CKCC Gene Expression And Their 95th Percentiles

Liam McKay
Treehouse Undergraduate Research

Overview

- In pediatric patients: we look for **genes highly expressed**

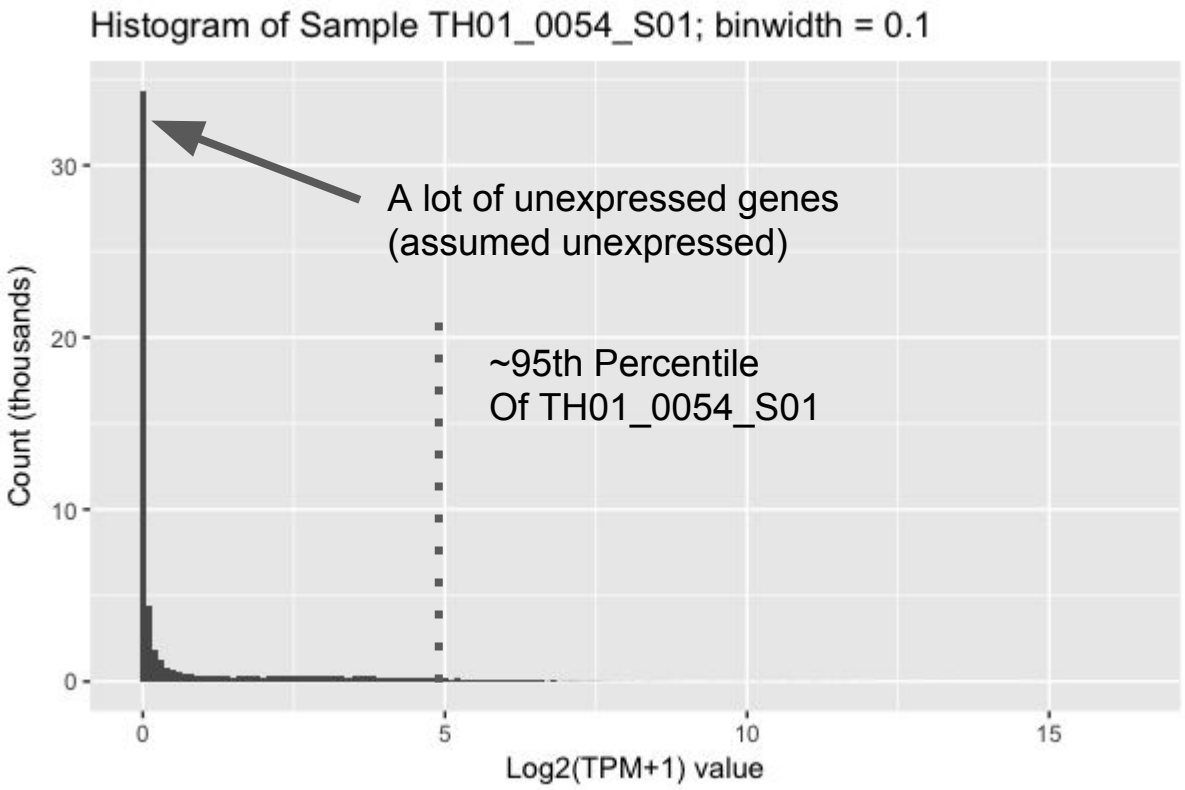
Slide Numbers

- We **cut off at 95 Percentile of gene expression** within the sample
  - What? - the genes with highest expression 5% in a sample
  - Why? - It’s a conservative measure with little doubt

Problems?

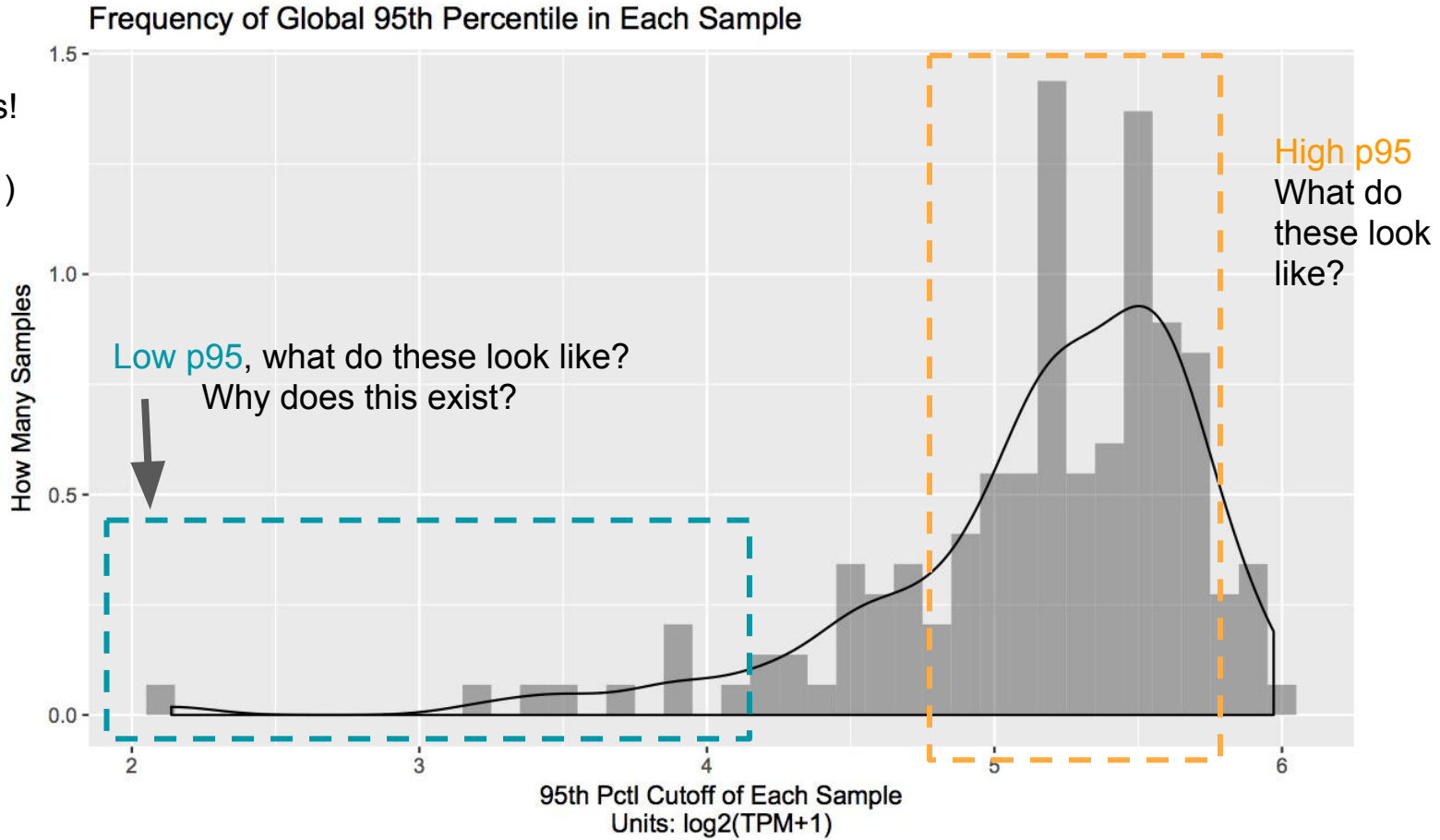
- Are we losing data?
- Can we reduce the stringency of 95 Percentile?

What does a single patient's gene expression look like?

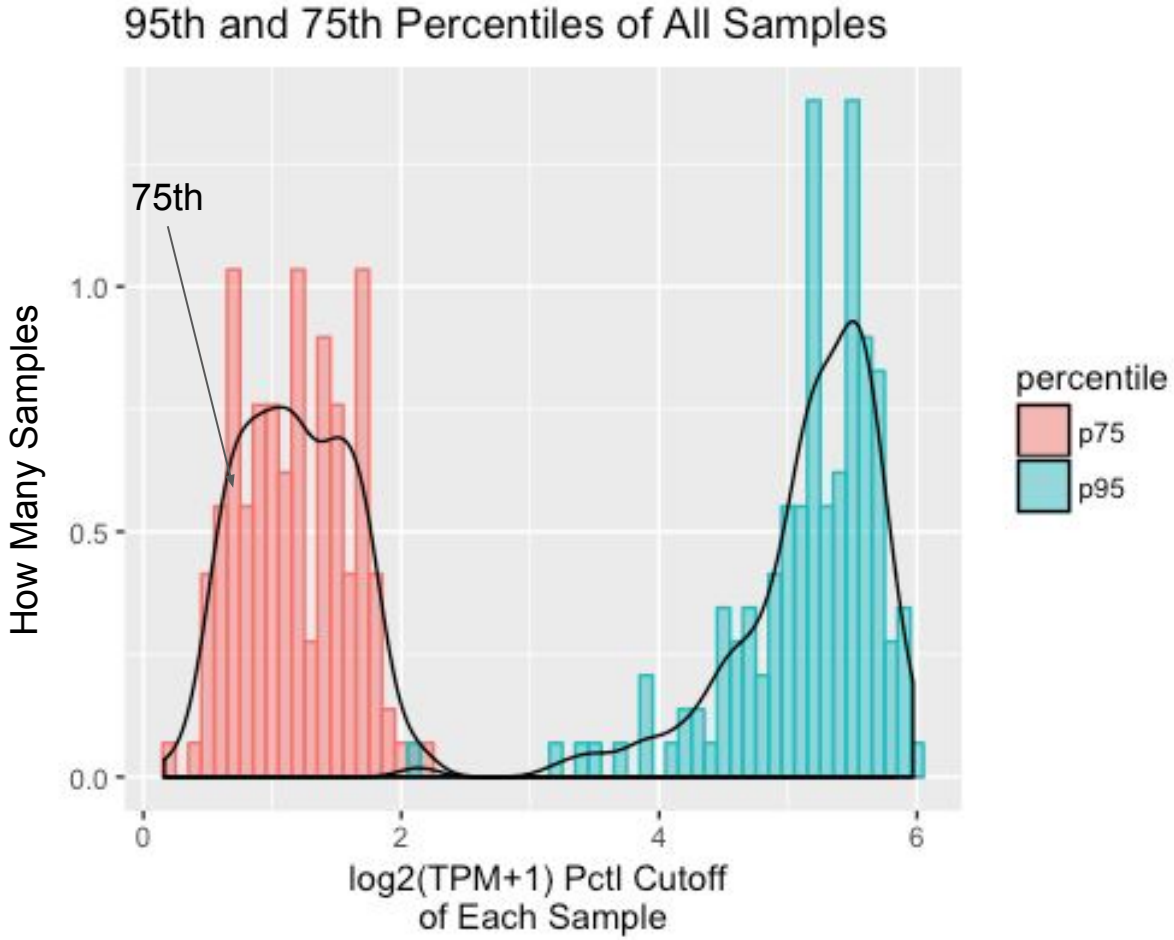


What does the 95th pctl look like across all patients?

- These are percentile cutoff values!
- Bin = 0.1 log2(TPM+1)



Skew is More Apparent in the 95th Pctl than 75th Pctl



Low vs. High 95th Percentiles

- Is the **quality** of the patient genes expression related to its 95th pctl?
- Is there something **characteristic** of a sample having high gene expression **other than the 95th pctl**

Time-out!

- What is causing this low p95?
- Is low p95 correlated to a major specific distribution change
- There is no major systematic differences between high and low p95 histogram (yet)

Where are we going?

- We need to find other ways to investigate low p95

How many unexpressed genes do samples have?

On average:

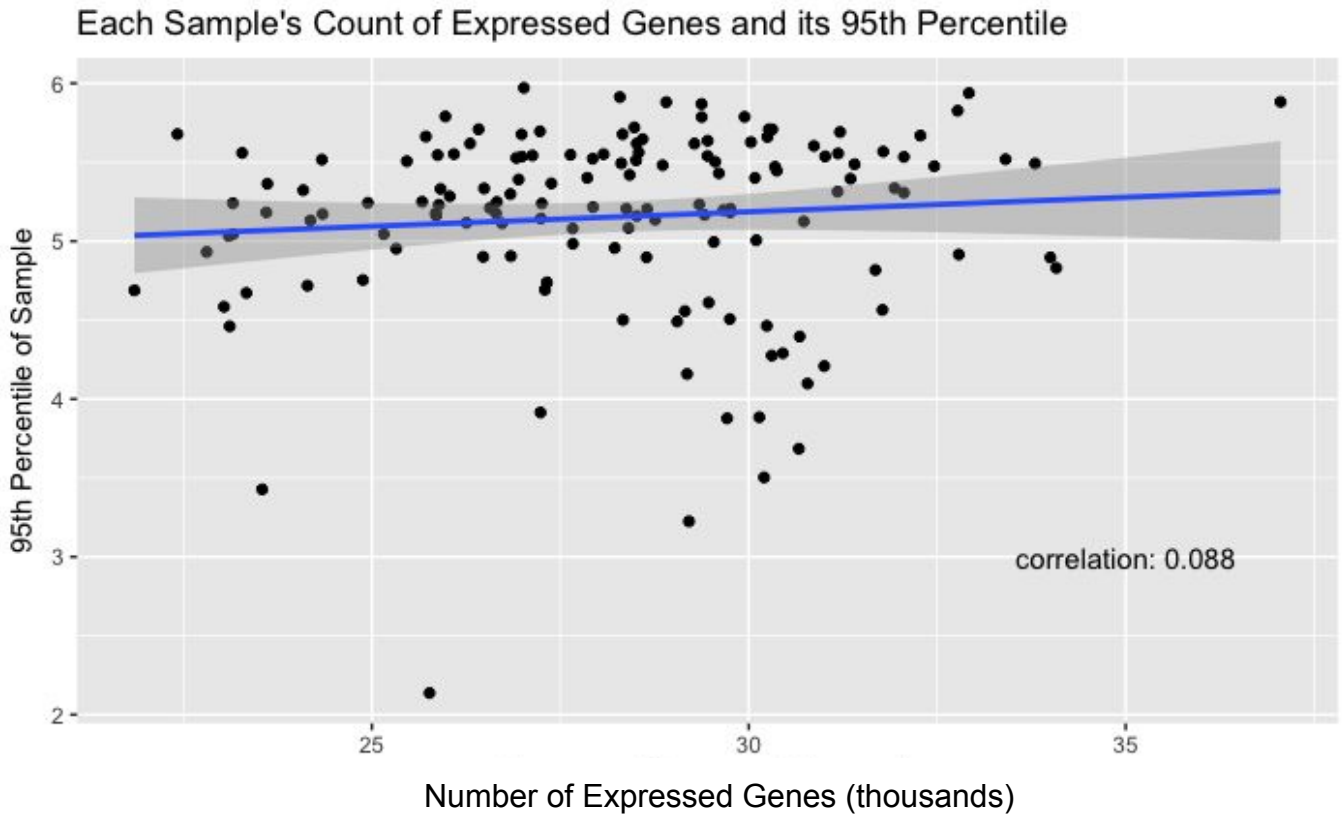
- ~50% of genes in each sample were unexpressed

- There are differences in number of genes expressed in histograms
- **Are fewer genes expressed** in samples with **low 95th pctl**
- Specifically, are unexpressed gene counts higher in samples with a low 95th percentile?

