Jennifer Yu, Bi183 HW #3

Problem 1:

a. A DNA library is a collection of DNA fragments cloned into vectors for scientific research and study. A cDNA library is described in the provided link, which is created with reverse-transcribed mRNA.

b. For single-end reads, a fragment is read from one end to another, which generates the base pairs. Pair-end reads starts one read at one end and then also reads from the other end of the fragment. You would want to use pair-end reads when you want to increase the mapping accuracy in resolving structural rearrangements from gene insertion/deletion/mutations. However, pair end reads are more expensive than single end reads, and shouldn't be used if the experiment doesn't necessitate a high degree of accuracy.

c. The workflow in an experimental RNA-seq protocol first involves experimental design. The researcher wants to determine if they wish to extract qualitative or quantitative data, what type of annotation should be used, how many samples to sequence, the depth of sampling, experimental complexity, and other considerations. Then, the second step is preparing RNA. This involves isolating and purifying cellular RNAs; then the mRNA is extracted and the RNA is fragmented to make it suitable for RNA-seq. Afterwards, the third step is to prepare a library. Here, the RNA is converted to cDNA, adaptors are added for sequencing and the DNA is amplified. It is also important to validate the library and quantify is accurately using methods such as PCR. Finally, the RNA is sequenced using a sequencing platform. Specifications on the sequencing, such as the length of reads, single-end or paired-end, etc should be provided. The last step is to analyze the sequences data. This involves filtering the sequencing reads, assembling them into transcripts, aligning them to other sequences, and looking at transcript abundance.

Problem 2:

a. FASTA

This format represents nucleotide sequences and/or amino acid sequences. It was developed by David J. Lipman and William R. Pearson in the FASTA suite of programs. The header of the FASTA file begins with ">", which gives a name/unique identifier for a sequence, along with other information. The sequence is represented after the header. These sequences can contain gaps or alignment characters, and the alphabet consists of amino acid/nucleic acid codes, U is allowed for amino acid sequences, and a hyphen/dash represents a gap.

```
ENST00000632963.1 cdna
chromosome:GRCh38:CHR_HSCHR14_3_CTG1:105884674:105884693:-1
gene:ENSG00000282674.1 gene_biotype:IG_D_gene
transcript_biotype:IG_D_gene gene_symbol:IGHD5-24
description:immunoglobulin heavy diversity 5-24 (non-functional)
[Source:HGNC Symbol;Acc:HGNC:5510] GTAGAGATGGCTACAATTAC
```

b. FASTQ

This format stores a biological sequence and corresponding quality scores. The sequence letter and the score are encoded with an ASCII character for brevity. It was originally developed at the Wellcome Trust Sanger Institute in order to bundle a FASTA sequence with its quality data. A

FASTQ file contains four lines per sequence. Each line begins with @ and is followed by a sequence identifier and other information, similar to a FASTA header line. The second line contains the sequence and the third line begins with "+", possibly following by a sequence identifier (again). Finally, the last line has the encoded quality values for the sequence, and contains the same amount of letters.

```
@cluster_2:UMI_ATTCCG
TTTCCGGGGCACATAATCTTCAGCCGGGCGC
+
9C;=;=<9@4868>9:67AA<9>65<=>591
```

c. GFF or GTF

This format is used for describing genes and other features of DNA, RNA and amino acid sequences. Some technologies use this format, such as GBrowse and IGB (Integrated Genome Browser). The general structure of GFF is that its tab delimited with 9 fields per line. The name of the sequence is in the first position, followed by the source, a feature type name, the start/end of the feature, its score, the strand of the feature, the phase, and other attributes (9th field).

```
transcribed_unprocessed_pseudogene  gene        11869 14409 . + .
gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source
"havana"; gene_biotype "transcribed_unprocessed_pseudogene";
1 processed_transcript               transcript  11869 14409 . + .
gene_id "ENSG00000223972"; transcript_id "ENST00000456328";
gene_name "DDX11L1"; gene_sourc e "havana"; gene_biotype
"transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-002";
transcript_source "havana";
```

