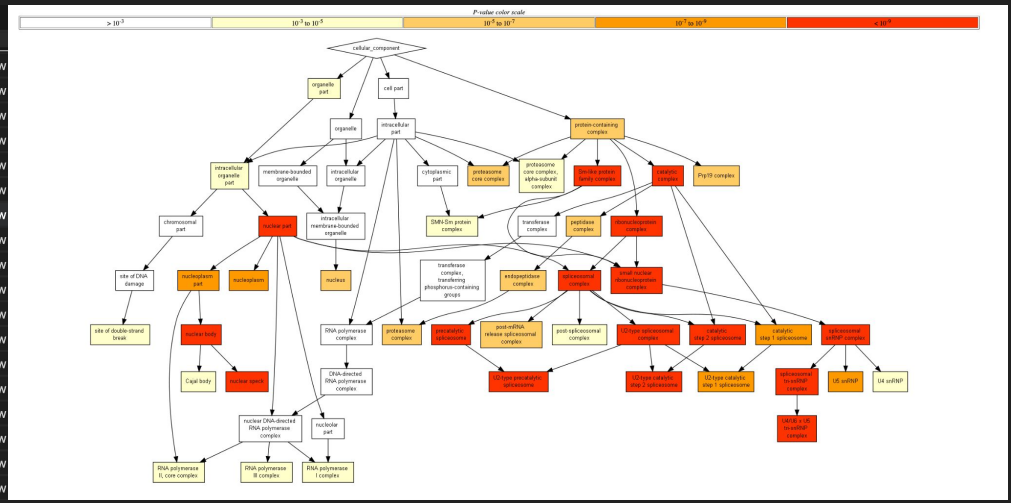
Week 4 - Visualizing GO (Gene Ontology) Analysis Output

I am looking at intron retention data and how it affects RNA metabolism. With the ReLiC screen data, I inputted the gene IDs based on its ranking relating to positive selection (increase of introns). I inserted the output into pandas and at the moment I am plotting a histogram to visually understand the data and how it answers the objective. Hopefully with Christine's and I's findings we can eventually create figures that can be beneficial to the project as a whole.

	index	GO term	Descrip...	P-value	FDR q-v...	*Enrichment (N, B, n, b)*	Genes
78	78	GO:00442...	cellular me...	0.000416	0.0387	1.29 (2149,1540,67,62)	[+] Show
59	59	GO:00081...	metabolic ...	0.000174	0.0213	1.29 (2149,1562,67,63)	[+] Show
37	37	GO:0071704	organic su...	0.0000162	0.00313	1.34 (2149,1504,67,63)	[+] Show
39	39	GO:00068...	nitrogen c...	0.0000304	0.0056	1.36 (2149,1424,72,65)	[+] Show
25	25	GO:00431...	macromol...	0.00000411	0.00116	1.38 (2149,1421,78,71)	[+] Show
33	33	GO:00442...	primary m...	0.0000155	0.00334	1.38 (2149,1440,67,62)	[+] Show
72	72	GO:0051179	localization	0.000296	0.0298	1.39 (2149,567,291,107)	[+] Show
86	86	GO:00512...	establishm...	0.000793	0.0671	1.40 (2149,513,291,97)	[+] Show
56	56	GO:1903311	regulation ...	0.000145	0.0187	1.53 (2149,250,439,78)	[+] Show
79	79	GO:1901360	organic cy...	0.000444	0.0408	1.55 (2149,1152,47,39)	[+] Show
70	70	GO:00464...	heterocycl...	0.000285	0.0295	1.57 (2149,1137,47,39)	[+] Show
71	71	GO:00067...	cellular ar...	0.000292	0.0299	1.57 (2149,1138,47,39)	[+] Show
67	67	GO:00061...	nucleobas...	0.000241	0.0261	1.58 (2149,1130,47,39)	[+] Show
51	51	GO:00903...	nucleic aci...	0.0000934	0.0132	1.63 (2149,1093,47,39)	[+] Show
77	77	GO:00301...	protein cat...	0.000404	0.0381	1.69 (2149,41,995,32)	[+] Show
87	87	GO:00485...	negative r...	0.000813	0.0679	1.82 (2149,155,313,41)	[+] Show
38	38	GO:0016070	RNA meta...	0.000028	0.00528	1.86 (2149,1016,33,29)	[+] Show
88	88	GO:00062...	nucleotide...	0.000842	0.0696	10.41 (2149,24,43,5)	[+] Show