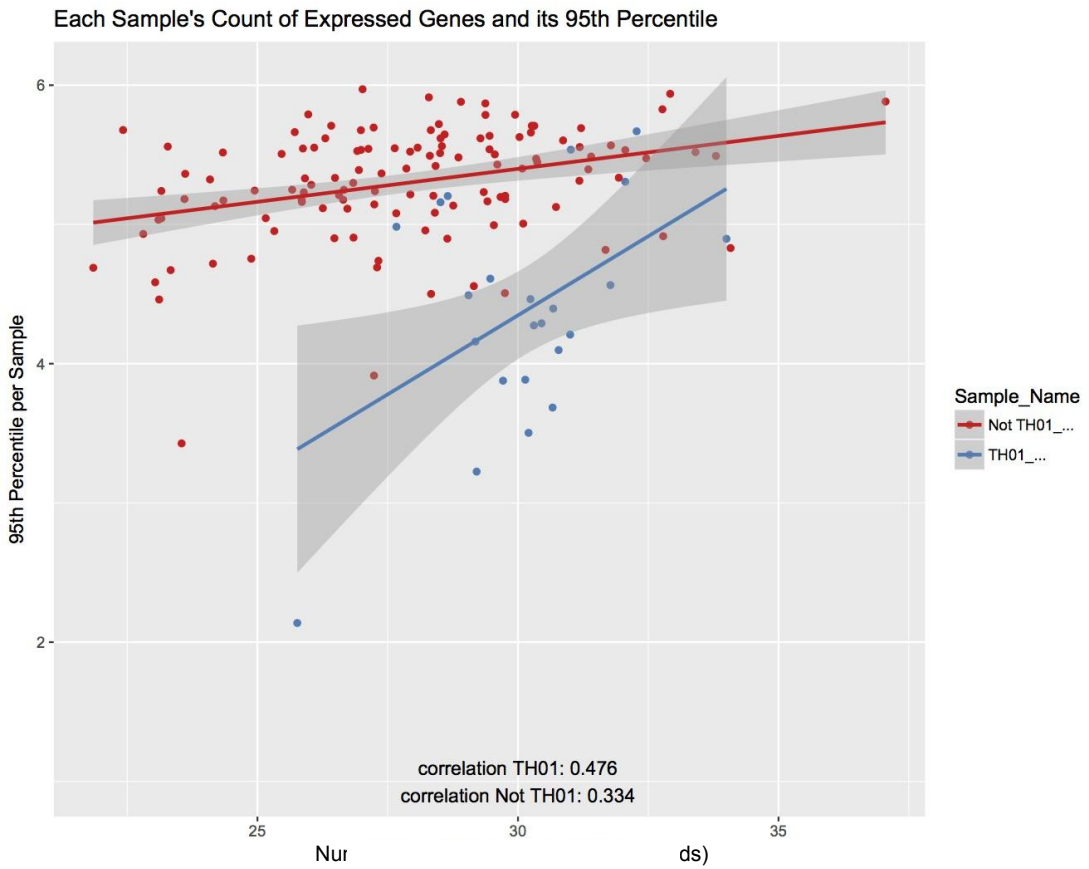
3. Across the cohort, the number of expressed genes **not correlated** to each patients' 95th pctl

- 8.8% correlation

- Best fit line shows flat correlation, however
- We cannot conclude anything



However, TH01*[RiboD] samples have **~50% correlation** between p95 and # of Expressed Genes

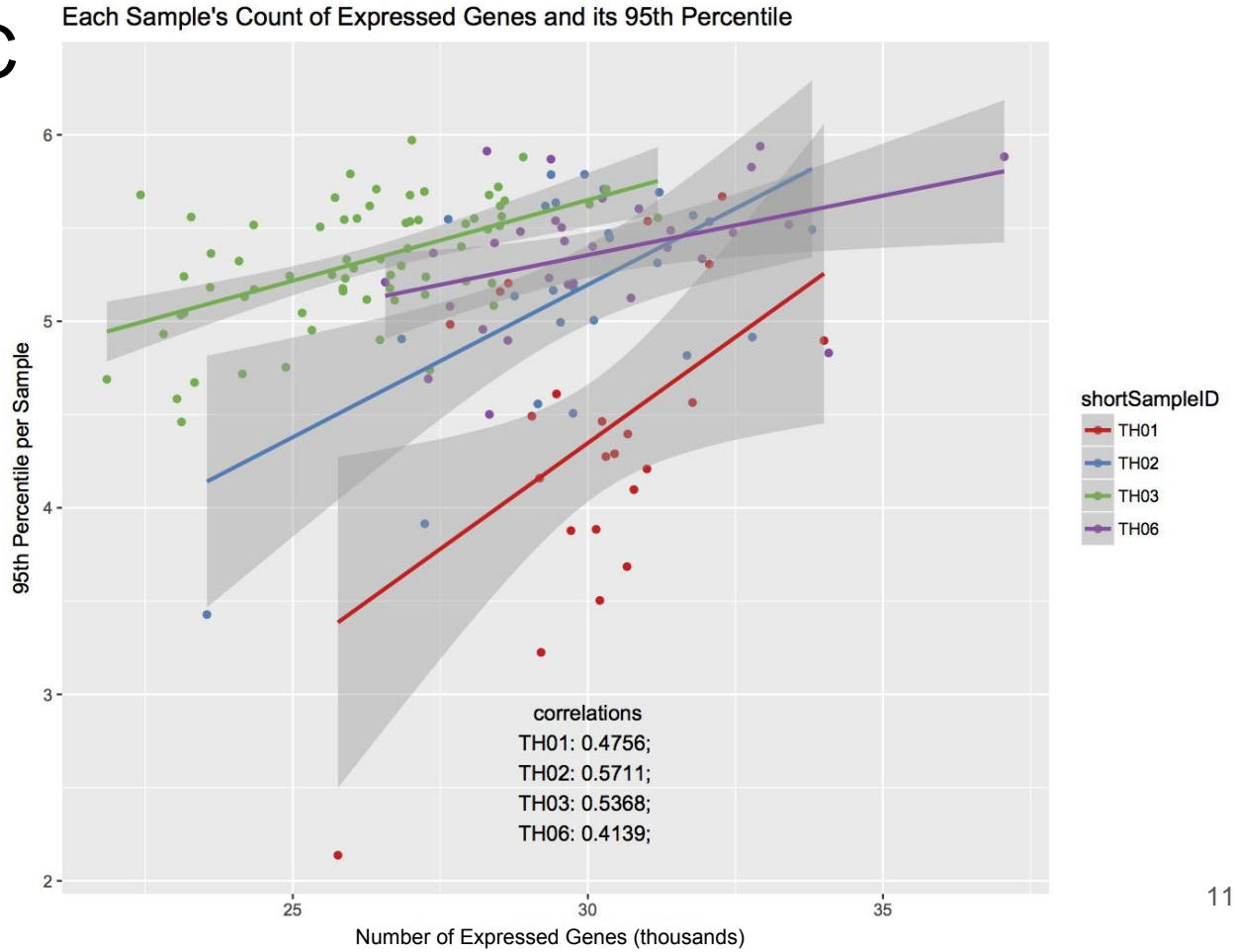


Blue = TH01
Red = everything else

- There is more correlation in TH01 than the rest
- 33.4% Not TH01 correlation
- 47.6% TH01 correlation

Substantial correlations are Present in the Data From Each CKCC partner

- TH02 has the highest correlation
- TH01 47.56%
- TH02 57.11%
- TH03 53.68%
- TH06 41.39%



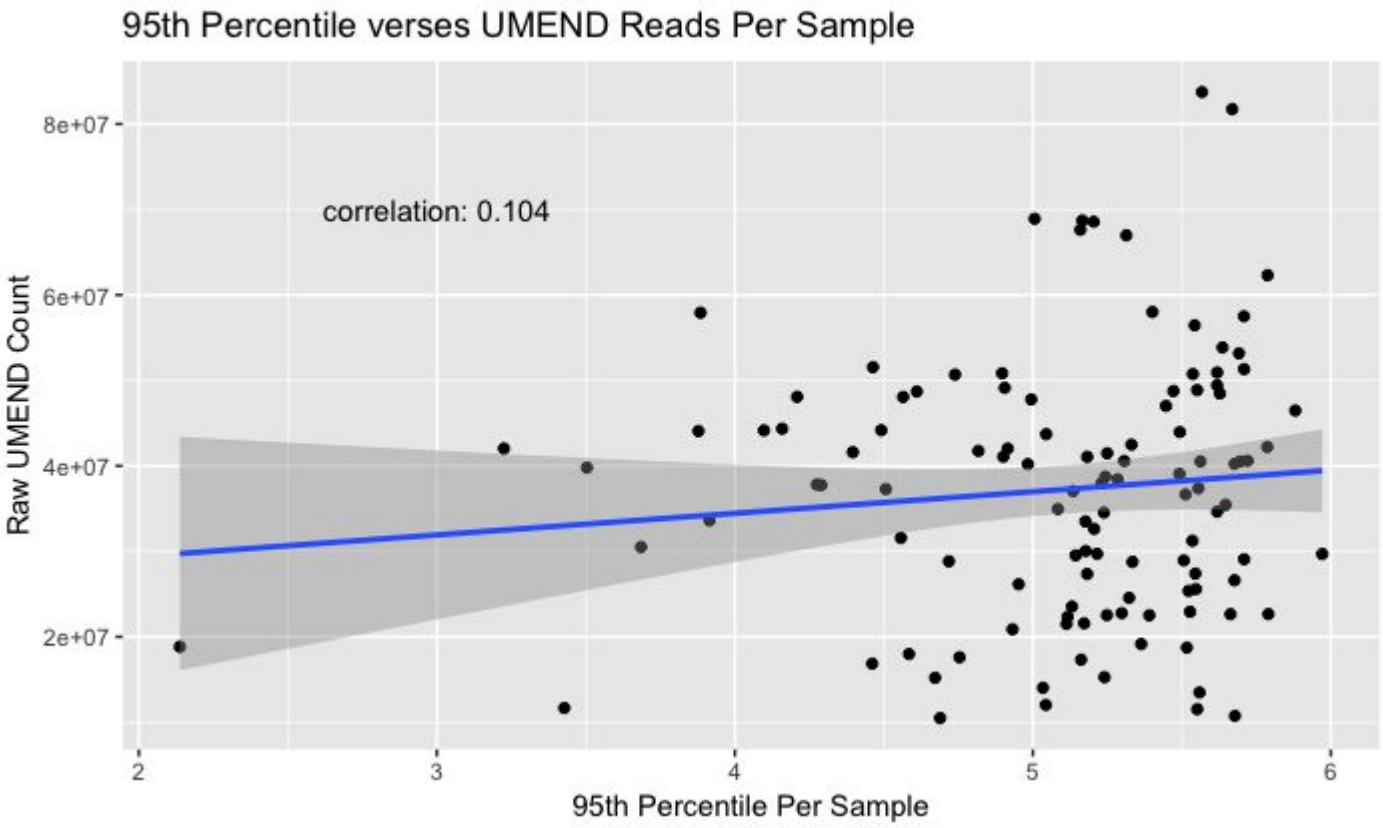
Conclusion

- There is a correlation between the **number of expressed genes** in a sample and **its p95**
- But only apparent in samples that were prepared and sequenced **by the same group**

Next

- We decided to investigate **UMEND reads and p95** to find what else is causing low p95