d. SAM and BAM

SAM stands for Sequence Alignment Map, a format used to store biological sequences aligned to a reference sequence. It is used for storing data generated by next generation sequencing technologies and has been broadened to include storage of unmapped sequences. It consists of a header and an alignment section. BAM stands for Binary Alignment Map and stores the same file as SAM does, except in a compressed binary representation. Headers in a SAM file begin with @ and distinguishes this from the alignment. The first field has the QName (query template name), followed by a bitwise flag, sequence name, leftmost mapping position, map quality, CIGAR (Concise Idiosyncratic Gapped Alignment Report) string, the reference name of the next read, the position of the next read, length of the template, the sequence (* if not stored) and an ASCII representation of quality.

```
@HD   VN:1.0      SO:coordinate
@SQ   SN:1  LN:249250621      AS:NCBI37
      UR:file:/data/local/ref/GATK/human_g1k_v37.fasta
      M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ   SN:2  LN:243199373      AS:NCBI37
      UR:file:/data/local/ref/GATK/human_g1k_v37.fasta
      M5:a0d9851da00400dec1098a9255ac712e
@SQ   SN:3  LN:198022430      AS:NCBI37
      UR:file:/data/local/ref/GATK/human_g1k_v37.fasta
      M5:fdfd811849cc2fadebc929bb925902e5
@RG   ID:UM0098:1 PL:ILLUMINA PU:HWUSI-EAS1707-615LHAAXX-L001     LB:80
      DT:2010-05-05T20:00:00-0400   SM:SD37743  CN:UMCORE
```

```
@RG   ID:UM0098:2 PL:ILLUMINA PU:HWUSI-EAS1707-615LHAAXX-L002    LB:80
      DT:2010-05-05T20:00:00-0400   SM:SD37743  CN:UMCORE
@PG   ID:bwa      VN:0.5.4
@PG   ID:GATK TableRecalibration    VN:1.0.3471
      CL:Covariates=[ReadGroupCovariate, QualityScoreCovariate,
CycleCovariate, DinucCovariate, TileCovariate],
default_read_group=null, default_platform=null, force_read_group=null,
force_platform=null, solid_recal_mode=SET_Q_ZERO, window_size_nqs=5,
homopolymer_nback=7, exception_if_no_tile=false,
ignore_nocall_colorspace=false, pQ=5, maxQ=40, smoothing=1
1:497:R:-272+13M17D24M  113   1     497   37    37M   15    100338662
      0     CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG     0;==-
==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>    XT:A:U      NM:i:0      SM:i:37
      AM:i:0      X0:i:1      X1:i:0      XM:i:0      XO:i:0
      XG:i:0      MD:Z:37
19:20389:F:275+18M2D19M 99    1     17644 0     37M   =     17919 314
      TATGACTGCTAATAATACCTACACATGTTAGAACCAT
      >>>>>>>>>>>>>>>>>>><<>>><<>>4::>>:<9     RG:Z:UM0098:1
      XT:A:R      NM:i:0      SM:i:0      AM:i:0      X0:i:4
      X1:i:0      XM:i:0      XO:i:0      XG:i:0      MD:Z:37
19:20389:F:275+18M2D19M 147   1     17919 0     18M2D19M    =     17644
      -314  GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT
      ;44999;499<8<8<<<8<<><<<<><7<;<<<><<     XT:A:R      NM:i:2
      SM:i:0      AM:i:0      X0:i:4      X1:i:0      XM:i:0
      XO:i:1      XG:i:2      MD:Z:18^CA19
9:21597+10M2I25M:R:-209 83    1     21678 0     8M2I27M     =     21469
      -244  CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT
      <;9<<5><<<<><<<>><<><<>><9>><>>>9>>><>     XT:A:R      NM:i:2
      SM:i:0      AM:i:0      X0:i:5      X1:i:0      XM:i:0
      XO:i:1      XG:i:2      MD:Z:35s
```

Problem 3:

a. PCA (Principal Component Analysis) is a statistical method for reducing the dimensionality of the data while retaining the variation in the data set. More specifically, given a set of n points, one finds, generally, the average or centroid of the points by minimizing the squared distance of the points to their orthogonal projection onto the space. PCA is similar to the linear regression method of fitting a line, except there is noise in both the x and y axes. The principal components generated through this analysis also retain the sample variance. Due to the Pythagorean theorem, the orthogonally projected points maximize the retained sample variance by maximizing the total squared distances of the projected points to the origin.
b. Multi-dimensional scaling is similar to PCA, except it converts the distances between the samples into lower dimensions, instead of correlations (as in PCA). PCA can be reduced to MDS if we interpret it as minimizing the total square distance between the original distance and the projected distance (thinking in terms of Pythagorean theorem, PCA involves maximizing one leg of the triangle to maximize the distance of the points from the origin; MDS would be minimizing the perpendicular squared component, e.g. squared distance between original and projected point). There are other methods of measuring distances between points which also fall under

MDS. In the case of minimizing the linear distance, this becomes essentially equivalent to maximizing the linear correlations.

c. See notebook.

