Differential Expression analysis

Harrison Smith

Outline

<u>Presentation Format</u>

1. Inquiries
2. Methodology
3. Differential Expression

   a. DESeq2/EdgeR

      i. DEG HeatMap

      ii. DEG barplots

   b. Summary

4. Pathway Analysis

   a. GO

   b. GSEA

5. Interpretation

<u>Outline of Contribution</u>

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

   h. Pathway Analysis

Experimental Questions

1) What is the gene expression profile of TGF-beta stimulation upon AMVF cells?

    a) Do we see Global changes mRNA expression?

    b) Do we see upregulation of genes in the canonical and non canonical TGF-beta pathways?

2) Do we see validation of the disease state based on select DE genes and pathway analysis?

3) What are the effects of HAT inhibition on TGF-beta treated AMVFs?

    a) Do we see global changes in mRNA expression
    b) Do we see a reduction in disease state pathways
    c) Can we isolate a gene list composed of all genes contributing to the disease.
    d) Can we isolate a gene list contributing to the reduction of ECM matrix development.
    e) What are the genetic side effects of HAT inhibition.
    f) Is HAT inhibition more or less broad than BRD4 inhibition

Experimental Design

Biological replicates of Adult Mice Ventricular Fibroblasts

4 Experimental Groups

3 replicates each

12 Samples total

DMSO	**TGFβ**	**TGFβ + A485**	**TGFβ + PFCBP1**
Control	Treated	HAT Inhibitor 1	HAT Inhibitor 2

Methods

RNA-seq preparation

Library Prep	:	NebNext Ultra directional RNA seq kit
Library quantification	:	NebNext library quant kit
Sequencing	:	Illumina
Read type	:	Paired-end (first stranded)
Fragment depth	:	5-12 million reads
Fragment length	:	~75 bp

Methods

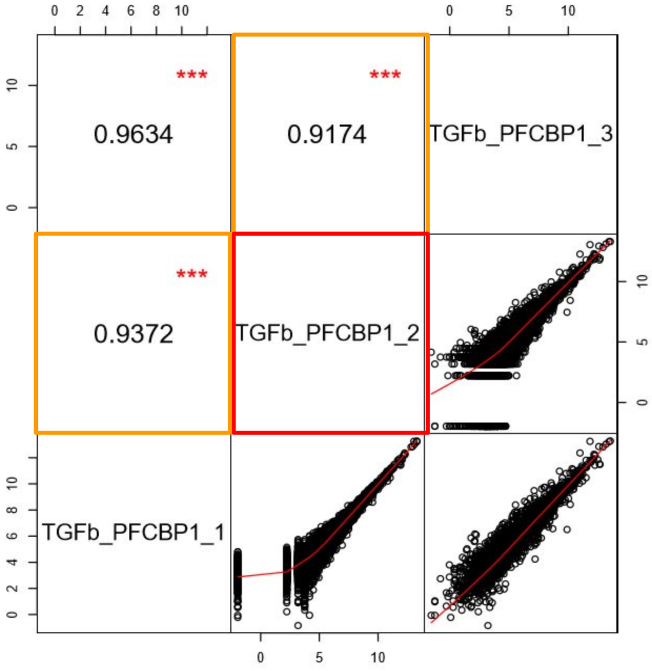
RNA-seq processing

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

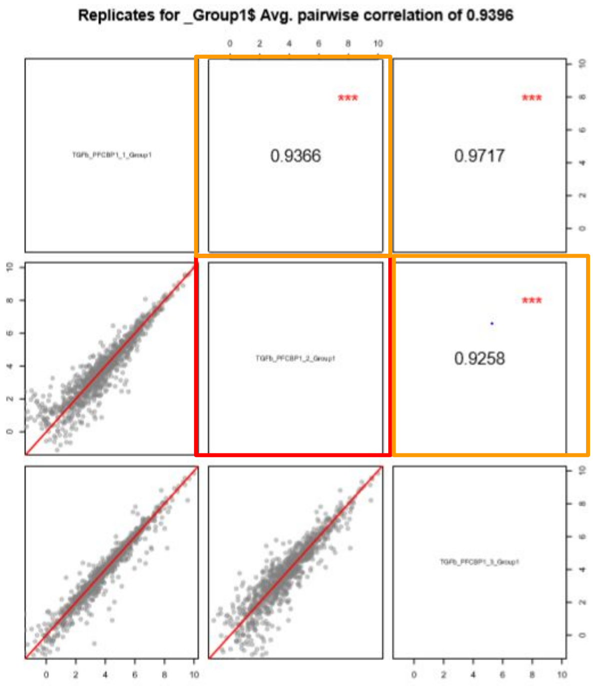
QC

Pairwise Replicate Correlations (PFCBP1)

logCPM



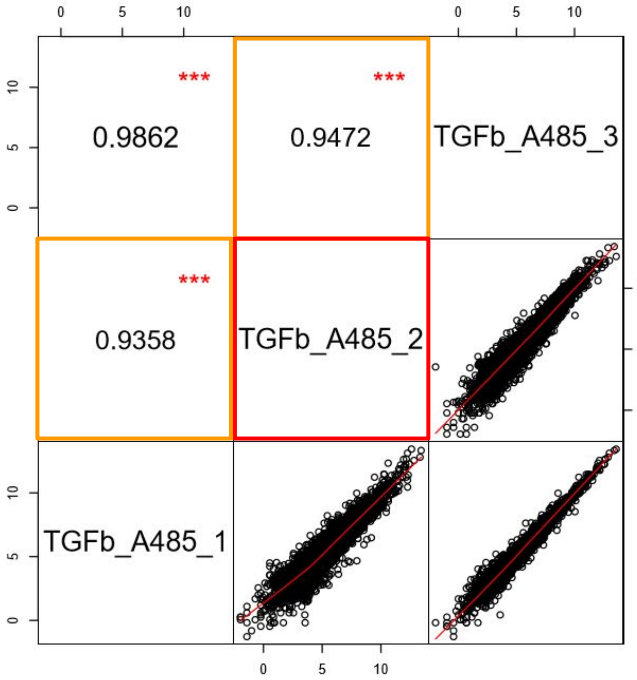
FPKM


Replicates for _Group1$ Avg. pairwise correlation of 0.9396

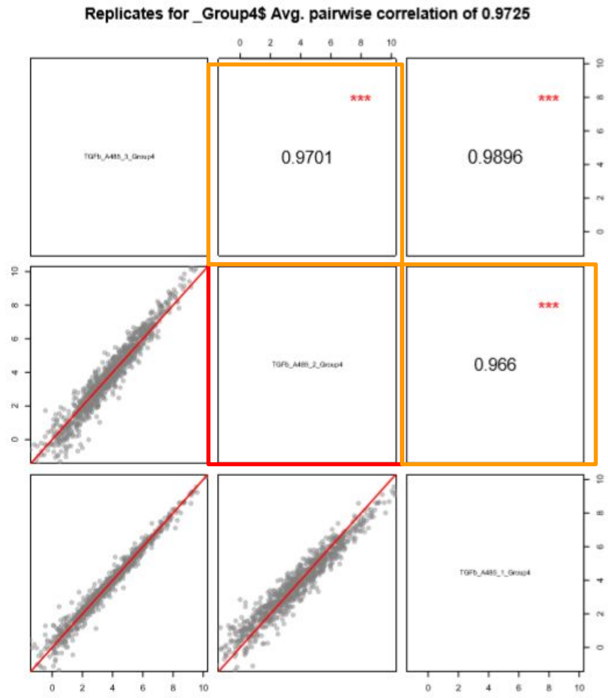
Pairwise Replicate Correlations (A485)

