

# Preliminary RNA-seq analysis (QC)

Harrison Smith

# Workflow

## Outline of Contribution

1. Data analysis
  - a. QC of sample
  - b. Mapping alignment
  - c. QC of alignment
  - d. Quantification
  - e. Normalization
  - f. QC of replicates
  - g. Differential expression

## Step in Project

1. Experiment Design
2. RNA isolation
3. Tapestation quantification
4. cDNA library prep
5. Next Gen Sequencing
- 6. Data analysis**

# Experimental Aims

1. Observe genetic effects of HAT inhibition upon TGFb stimulation in Adult Mice Ventricular Fibroblasts (AMVFs)
  - a. Deduce if HAT inhibition has global effects
  - b. Isolate all genes contributing to the cause and prevention of Cardiac Fibrosis.
  
2. Examine the dichotomy between histone acetyltransferase inhibition and histone reader inhibition.
  - a. Contrast efficacy between Brd4 and CBP+ p300 inhibition
  - b. Determine which inhibitor has a less broad effect on genes

# Experimental Design

Analyze bulk RNA samples from Adult Mice Ventricular Fibroblasts (AMVFs)

4 Experimental Groups

3 replicates each

12 Samples total

**DMSO**

**TGF $\beta$**

**TGF $\beta$  + A485**

**TGF $\beta$  + PFCBP1**

Control

Treated

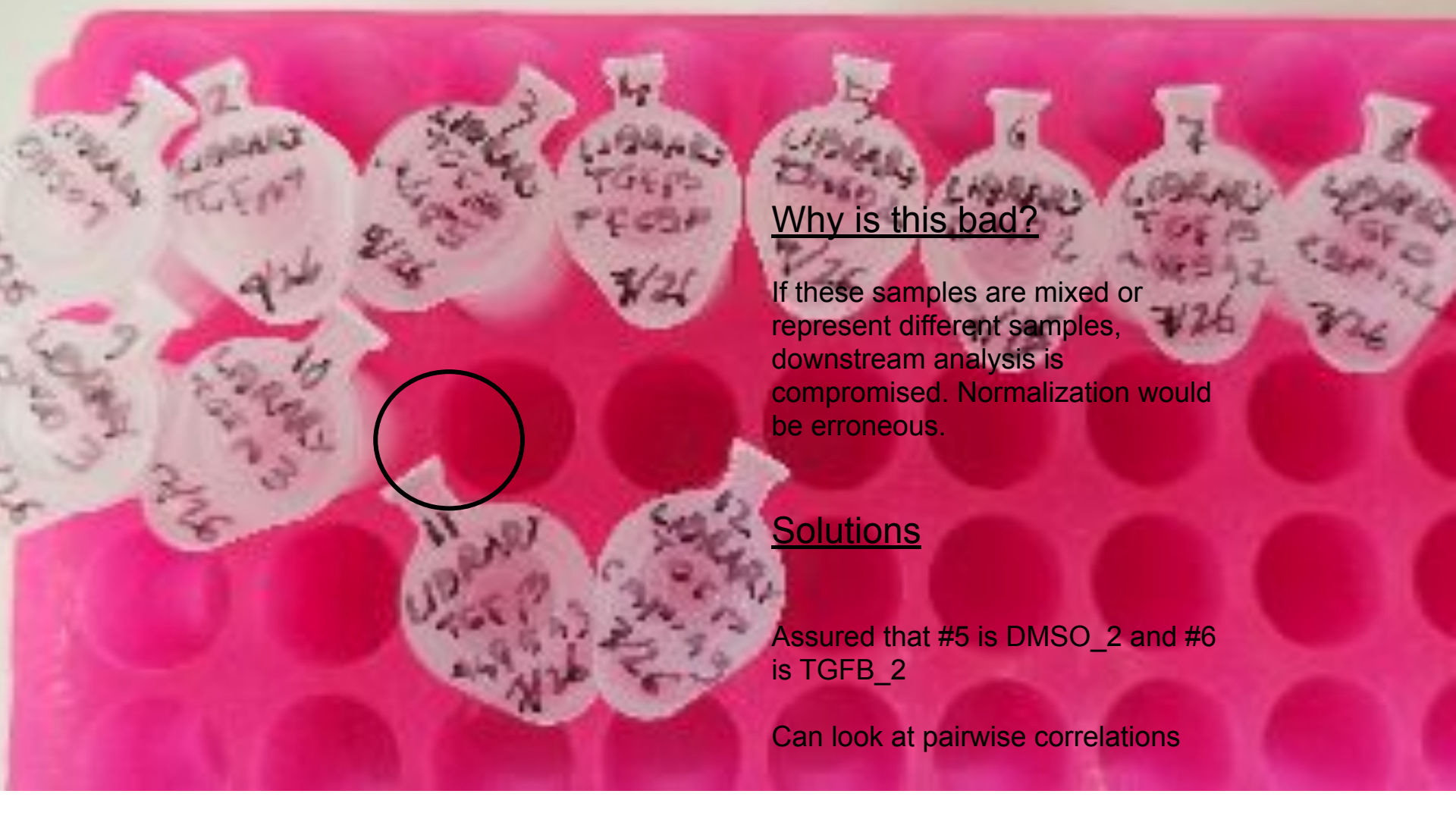
HAT Inhibitor 1

HAT Inhibitor 2

# Methods

## RNA-seq processing

Adapter Trimming	:	BBduk
Alignment	:	STAR
Quantification	:	featureCounts