Problem 4:

1. PCA - Linear dimensionality reduction technique to visualize similarity between samples
This technique is used to identify global patterns in a high dimensional dataset. A scatterplot is used to represent the Principal Components, which is a set of uncorrelated features that represent the most relevant sources of variance in the data.

2. Clustergrammer - Interactive hierarchical clustering heatmap visualization
This tool helps to visualize and analyze high dimensional data as heatmaps. Each row of the heatmap represents a gene and every column is a sample. This helps to identify which genes contribute to clustering.

3. Library Size Analysis - Analysis of readcount distribution for the samples within the dataset
This quantifies gene expression by mapping reads generated from sequencing to a reference genome, and then obtaining the gene counts. The bar chart displays the total number of reads mapped to each RNA-seq sample in the dataset.

4. Differential Expression Table - Differential expression analysis between two groups of samples
Differential gene expression methods characterize gene expression signatures between two groups of samples. This identifies which genes' expressions are altered significantly due to perturbation. In the table, each gene has different estimated measures of differential expression.

5. Volcano Plot - Plot the logFC and logP values resulting from a differential expression analysis
This scatter plot is used to display the results of a differential gene expression analysis. Each point represents a gene and the axes show the significant vs. fold-change estimated by the analysis. Red points indicate upregulated genes and blue points indicate down-regulated ones.

6. MA Plot - Plot the logFC and average expression values resulting from a differential expression analysis
This plot shows the average expression of each gene calculated. The MA Plot is similar to the Volcano Plot, in that it measures the result of a differential gene expression analysis and assesses the global similarity of gene expression in two groups of biological samples.

7. Enrichr Links - Links to enrichment analysis results of the differentially expressed genes via Enrichr
The Enrichr analysis shows biological terms that are overrepresented in a gene set. These are based off prior knowledge of the gene and include signaling pathways, diseases, molecular functions, etc. The links contain results of the analyses generated by analyzing up-regulated and down-regulated genes.

8. Gene Ontology Enrichment Analysis - Identifies Gene Ontology terms which are enriched in the differentially expressed genes
Gene Ontology unifies the representation of gene attributes. The information is used by Enrichr to identify the overrepresented biological processes in genes identified by comparing two groups of samples. The bar charts show the results of Gene Ontology analysis and which genes are up-regulated or down-regulated in perturbation.

9. Pathway Enrichment Analysis - Identifies biological pathways which are enriched in the differentially expressed genes
The interactions between biochemical compounds are important in determining cellular behavior, and this analysis uses information from databases such as WikiPathways, that contain associations between biological pathways and genes. Similar to the GO analysis, there are bar graphs showing the result of this analysis and biological pathways which are overrepresented in the up-regulated and down-regulated genes are identified by comparing two groups of samples.

10. Transcription Factor Enrichment Analysis - Identifies transcription factors whose targets are enriched in the differentially expressed genes
The transcription factors are proteins involved in transcriptional regulation of gene expression. Databases like ENCODE have associations between these factors and their targets, which are then used by Enrichr to identify which targets are overrepresented in the up-regulated and down-regulated genes.

11. Kinase Enrichment Analysis - Identifies protein kinases whose substrates are enriched in the differentially expressed genes
Protein kinases modify other proteins by adding phosphate groups, and databases like KEA have associations between kinases and their subtrates. Enrichr uses this information to find the protein kinases which are overrepresented in the genes.

12. miRNA Enrichment Analysis - Identifies miRNAs whose targets are enriched in the differentially expressed genes
microRNAs are non-coding RNAs which are involved in post-transcriptional regulation of gene expression. Databases contain associations between miRNA and their target, which is used by Enrichr to find which miRNA has overrepresented targets in the genes.

