Final Assessment, Task 2

Theme: CRISPR knock down screens – what is in there by GO analysis?

Introduction: The Data-Set:

The two data files provided (set1.tsv and set2.tsv) are pooled knock-down experiments using the AVANA knock-down library, collected from the Pickles 3.0 database (https://pickles.hart-lab.org/). The data has been scored using two different systems, Avana Z-Score and Avana Bagel Bayes Factor scoring. Each gene has been scored in both scoring schemas to determine which of them are essential in the respective cancer cell-line. (See 'About' Tab in the Pickles Web-page.)

Task 1) Exploring the input data (6 points):

A) Summarise the content of the file set1.tsv. How many genes are there? Is the data sorted, are there missing data-points (i.e. missing values for one of the enrichment scores)? Write 3-5 sentences *(1 point)*

The set1.tsv has 18578 CRISPR-Cas9 Knockdown genes from the cell line ACH-000036, which is a glioblastoma cancer cell line. Avana Z scores reflect the impact of gene knockout on cell growth with a range of -12.42 and 8.69 and a mean of -0.7308. The Avana BF score indicates the strength of evidence for an effect with a range of -80.43 and 67.75 and a mean of -13.15. From the negative mean of the BF and Z scores it is plausible that on average, the gene knockouts could have less evidence to support an effect as well as reduced cell growth; however more screening or data analysis is required to support this notion. There is approximately 2.51% of data missing from the AVA Z score and 2.59% missing from the AVA BF score.

B) How could you check with GO enrichment analysis if the Avana library is unbiased? Write 2-3 sentences, no actual check needed! *(1 point)*

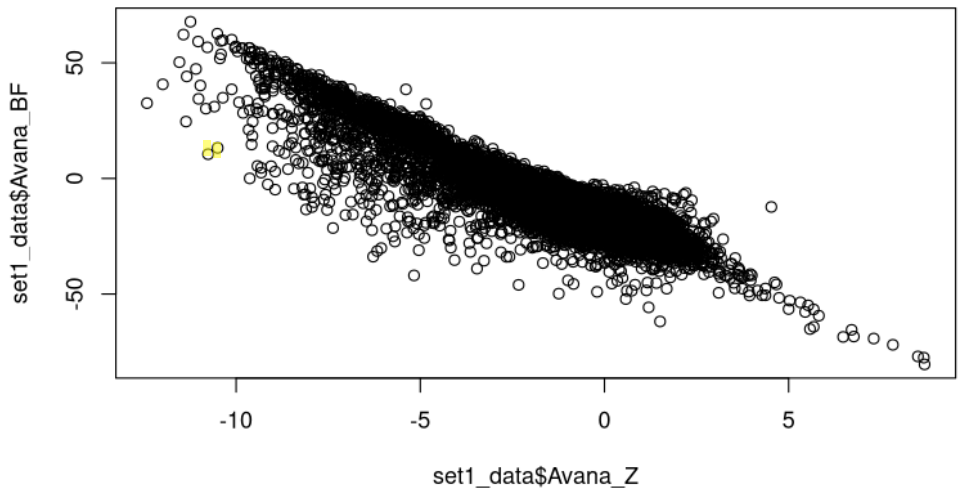
One check you can perform to see if the Avana library is unbiased is by analysing the max and minimum Z scores and comparing them to the GO terms enriched in the total gene sets. If the gene sets are unbiased the distribution between the enriched GO terms and min and max z score sets should be similar.