Normalization	:	FPKM, TPM, CPM
Differential expression	:	EdgeR, DESeq2
Quality Control	:	MultiQC



### Why is this bad?

If these samples are mixed or represent different samples, downstream analysis is compromised. Normalization would be erroneous.

### Solutions

Assured that #5 is DMSO\_2 and #6 is TGFB\_2

Can look at pairwise correlations

# Issues

## 2. Low quality of sample #8

- a. Not enough sample for effective library size.



## Why is this bad?

This sample represents  $\frac{1}{3}$  of the replicates for an experimental group. Analysis of TGFb + PFCBP1 Treatment is less reliable.

## Solution

The replicate can be dropped as two replicates are better than one but accuracy takes a hit.

# Issues

## 3. Mislabeled Fastq files

- a. The order in which files were labeled did not match the index proposed in experiment

## Why is this bad?

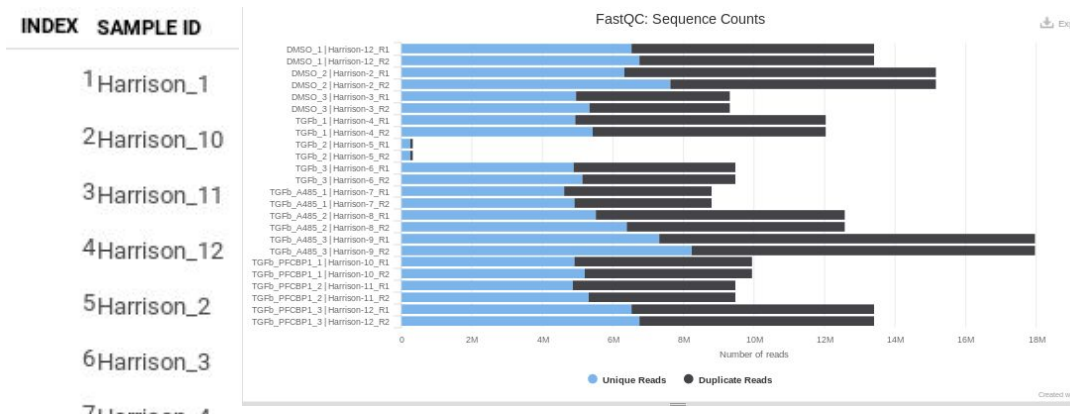
Analysis would be completely inaccurate

## Solution

Downstream pairwise quality control would have identified this, but the low read sample # changed. This was a red flag.

Renaming while downloading and uploading corrected this.

From Biospace :





# Issues

## 4. Duplicate sample #12

- a. When downloading samples from source. Technical error causes duplication of sample # 12.

## Why is this bad?

Sample #1 was replaced with a copy of sample #12. Data is skewed and unreliable.