13. L1000CDS2 Query - Identifies small molecules which mimic or reverse a given differential gene expression signature
L1000CD52 is a tool that queries gene expression signatures against signatures from human cell lines treated with 20,000 small molecules and drugs. It is used to identify small molecules that mimic or reverse the effects of a gene expression signature. The bar chart displayed show the molecules which mimic the gene expression and those which reverse it.

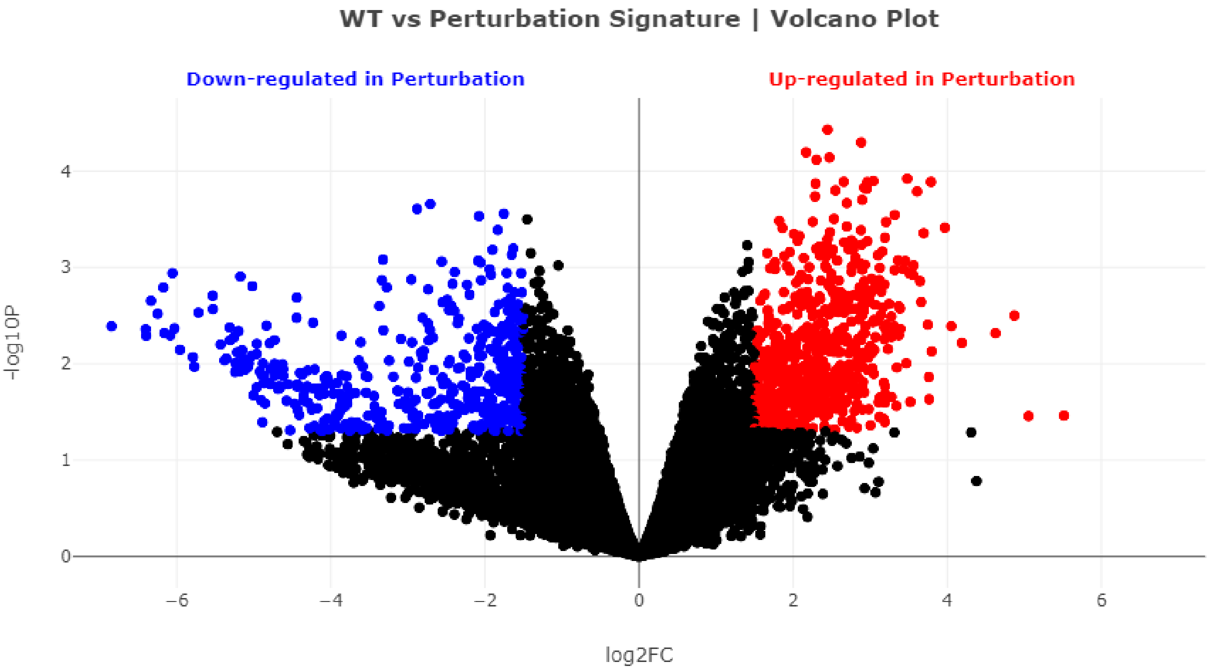
14. L1000FWD Query - Projects signatures on a 2-dimensional visualization of the L1000 signature database
This query also queries gene expression signatures against those from human cell lines treated with over 20,000 small molecules and drugs. You can search for molecules with similar signatures or opposite signatures, and each has a similar score and other information associated with it, obtained from the database.