Week 4

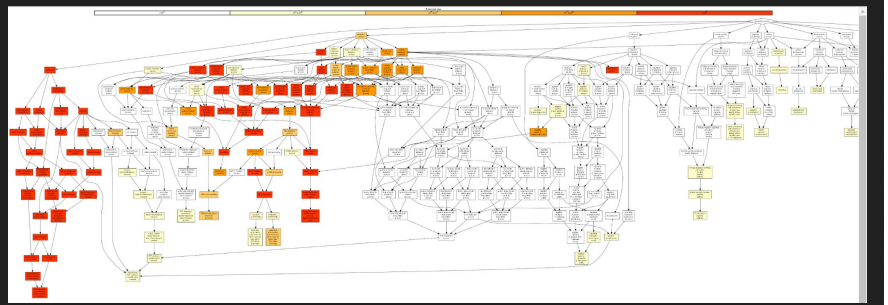
Also starting to developing a rough draft of a presentation into markdown by noting all the significant finding along my studies.

notes > gq_methodology.md > # Objective > ## Output

```
# Objective
- RNA metabolism is influenced by various factors, including processes like splicing and nonsense-mediated decay (NMD). Assessing the impact of these factors is crucial. In the context of CRISPR screens, direct measurement of RNA levels becomes important. RNA sequencing provides a quantitative approach to evaluate RNA molecules, offering insights into the effects of gene perturbations on transcriptional regulation, key genes, and pathways. The ReLic screening strategy, focused on splicing and NMD, holds promise in identifying functional consequences resulting from CRISPR-mediated gene modifications. Such investigations contribute to a deeper understanding of the impact of RNA-related processes on gene function.
- **GQ** - What are the factos that affects RNA metabolism specifically spilicing?
## Research
- **What is CRISPR Screening?**
    - CRISPR screening is a technique used to identify and understand the function of specific genes in an organism. It involves using CRISPR-Cas9, a powerful gene editing tool, to systematically target and modify different genes within a cell or organism. By observing the resulting changes in cellular behavior or characteristics, researchers can gain insights into the roles and importance of different genes in various biological processes.
- **ReLiC Screening** (Notes from Presentation)
    - CiBER-seq is a powerful tool for studying gene expression and identifying gene pathways related to protein synthesis. It utilizes a barcode system to measure the effects of gene changes, revealing varying impacts on cells. By integrating CRISPR interference and barcoded expression reporters, it allows measurement of gene expression and understand how different genetic changes influence gene interactions. CiBER-seq uncovers the connections between genes and their roles in biological processes.
    - Splicing ReLiC: As learned in started Biology classes, the relation between DNA and RNA is known as the central dogma of molecular biology where DNA makes mRNA makes protein through a strategic process. The focus of this process includes a complex and regulated step that includes transcription, pre mRNA splicing, editing, degradation, polyadenylation, etc. known as RNA metabolism. Splicing plays a vital role in which the cutting of introns, noncoding, for exons, coding, to covalently connect for the exons to exit the nucleus and be expressed. The exons acts as a random barcode that is counted by a deep screening to find the effect of sgRNA on reporter mRNA levels. Pooled CRISPR screens are widely used for forward genetic screening in human cells; however, the disrupted RNA metabolism does not manifest with growth and viability defects. For instance, the presence of a mutation can change a phenotype completely or not change it at all. FACS-based CRISPR screening uses protein level readouts for mRNA phenotypes; however, RNA molecules may not alter protein levels. RNA molecules can undergo modifications with specific proteins and the production can be independently performed.
    - **Insert Figures**
- **MaGeCK**
    - Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout: a method used to prioritize single guide RNA, genes, and specific pathways in CRISPR screenings. This method is commonly used in comparison to other methods, for it identifies positive and negatively selected genes under robust and extreme conditions, noting which genes survived and which genes didn't. Though the use and integration of public data sets knocked out data sets are easily identifiable.
    - Pandas is a powerful Python library that specilizaes in data manipulation and analysis was used to focus on the genetic screening data analysis. Pandas is also used to manage tabular data, which helped with analyizing which column to use for the input on GOrilla.
    - Utilizing pandas, I compared the variables from MaGeCK and visualized the data as a graph. Through this analysis, I identified the most suitable column for GOrilla.
- **GO(Gene Ontology)rilla**
    - GOrilla is a web-based application that performs gene ontology (GO) enrichment analysis on ranked gene lists, without the need for explicit target and background sets. GOrilla provides efficient and rigorous statistical analysis, taking threshold multiple testing into account without the need for simulations. The results are presented in a hierarchical structure, offering a clear visualization of the relationships between enriched GO terms. GOrilla is a valuable addition to the existing repertoire of GO enrichment tools, offering unique features, fast running time, and effective graphical representation.
```

