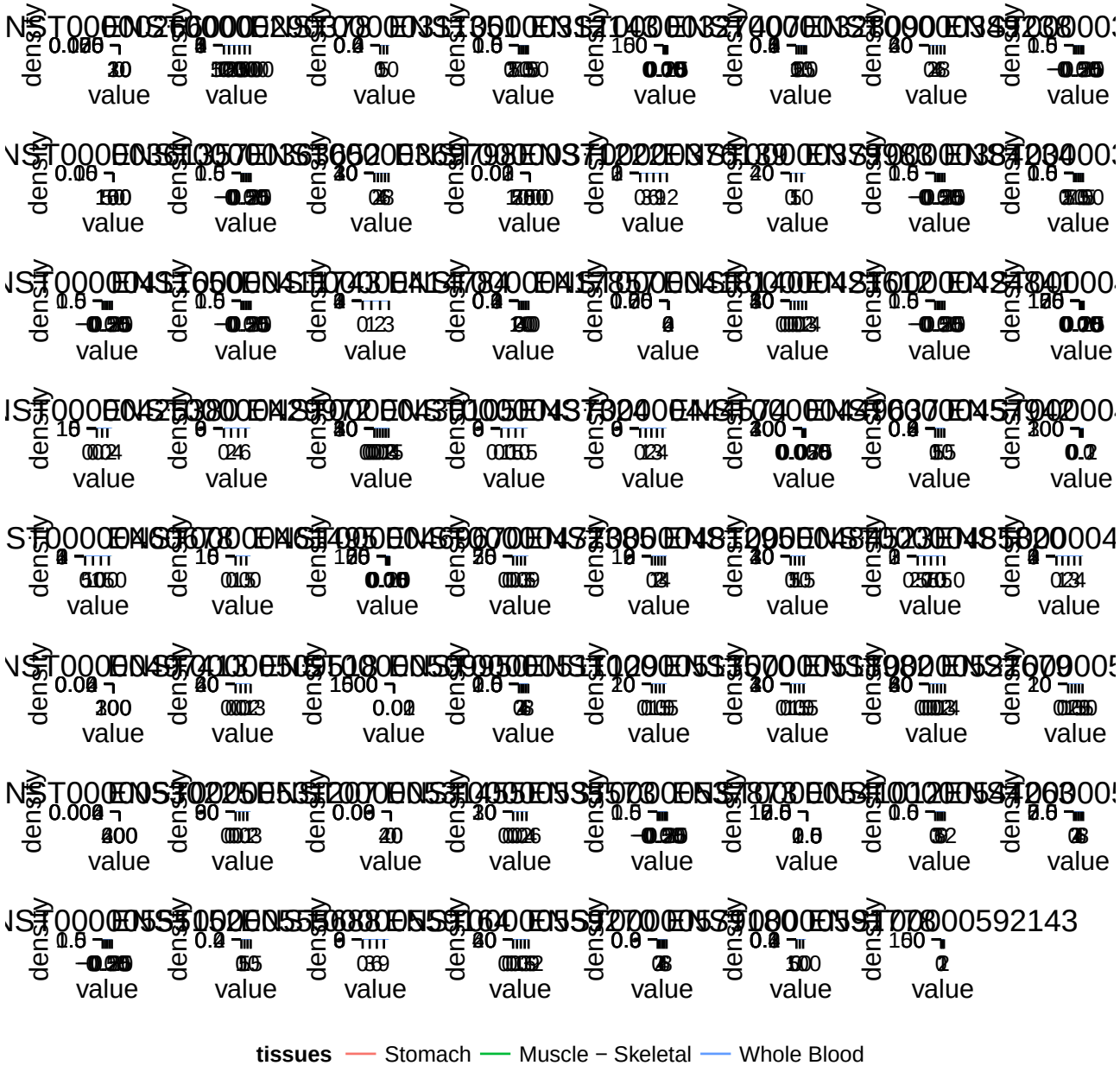
```
## [1] "ALL SAMPLES: summary(expression_df$value[expression_df$variable == Muscle  Skeletal]"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
## [1] "ALL SAMPLES: summary(expression_df$value[expression_df$variable == Fallopian Tube]"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```



We can see that the number of samples does not drastically change the overal distribution of expression per tissue, only that the number of extreme values (i.e. the black dots) are decresed 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 comparibility of the average expression value.