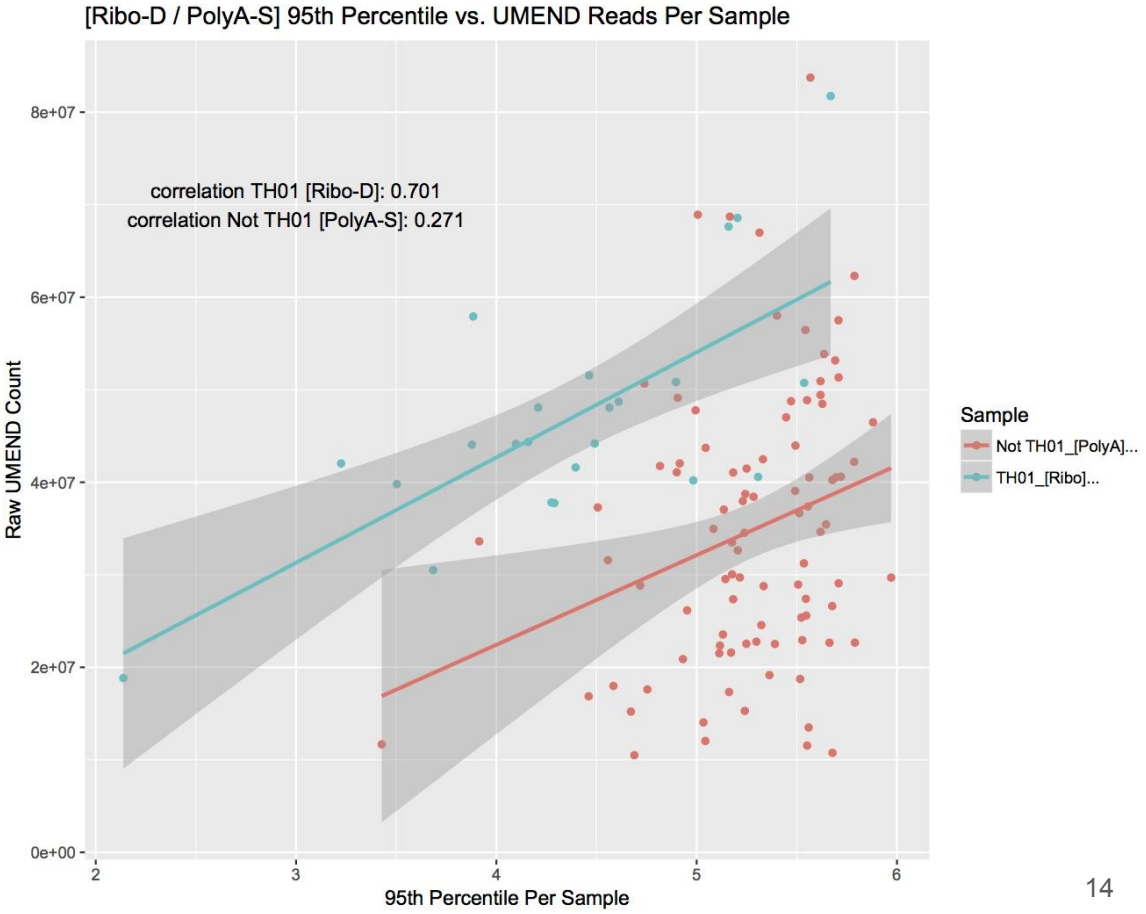
4. Across the cohort, the number of UMEND reads is **not correlated** to each patients' 95th pctl

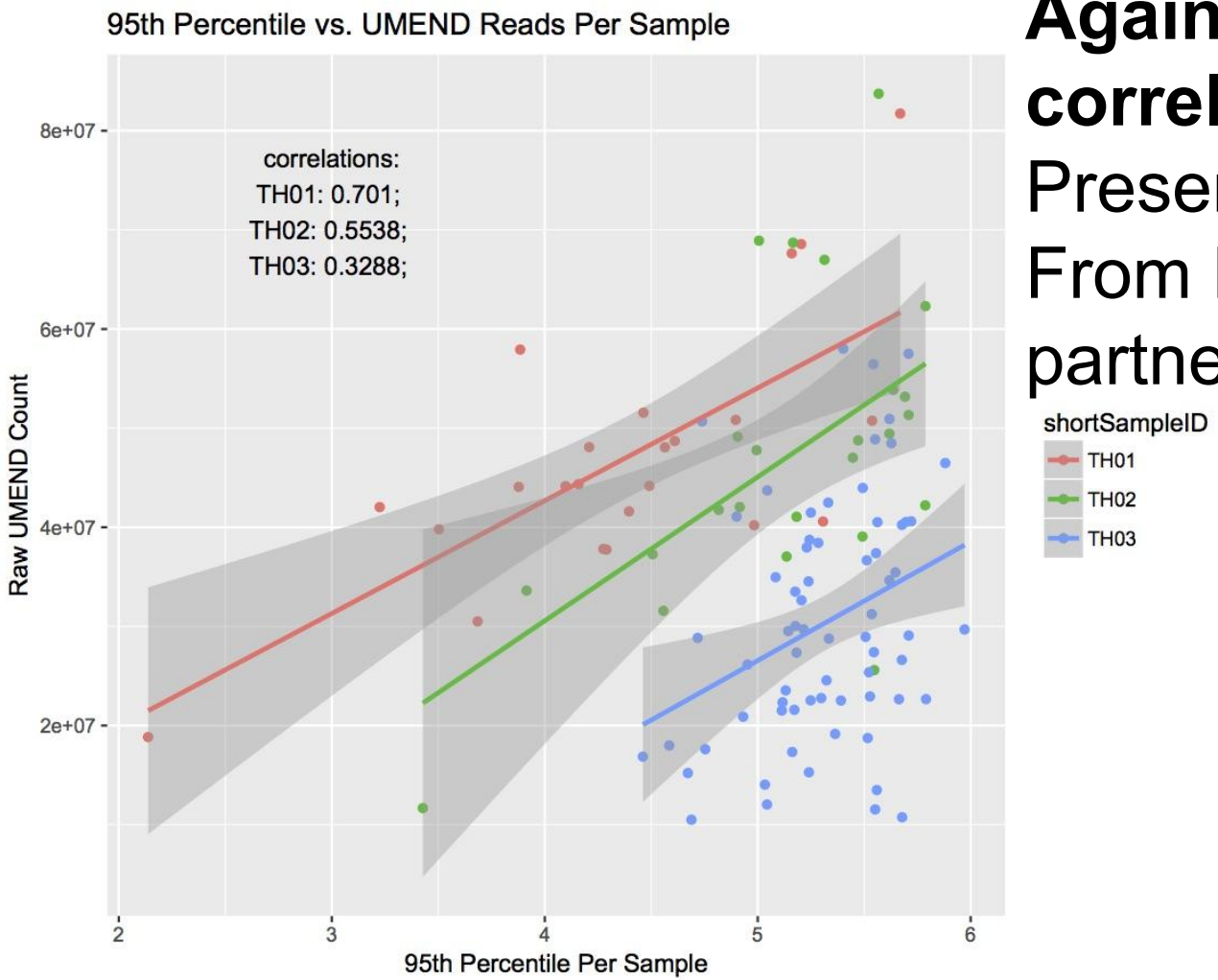


- **10.4% correlation**

However, TH01*[RiboD] samples have **~70% correlation** between p95 and UMEND

- riboD samples especially
- RiboD has **70.1% correlation**
- PolyA-S 27.1%



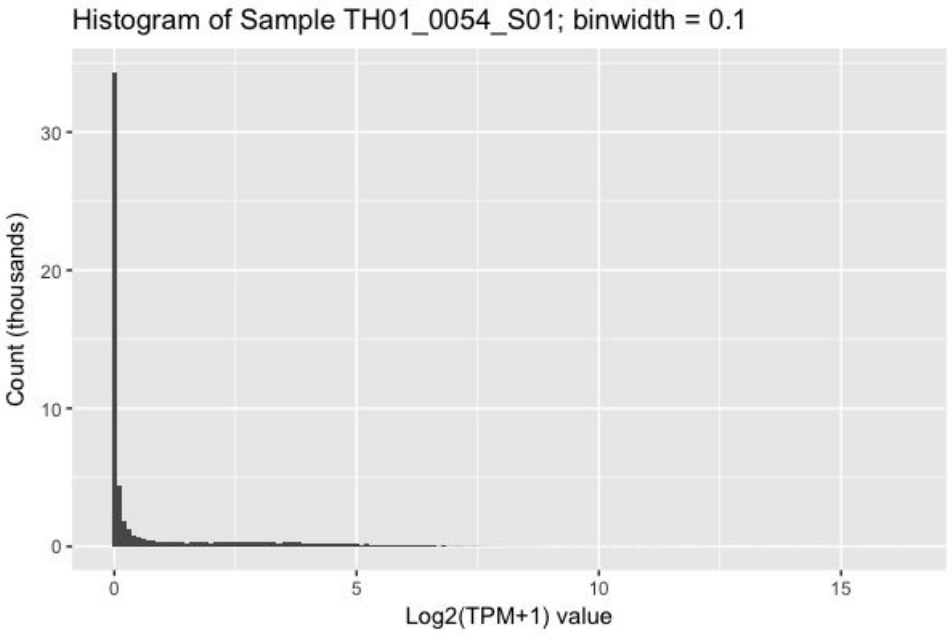


Again, Substantial correlations are Present in the Data From Each CKCC partner

- Separation by sample center
- Shows TH01 has highest correlation
- TH01 70.1%
- TH02 55.38%
- TH03 32.88%

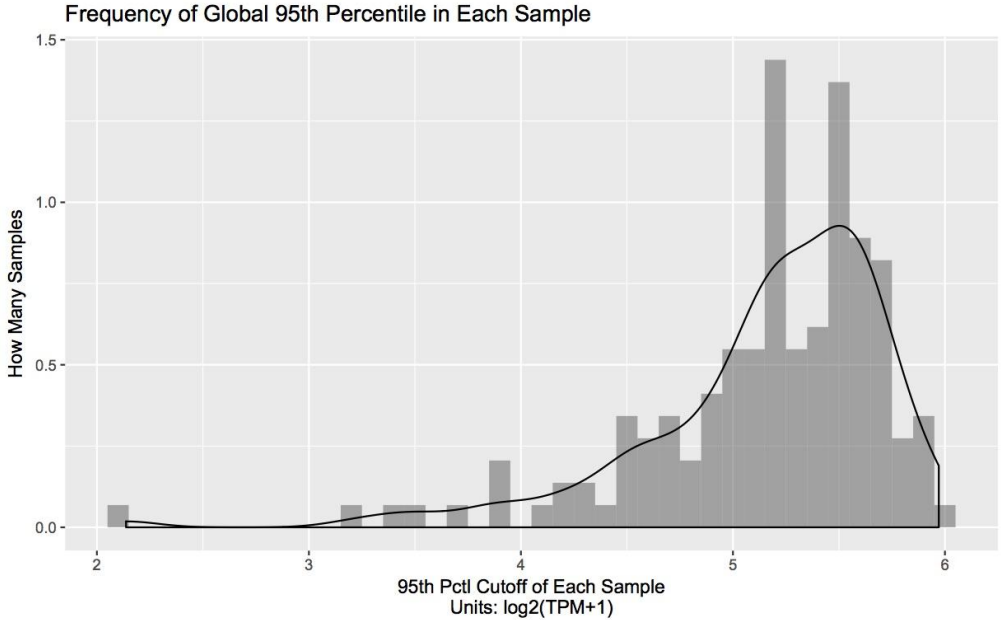
Results

1. Plot every patient's expression values (histograms)
   a. There was half of the genes expressed (we assumed other half was unexpressed)



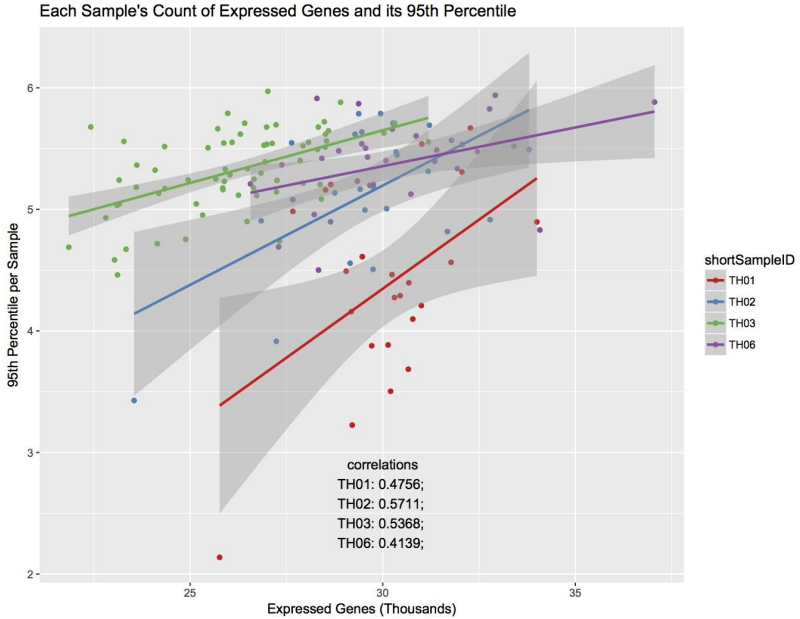
Results

2. Plot the distribution of 95pctl per sample
   a. There are high and low 95th pctls
   b. Why? because of UMEND / # of Expressed Genes



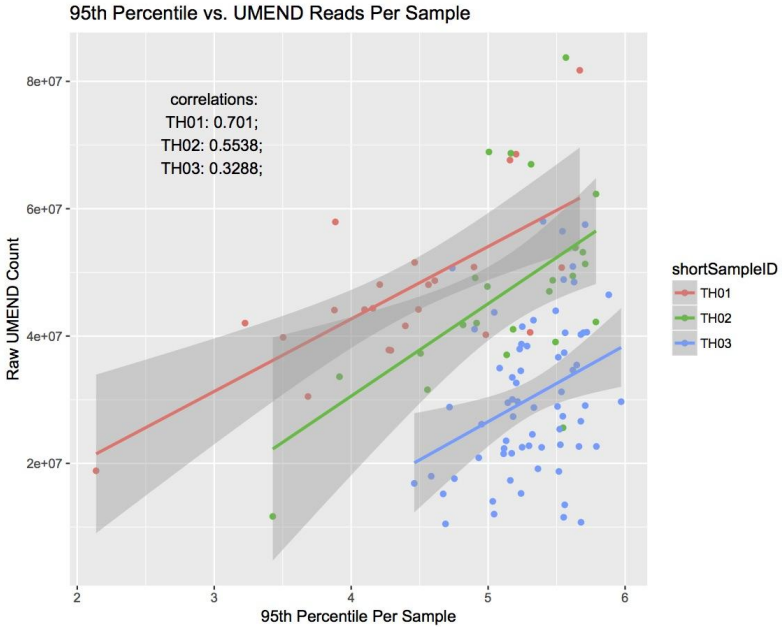
Results

3. Plot # of genes expressed vs. 95th pctl per sample
   a. Looking at total ungrouped data there is no correlation
   b. Each sample center = positive correlation p95/expression count (TH01: 47.6%)



Results

4. Plot UMEND count vs. 95th pctl pan-sample (scatter)
   a. Looking at total ungrouped data there is no correlation
   b. Each sample center = positive correlation p95/UMENDread (TH01: 70.1%)



What does it all mean?