C) Plot the two scoring schemes against each other for set1.tsv (x-axis Avana_Z, y-axis Avana_BF) and answer:

1. How are the scores generally correlated? Does it match with the information on the scores in the 'About' tab of the web-site? *(0.5 point)*

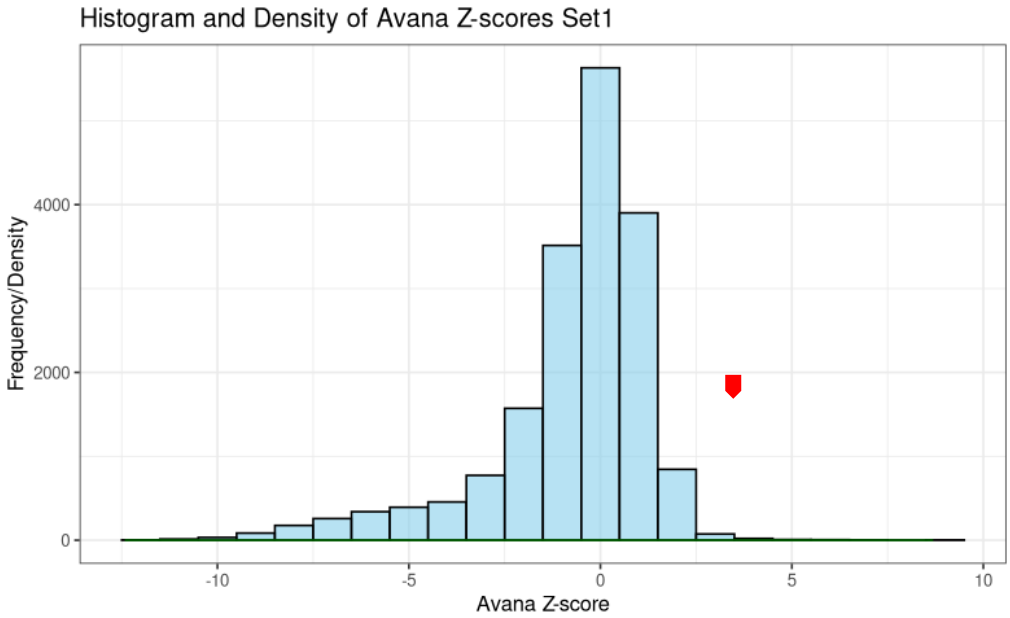
As the Z score increases, the BF score decreases. The website states that a negative Z score is essential for survival and vice versa with a positive BF score. The data does agree, a more negative z score causes the spread of datapoints while the BF score is increasing.

2. Are there any data-points which are in total disagreement between the two scoring schemes? Highlight those in the plot. *(0.5 point)*



The top left highlighted points disagree with the scoring scheme. The Z score indicates strong essentiality however the BF score does not support this. There could be weak evidence for essentiality, complex biological inference than being essential or non-essential, or experimental noise causing a false positive or negative Z score.

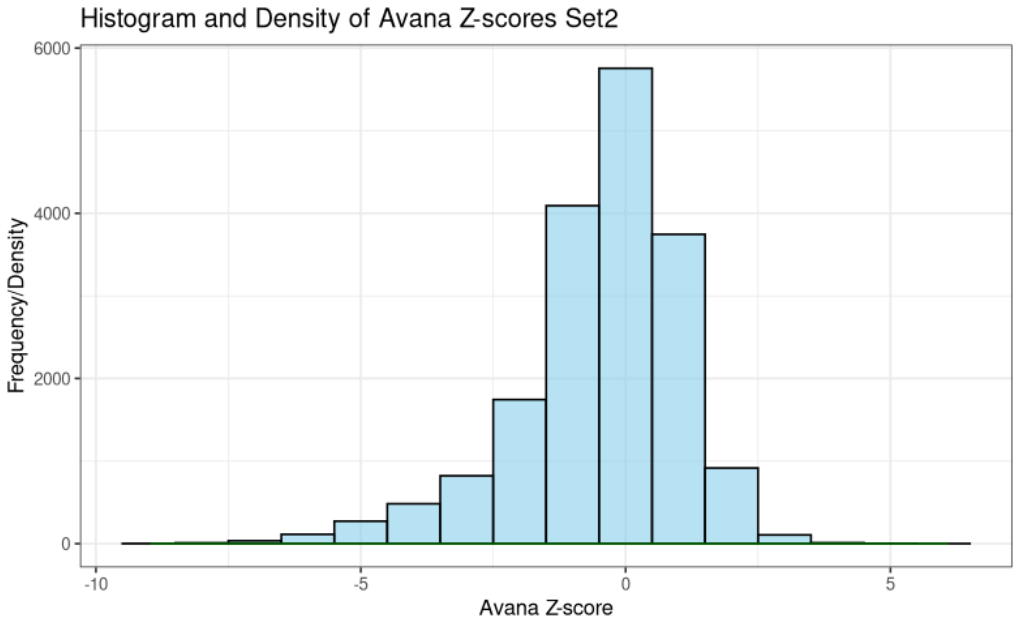
3. Plot a distribution for Set1 of scores for the *Avana_Z* (Density or Histogram) *(1 point)*