## Solution

Downstream pairwise quality control would have identified this, but the low read sample # changed while viewing on Lin Lab's platform, Genialis.

Another red flag

Renaming while downloading and uploading corrected this.

# Quality Control

## Library Size

Good Sample examples

<u>DMSO 1   Harrison-1 R1</u>			
Unique Reads:	7 251 029	(43.5%)	
Duplicate Reads:	9 428 103	(56.5%)	

<u>DMSO 2   Harrison-2 R1</u>			
Unique Reads:	6 332 118	(41.7%)	
Duplicate Reads:	8 839 300	(58.3%)	

<u>TGFB 2   Harrison-3 R1</u>			
Unique Reads:	4 957 856	(53.1%)	
Duplicate Reads:	4 377 175	(46.9%)	

<u>TGFB 3   Harrison-7 R1</u>			
Unique Reads:	4 637 906	(52.6%)	
Duplicate Reads:	4 172 806	(47.4%)	

<u>TGFb A485 3   Harrison-8 R1</u>			
Unique Reads:	5 526 758	(43.9%)	
Duplicate Reads:	7 050 533	(56.1%)	

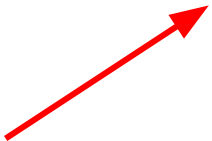
<u>TGFb A485 1   Harrison-11 R1</u>			
Unique Reads:	4 869 319	(51.3%)	
Duplicate Reads:	4 620 462	(48.7%)	

<u>TGFb PFCBP1 3   Harrison-9 R1</u>			
Unique Reads:	7 330 041	(40.8%)	
Duplicate Reads:	10 647 632	(59.2%)	

<u>TGFb PFCBP1 1   Harrison-12 R1</u>			
Unique Reads:	6 544 152	(48.8%)	
Duplicate Reads:	6 877 510	(51.2%)	

Sample #8

<u>TGFb PFCBP1 2   Harrison-5 R1</u>			
Unique Reads:	270 696	(81.0%)	
Duplicate Reads:	63 346	(19.0%)	



Here we can see the library size of sample #8 is drastically lower than that of other samples.

This would fail to show accurate normalized counts due to insufficient sequencing depth.

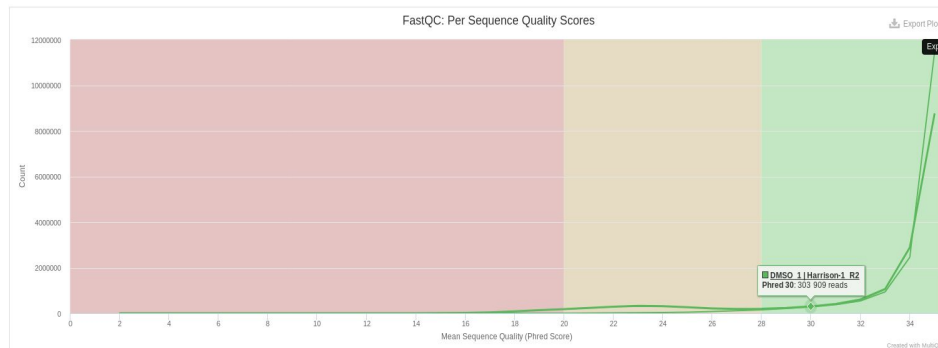
# Quality Control

## Read/ Fragment quality

In addition to fastqc, **multiQC** covers additional quality control aspects over read quality.

Here we can see the mean quality score per base of all the fragments of this particular sample. This is prior to BBDuk's adapter trimming. The Next Gen sequencing step can take care of adapter trimming prior.

