

Data Analysis & Visualization Portfolio

Justin Jao

jao.c.justin@gmail.com

+17786812800

Table of Contents

<i>Ciernia Lab Projects</i>	3
Microglia Enrichment Calculator Project	4
Gene Expression Project	6
ChIP-seq Workflow Diagram.....	9
<i>BC Cancer Projects</i>	12
Cancer of Unknown Primary Nomogram	13
BRCA Analysis Visualizations.....	14
<i>Genome Institute of Singapore Projects</i>	18
RGB Analysis Project	19

Ciernia Lab Projects

Ciernia Lab (Djavad Mowafaghian Centre for Brain Health, UBC)

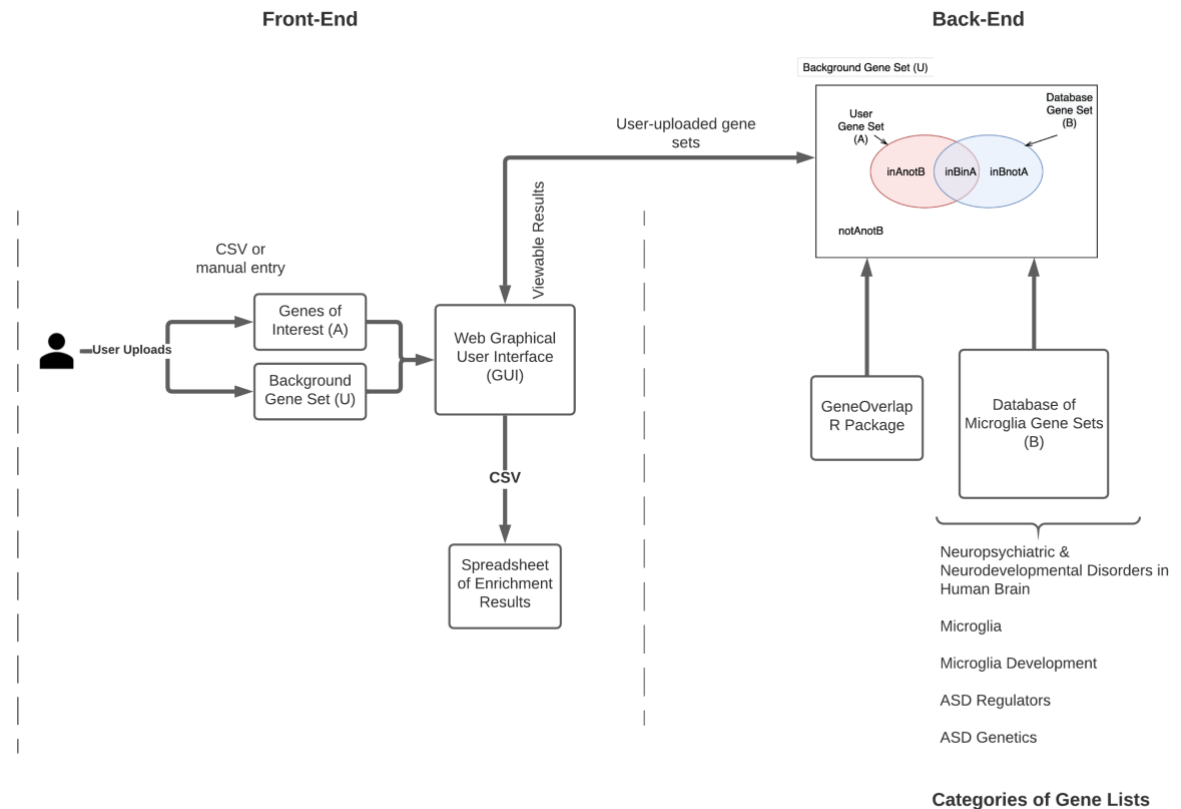
Period of Involvement: September 2019 - Present

Microglia Enrichment Calculator Project

An interactive dashboard developed to allow non-technical lab members to perform basic statistical analysis. Outputs results to an interactive table.

Built using the [Shiny](#) package and the [GeneOverlap](#) package in R, and hosted on shinyapps.io.