Here we can see the A485 is slightly more distince in with the log(CPM) counts but not quite as drastic as the FPKM normalized reads.

logCPM



FPKM

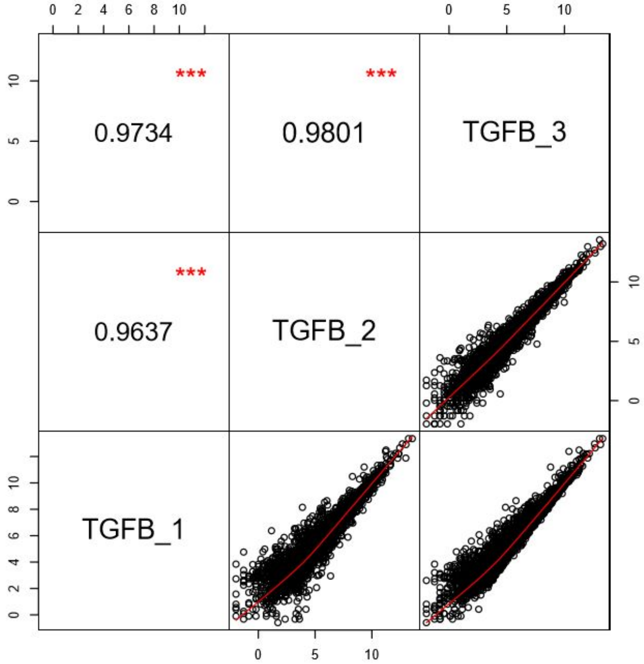


Pairwise Replicate Correlations (TGF-beta)

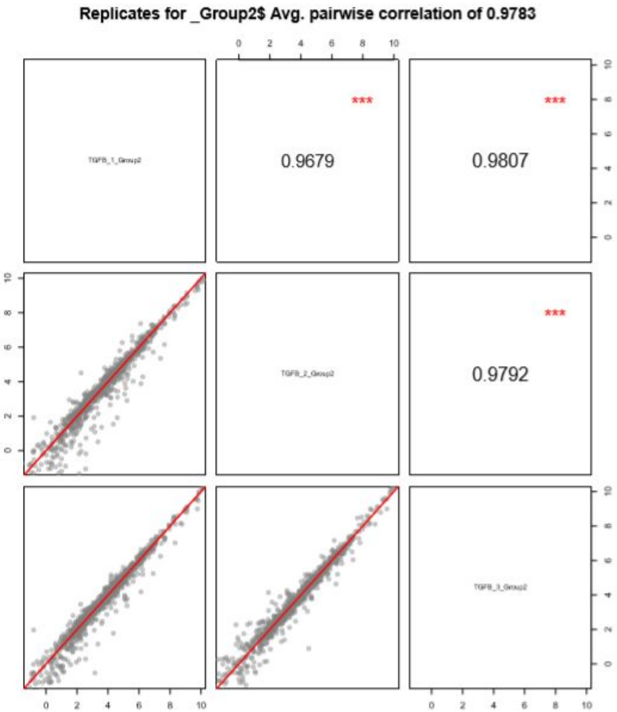
Nothing too drastic about TGFb1 which makes me question the TGFb group but hesitant to dropping.

We do see a stronger correlation from TGFb in other groups though.

logCPM



FPKM

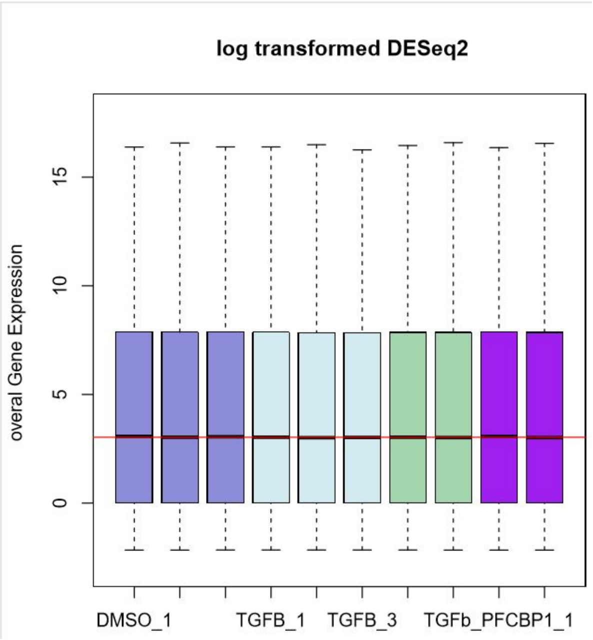

Replicates for _Group2$ Avg. pairwise correlation of 0.9783

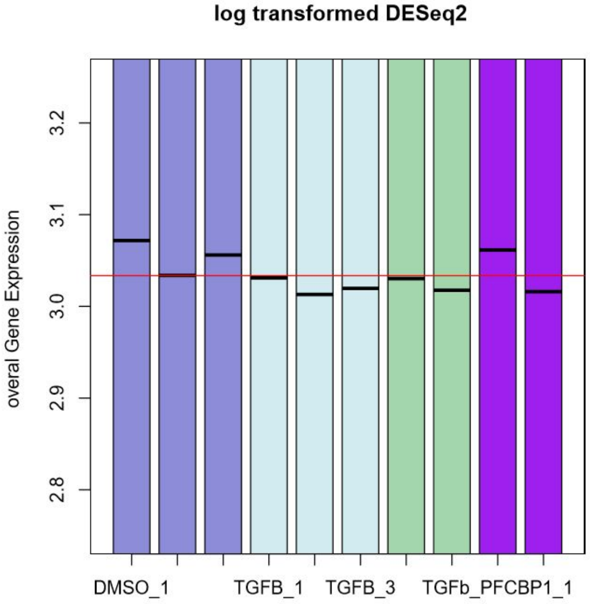
Overall Expression boxplots

Here we are viewing the total expression average of DESeq2 normalized reads across the replicates post dropping of sample #7 and #8

Differences are small yet we see lower overall average expression among the TGFb stimulated group.

Shouldn't stimulation induce more expression?

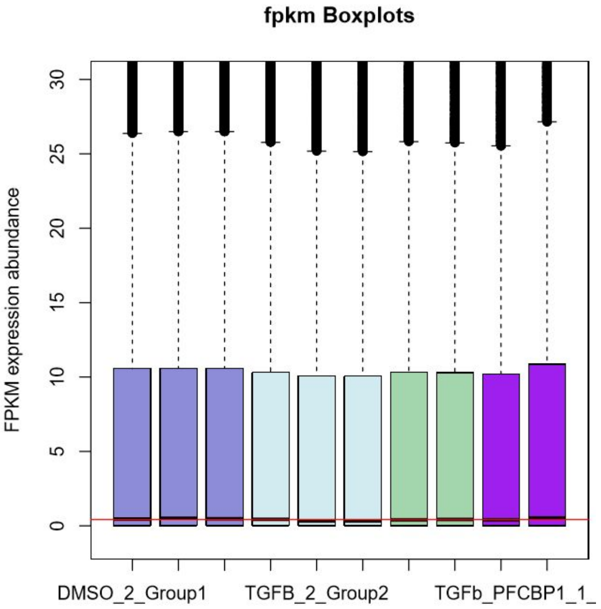


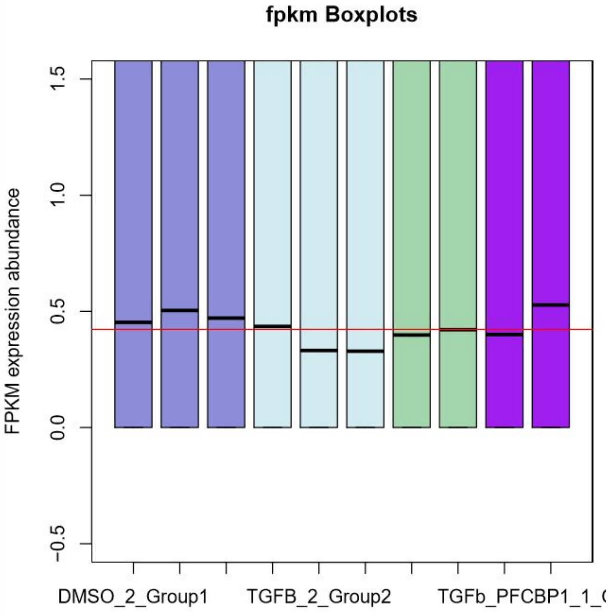


Overall Expression boxplots

Again, We see similar characteristics among FPKM expression.