Problem 5

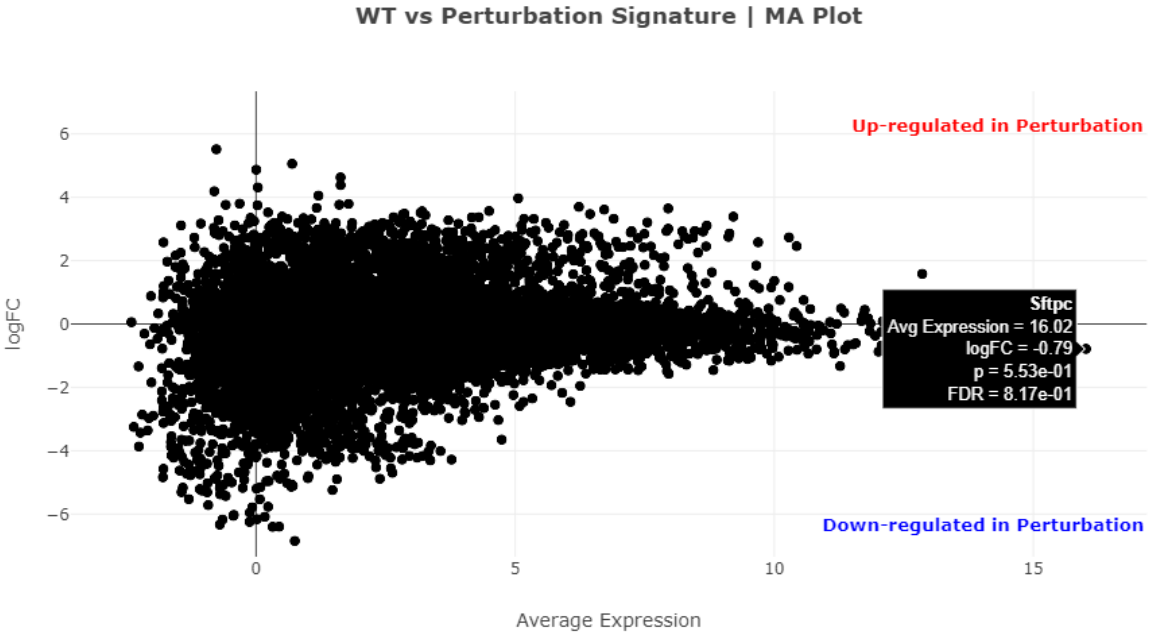
a. This experiment focused on the role/effect of SIRT1 on KRAS driven lung carcinogenesis. Sirtuins are a family of protein deacylases and mono-ADP-ribosyltransferases that re active in low energy cellular scenarios. Many sirtuins protect against tumor development, and SIRT1 has been reported to play tumor suppressive roles. This study focuses on the functional interaction between SIRT1 and K-RAS. It was found that in mice, SIRT1 acts as a tumor suppressor and these results were verified in human lung tumors. Within this experiment, since K-Ras is the most frequently mutated member of the Ras family, mice with an extra copy of the Sirt1 gene were crossed with those expressing an oncogenic K-Ras mutant. After extracting the MEFs from these mice, it was found that MEFs overpressing the KRas-KI oncogene (Cre-inducible) had a more intense reduction in Sirt1 protein levels than the control group of K-Ras wildtype expressing MEFs.

In investigating the mechanisms driving Sirt1 mediated protection, the pneumocytes where K-Ras-KI is activated in Sirt1-WT and Sirt1-Tg were characterized. Cre was activated by a pulse of tamoxifen during two weeks, which activated the expression of the Katushka reporter gene. A chase of 2 weeks reduced the amount of active pneumocytes. The assumption is that tumors arise from a few long-term surviving K-Ras-KI-activated pneumocytes (tumor-initiating cells). During the pulse phase, genes involved in antioxidant defences such as members of the glutathione S-transferase family and Sod3 were upregulated, suggesting that overexpressing Sirt1 is associated with stronger oxidative defence. There was also increased Pygm and decreased expression of Gftp1 and Hk2. Activators of apoptosis and inhibitors of proliferation were also overexpressed, such as Hoxa5, Fas, Timp2, etc. In the pulse + chase experiment, many of the genes upregulated during the pulse experiment (Gstm5, Sod3, Pygm, Timp2, Php3, etc) were downregulated. There was also increased expression of both oncogenes and tumor suppressors, though there is a net balance for tumor suppressive effects.