Workflow Overview



You can access the app at: <https://ciernialab.shinyapps.io/MGEnrichmentApp/>
GitHub repository with more details can be found at: <https://github.com/ciernialab/MGEnrichmentApp>

Screenshot of Dashboard:

Microglia Gene Set Enrichment Calculator

Input your genes of interest here (must all be the same gene ID format)

ENSMUSG00000067879
ENSMUSG00000070576

or upload your gene list here

Browse... No file selected

Which gene ID are you using?

☒ Ensembl
☐ Entrez
☐ MGI Symbol

Which gene list groups are you interested in?

☒ Neuropsychiatric & Neurodevelopmental Disorders human brain
☒ Microglia Development ☒ Microglia ☒ ASD regulators
☒ inflammation ☒ ASD genetics

Set the background query:

☒ All mm10 Genes
☐ All Genes in the Database
☐ Custom

Disable Intersection Gene IDs?

☐ Intersection IDs ☐ Ensembl ☐ MGI Symbol ☐ Entrez

Change Minimum FDR-value (1.0 means no filtering):

0.01 0.11 0.21 0.31 0.41 0.51 0.61 0.71 0.81 0.91 1

Query Genes Download Results

Table Help

Show 10 entries

Search:

	listname	pvalue	OR	notAnotB	inAnotB	inBnotA	inBinA	intersection_IDs	intersection_ensembl	intersection_mgi_symbol	intersection_er
1	Acute EAE vs. CTRL spinal cord MG	5.7565e-7	74.5098131940198	56116	18	167	4	ENSMUSG000000029581, ENSMUSG000000005583, ENSMUSG000000036353, ENSMUSG000000070348	ENSMUSG000000005583, ENSMUSG000000029581, ENSMUSG000000070348, ENSMUSG000000036353	Mef2c, Fscn1, Ccnd1, P2ry12	17260, 14086, 1270839
2	adult MG cluster 1	0.00041723	23.5867213742599	55909	19	374	3	ENSMUSG000000029919, ENSMUSG000000069662, ENSMUSG000000036353	ENSMUSG000000069662, ENSMUSG000000036353, ENSMUSG000000029919	Marcks, P2ry12, Hpgds	17118, 70839, 544
3	adult MG cluster 2	0.0004674	22.6776087644377	55894	19	389	3	ENSMUSG000000029581, ENSMUSG000000021224, ENSMUSG000000038175	ENSMUSG000000029581, ENSMUSG000000038175, ENSMUSG000000021224	Fscn1, Mylip, Numb	14086, 218203, 16
4	Amoeboid > Ramified MG	0.04898	6.07758598273977	55372	20	911	2	ENSMUSG000000038331, ENSMUSG000000069662	ENSMUSG000000069662, ENSMUSG000000038331	Marcks, Satb2	17118, 212712
5	Apoe KO vs	1	0	0	0	0	0				

You can access the app at: <https://ciernialab.shinyapps.io/MGEnrichmentApp/>

GitHub repository with more details can be found at: <https://github.com/ciernialab/MGEnrichmentApp>