Drastic differences among replicates are due to a different order than the previous plot but overall we still see similarity.



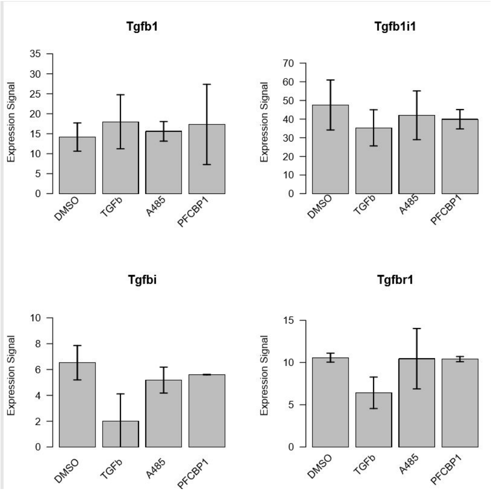


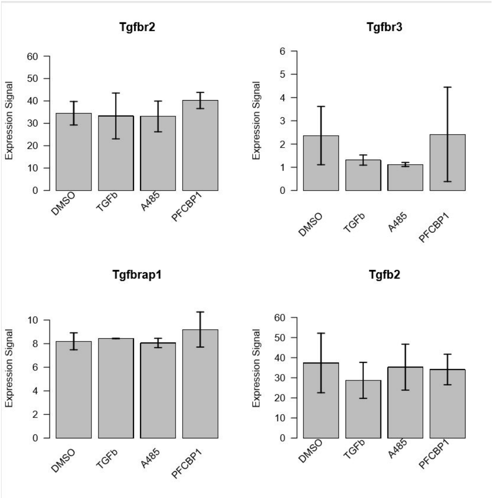
Validation of TGF-beta stimulation (FPKM)

Upon stimulation of TGFb I would expect to see more expression of the TGFb gene group and more expression of down-stream genes in all related pathways. Especially the canonical pathway. However, it seems the opposite is true. Which is surprising to me.

Here we explore the expression levels of TGF-beta proteins, Smad proteins and some non-canonical pathway proteins.

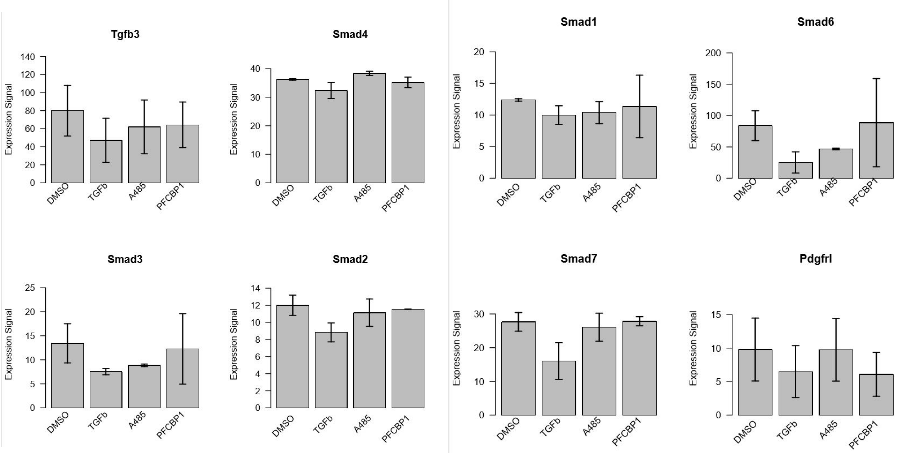
7 out of 9 TGF-beta related genes are distinctly down regulated in the supposed TGFb stimulated group



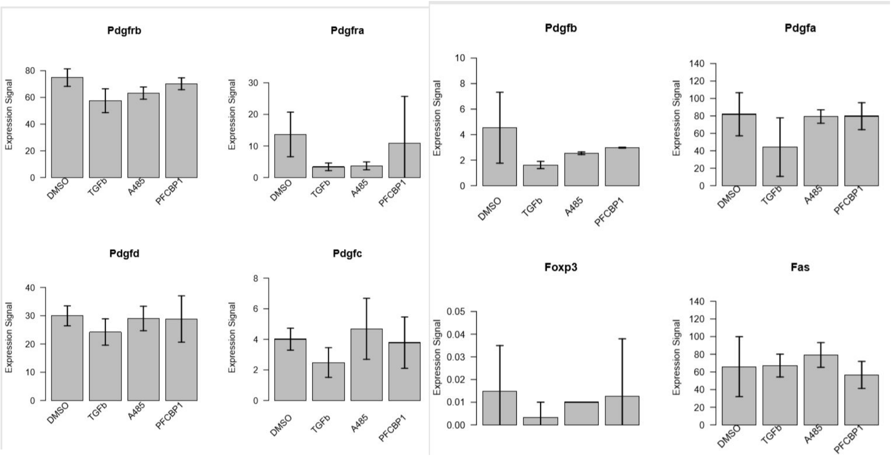


Gene expression

Here we can see the Smad proteins (canonical pathway) Are also down regulated in the TGF-b stimulated group.

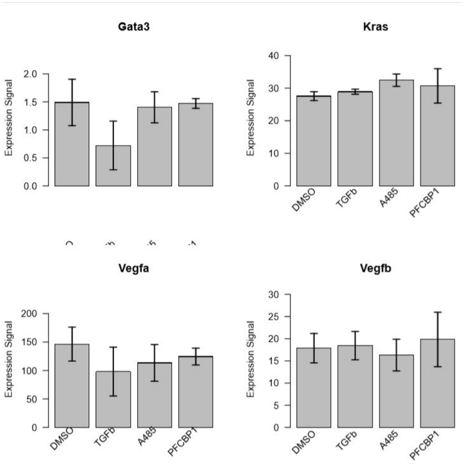


Here we see down regulation of cell proliferating genes and Immune response suppressors.