The bottom figure represents overall quality of total fragments in a sample. The average Phred score is in the green zone. Good

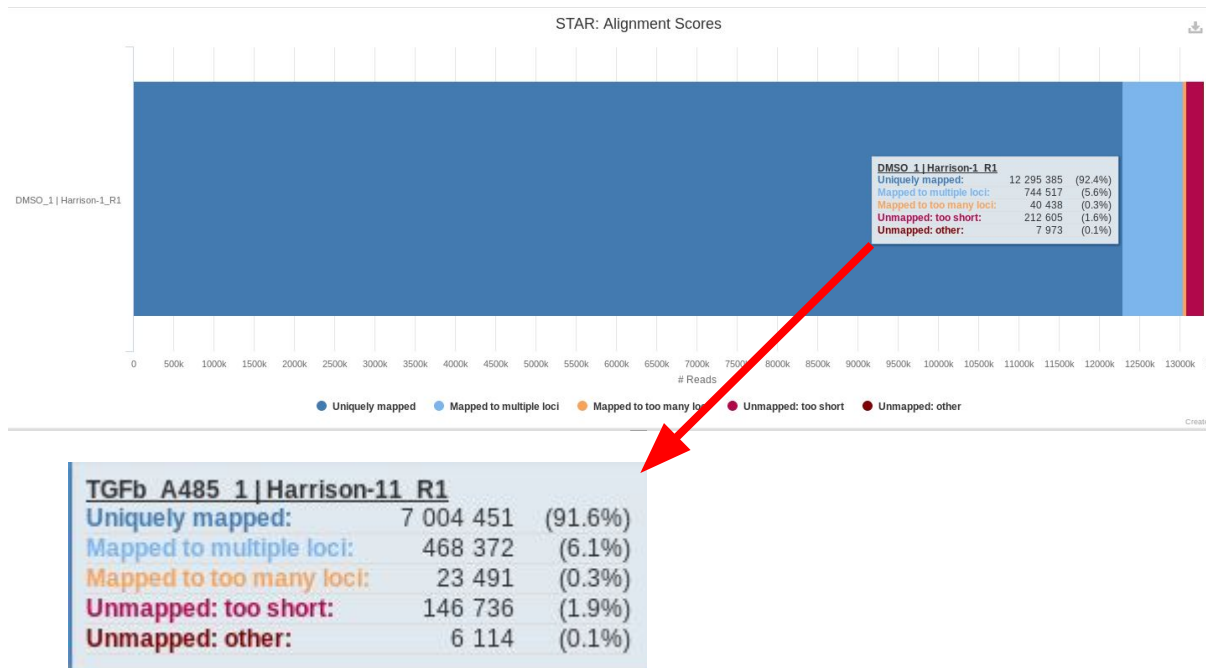


# Quality Control

## Alignment mapping

This multiQC plots visualises the distribution of reads mapped to the mm10 genome using the STAR aligner.

All samples had consistent mappability.

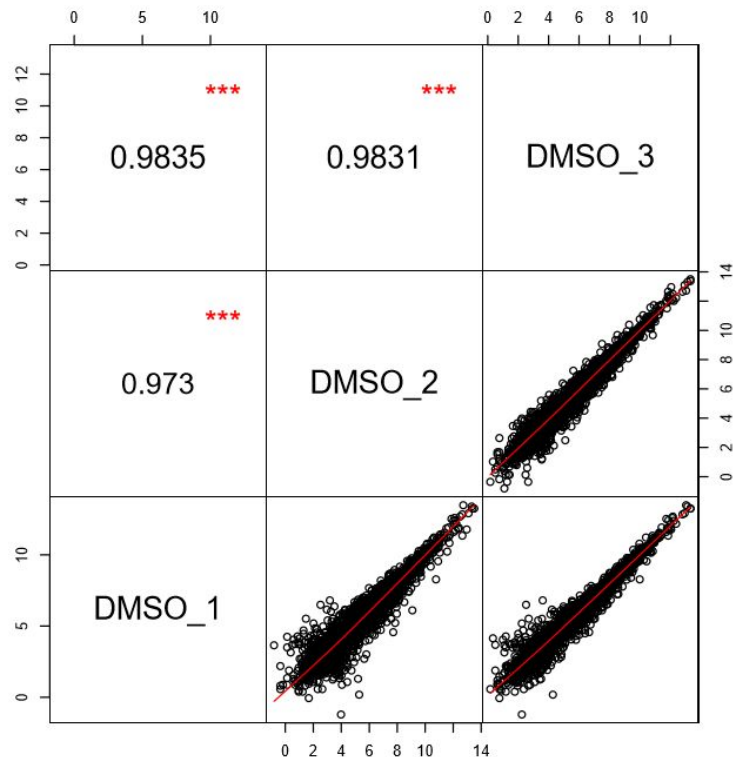


# Pairwise Replicate Correlations (logCPM)

Explains some similarity among replicates.