4. Discuss in 1-3 sentences where you would put a cut-off to delineate essential genes for *Avana_Z* given the density plot under the assumption that most genes are not essential (and the information from the web-page) *(1 point)*

   Since most genes are not essential in the dataset and a negative Z score indicates essentiality. From the density plot, the bulk of the genes are near 0 from -1.25 to 1.25. We want negative Z data points, so the cutoff to delineate essential genes would be -1.25 on the left tail of the distribution (<=-1.25).

5. Plot the same plot for set2.tsv – do you see the same picture and would you set the cut-off the same? Answer in 1-3 sentences *(1 point)*



The density/ histogram plot for Set2 is roughly the same distribution as Set1. Most of the genes have a nonessential distribution around -1.25 to 1.25, therefore the cutoff for the Z score will be approximately the same at -1.25 (<=-1.25).

Task 2) Preparing the data for STRING based enrichment analysis and initial analysis (6 points)

For the purpose of this assessment, we decide on a cut-off of -5 (<=-5) on *Avana_Z to delineate essential genes from Set1. You do not need to upload the resulting file to Canvas.*

A) How many genes make the cut-off and are reported as essential for each set? Save the essential genes in a file for further investigation *(1 points)*

For set 1 there are 1095 and in set 2 there are 273 genes that make the cut-off and are reported as essential for each set.

B) In addition, for Set1, generate files containing the 500 least essential genes excluding instances with no measurements ('NA'). *You do not need to upload these to Canvas. (1 point)*

C) With STRING (https://string-db.org) in multiple protein mode and subsequent 'Analysis' function, find the 5 most enriched Biological Processes (Gene Ontology) for both essential data-sets separately. Report and interpret the results (enriched pathways, edge enrichment, etc). *(2 points)*

STRING-db analysis was performed in multiple protein mode to identify enriched Biological Processes for the gene sets from Set 1 and Set 2.

For Set 1 and 2, the top 5 enriched BP terms (ranked by FDR) were the same:

- gene expression (GO:0010467)
- cellular nitrogen compound metabolic process (GO:0034641)
- nucleic acid metabolic process (GO:0090304)
- macromolecule metabolic process (GO:0009058)
- nucleobase-containing compound metabolic process (GO:0006139)

The PPI network for Set 1 showed a significant enrichment of interactions. The observed number of edges (11,150) was higher than the expected number (5,055), indicating that the proteins in Set 1 interact more than would be expected. This suggests functional linkages among the proteins. The average local clustering coefficient of 0.6 further supports this, suggesting that the proteins tend to cluster together within the network, implying that they participate in common biological processes or pathways.

For Set 2, the PPI network had more interactions than random. The observed number of edges (1157) was higher than the expected number of edges (496). With a average local clustering coefficient of 0.544 the data also suggests that the there are functional linkages and that the proteins are involved in common pathways.

The interconnectedness of the proteins in Set 1 and 2, along with the enrichment of the identified BP terms, suggests that these processes are likely related. The top enriched term, gene expression, may play a regulatory role in the other four metabolic processes. Further investigation is required to explore these relationships in more detail.