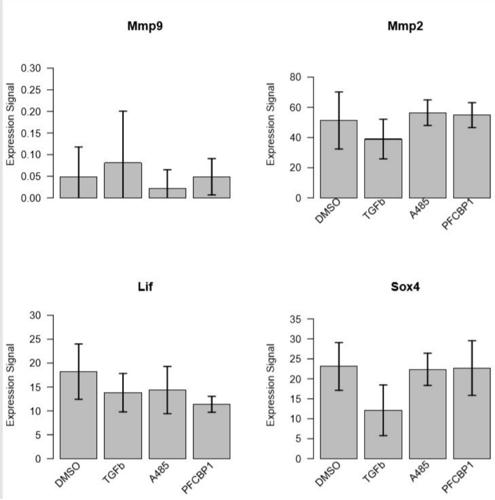


Gene expression

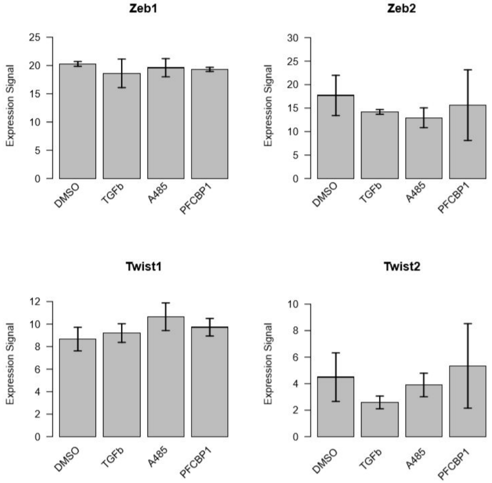
Immune response suppression and angiogenesis promotion.



Angiogenesis promotion and cancer stem cell self-renewal



Mesenchymal Markers



Methodology (normalization)

DESeq2s “log transformed”

- This normalizes using a geometric averaging using logs which reduces the effect that outliers have on the rest of the data. Estimates depth based on the count of the gene with the median count ratio across all genes.

EdgeRs “TMM normalization”

- Trimmed mean of M values are estimates of sequencing depth after excluding genes for which the ratio of counts in different samples is too extreme or if average expression is too extreme. Removes genes with extreme bias.

DESeq2 and EdgeR normalize for both sequencing depth and library compositions. Where as FPKM normalizes for sequencing depth and for gene length

FPKM values are typically used to measure expression within the same sample type. While on the surface it appears that this allows us to view change in expression it is not necessarily the case. DESeq2 and EdgeR use more rigorous and stringent methods that help determine, more accurately, the actual change in expression by a more reasonable statistical method. It seems that the general consensus is that FPKM is not ideal for differential expression but is still used in some cases.

Methodology (DE parameters)

Typically we use an adjusted P value (FDR) < .05 to select the most reasonably significant genes we can.

We also select a log2 Fold Change cut off. This is arbitrary as even slight changes may have large impact. Here we use a log2fold change > |1| but even > |.58| can show small change and be used if DE analysis shows too few significant genes.

This can be modified very slightly to select for more genes but will be less reliable and is not recommended.

Replicate dropping or reproducing would change a lot of the results or at least expand our selection of genes of interest.

DESeq2 analysis

Filtered genes for at least 10 counts in ⅓ reps

Post dropping of sample #7 and #8

Log Transformed counts

FDR < .05

Log2FC > |1|

DMSO vs TGFb

548 DEG

50 up regulated 498 downregulated

TGFb vs PFCBP1

51 DEG

42 up regulated 9 downregulated

DMSO vs A485

63 DEG

1 up regulated 62 downregulated

DMSO vs PFCBP1

0 DEG

0 up regulated 0 downregulated

TGFb vs A485

0 DEG

2 up regulated
0 downregulated

A485 vs PFCBP1

0 DEG

0 up regulated 0 downregulated

DESeq2 analysis

Filtered genes for at least 10 counts in ⅓ reps

Post dropping of sample #8 but retaining #7

Log Transformed counts

FDR < .05

Log2FC > |1|

DMSO vs TGFb

548 DEG

50 up regulated 498
downregulated

TGFb vs PFCBP1

51 DEG

42 up regulated 9
downregulated

DMSO vs A485

63 DEG

1 up regulated 62
downregulated

DMSO vs PFCBP1

0 DEG

0 up regulated 0
downregulated

TGFb vs A485

0 DEG

2 up regulated
0 downregulated

A485 vs PFCBP1

0 DEG

0 up regulated 0
downregulated

EdgeR

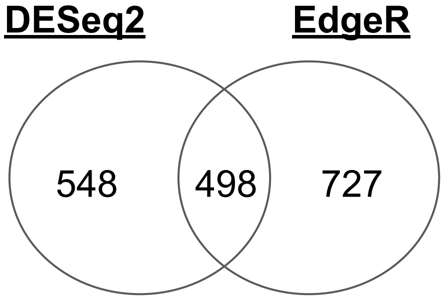
TMM normalization

Unfortunately only significant in one comparison

Log2fc	> |1|
FDR	< .05

DMSO vs TGFb	TGFb vs A485	TGFb vs PFCBP1
727 DEG	0 DEG	0 DEG
79 up regulated and 648 downregulated.	0 up regulated 0 downregulated.	0 up regulated 0 downregulated.

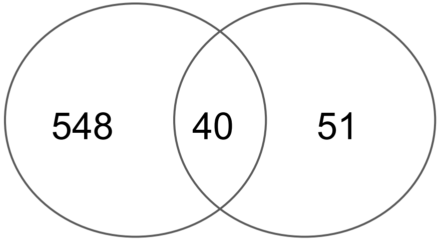
EdgeRs genes that overlap with DESeq2



DESeq2 analysis

Gene selection

DMSO vs TGFb	TGFb vs TGFb +PFCBP1
548	51



Log2FC	> |1|
FDR	< .05

Heatmap generation

Clustering method	=	Complete
Distance measurement	=	Euclidean
Normalization	=	DESeq2 log transformed
Standardization	=	Z -score

Note : These 40 genes also overlap with EdgeRs DE genes

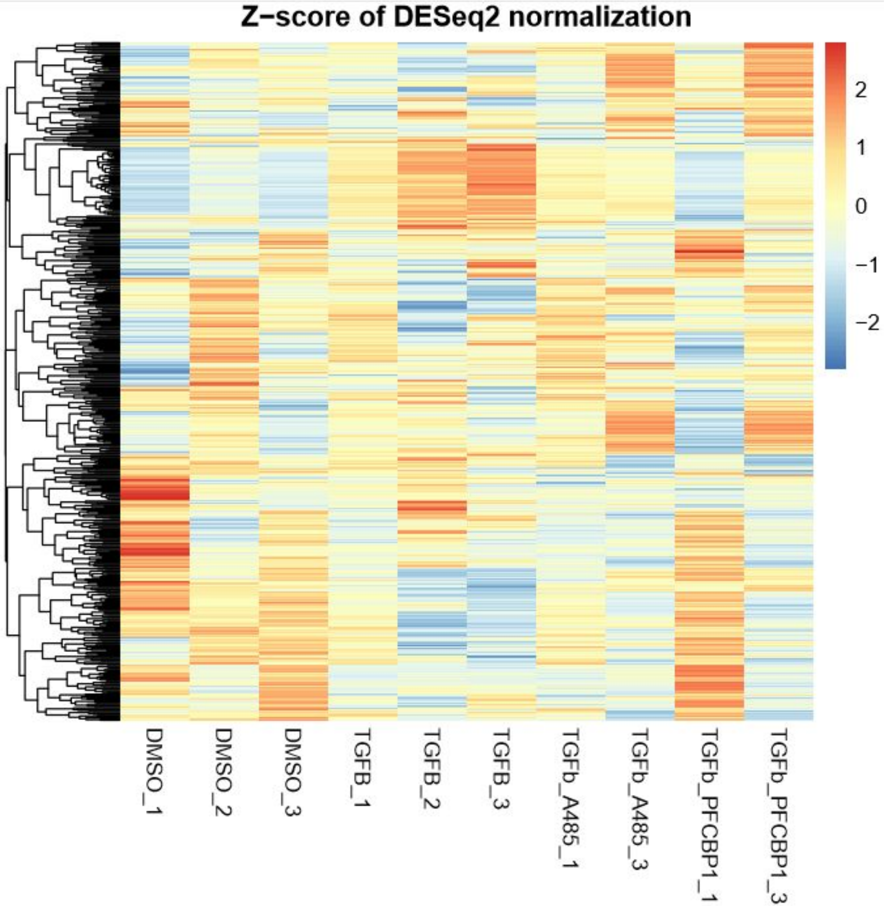
DESEq2 analysis

DMSO --- TGFb --- TGFb+PFCBP1

548 DEG per DESeq2 analysis

These are the genes that are most likely affected by the TGFb treatment compared to the DMSO group.