Week 1-4 Update

- Previous weeks focus: python, data interpretation/organization, and gene ontology
- NMD ReLiC: The analysis of NMD (nonsense-mediated mRNA decay) insertion and deletion events in CRISPR data screening yields valuable insights into the functional consequences and potential regulatory effects of gene modifications within the nonsense-mediated decay pathway.
- This week we are analyzing, interpreting, and putting the gene ontology data into a figure.

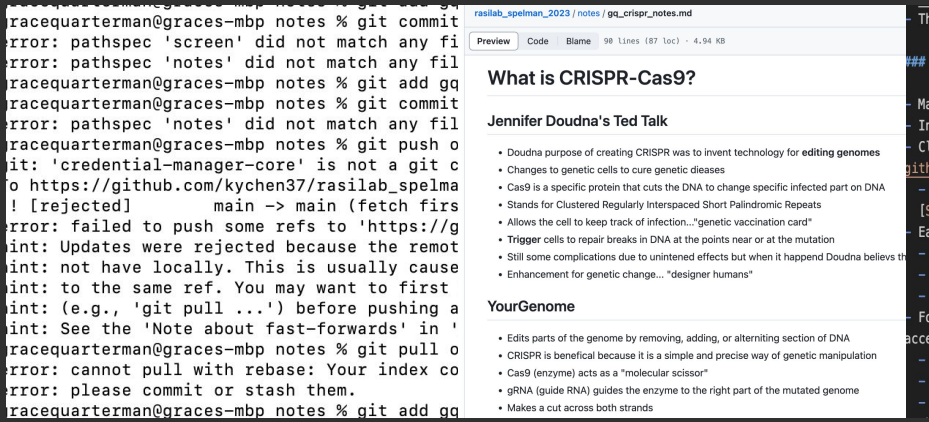


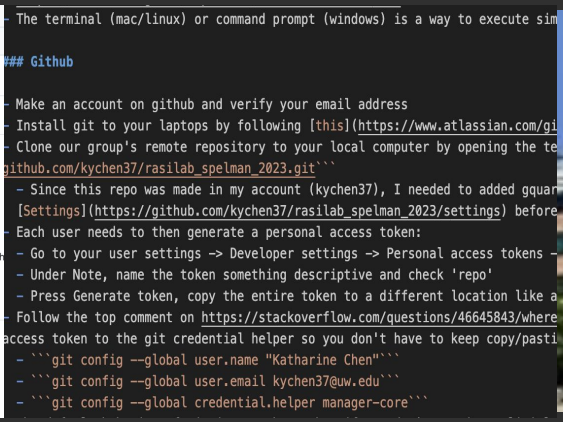
GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)	Genes
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	2.85E-28	2.09E-24	4.48 (2150,115,263,63)	[+] Show genes
GO:0019083	viral transcription	2.1E-27	7.72E-24	5.02 (2150,88,263,54)	[+] Show genes
GO:0006613	cotranslational protein targeting to membrane	2.65E-27	6.5E-24	4.63 (2150,90,294,57)	[+] Show genes
GO:0006413	translational initiation	2.99E-27	5.5E-24	4.12 (2150,133,263,67)	[+] Show genes
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	5.96E-27	8.76E-24	4.65 (2150,88,294,56)	[+] Show genes
GO:0070972	protein localization to endoplasmic reticulum	6.37E-27	7.8E-24	4.58 (2150,91,294,57)	[+] Show genes
GO:0072599	establishment of protein localization to endoplasmic reticulum	6.37E-27	6.69E-24	4.58 (2150,91,294,57)	[+] Show genes
GO:0045047	protein targeting to ER	6.37E-27	5.83E-24	4.58 (2150,91,294,57)	[+] Show genes
GO:0006612	protein targeting to membrane	1.5E-26	1.23E-23	4.53 (2150,92,294,57)	[+] Show genes
GO:0006605	protein targeting	3.26E-25	2.4E-22	4.23 (2150,102,294,59)	[+] Show genes
GO:0090150	establishment of protein localization to membrane	7.72E-25	5.16E-22	4.24 (2150,100,294,58)	[+] Show genes
GO:0072657	protein localization to membrane	7.58E-22	4.64E-19	3.86 (2150,110,294,58)	[+] Show genes
GO:0072594	establishment of protein localization to organelle	2.24E-21	1.27E-18	3.52 (2150,133,294,64)	[+] Show genes
GO:0000956	nuclear-transcribed mRNA catabolic process	5.08E-20	2.67E-17	3.22 (2150,175,263,69)	[+] Show genes
GO:0006412	translation	6.13E-19	3.02E-16	3.02 (2150,172,294,71)	[+] Show genes