D) Instead of using the essential genes, also test the top 500 non-essential ones from Set1. Report and interpret the results (enriched pathways, edge enrichment, etc). Discuss if the result is expected and give a biological interpretation (100-200 words)? (*2 points)*

Results:

STRING-db analysis was performed to identify enriched Biological Processes for the non essential gene set from Set1:

For Set-1 Non-Essential Genes enriched BP terms (ranked by FDR) were:

Nucleic acid metabolic process (GO:0090304)

Nucleobase-containing compound metabolic process (GO:0006139)

Cellular nitrogen compound metabolic process (GO:0034641)

RNA metabolic process (GO:0016070)

Heterocycle metabolic process (GO:0046483)

In terms of the PPI network the observed interactions were significantly more than random estimated interactions, suggesting that some enrichment of interactions exist. The number of edges in the network (1023) were higher than the estimated number of edges (481) with an average cluster coefficient of 0.482. The cluster coefficient is slightly less than 0.5

Discussion:

The enrichment of some of the metabolic processes in the non-essential gene set-1 is somewhat expected. Even though non-essential genes are not critical of cell survival in ideal conditions, they are still important for cellular functionality. For example, non-essential metabolic genes may contribute to the efficiency of metabolism. The non-essential genes in this set may play a role in surviving under certain conditions or adjustment of metabolic pathways.

The modest degree of clustering is not unexpected. The non-essential genes might be involved in diverse or less tightly knit pathways when compared to essential genes. The lower clustering coefficient can reflect this wider range of function compared to the essential gene set-1.

Task 3) Analysing the data (6 points)

The resulting enrichments are potentially very cancer unspecific since many genes are generally 'essential' and have an aggravating effect on the cell, irrespectively of the cancer type. The two data-sets are from two different cancers. We are interested in the communalities and differences between the two sets now. The file 'both_sets.tsv' prepared for you has the Avana_Z scores from both screens.

A) How many essential genes (cut-off -5: the 'top-essential genes') are shared between the two? Write 1-2 sentences interpreting the result *(1 point)*

There are approximately 230 genes that meet the cutoff of a -5 gene score shared between set-1 and set-1.

B) Investigate this shared subset using STRING in multi-protein mode and subsequent 'Analysis' (https://string-db.org). Write 5-10 sentences on your observations concerning:
- are there more edges than expected?
- which are the top enrichments for the Gene Ontology Biological Process?
Discuss potential biological interpretation in 3-5 sentences. (2 points)

Essential genes shared for both data sets were analysed using StringDB with the highest confidence settings (>=0.9). The number of edges (1055) were higher than the expected number of edges (424) and the network has more

interactions than random; this suggests that the proteins might be connected via biological pathways.

The top 5 genes in the analysis which were ranked by the FDR:

- Nucleic acid metabolic process (GO:0090304), Increased DNA replication and transcription are hallmarks of cancer. Defects in DNA repair can cause increased cancer progression and genomic instability.
- Nucleobase-containing compound metabolic process (GO:0006139), control of nucleobases which are involved in cell growth, repair, and differentiation which supports rapid DNA synthesis observed in cancer cells.
- Cellular nitrogen compound metabolic process (GO:0034641), regulation of amino acids for protein synthesis to support growth of cancer cells
- RNA metabolic process (GO:0016070), regulation of gene expression via noncoding RNA which control cell activity like tumor suppressors. Deregulation of tumor suppressors can contribute to cancer growth and differentiation.
- Heterocycle metabolic process (GO:0046483), breakdown and utilization of heterocycles which are building blocks of proteins, nucleotides, and other essential biomolecules for cancer survival

C) Create a scatter-plot (x-axis set1, y-axis set2). Which areas contain potentially interesting data-points to learn about differences between the two cell-lines (write 3-5 sentences interpretation) *(1 points)*

A scatter plot of the Z-scores from the “both-sets” file after filtering was created with a linear line (in green) to help observe a probable correlation between sets. The density of points seems to cluster more on the middle-top right of the graph (set-2 -5:-7, set-1 -9:-7) as opposed to the bottom left which has less points and they are further apart. Correlation was calculated to be 0.25 which suggests a weak positive correlation between datasets. Some genes have substantially different Z scores between the two set as they are further away from the linear line. Further statistical analysis is required to identify a clear trend when comparing set-1 to set-2.