The inhibitor included groups are mostly not significant. Only 40 out the 548 are considered “significant’ from the PFCBP1 inhibitor.



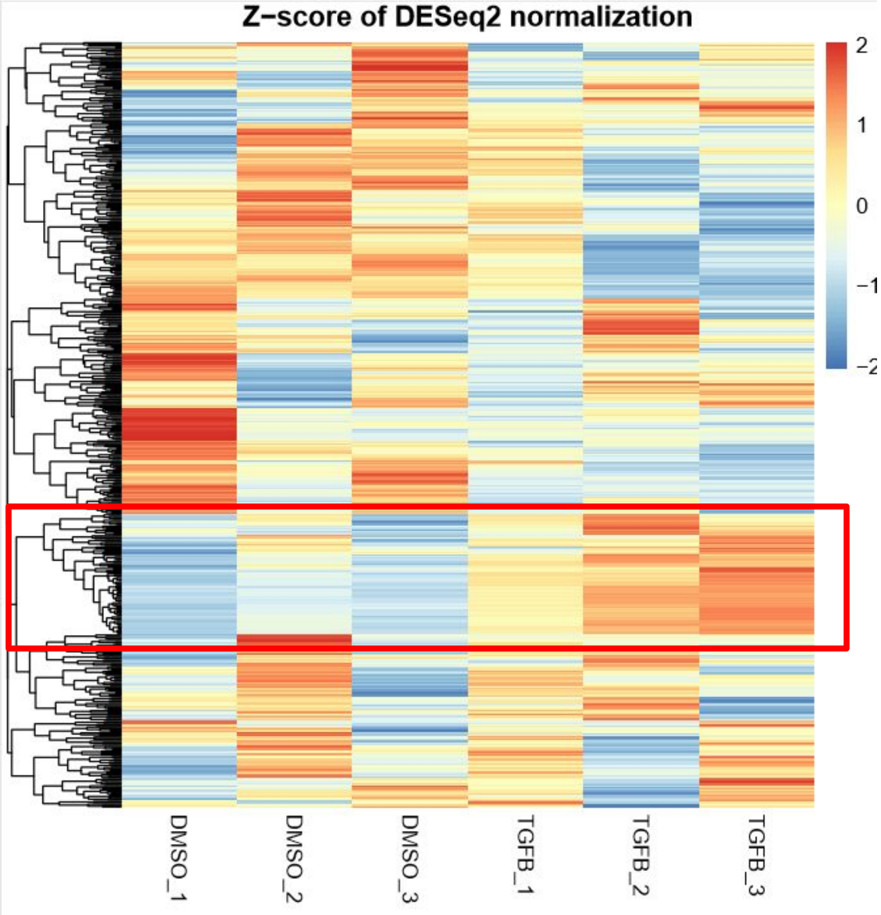
DESEq2 analysis

DMSO --- TGFb

548 DEG per DESeq2 analysis

This is the same heatmap as before but without the inhibitor groups. I did this simply because most of these genes are not regulated significantly in those groups.

We can cut clusters at different branches for gene identification and determine if they overlap with other genes of interest.



DESEq2 analysis

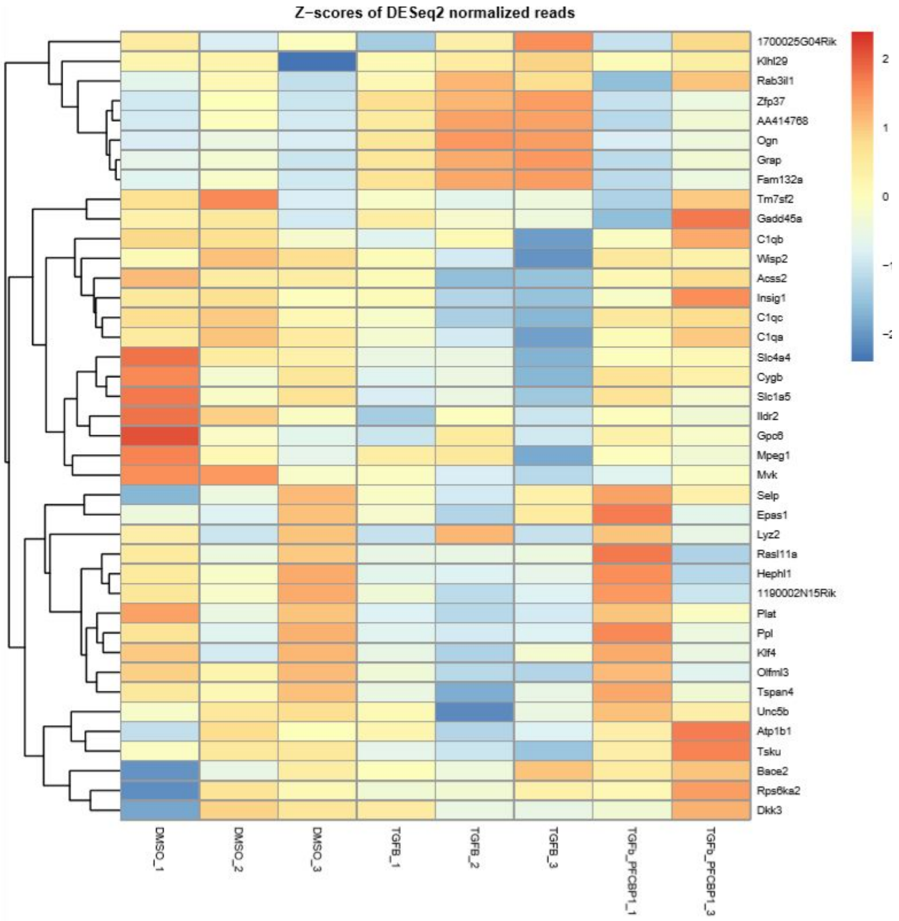
DMSO --- TGFb --- TGFb+PFCBP1

40 DEG per DESeq2 analysis

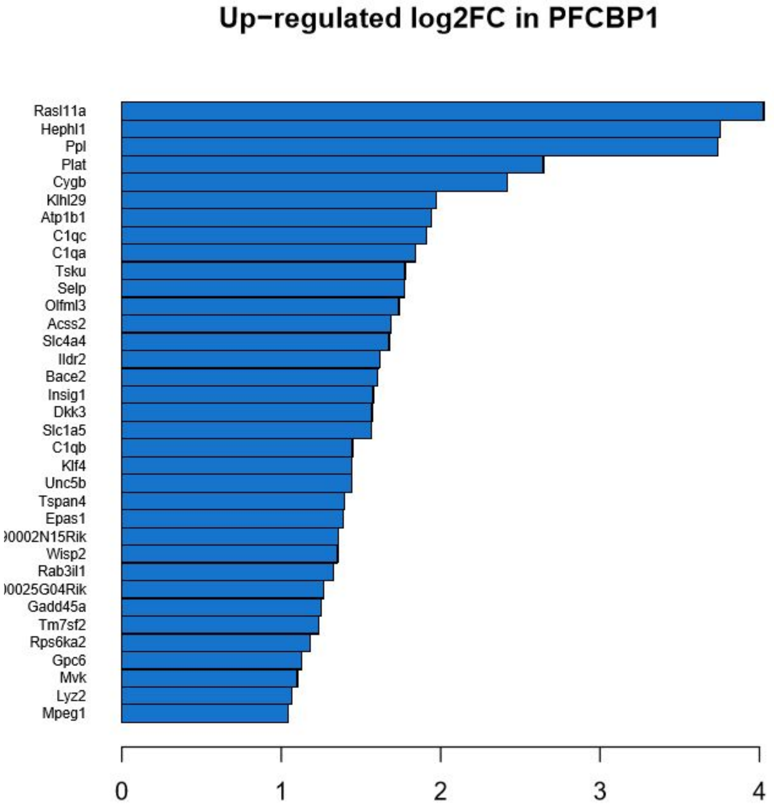
These are the genes that we can most likely assume are affected by the TGFb treatment and at the same time also altered by the inhibitor.