- **There is a systematic difference between** low p95 and high p95 using other investigations than just sample 95th percentile

- UMEND, and number of expressed genes are correlated to the 95th percentile
- More robust, less stringent statements about a sample's p95
- However, this systematic difference should be investigated with significance. ($p<0.0005$)

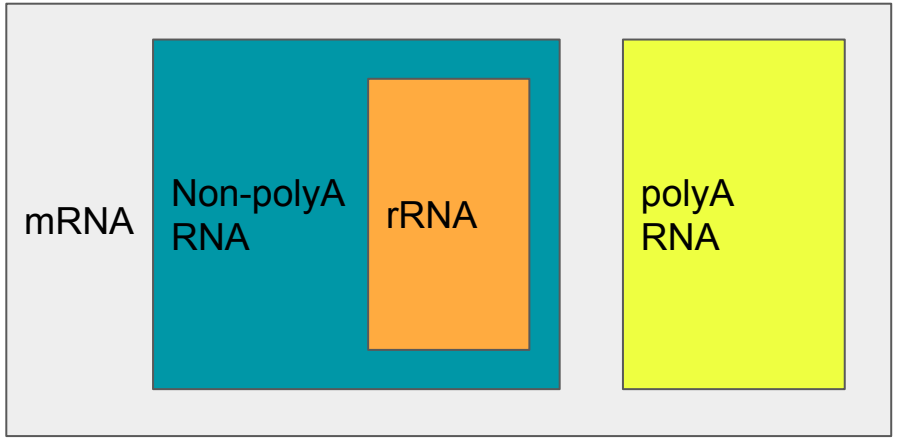
The End!
Questions?

Supplemental

RiboD vs. PolyA Selection

Non poly a doesn't need to make a ton of copies for g
because its not going to be expressed
Poly A 95th percentile is therefore higher than non
Poly A

- All rna is PolyA or non PolyA
  - PolyA has no ribosomal
  - Non PolyA has ribosomal
- RiboD → sequence more of the total RNA than polyA



Ribonuclease Depletion [RiboD]
Everything Not rRNA
mRNA, Non-PolyA-rRNA, polyA

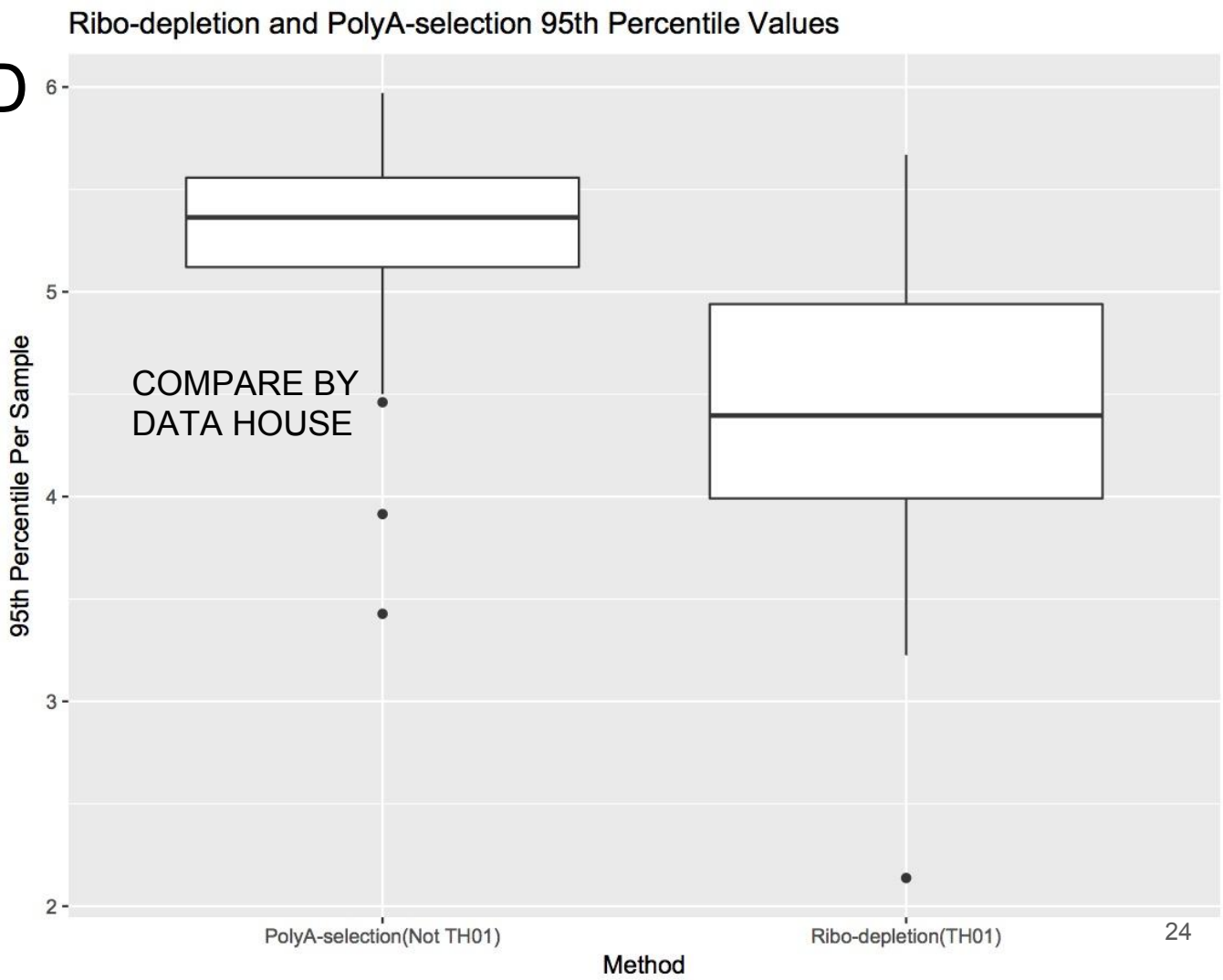
Poly-A Selection [PolyA]
polyA

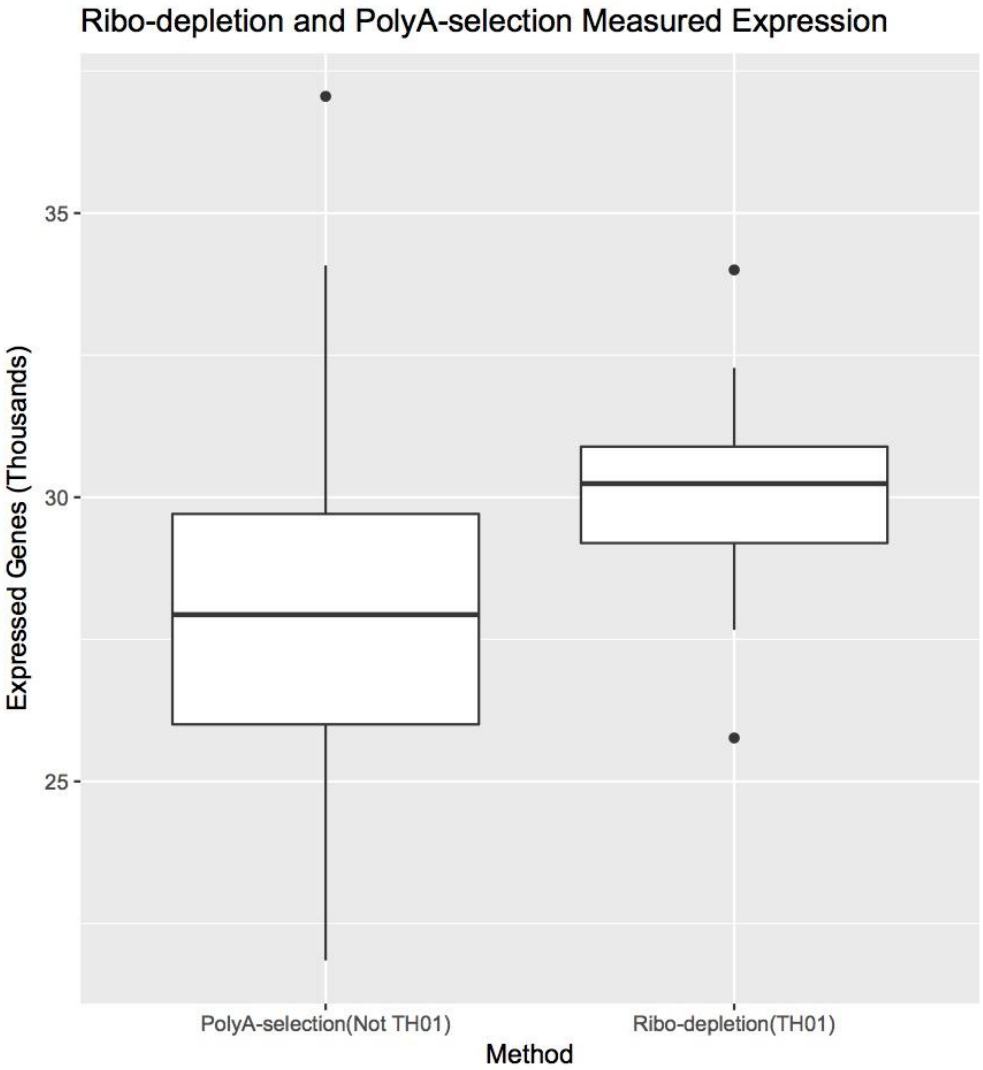
5. PolyA > RiboD rel. p95

Welch Two Sample t-test

- t = -5.3986,
- df = 24.118,
- **p-value = 1.496e-05**
- alternative hypothesis: **true difference in means is not equal to 0**

95 percent confidence interval:
-1.2779641 -0.5712024
sample estimates:
mean of x mean of y
4.374965 5.299548