D) Which are essential only in one but not the other set? To get an initial answer to this, compute the absolute difference (called here DeltaAbs) between scores from Set1 and Set2 and report the top 15 most differently essential genes and their individual scores, as well as the derived DeltaAbs value and a short description of each gene (use public databases for the latter information). Give a 2-5 sentence interpretation of the result and report the top 15 table.
Tip 1: you may also use online tools to gain further insights, perhaps one you used here already). Tip 2: a simple subtraction will not work to compute the distance since, for example, some scores are negative in one cell-line but positive in the other *(2 points)*.

Gene	Set1	Set2	Delta_Abs
POLR3H	-11.99	-5.36	6.63
IFITM3	-11.36	-5.06	6.3
RPL12	-11.54	-5.7	5.84
SS18L2	-11.43	-6.16	5.27
RPL10A	-10.78	-5.67	5.11
TIMM10	-11.24	-6.37	4.87
RPS6	-9.69	-5.05	4.64
RPS20	-9.81	-5.2	4.61
RPS8	-10.59	-6.12	4.47
RPL23	-10.44	-5.99	4.45
NOC3L	-9.65	-5.21	4.44
NIP7	-10.36	-5.94	4.42
PRPF19	-9.4	-5.05	4.35
RPS28	-9.27	-5.07	4.2
RPL17	-9.32	-5.13	4.19

Figure 1: Ranked Top 15 Delta_Abs Shared Genes

The top 15 genes show a trend in which set-1 is significantly more essential by a factor of almost 2.0 compared to set-2. StringDB analysis was then performed using the top 15 delta_abs value genes, with the top 5 biological pathways ranked by FDR with a confidence score of >=0.9.