If any desired phenotype is observed in TGFb_PFCBP1 group, these genes most likely to at least play a role.

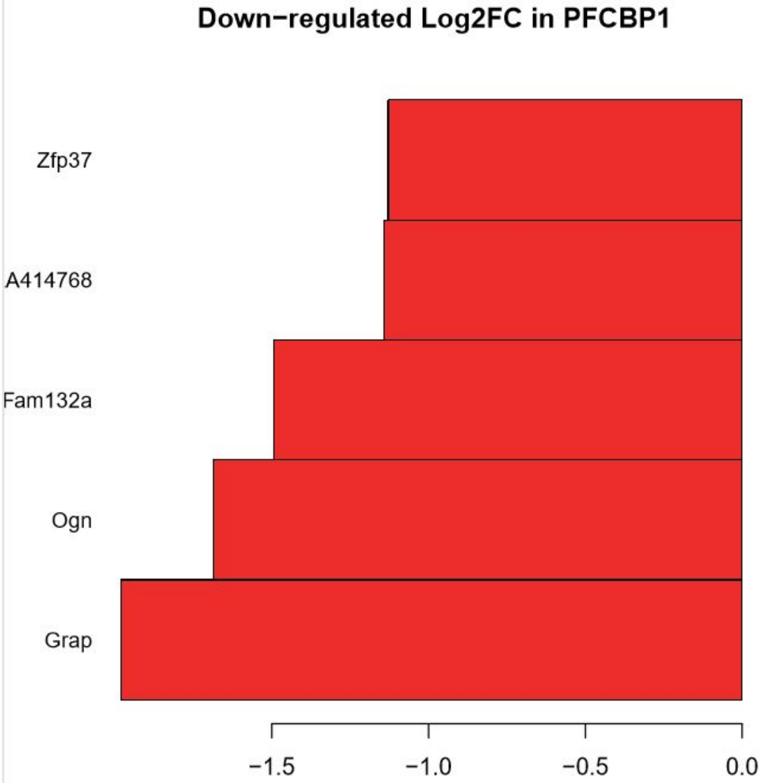
These should not be considered the sole contributors but rather my best estimate.



DE genes (40 DE genes that are most significant)



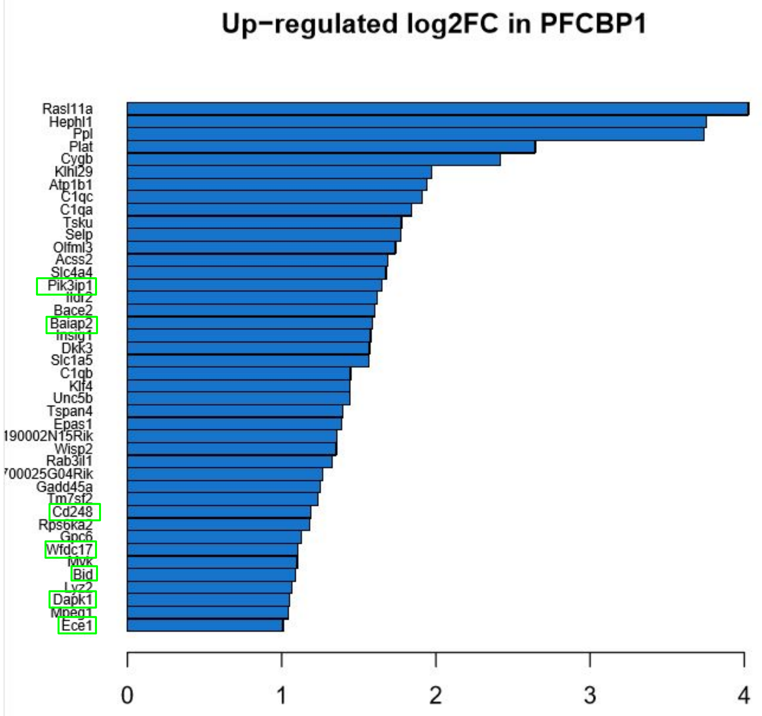
Up−regulated log2FC in PFCBP1

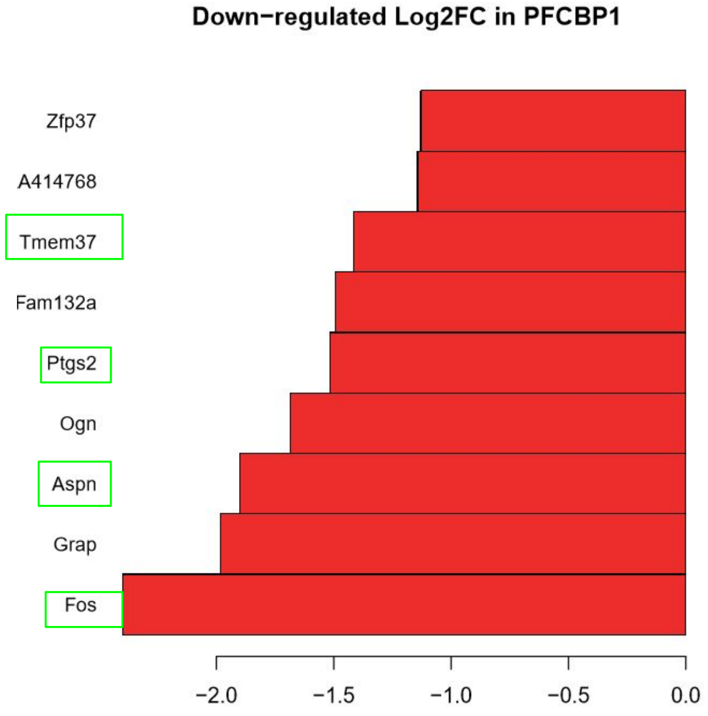


Down−regulated Log2FC in PFCBP1

DE genes (All 51 altered upon PFCBP1 treatment)

Green boxes are the 11 genes changing upon HAT inhibition that were not altered by TGFb treatment.





DESEq2 analysis

Summary

I believe these 40 genes are significant enough for consideration despite weight of replicate inconsistencies. There may be batch effects that will still sway and perhaps even skew the real results at least slightly. DESeq2 normalization method attempts to correct the results by minimizing the effects of outliers and batch effects but the pairwise replicate correlation of TGF-beta replicate 1 with the DMSO group is a bit concerning.