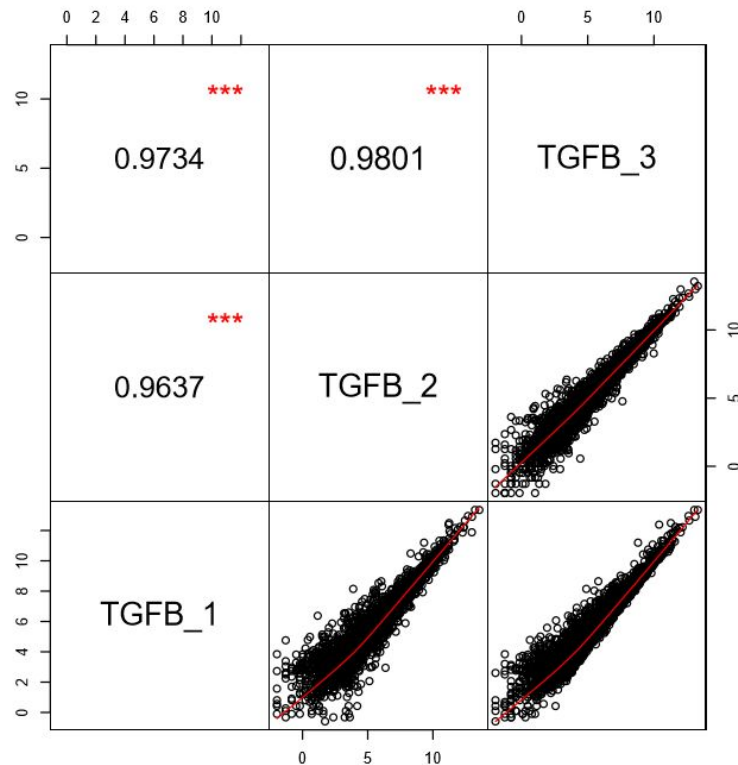
X < 90% is typically concerning but the lowest score of a group is also a good indicator if it's slightly above 90.

Here we can see the DMSO group is pretty similar



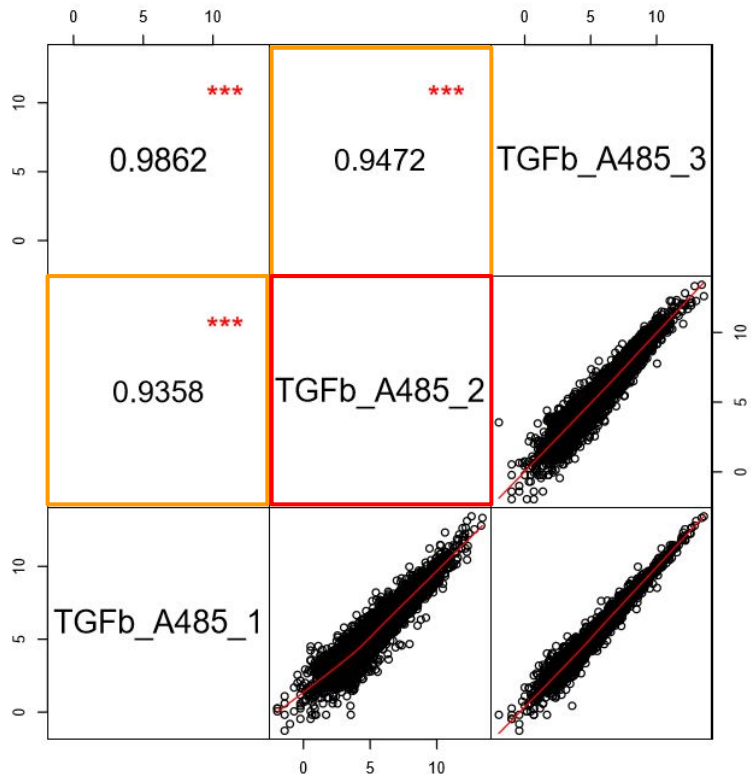
# Pairwise Replicate Correlations (logCPM)

Here we can see the TGFb stimulated group is also seems fairly similar.