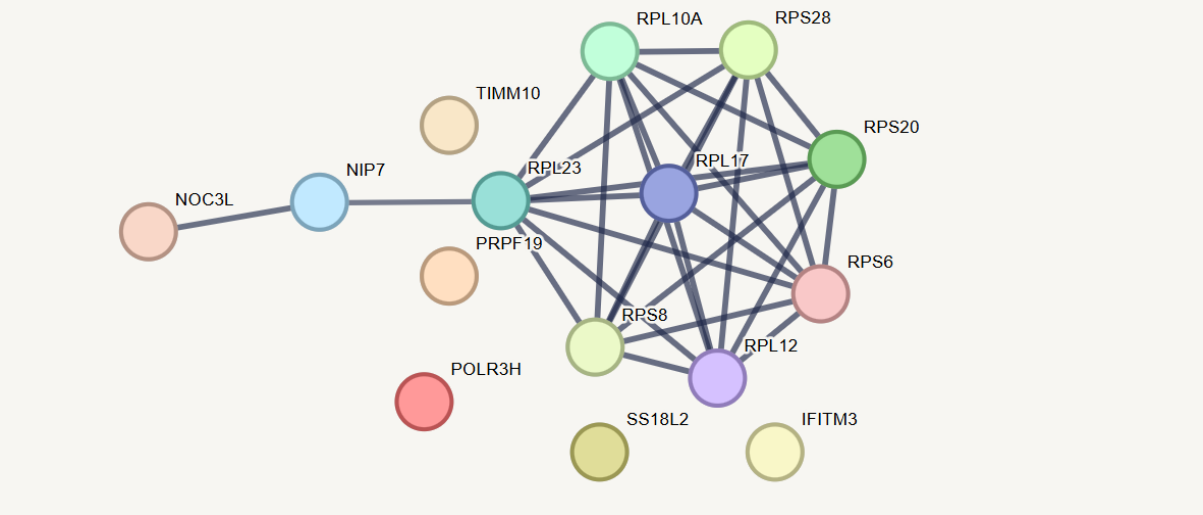


Figure 2: Network of Top 15 Enriched Proteins via StringDB

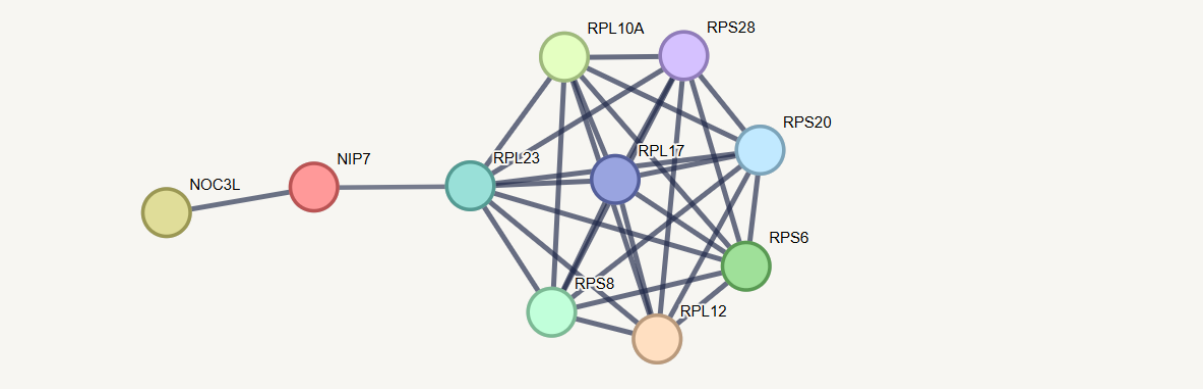
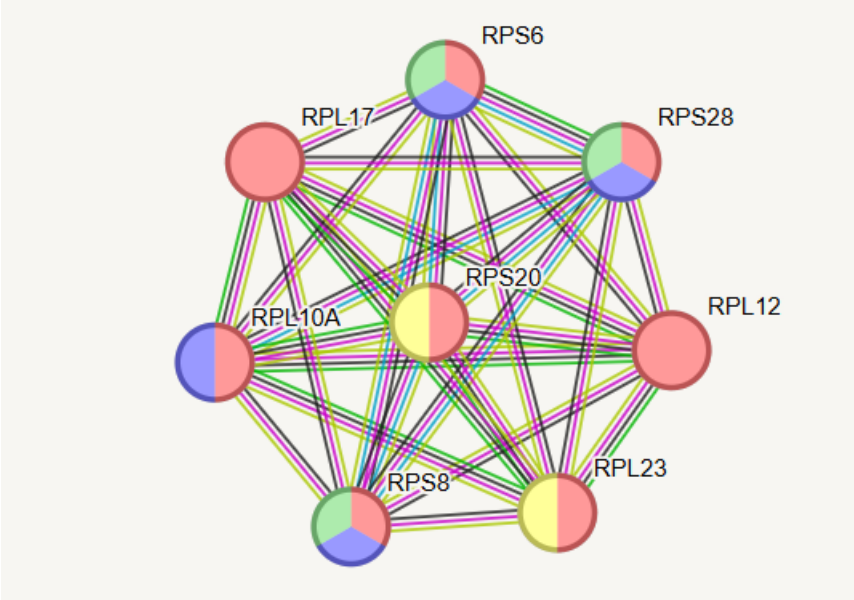


Figure 3: Network of Top 10 Enriched Proteins after Removal of Unconnected Proteins (TIMM10, PRPF19, POLR3H, SS18L2, and IFITM3) (See Figure 2)