GO:0043604	amide biosynthetic process	8.29E-19	3.81E-16	2.95 (2150,181,294,73)	[+] Show genes
GO:0006518	peptide metabolic process	1.24E-18	5.36E-16	2.93 (2150,182,294,73)	[+] Show genes
GO:0043043	peptide biosynthetic process	1.41E-18	5.76E-16	2.98 (2150,174,294,71)	[+] Show genes
GO:0043603	cellular amide metabolic process	1.54E-18	5.95E-16	2.87 (2150,191,294,75)	[+] Show genes
GO:0006402	mRNA catabolic process	3.95E-18	1.45E-15	3.00 (2150,191,263,70)	[+] Show genes
GO:0033365	protein localization to organelle	5.35E-18	1.87E-15	2.99 (2150,169,294,69)	[+] Show genes
GO:1901566	organonitrogen compound biosynthetic process	1.17E-16	3.9E-14	2.69 (2150,217,280,76)	[+] Show genes
GO:0009057	macromolecule catabolic process	1.87E-16	5.99E-14	2.32 (2150,279,315,95)	[+] Show genes
GO:0044265	cellular macromolecule catabolic process	1.97E-16	6.03E-14	2.33 (2150,270,315,93)	[+] Show genes
GO:0006886	intracellular protein transport	6.21E-16	1.83E-13	2.92 (2150,182,263,65)	[+] Show genes
GO:0006401	RNA catabolic process	7.78E-16	2.2E-13	3.04 (2150,215,294,62)	[+] Show genes
GO:0034613	cellular protein localization	2.5E-15	6.81E-13	2.68 (2150,194,294,71)	[+] Show genes
GO:0070727	cellular macromolecule localization	5.85E-15	1.54E-12	2.62 (2150,201,294,72)	[+] Show genes
GO:1901361	organic cyclic compound catabolic process	1.74E-14	4.42E-12	2.51 (2150,248,263,76)	[+] Show genes
GO:0046700	heterocycle catabolic process	1.74E-14	4.28E-12	2.51 (2150,248,263,76)	[+] Show genes
GO:1901575	organic substance catabolic process	2.35E-14	5.58E-12	2.11 (2150,326,319,102)	[+] Show genes
GO:0016032	viral process	2.76E-14	6.35E-12	2.47 (2150,238,263,72)	[+] Show genes
GO:0044403	symbiont process	2.76E-14	6.16E-12	2.47 (2150,238,263,72)	[+] Show genes
GO:0046907	intracellular transport	3.65E-14	7.9E-12	3.08 (2150,317,110,50)	[+] Show genes
GO:0044270	cellular nitrogen compound catabolic process	3.86E-14	8.11E-12	2.48 (2150,251,263,76)	[+] Show genes
GO:0034655	nucleobase-containing compound catabolic process	5.42E-14	1.11E-11	2.48 (2150,247,263,75)	[+] Show genes
GO:0019439	aromatic compound catabolic process	5.42E-14	1.08E-11	2.48 (2150,247,263,75)	[+] Show genes