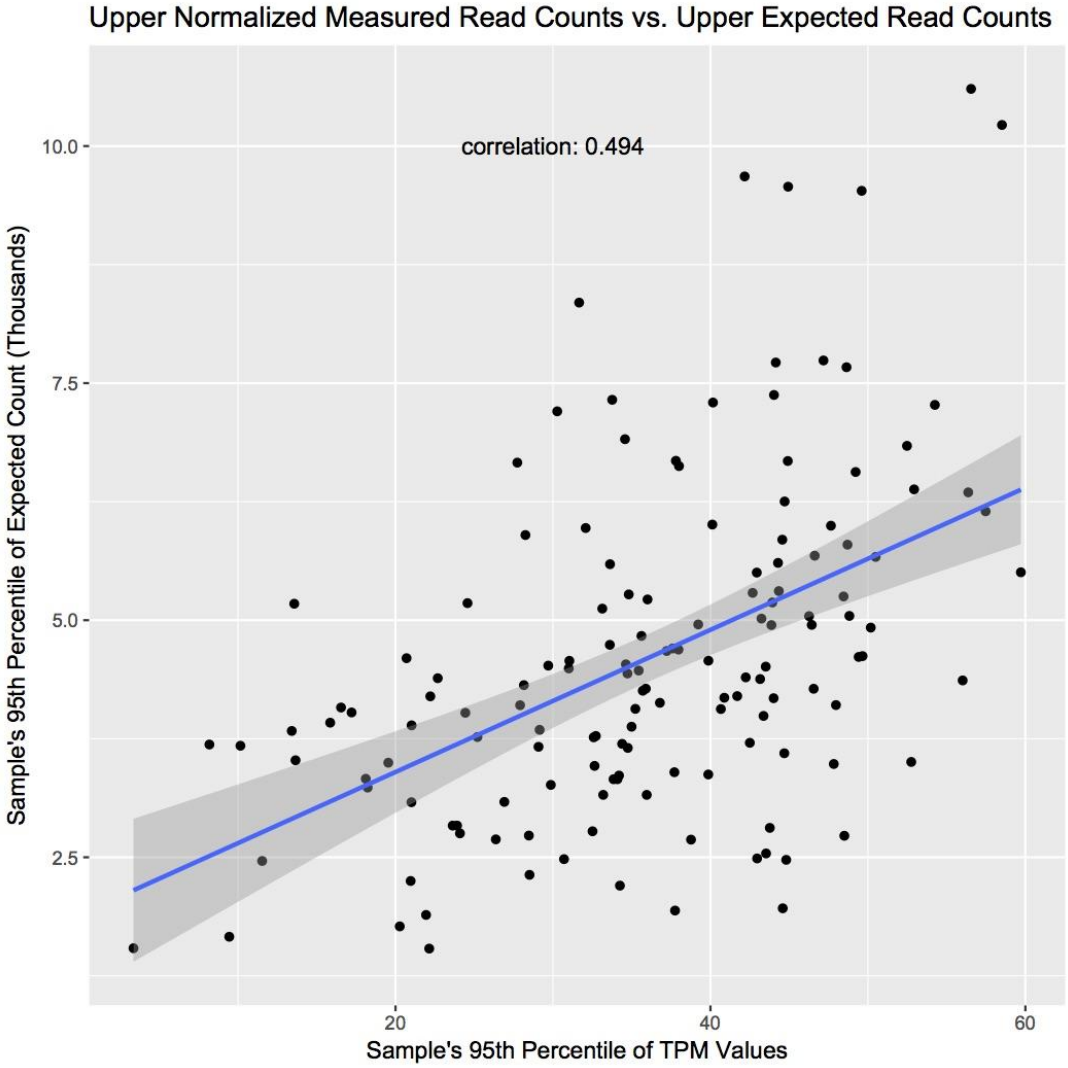




6. Ribo D > PolyA rel. Exp Genes

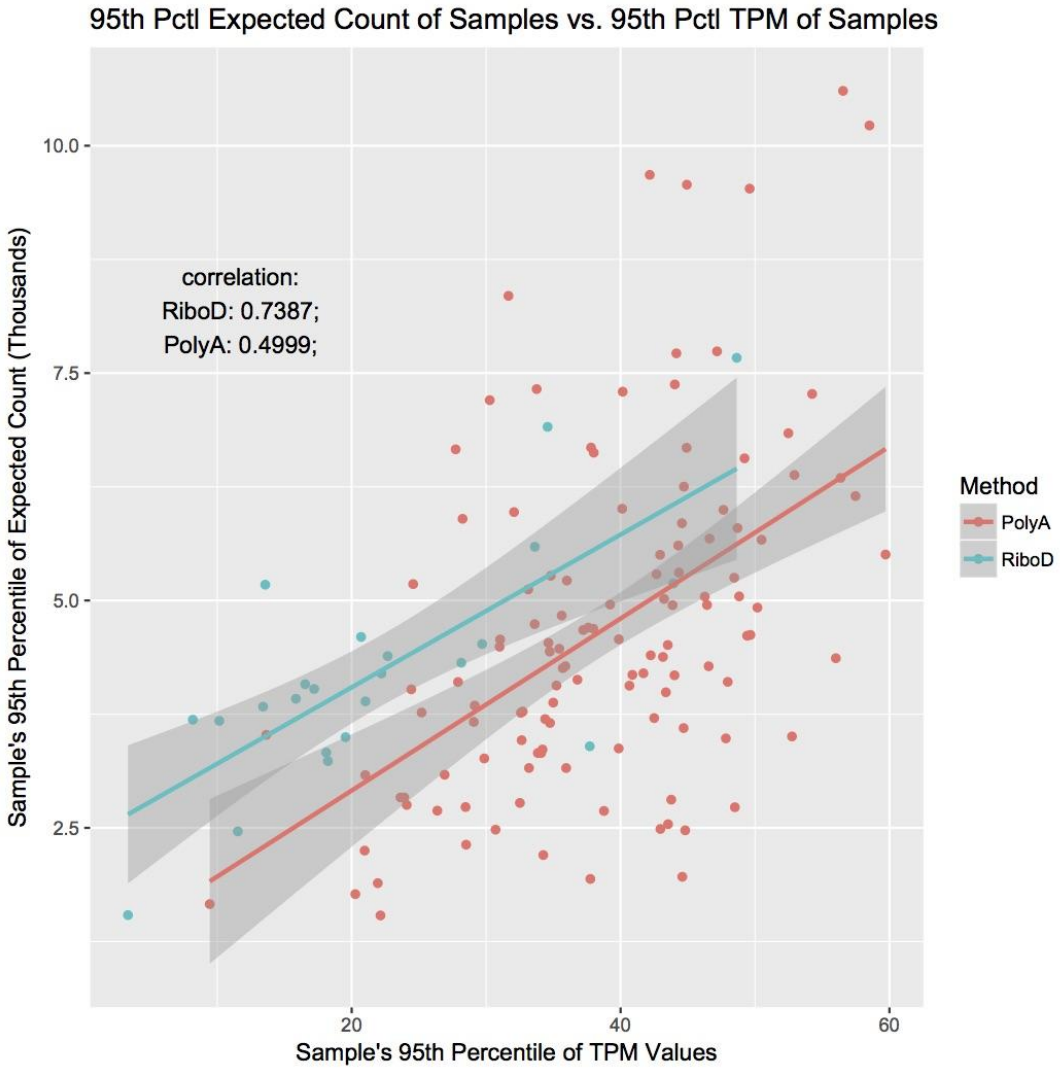
- Box Plot Ribo-/- | PolyA +
- t = 5.121, df = 49.389,
- **p-value = 5.016e-06**
- Therefore, true difference in means is not equal to 0
- 95 percent confidence interval: 1352.013 3097.878

sample estimates:
mean of x 30121.91
mean of y 27896.97



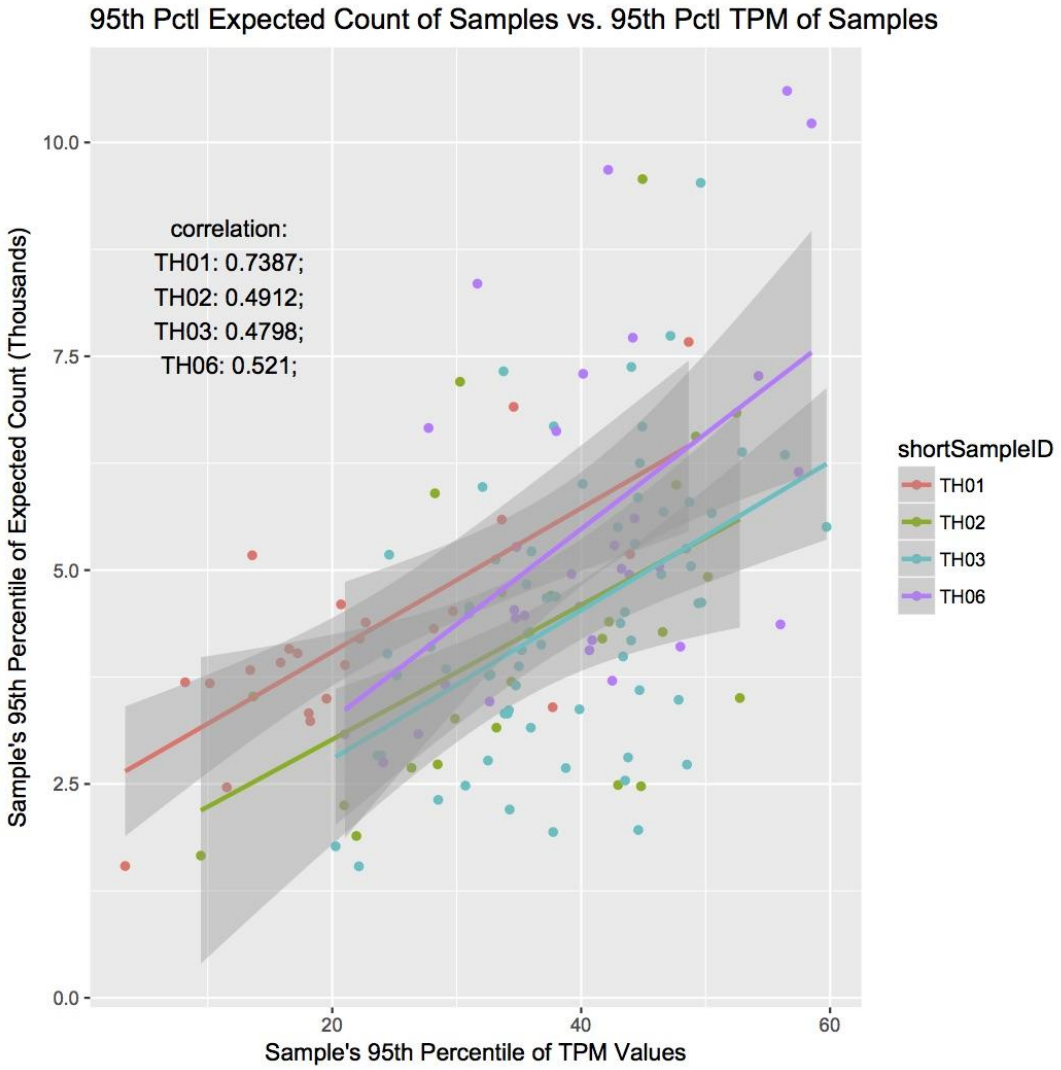
7. p95 expected count vs. p95 TPM

- Actual Count
- Expected count is not read length normalized, it's what we expect to see
- TPM is read length normalized



RiboD has less variance of expected and actual read count

- RiboD cor = 0.7387
- PolyA cor = 0.4999



RiboD has the least variance of high expected counts vs TPM

- RiboD has the least variance even when comparing against other sample centers

8. Let's analyze by Gene!

- Method:
  - Take most **variably** expressed genes
  - Compare the same in each sample.
  - Variance = SD^2

Standard Deviation

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$n =$ The number of data points

$\bar{x} =$ The mean of the x_i

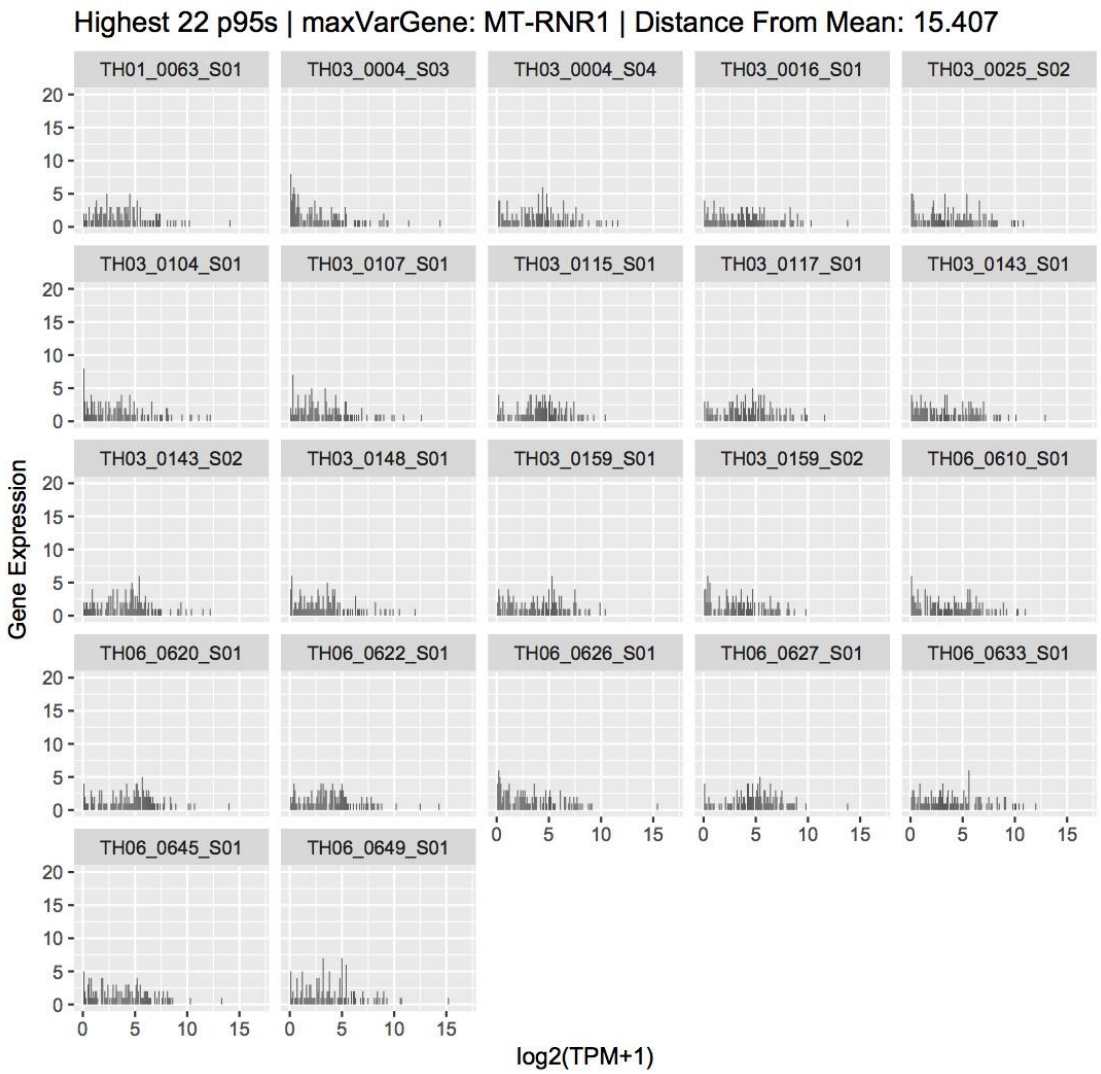
$x_i =$ Each of the values of the data

Variance of Genes Calculations

- Top 5% of the data variates from the mean by 2.56
- the overall mean is 1.019 log2(TPM+1) pan-sample
- average standard deviation of **genes differing from the mean** is 0.56

```
dfGeneVar <- outlierResults %>%
        group_by(Gene) %>%
        summarize(variation = var(sample)) %>%
        arrange(desc(variation))
```