# Pairwise Replicate Correlations (logCPM)

Here you can see that replicate #2 is questionably lower than the others and may be excluded from analysis for an increase of accuracy but this may be hasty prior to viewing all correlations.

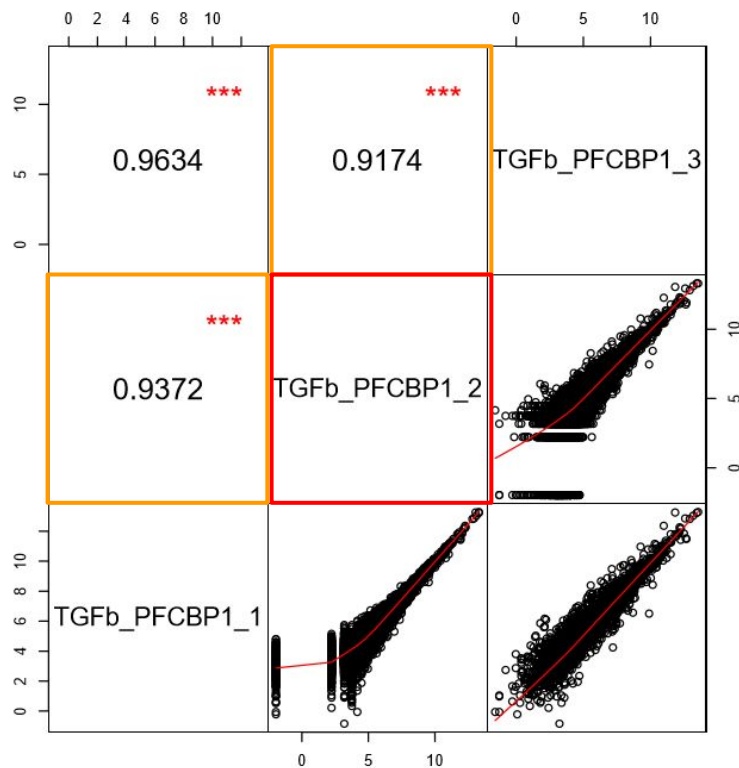


# Pairwise Replicate Correlations (logCPM)

Here you can see that replicate #2 is significantly less similar than the rest.

This is due to the small library size of sample #8

Should definitely be dropped.

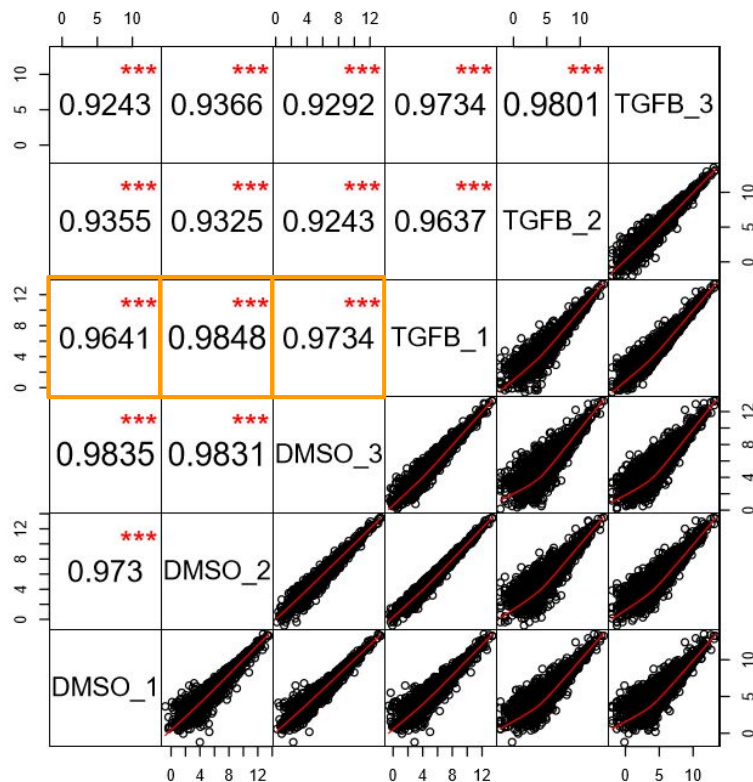




# Pairwise Replicate Correlations (logCPM)

Let's view correlation scores of mixed experimental groups to validate scores.