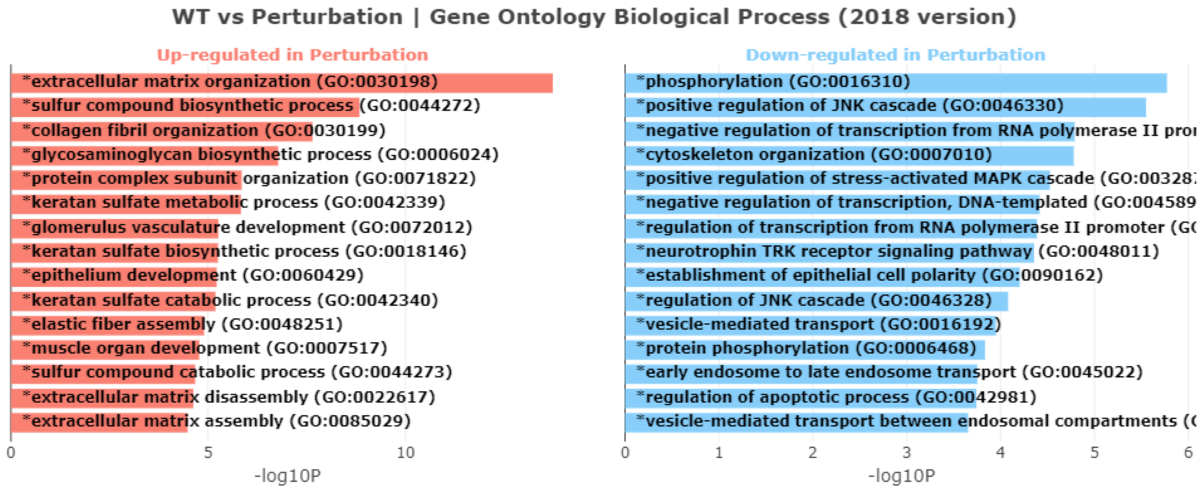
b. In the BioJupies for the pulse dataset, we are given a volcano plot denoting a gene expression analysis:



Here, there is a slight difference in the statements made in the paper and what is shown on this scatterplot. One of the most significantly upregulated genes was Adrenomedullin (ADM), which was not mentioned in the paper. This hormone normally functions in circulation control and stimulates the growth of new blood vessels and increases the tolerance of cells to oxidative stress. Though these are positive attributes when dealing with diseases such as hypertension, they are negative in potentiating the ability of cancerous cells to extend their blood supply and proliferate.

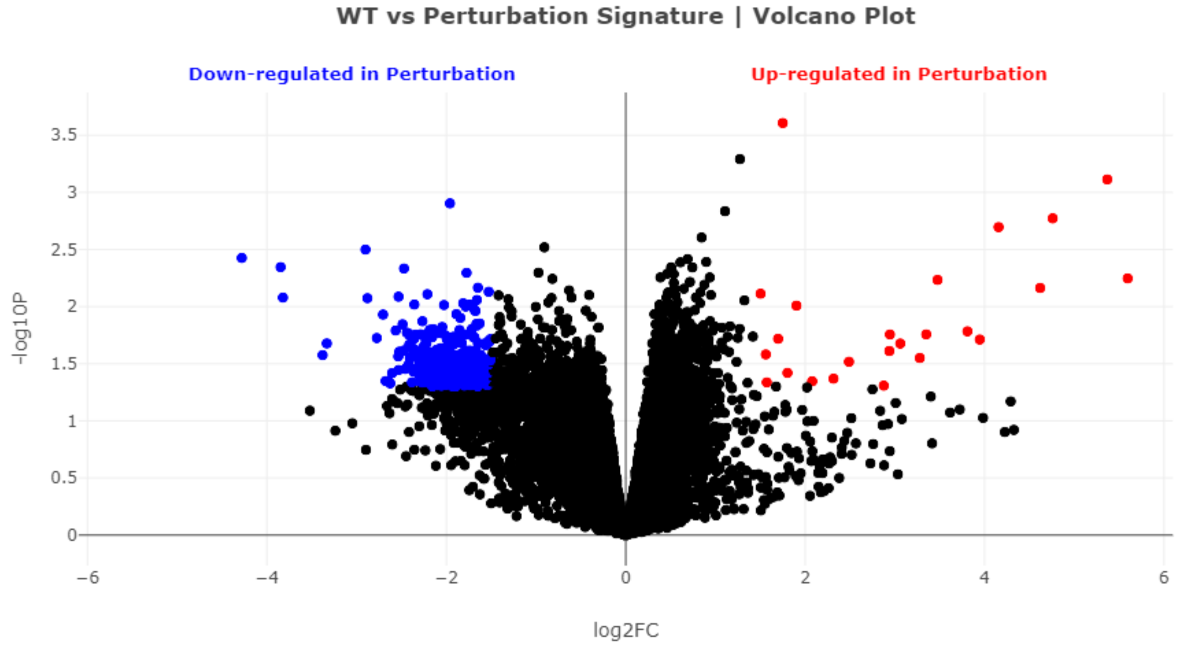


This scatterplot aligns with analysis done on last weeks problem set, where we found that Sftpc was the most abundant transcript in this dataset.