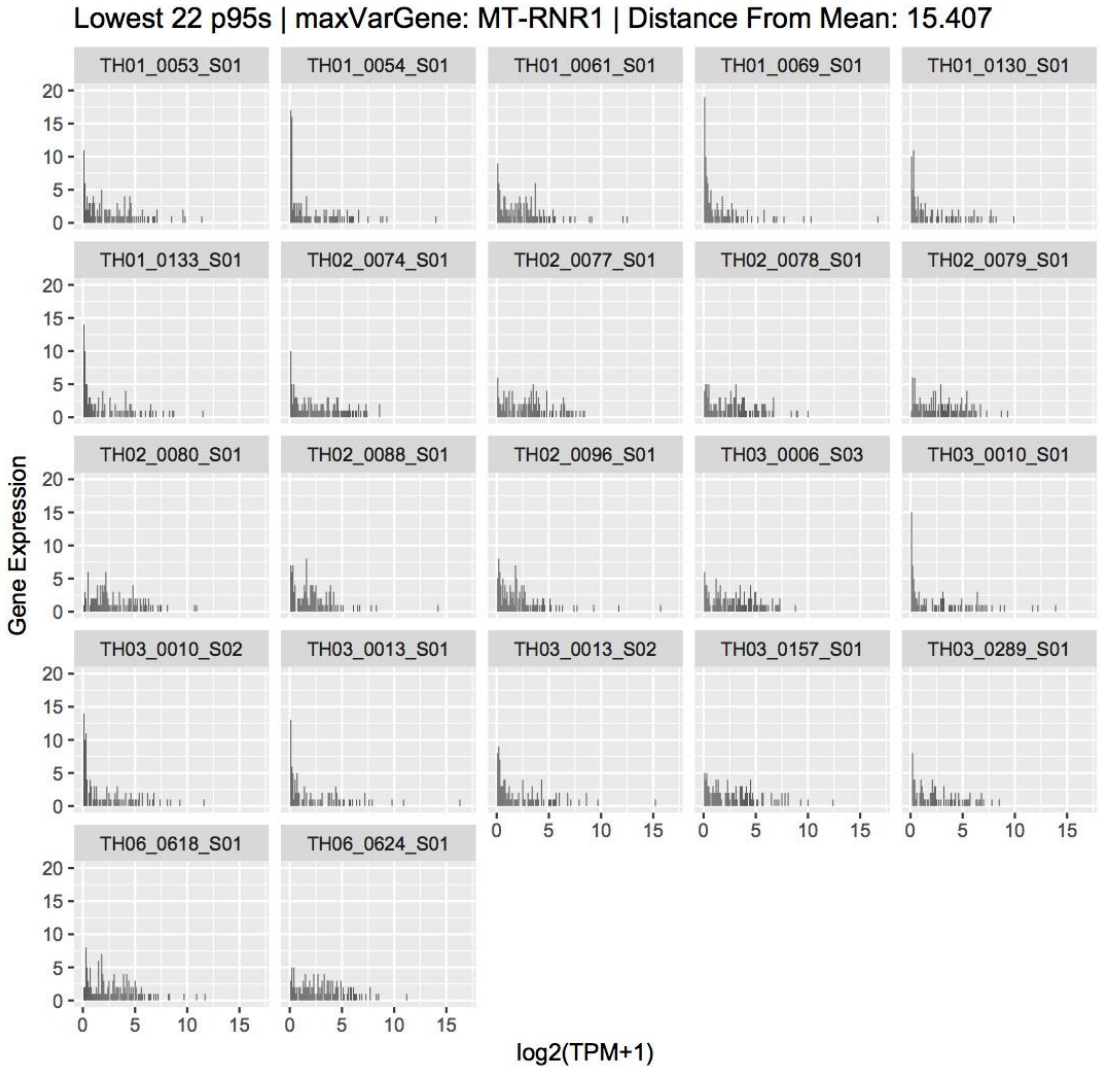
Gene	variation
GFAP	28.6
COL1A1	20.6
TMSB4XP6	19.5
COL3A1	18.7
COL1A2	17.7
SNORA73A	16.9
AP003041.1	16.4
SNORD13	15.9
SNORD3A	15.7
AL162151.3	15.6



← Samples with High gene expr variance

22 Highest 95 pctls

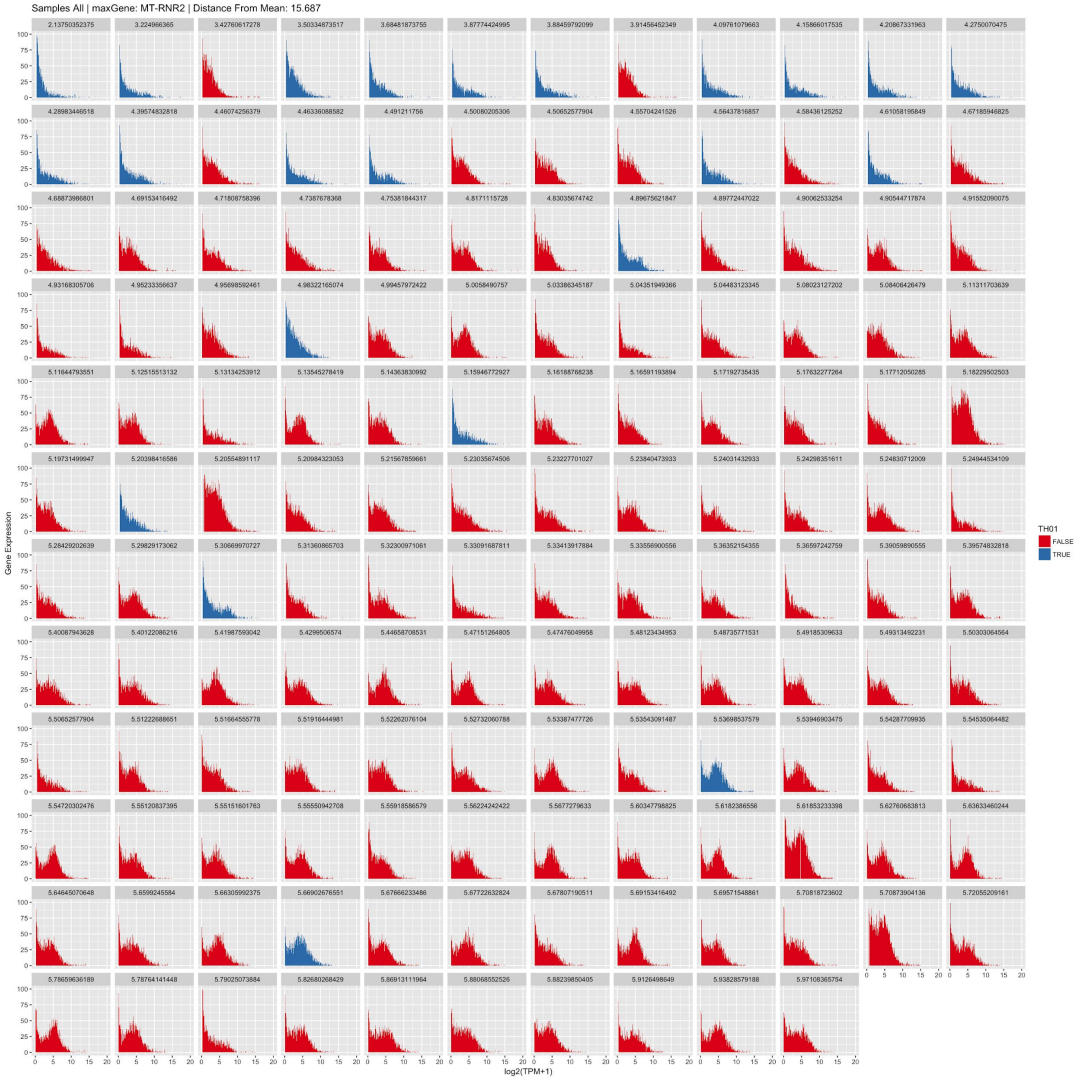
-



← Samples with High gene expr variance

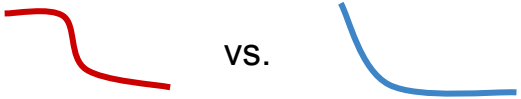
22 Lowest 95th Pctls

- Slightly lower log2(TPM+1) than highest 22 95th pctls

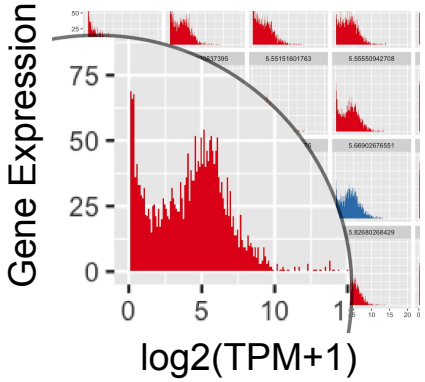


Highest Gene Variance pan-sample

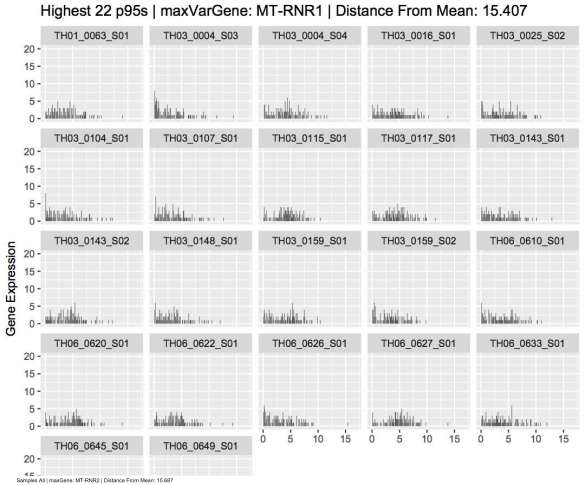
- Non TH01 have mostly square distributions

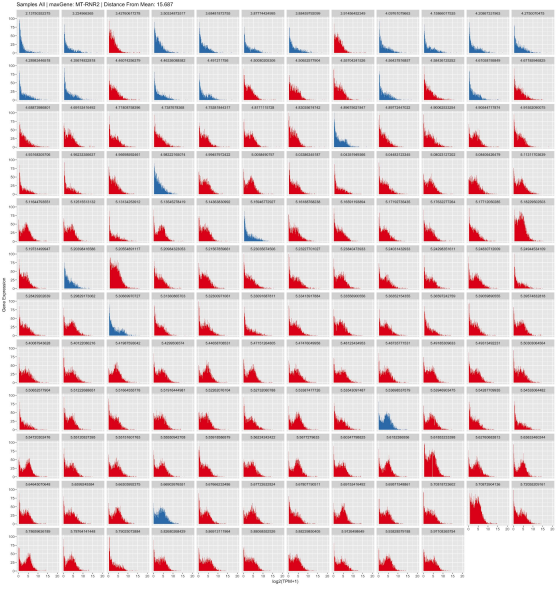


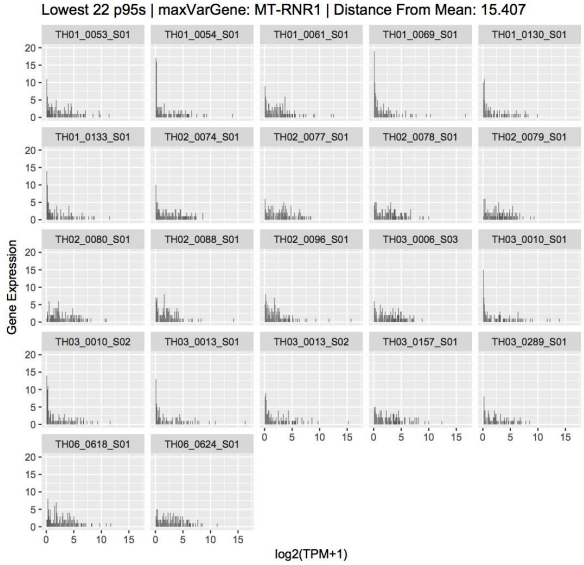
- 142 samples plotted



No Difference in low p95 and high p95 genes with high variability







Results full

1. Plot every patient’s expression values (hists)
    a. There was half of the genes expressed (we assumed other half was unexpressed)
2. Plot the distribution of 95pctl per sample (hist)
    a. There are high and low 95th pctls
    b. Why? because of RiboD / PolyA difference
3. Plot gene expression vs. 95th pctl per sample (scatter)
    a. expression count doesn’t affect patients 95th pctl pan-sample and pan-center
    b. TH01 sample center = positive correlation p95/unexpression count (47.6%)
4. Plot UMEND count vs. 95th pctl pan-sample (scatter)
    a. UMEND read doesn’t affect p95 pan-sample and pan-center
    b. TH01 sample center = positive correlation p95/UMENDread (70.1%)
5. Method RiboD vs. PolyA 95th pctl (boxplot)
    a. RiboD 95th Pctl pan-sample < PolyA 95th Pctl pan-sample

Results full

6. Method RiboD vs. PolyA Expression (boxplot)
    a. RiboD Expressed Genes > PolyA Expressed Genes
7. Plot 95th pctl of expected_count vs. 95th pctl of TPM (scatter)
    a. RiboD has less variance comparing actual count to depth of sequencing
8. High variance genes (hists)
    a. Samples with high gene variance have similar distributions no matter their 95th pctl