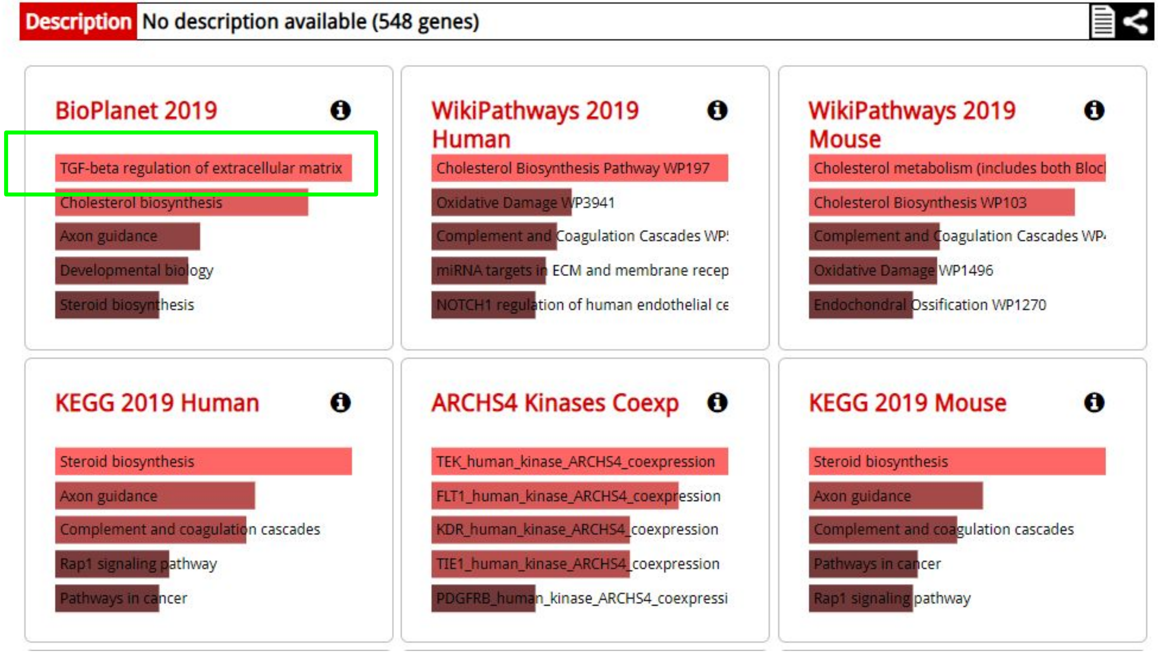
If need be we can remove this replicate and try another DE analysis or we can get a whole new replicate group, but between procedures that gap may be too great to compare or it may prove to be consistent.

Pathway Analysis

A brief look at pathway analysis was performed using Enrichr which links the DE genes to pathways from various databases and ranks them according to enrichment.

The pathways are ordered by P_values.

GSEA will be performed for further enrichment analysis but this provides insight as well.



Description No description available (548 genes)

BioPlanet 2019

TGF-beta regulation of extracellular matrix
Cholesterol biosynthesis
Axon guidance
Developmental biology
Steroid biosynthesis

WikiPathways 2019 Human

Cholesterol Biosynthesis Pathway WP197
Oxidative Damage WP3941
Complement and Coagulation Cascades WP!
miRNA targets in ECM and membrane recep
NOTCH1 regulation of human endothelial ce

WikiPathways 2019 Mouse

Cholesterol metabolism (includes both Bloc
Cholesterol Biosynthesis WP103
Complement and Coagulation Cascades WP
Oxidative Damage WP1496
Endochondral Ossification WP1270

KEGG 2019 Human

Steroid biosynthesis
Axon guidance
Complement and coagulation cascades
Rap1 signaling pathway
Pathways in cancer

ARCHS4 Kinases Coexp

TEK_human_kinase_ARCHS4_coexpression
FLT1_human_kinase_ARCHS4_coexpression
KDR_human_kinase_ARCHS4_coexpression
TIE1_human_kinase_ARCHS4_coexpression
PDGFRB_human_kinase_ARCHS4_coexpressi

KEGG 2019 Mouse

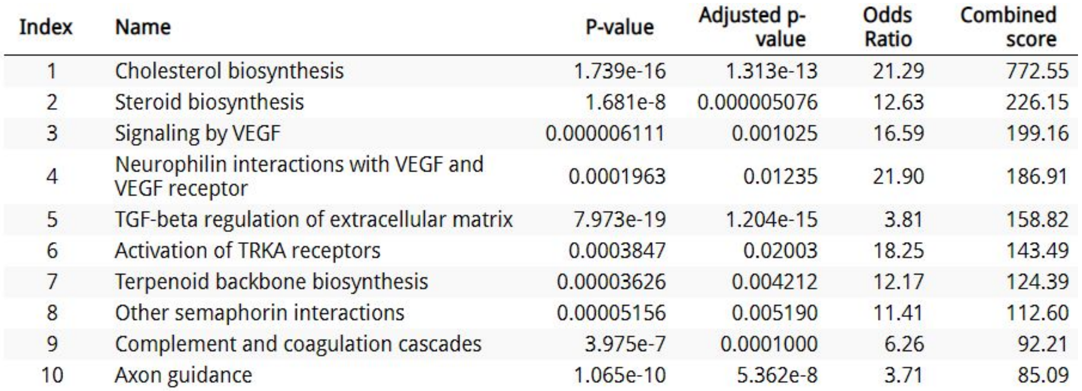
Steroid biosynthesis
Axon guidance
Complement and coagulation cascades
Pathways in cancer
Rap1 signaling pathway

Pathway Analysis

Here we can see the top enriched pathways according the genes found in the DMSO_vs_TGFb group.

This can at least validate ECM activation as a result of TGF-beta regulation.

We can also view all enriched pathways.



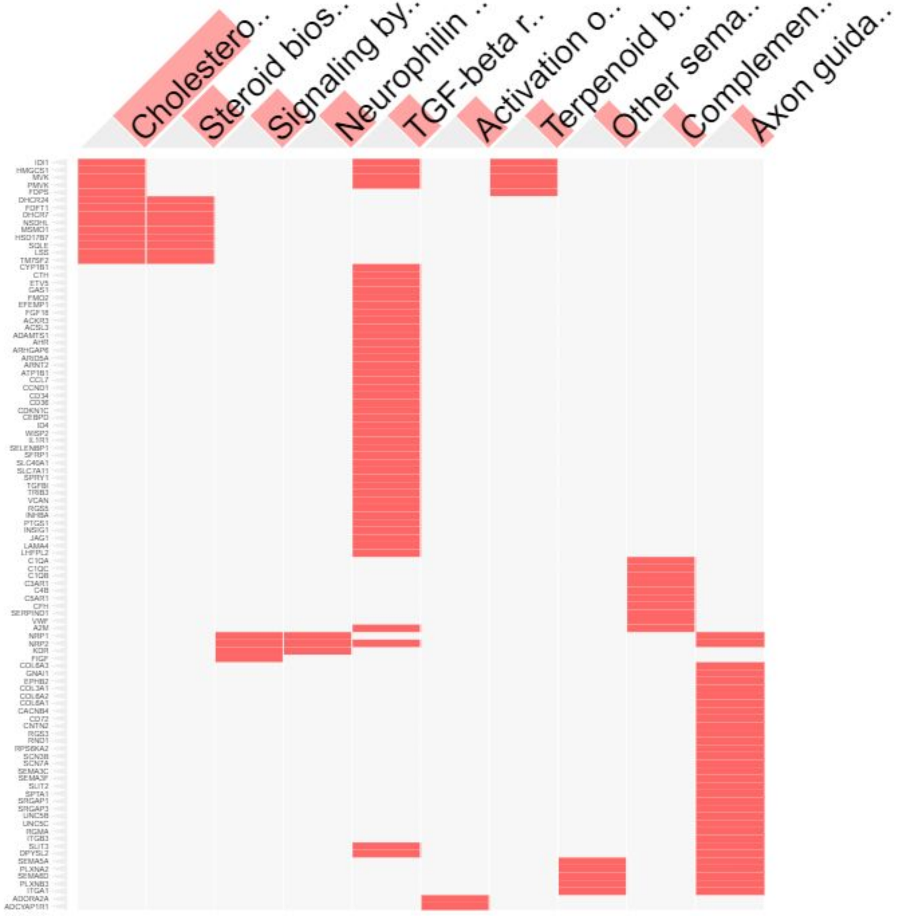
Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Cholesterol biosynthesis	1.739e-16	1.313e-13	21.29	772.55
2	Steroid biosynthesis	1.681e-8	0.000005076	12.63	226.15
3	Signaling by VEGF	0.000006111	0.001025	16.59	199.16
4	Neurophilin interactions with VEGF and VEGF receptor	0.0001963	0.01235	21.90	186.91
5	TGF-beta regulation of extracellular matrix	7.973e-19	1.204e-15	3.81	158.82
6	Activation of TRKA receptors	0.0003847	0.02003	18.25	143.49
7	Terpenoid backbone biosynthesis	0.00003626	0.004212	12.17	124.39
8	Other semaphorin interactions	0.00005156	0.005190	11.41	112.60
9	Complement and coagulation cascades	3.975e-7	0.0001000	6.26	92.21
10	Axon guidance	1.065e-10	5.362e-8	3.71	85.09