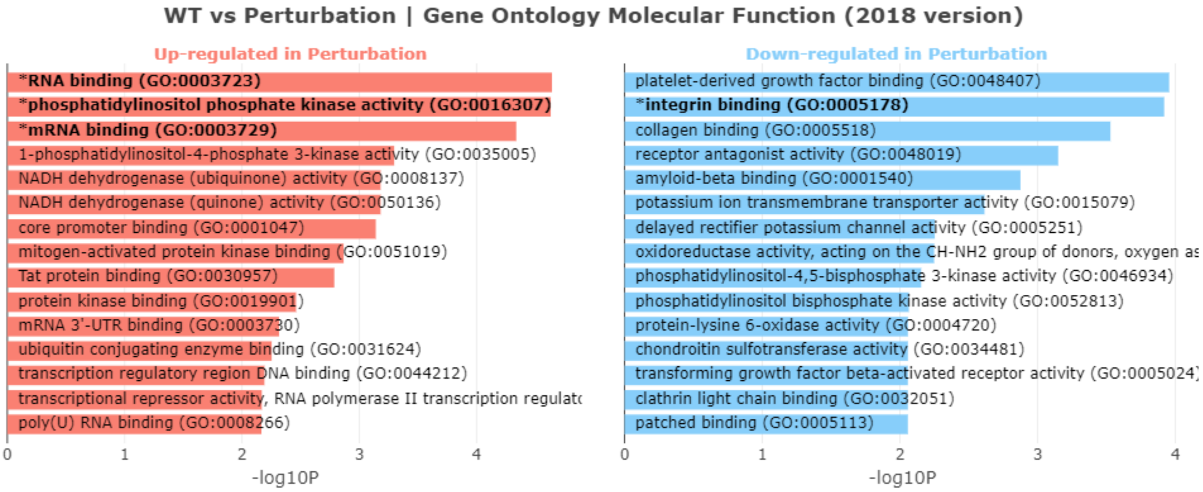


We see matching data from the paper in these graphs detailing biological pathways. In the second most upregulated association, the sulfur compound biosynthetic process is responsible for forming compounds such as glutathione, which was described as being upregulated during the pulse experiment. Additionally, there is downregulation in phosphorylation, which is an important cellular regulatory mechanism. Overexpression of kinases can lead to tumor development.

c. In the pulse and chase dataset, we see in the volcano plot amongst the red points that Sirt1 is very highly expressed (up-regulated) during perturbation, which was one observation made in the paper about the pulse + chase period. Additionally, many of the genes that were upregulated in the pulse period are downregulated in this data, as shown in the volcano plot, such as Timp2.



In the MA plot, Sftp-c is shown to be the most abundant transcript again.



As shown here, there is upregulation in kinase activity, which can be responsible for triggering tumor growth. At the same time, there is downregulation in multiple types of binding, significantly in platelet-derived growth factor binding and integrin binding, which can restrict cell growth and division.

d. I expected the data to be pretty nonsensical, which is essentially what I got. One of the most upregulated genes in the volcano plot ended up as Emilin2, which is involved in the extracellular matrix (a biological pathway that was upregulated in previous datasets), but hasn't occurred in previous analysis as a highly upregulated gene. The most abundant transcript is still Sftp-c, which makes sense, since that isn't dependent on which groups are control/mutated. In the GO Enrichment Analysis, many of the biological processes and molecular functions have changed.

Currently, establishment of skin barrier is the most downregulated biological process, though extracellular matrix organization is still the most upregulated biological process.

e. Again, most of the data analysis is not matched accordingly to the results in the paper, mainly because the control and mutated groups were scrambled. The volcano plot has very few red/blue points, and most downregulated gene is Spink5, which provides instructions fro making a serine peptidase inhibitor. Again, Sftp-c is found to be the most abundant transcript, because the arrangement of the groups doesn't affect this fact. In the Enrichr GO analysis, some of the binding activities that were upregulated before are now placed in the down-regulated column. Many of the biological processes also vary in their detected level of regulation. For example, cilium assembly is the most upregulated biological process, but it doesn't make a lot of sense in the context of this study.