Code to calculate Variance and Mean

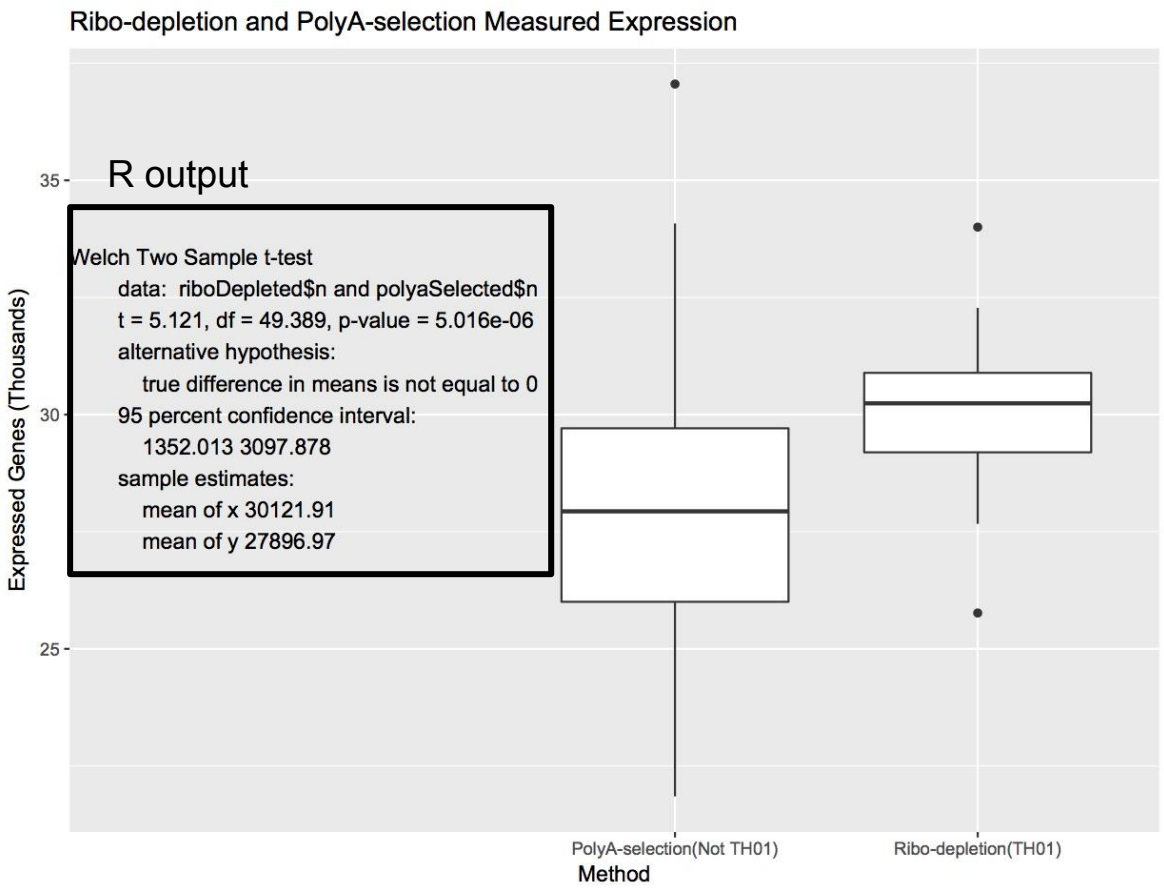
```
mean(dfGeneMean$mean)
# the overall mean is 1.019 sample

dfGeneVar <- outlierResults %>%
        group_by(Gene) %>%
        summarize(variation = var(sample)) %>%
        arrange(desc(variation))

mean(dfGeneVar$variation) # 0.56
# so most genes differ from the norm on average by 0.56

sd(dfGeneVar$variation) # standard deviation = 1.15

summary(dfGeneVar)
 #     Gene             variation
 # Length:58581       Min.   : 0.000000
 # Class :character   1st Qu.: 0.001579
 # Mode  :character   Median : 0.069611
 #                    Mean   : 0.565726
 #                    3rd Qu.: 0.707318
 #                    Max.   :28.579563
```

```
quantile(dfGeneVar$variation, 0.95)
# > 95% of the data variates from the mean by 2.56
```

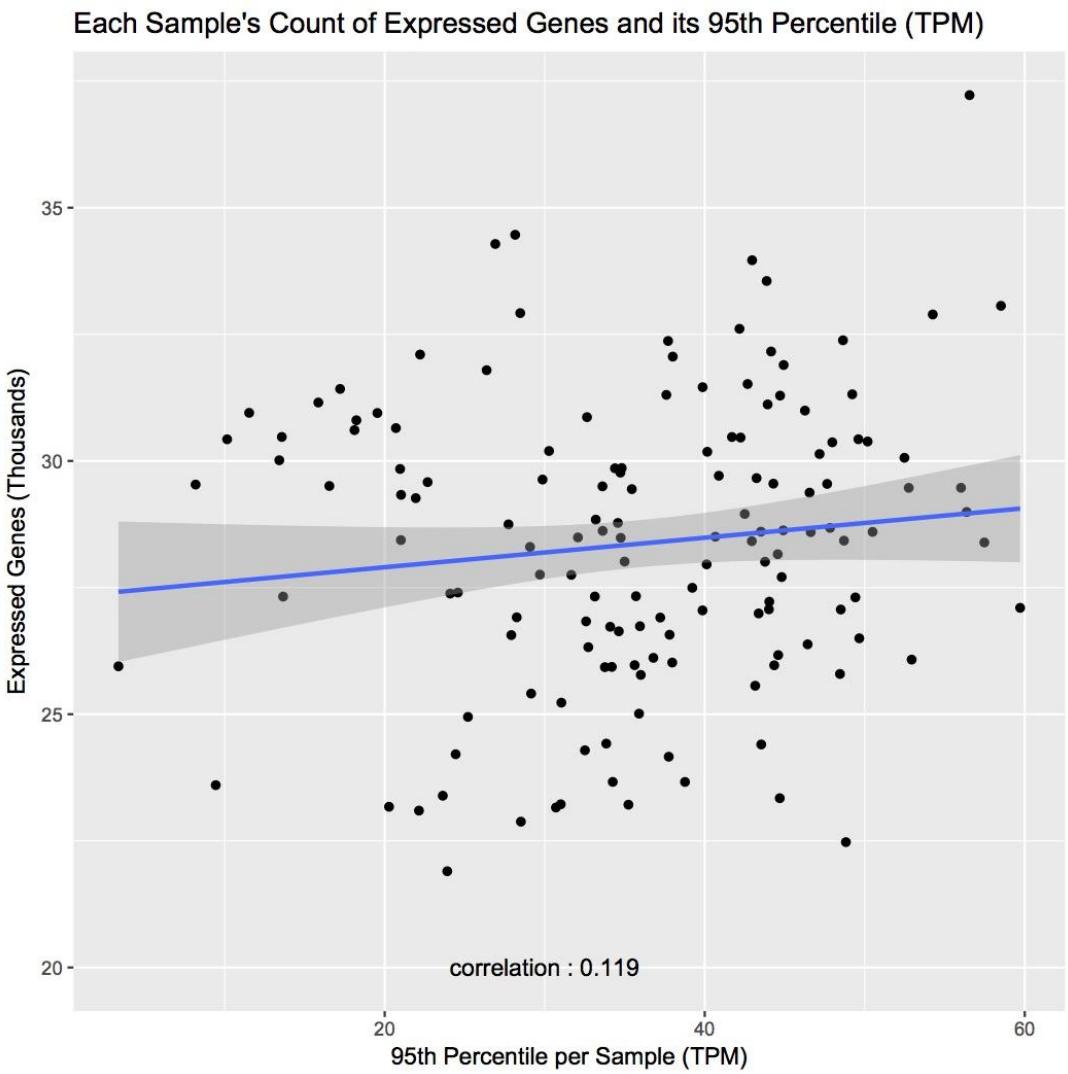
```
geneList <- dfGeneVar %>% filter(variation >
quantile(dfGeneVar$variation, 0.95))
# get names of genes p95 of variation and up
```

```
dfSamples <- outlierResults %>% group_by(sampleID) %>% filter(Gene
%in% geneList$Gene)
        # match names to all of their th01 th02 etc...
```

Reference In R Code




```
ggplot(dfBox, aes(x=Method, y=n/1000)) + geom_boxplot()+
        ylab('Expressed Genes (Thousands)') + xlab('Method') +
        ggtitle('Ribo-depletion and PolyA-selection Measured Expression')
        annotate("text",x = -1,y = 30 , hjust =0,
                label = paste0(  "Welch Two Sample t-test...)

                )
                riboDepleted <- filter(dfBox,
                        Method=="Ribo-depletion(TH01)")
                polyaSelected <- filter(dfBox,
                        Method=="PolyA-selection(Not TH01)")
                t.test(riboDepleted$n, polyaSelected$n,
                        alternative = "two.sided",
                        mu = 0, paired = FALSE, var.equal = FALSE,
                        Conf.level = 0.95)
```