Biological Process (Gene Ontology)					
GO-term	*description*	*count in network*	*strength*	*signal*	*false discovery rate*
GO:0002181	Cytoplasmic translation	8 of 123	2.11	5.11	2.14e-12
GO:0042254	Ribosome biogenesis	5 of 299	1.92	1.5	0.00026
GO:0006364	rRNA processing	4 of 220	1.55	1.19	0.0027
GO:0042274	Ribosomal small subunit biogenesis	3 of 78	1.88	1.16	0.0059
GO:0044260	Cellular macromolecule metabolic process	9 of 2512	0.85	0.78	0.00011

Figure 4: Background Reference Pathways Used for StringDB Analysis of the 10 Proteins

Two genes, NOC3L and NIP7 were removed from further analysis due to a lack of connectivity to other proteins at confidence scores above 0.5 and the absence of independent functional enrichment. The confidence threshold in STRING-db was lowered to 0.7 since any value greater resulted in no significantly enriched pathways. While this allows for the inclusion of less confident interactions, it revealed a tightly connected network of 8 proteins (RPL10A, RPS28, RPS20, RPL17, RPS6, RPS8, RPL23, RPL12). Functional enrichment analysis of these 8 interconnected proteins, using a background dataset (Figure 4) obtained from QuickGO, revealed a strong enrichment for GO terms related to cytoplasmic regulation, rRNA processing, ribosomal small subunit biogenesis, and positive regulation of signal transduction by p53 class regulator.



	Biological Process (Gene Ontology)				
GO-term	description	count in network	strength	signal	false discovery rate
GO:0002181	Cytoplasmic translation	8 of 123	2.2	6.23	4.82e-14
GO:0006364	rRNA processing	4 of 220	1.65	1.47	0.0010
GO:0042274	Ribosomal small subunit biogenesis	3 of 78	1.98	1.42	0.0029
GO:1901798	Positive regulation of signal transduction by p53 class mediator	2 of 30	2.22	0.86	0.0372

Figures 5 and 6: Network of Remaining 10 Proteins after Using Background Dataset and the Biological Processes Table Associating each Protein with Separate Colors.

Probable Effects of Enriched Biological Processes Relating to Cancer:

Cytoplasmic Translation: translation plays an important role in gene expression in which cancer cells can reprogram translation to create proteins that promote growth, survival, and metastasis

rRNA Processing: Ribosomal protein mutations can affect the amount of translation by changing the rate of protein synthesis; in cancer cells the increased rate of protein synthesis can be used to influence cell growth

Ribosomal Small Subunit Biogenesis: The small subunit (40S Subunit) biogenesis is responsible for starting protein synthesis and initiating translation. In cancer cells the 40S Subunit can lead to an increased amount of protein synthesis and capacity, loss of tumor suppressor genes, evasion of apoptosis via the p53 pathway, and the translation of oncoproteins or proteins that promote cancer.

Positive Regulation of Signal Transduction by P53 Class Mediator: P53 is a tumor suppressor protein that can activate cell arrest and apoptosis pathways. The positive regulation of signal transduction of P53 may be counterintuitive; however, the increased P53 activity may be a response to the loss of other tumor supressors or onco gene activation.

Task 4) Discussion of future steps (mini-essay)

The lab has a hand-curated list of signalling pathways which they think are enriched in the essential genes in sets 1 and 2. They have asked you to propose a set of

computational analyses to test if this is true. Write a short (**maximum 500 words**) paragraph on how you would conduct this analysis. Your answer should consider the following points:

- What sort of analysis would you recommend, and what are the statistical principles that underpin it?
- What are the limitations of this dataset? Can it be used effectively to learn about kinases in cancer with the data given?
- What are the potential sources of biases in this analysis? Can these be accounted for?
- What would the output look like? How would you present the result to a non-computational scientist?
  (6 points)

Gene Set Enrichment Testing:

The first step in testing would be to create a list of related proteins for each pathway given by the lab by utilizing online databases like QuickGO. Next, a cutoff would be determined for the given Z scores from both datasets (e.g Z>=-5) to determine the essential genes. The lesser the Avana Z-score the more essential a gene is. The essential genes from both sets will be put into a list to be used in the analysis.