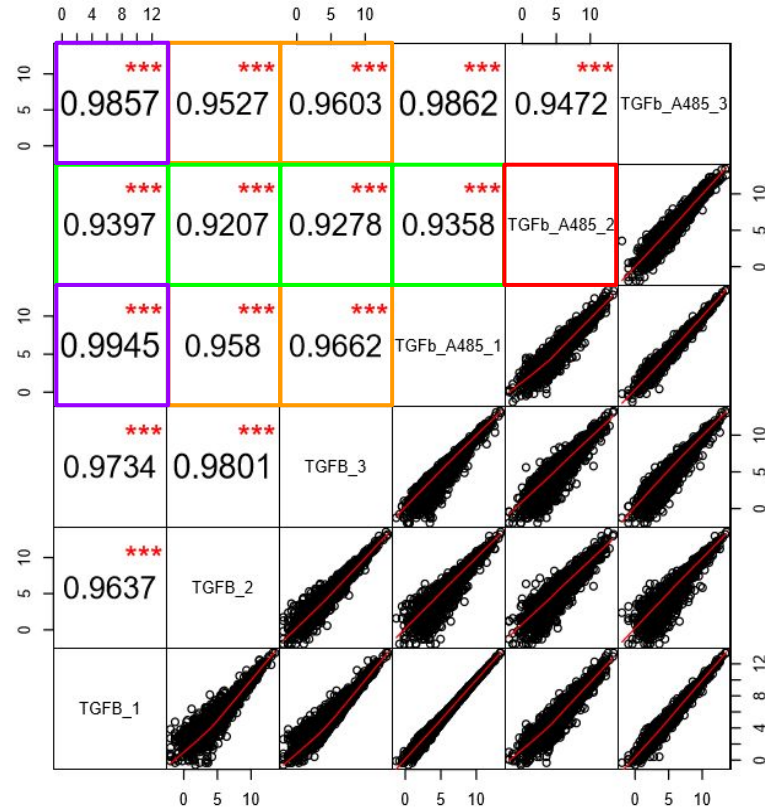
As we would expect, the groups are mostly dissimilar. However, TGFb rep #1 strangely correlates with the DMSO group. Even more so than its respective replicates.



# Pairwise Replicate Correlations (logCPM)

Again, we see more similarity among separate groups. I use purple as extreme examples :(

At least A485 rep #2 does not correlate more strongly in a different group.



# Pairwise Replicate Correlations (logCPM)

## Conclusion

- Sample TGFb + PFCBP1 replicate #2 will be excluded.
- Sample TGFb + A485 replicate #2 can also be excluded.
- Similarity between experimental groups can indicate weaker or stronger administration of drug but also questions the integrity of the results.