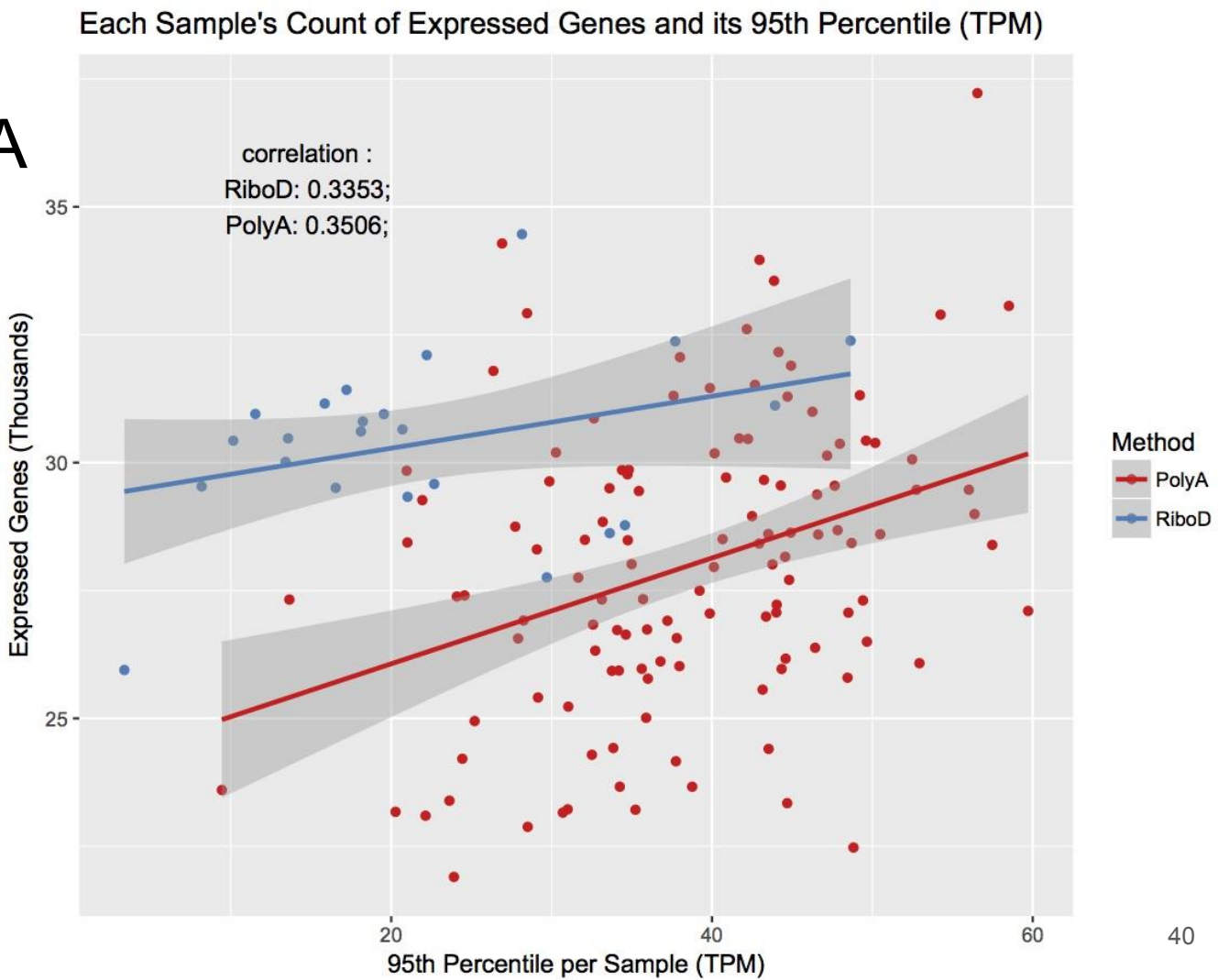


TPM increases,
p95 doesn't change

RiboD vs PolyA TPM

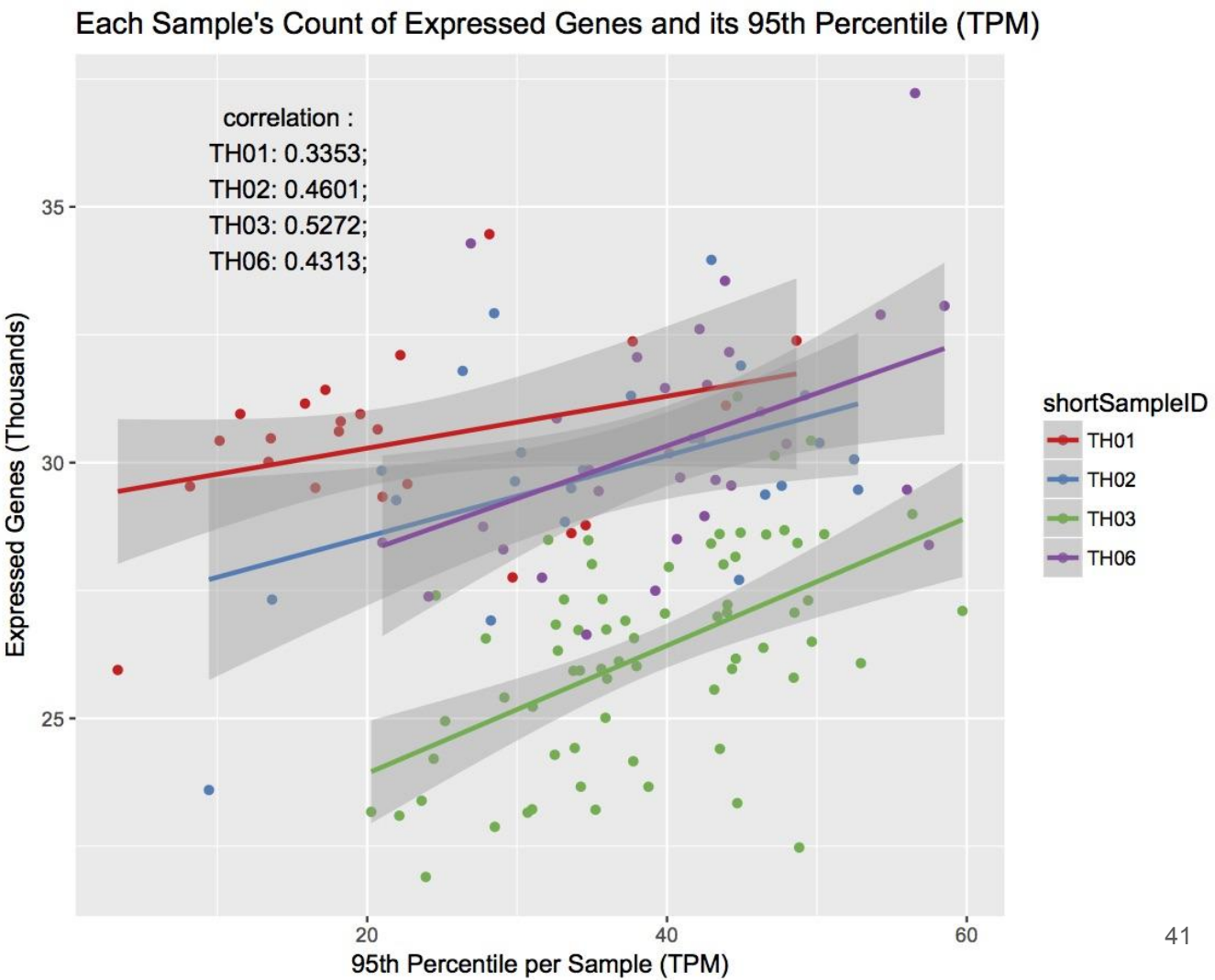
Same correlation

RiboD has more expressed genes

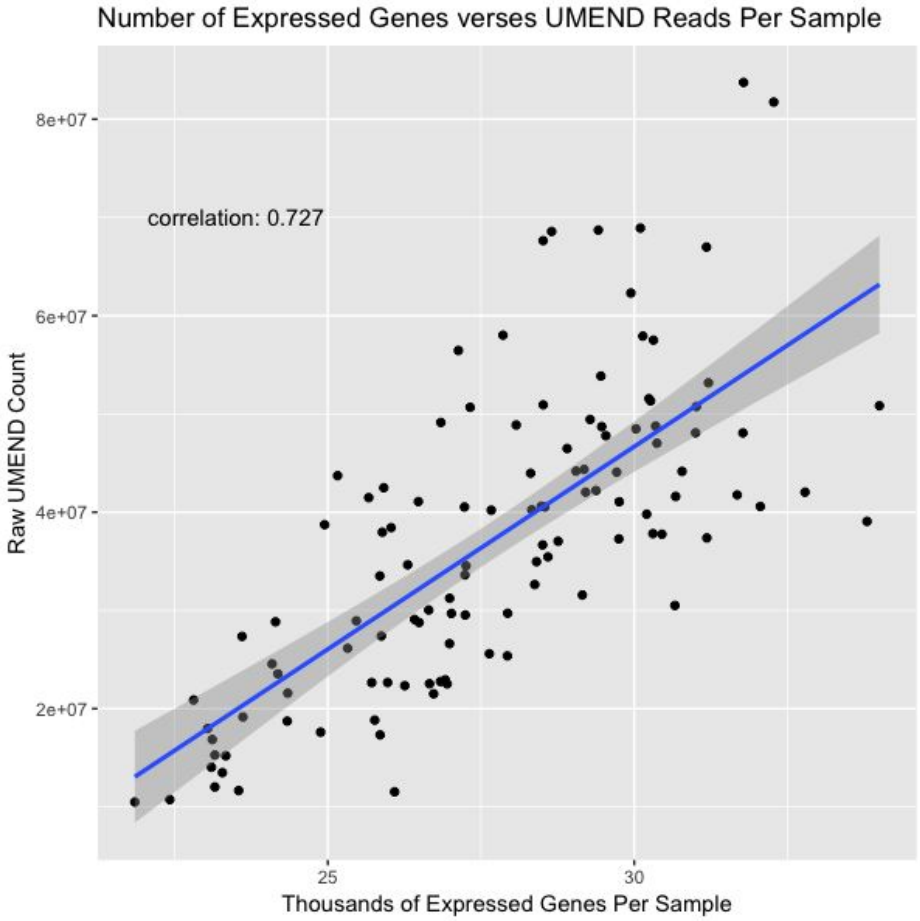


TH03 has the Most correlation

TH03- low expression, high 95th pctls



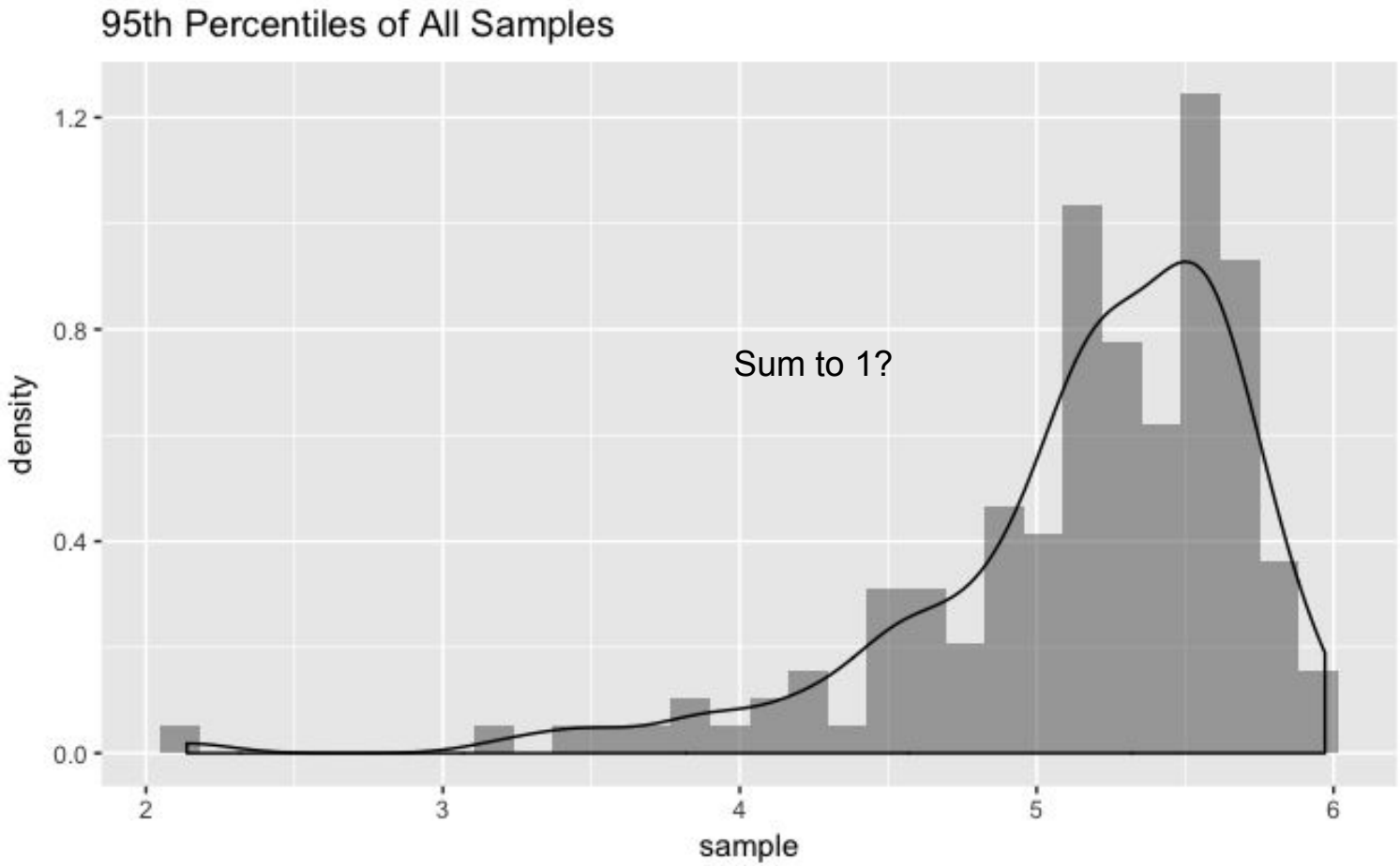
More UMEND reads = More Expressed Genes

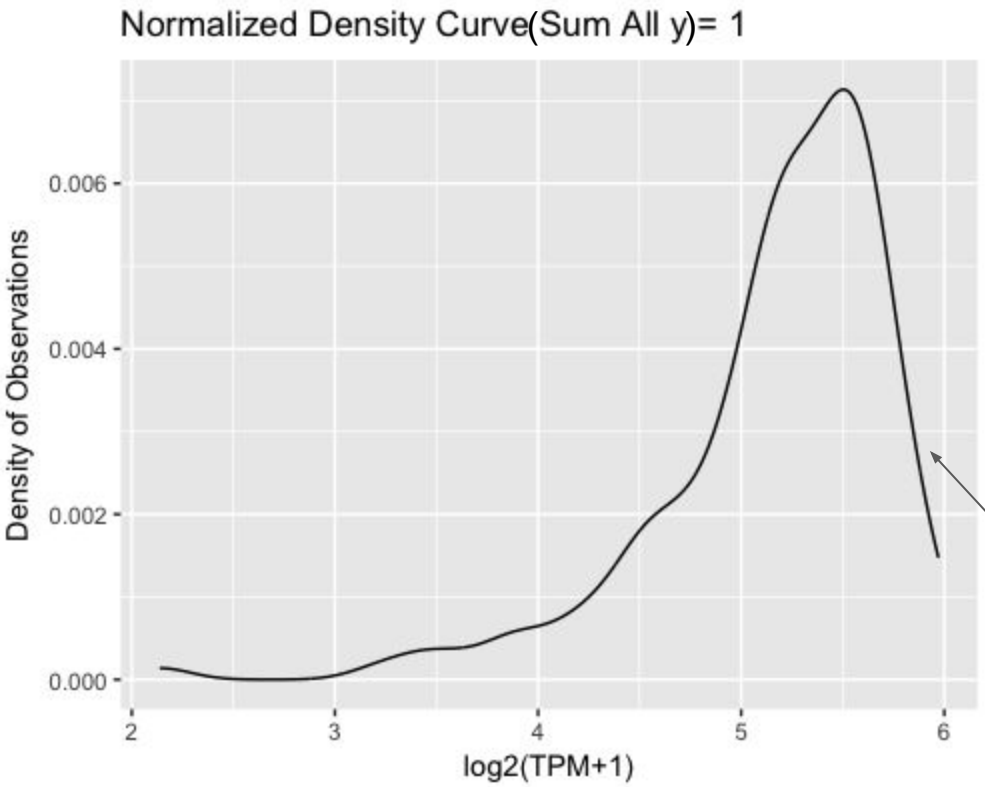


- More UMEND = more expressed Genes per sample
- 72.7% Correlation

Ideas (old)

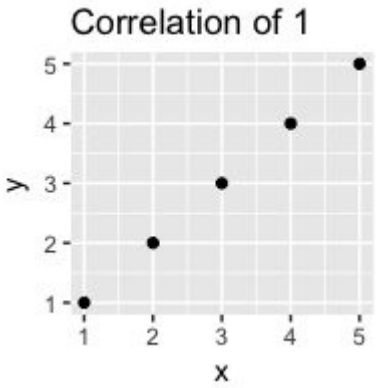
- 22 best/worst counts histogram for expected_count
- Get back values from ggplot
- Get values its going to plot
- Pick max y val greater than 2 log2(TPM+1)
- Find max using calculus with density points
-


95th Percentiles of All Samples
1.2
0.8
0.4
0.0
density
Sum to 1?
2
3
4
5
6
sample




```
m <- ggplot(percentileOfEachSampleDf, aes(p95))
m <- m + geom_density()sum(p$data[[1]]$y)
# sum of data is 173.9702
# I want to divide the points of the density
p <- print(m)

head(p$data[[1]]$y)
by the sum of all to get a normalized curve that adds
to one

normalized <- data.frame(p$data[[1]]$y,p$data[[1]]$x)
sum(p$data[[1]]$y/sum(p$data[[1]]$y))
# this sum is 1

ggplot(normalized, aes(p$data[[1]]$x,
p$data[[1]]$y/sum(p$data[[1]]$y))) + geom_line() +
  ggtitle("Normalized Density Curve Sum All y = 1") +
  ylab("Density of Observations") +
  xlab("log2(TPM+1)")
```

```r
cor(dfScatter$n,dfScatter$p95)
# 0.08755389
# about 8.8% correlation

# REFERENCE
x <- seq(1,5)
y <- x
df = data.frame(x,y)

cor(x,y)
ggplot(df, aes(x,y)) + geom_point() +
ggtitle("Correlation of 1")
# correlation of 1 (when x = y)
```



Cohorts Used

- Ckcc rsem genes → Raw TPM values of each sample
- Ckcc comp4.3 tert8 → Log_2(TPM+1) values of each sample

- TPM
  - Normalizes all gene expression from RNA seq reads
  - Transcripts Per Kilobase Million
    - Divide the read counts by length of each gene in Kb
    - Add all of these up and divide by 1,000,000 (scaling factor)

Sum of TPM Values

```
sum_TPMDf <- rawTPMDf %>%
  group_by(sampleID) %>%
  summarise(sum = sum(TPM))
   sampleID                           sum
   <chr>                            <dbl>
 1 TH01_0053_S01_rsem_genes.results 1000000
 2 TH01_0054_S01_rsem_genes.results 1000000
 3 TH01_0055_S01_rsem_genes.results 1000000
 4 TH01_0061_S01_rsem_genes.results  999998
 5 TH01_0062_S01_rsem_genes.results 1000000
 6 TH01_0063_S01_rsem_genes.results 1000000
 7 TH01_0064_S01_rsem_genes.results 1000000
 8 TH01_0069_S01_rsem_genes.results  999998
 9 TH01_0120_S01_rsem_genes.results  999999
10 TH01_0121_S01_rsem_genes.results 1000001
# ... with 136 more rows

min(sum_TPMDf$sum)
 [1] 999997.7
max(sum_TPMDf$sum)
 [1] 1000001
```

- Sum of all TPM values for each sample are 1,000,000 ± 1 TPM