A hypergeometric test would then be performed to determine if the overlap between the essential genes and pathway genes is statistically significant. The test is used to calculate the probability of the observed overlap between the data due to chance, given the pathway gene list and essential genes from Set-1 and Set-2. The Benjamini-Hochberg method will be applied to the p values to correct for multiple testing and false discovery rate (FDR). Genes enriched for a certain pathway will have a p value <=0.05.

Protein-Protein Interaction (PPI) Network Analysis using StringDB:

For each pathway given by the lab, a pathway specific background set will be made including the essential genes from both Set-1 and 2. PPI Analysis is then performed using the background set explained above and the essential genes from set-1 and 2 as a foreground. A confidence threshold will start with 0.9, but a lower threshold of 0.7 can be used if necessary to explore potential interactions. The metrics that will be used for this analysis are a low FDR of enriched GO terms and pathways (>=0.05), high cluster coefficient indicating network interconnectivity (>0.5), and the high number of observed edges vs expected edges which represent the interaction strength.

Limitations of the Dataset:

- In-Vitro Data: The pooled CRISPR knockout screens are in-vitro data which isn't fully representative of in-vivo processes or disease mechanisms.

- Z-Score Cutoff: The cutoff choice of the Z-score can introduce bias since it can affect the essential gene set. A good Z-score cutoff is important for the analysis.
- Kinase Interference: Since additional data from kinases such as kinase activity assays aren't provided, inference about specific kinase roles in cancer is difficult.

Potential Biases:

- Gene Length and Expression Bias: Gene length and expression can add bias to CRISPR screens. However, pathway specific backgrounds can eliminate this bias.
- Background Gene Set: Using the whole genome as a background set can introduce bias, so pathway-specific background sets will be used
- Multiple Testing: Benjamini-Hochberg correction controls the false discovery rate accounting for multiple correction error.

Presentation of Results:

Pathway Comparison Table		
Analysis Metrics	Pathway 1	Pathway 2
No. Pathway Genes		
No. of Overlapping Genes		
Hypergeometric Test Value		
FDR of Hypergeometric Test Values		
String-DB FDR		
Cluster Coefficient		
Observed vs Expected Edges Ratio		

Figure 1: A table with will be created to present the results. Each pathway (1,2, etc) will have specific information such as the number of pathway genes and hypergeometric test values. This information will be used to compare each pathway given by the lab and if the data from each set (1 and 2) agree with each pathway.

Visualization:

A bar chart will then be created displaying the pathways given by the lab on the x axis and -log10(FDR of Hypergeometric Test Values) on the y axis. Another bar

chart will be a grouped bar chart. The x axis would still be the pathways given by the lab, however a weighted and normalized score between the analysis metrics in the pathway comparison. Each analysis metric will have a different shade of color based on their contribution or value. Each analysis has it's pros and cons, the single bar chart is useful for a simple yet effective way to recognize enriched pathways; but the multiple shaded bar chart can display multiple metrics per pathway leading to more information.

References:

- The European Bioinformatics Institute. (2025). *QuickGO*. [online] Available at: https://www.ebi.ac.uk/QuickGO/ [Accessed 1 Mar. 2025].

- National Library of Medicine. (2025). *PubMed*. [online] Available at: https://pubmed.ncbi.nlm.nih.gov/ [Accessed 1 Mar. 2025].

- The Gene Ontology Consortium. (2025). *Gene Ontology Resource*. [online] Available at: https://geneontology.org/ [Accessed 1 Mar. 2025].

- Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J. and Mering, C. von (2019). STRING v11: protein--protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1), pp.D607–D613.