What have I learned during my first week at Fred Hutch?

I learned ...

- About the application **terminal** on my mac computer and the similarities and difference between using the app on a mac computer and windows computer.
- **Markdown** and its markup language to transfer to a text document, website, etc.
- **Github** and **VSCode** and the connection to transfer text using the terminal on my mac
- About **Slack** and its use for messaging
- And refreshed my memory on **CRISPR** and its strategy to edit genomes
- The routes of the campus





Grace Quarterman

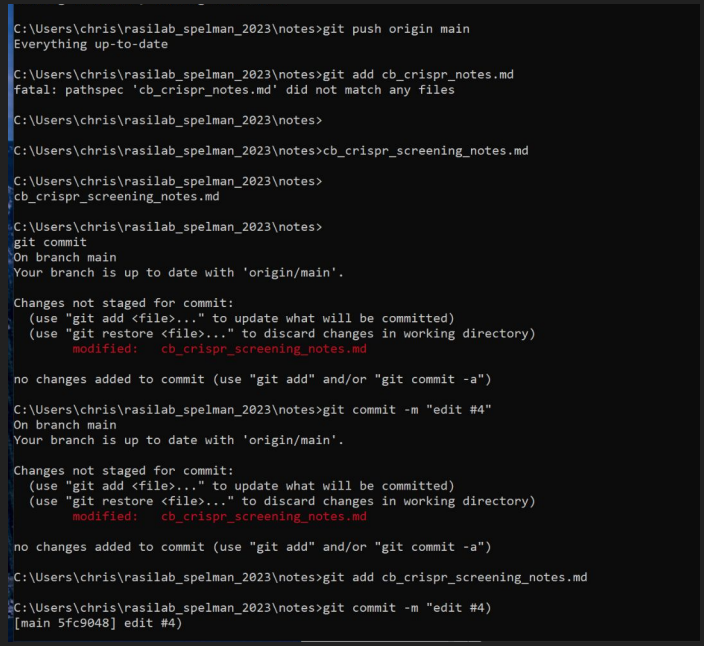
WEEK 1 AT FRED HUTCH CANCER CENTER IN THE SUBRAMANIAM LAB

GITHUB, COMMAND PROMPTS, VSCODE AND MARKDOWN

- LANGUAGE BETWEEN WINDOWS AND MAC COMPUTERS
- COMPUTER INTERACTIONS AND NAVIGATIONS
- TROUBLESHOOTING ERROR MESSAGES WITHIN OUR CODE
    - THERE WERE A LOT OF THEM

CRISPR

- IT USES A SPECIALIZED PROTEIN CALLED CAS9, GUIDED BY A SMALL RNA MOLECULE, TO TARGET SPECIFIC DNA SEQUENCES AND INTRODUCE MODIFICATIONS, WITHIN LIVING ORGANISMS



CHRISTINE M. BYNUM