4 Expression distribution per transcript

Selected 5 tissues with different sample numbers (2 with highest, 1 medium, 2 with lowest sample numbers), randomly selected 12 transcripts and plotted their RPKM value distribution, using all available samples. Note that many transcript have 0 or close to 0 RPKM (*Bug in graphics when all RPKMs are zero*) in all addressed tissues. We plan to average the RPKM values of a transcript in a given tissue.



The distributions are highly variable between transcripts.

5 Average expression of transcripts per tissue

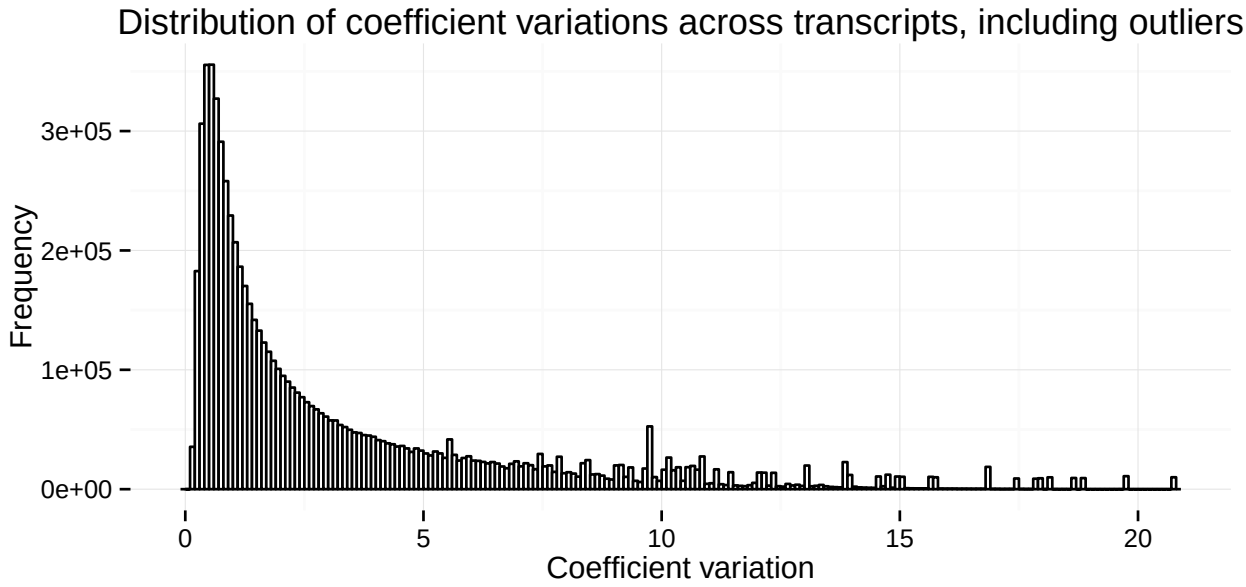
To merge sample values within the same tissue and transcript, we can use mean or median. Below are the distributions of averaged RPKM values among all transcripts and all tissues.