Pathway Analysis

Here are the top 100 genes associated with enrichment.

We can see a strong association with the TGF-beta pathway.

We can look at all enriched genes associated with TGFb regulations and view their expression separately if need be.



GSEA

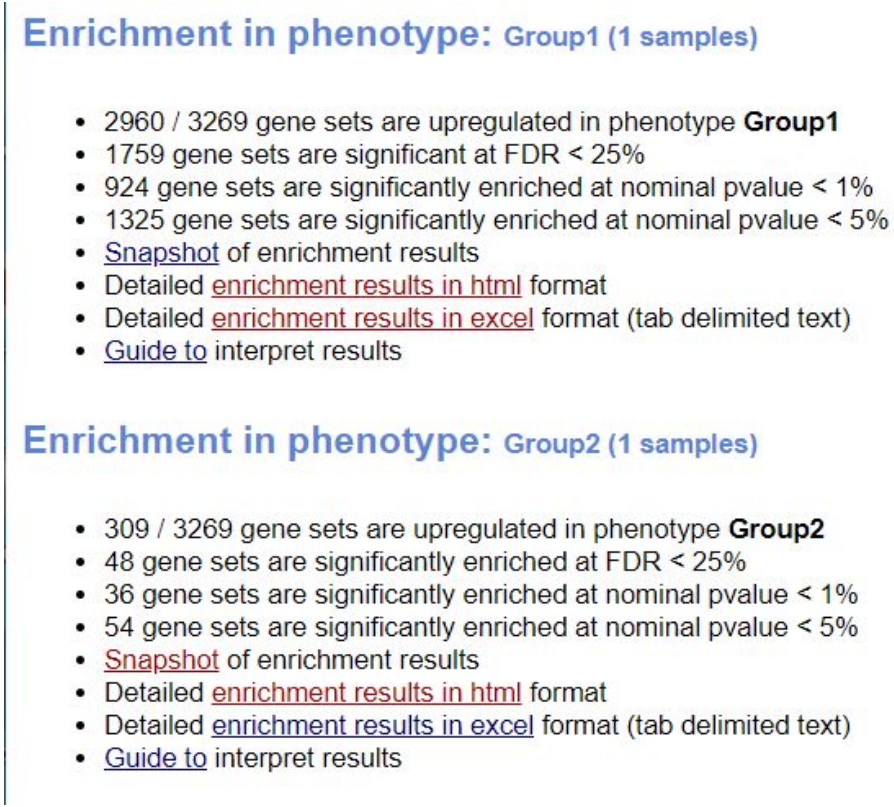
DMSO_vs_TGF-beta

Group 1 = DMSO

Group 2 = TGFb

Oddly we see much more genesets enriched in the Group 1. This perhaps contains many typical pathways we would see enrichment for.

While TGFb may be enriched in more undesirable pathways.



Enrichment in phenotype: Group1 (1 samples)

- 2960 / 3269 gene sets are upregulated in phenotype **Group1**
- 1759 gene sets are significant at FDR < 25%
- 924 gene sets are significantly enriched at nominal pvalue < 1%
- 1325 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: Group2 (1 samples)

- 309 / 3269 gene sets are upregulated in phenotype **Group2**
- 48 gene sets are significantly enriched at FDR < 25%
- 36 gene sets are significantly enriched at nominal pvalue < 1%
- 54 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

GSEA

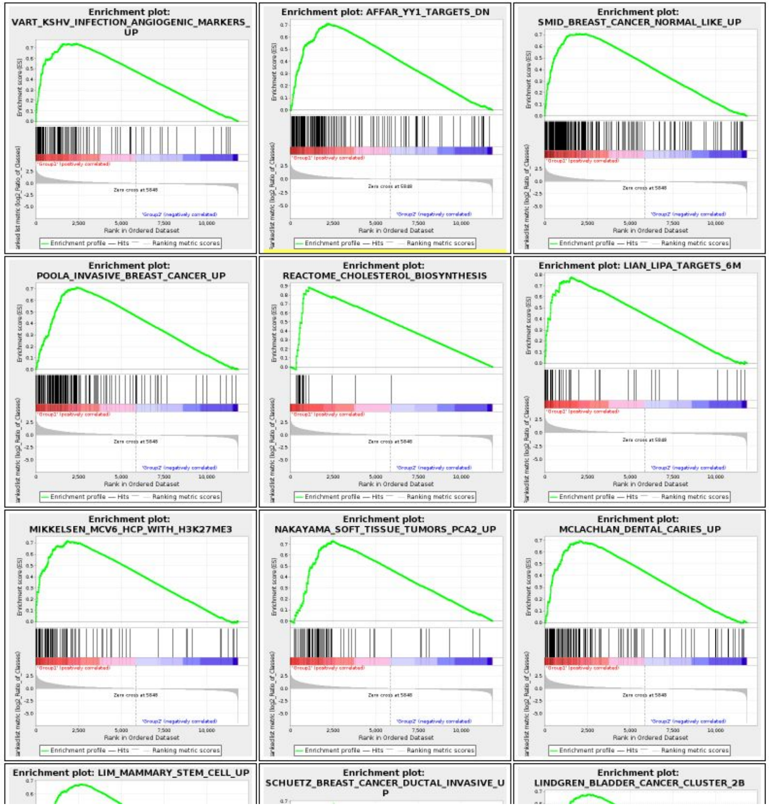
Up in **DMSO** vs TGFb

These are the top enriched pathways in Group1.

Oddly, just on the surface, it appears that the DMSO group is clearly enriched in more inflammatory and invasive pathways.

I don’t understand how this could be. I double checked the dataset throughout processing and nothing appears to be mislabeled.

The Enrichr analysis states enrichment but perhaps it was correlated with the DMSO group and I hastily assumed otherwise.



Summary

I don’t feel as though further analysis can be done. It’s clear that something is just not right. Could be contamination or it could be that DMSO and TGFb were switched somewhere along the analysis.

Shouldn’t the negative control (DMSO) have less enriched pathways? And if so , especially ones related to cell proliferation, tumor development, and inflammation?

I immediately see pathways that were found when studying the DOCA group in the Givinostat project. Seems unusual.