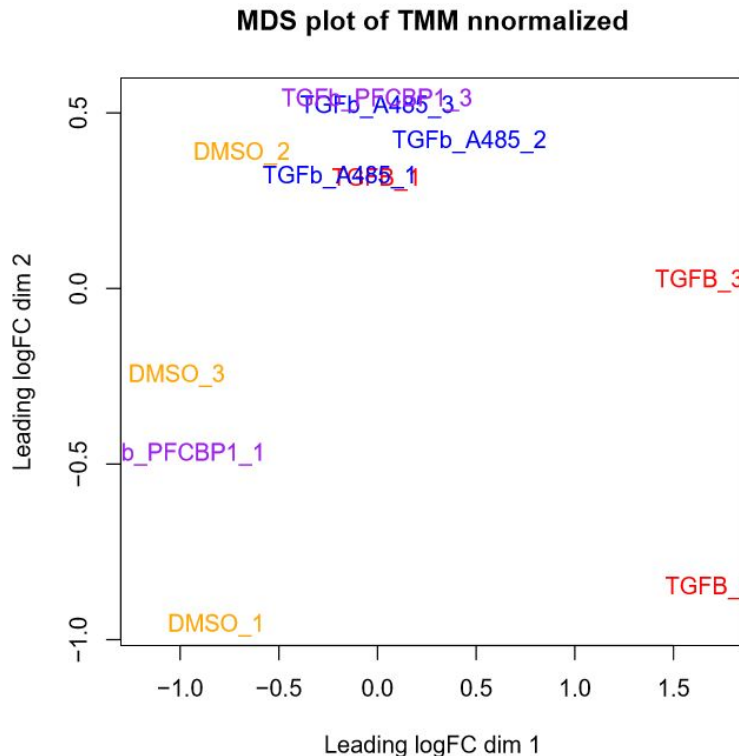
# Multidimensional Scaling (MDS)

Here we are viewing the TMM normalized reads of the leading logFC genes among replicates to determine batch effects that may skew results.

Similarly to PCA, similar samples should cluster.

The distance between each pair of samples can be interpreted as the leading log-fold change between the samples for the genes that best distinguish the pair of samples. By default, leading fold-change is defined as the root-mean-square of the largest 500 log2-fold changes between that pair of samples.

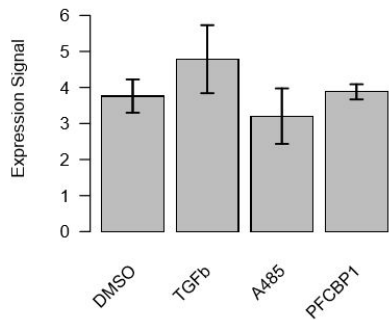
There's quite the cluster among different replicate groups which is concerning.



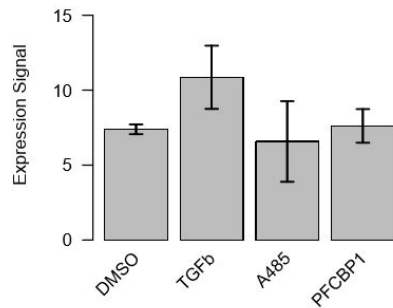
# Expression Graphs

Here we are viewing the expression of genes post FPKM normalization and post dropping of sample #8

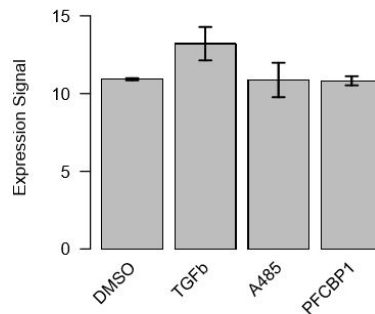
**Crebbp**



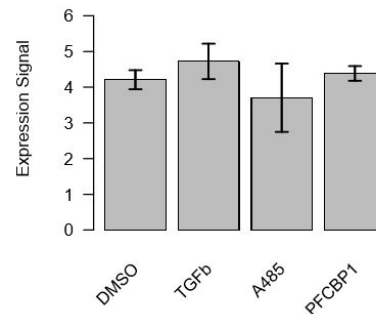
**Ep300**



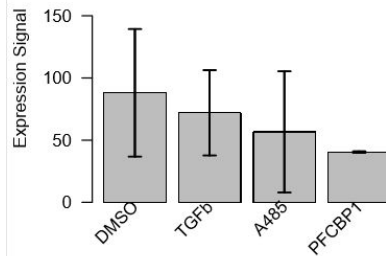
**Map3k7**



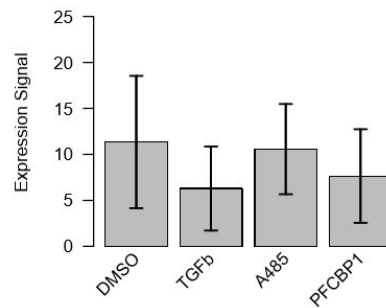
**Brd4**



**Postn**



**Sertad4**





# Expression Graphs

**Polr2a**