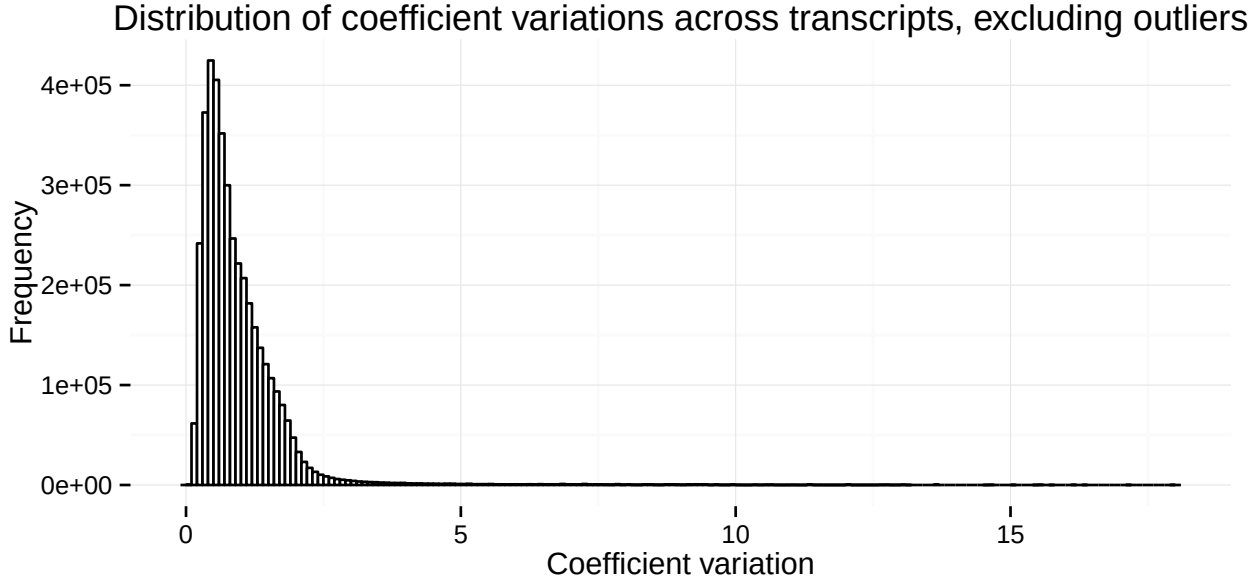
The blue horizontal lines represent either the mean or the median, the red lines represent the noise cutoff of 0.1 RPKM suggested in the GTEx papers.

The coloured plots separate the RPKM distribution per tissue. Also note that in these plots do not display values equal to zero. Note: using xlimit at 5 RPKM.

```
##
Read 16.6% of 9629480 rows
Read 32.4% of 9629480 rows
```

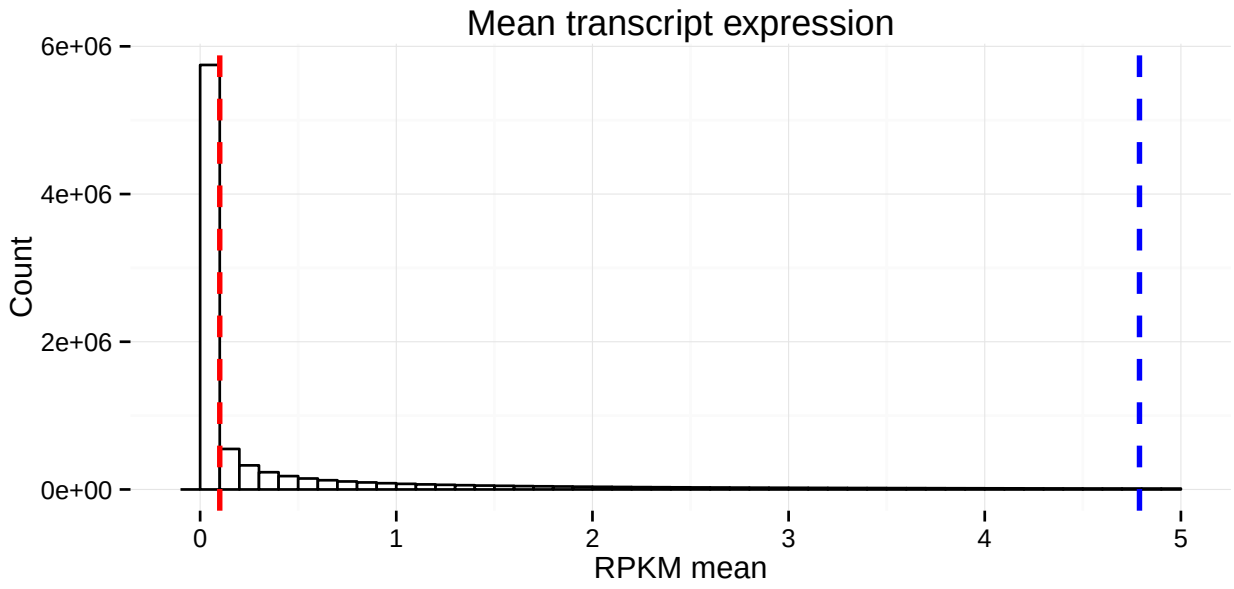
```
Read 48.2% of 9629480 rows
Read 63.9% of 9629480 rows
Read 79.8% of 9629480 rows
Read 95.5% of 9629480 rows
Read 9629480 rows and 7 (of 7) columns from 0.641 GB file in 00:00:08
##
Read 14.7% of 9629480 rows
Read 34.1% of 9629480 rows
Read 53.5% of 9629480 rows
Read 72.9% of 9629480 rows
Read 92.3% of 9629480 rows
Read 9629480 rows and 7 (of 7) columns from 0.609 GB file in 00:00:07
```

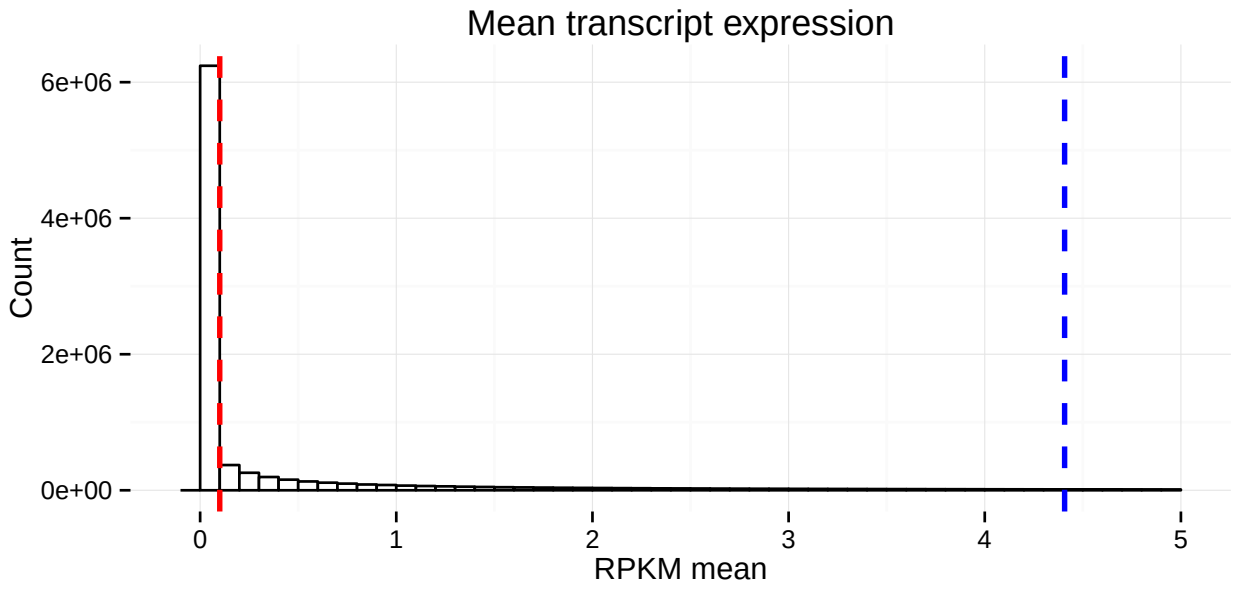





```
## [1] "#### Summary without removing outliers ####"
## [1] "summary(expression_df1$rpkm_mean)"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      0.0      0.0      0.0      4.8      0.6 702300.0
## [1] "#### Summary removing outliers ####"
## [1] "summary(expression_df2$rpkm_mean)"
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##       0.0       0.0       0.0       4.4       0.5 406800.0
## [1] "Plots with xlim: 5"
```





From the RPKM value distributions we see that applying cutoff of 0.1 RPKM (as used in the GTEx papers) will render most transcripts in tissue as non-expressed. Also evident is the spread of the RPKM values is very high, as the mean and median are well above the values of the vast majority of transcripts. This suggests our mean is being driven by outliers with very high expression values.

The coloured plots were produced to see if different tissues would display different distributions (density curves), however these are very similar to the above histogram distributions.

6 Conclusion

We intend to use this expression data not for an expression analysis, but only as a filter layer to ensure lncRNA-protein co-existence in a cell. After setting a rpkm cutoff that distinguishes expression noise from real expression, we will turn our values into binary presence/absence.

Because we want to keep the maximum data possible, we do not need to filter out tissues that have low sample numbers, or to normalise data in tissues with a vast number of samples.

Parameters for filtering:
- minimum rpkm value
- interacting pair having to be present in X% tissues