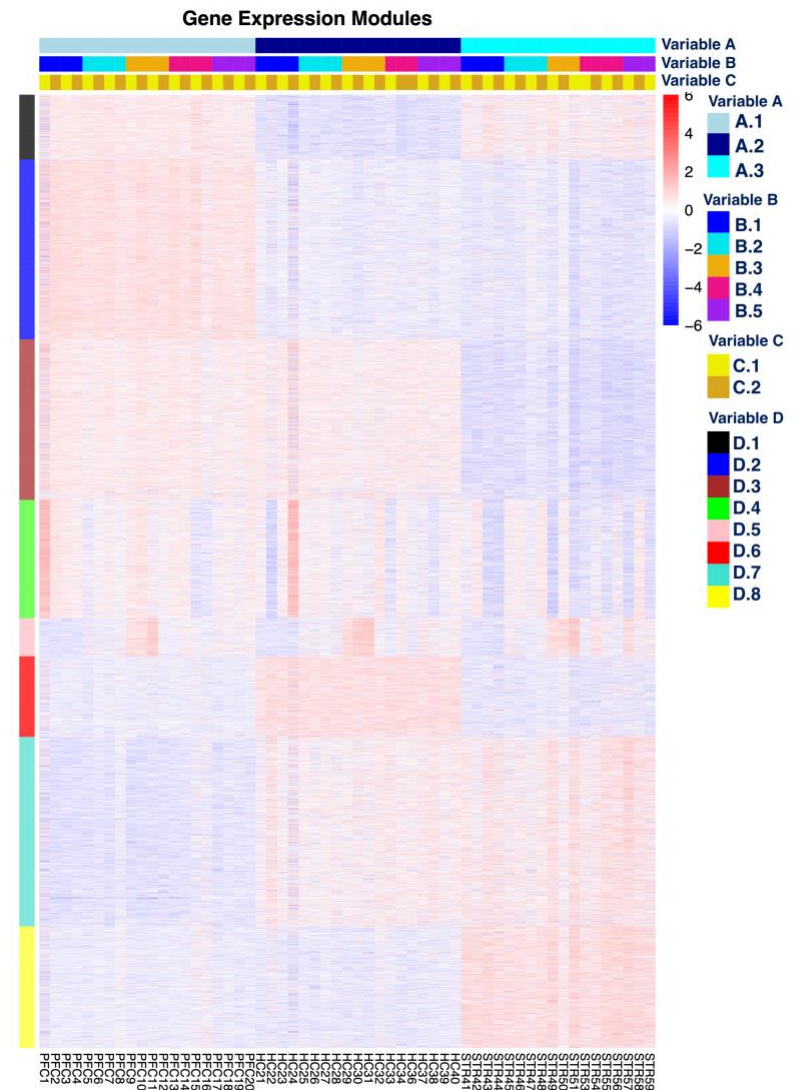
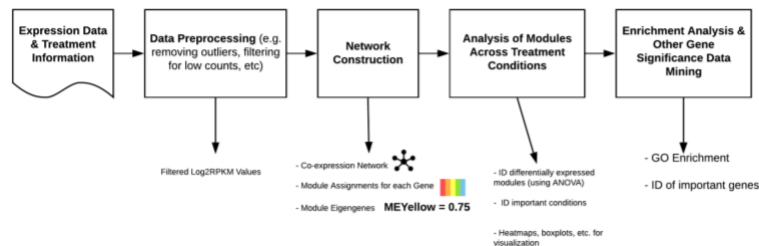
Gene Expression Project

A data analysis project I worked on as a bioinformatics research assistant in the Ciernia Lab.

This heatmap figure on the right was the key visual I made, representing changes in gene expression of a certain category (module) of genes across different samples and variable conditions. Each column is a sample, and each row is a different gene.

Made using the [WGCNA](#) package in R.

Workflow Overview:



Side note: I know the heatmap is very busy, and the figure we actually used was a smaller, more concise version of this, which I can't show since the data is currently unpublished.

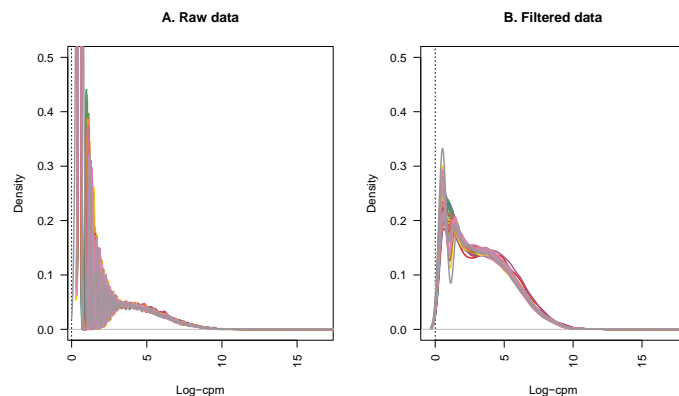
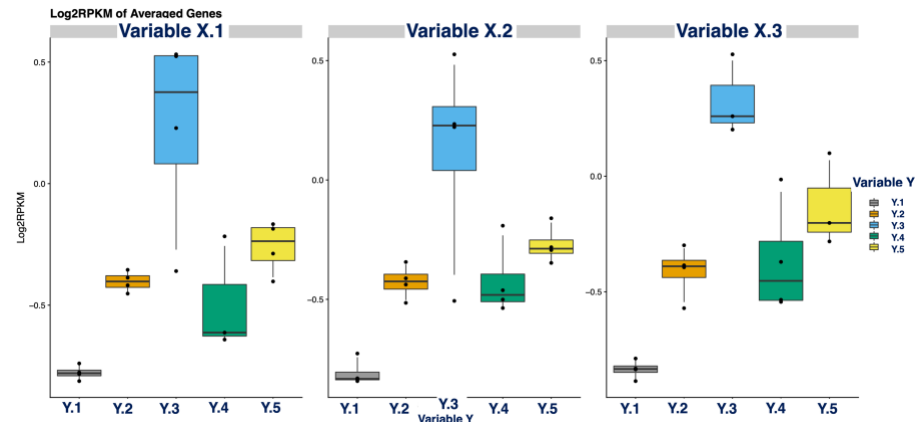
GitHub repository with more details about the analysis and background of the project can be found at:

https://github.com/justinjao/Gene_Expression_Project

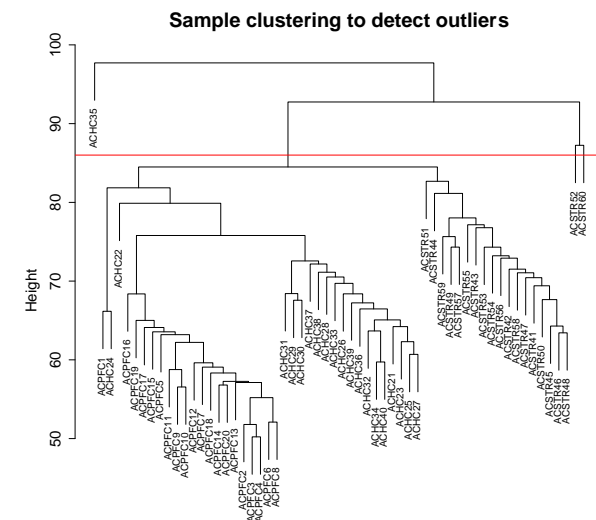
Some other visualizations I made in the intermediate steps of the analysis:

Boxplots showing changes in gene expression values in different conditions.

Made using [ggplot2](#) package in R.



Both made using [WGCNA](#) package in R.

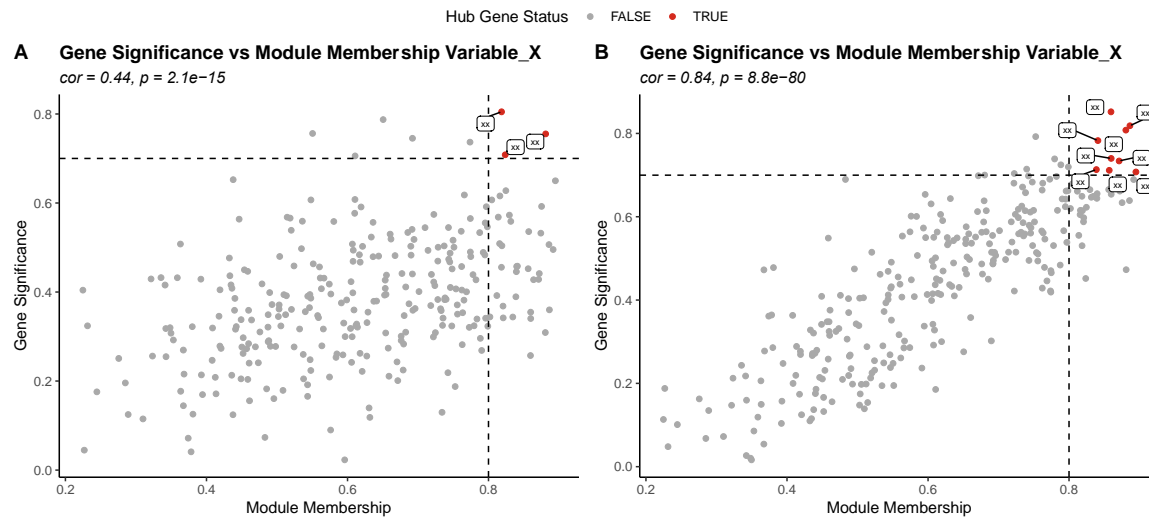


Expression values before and after filtering

Hierarchical clustering to visually inspect samples for outliers

GitHub repository with more details about the analysis and background of the project can be found at:

https://github.com/justinjao/Gene_Expression_Project

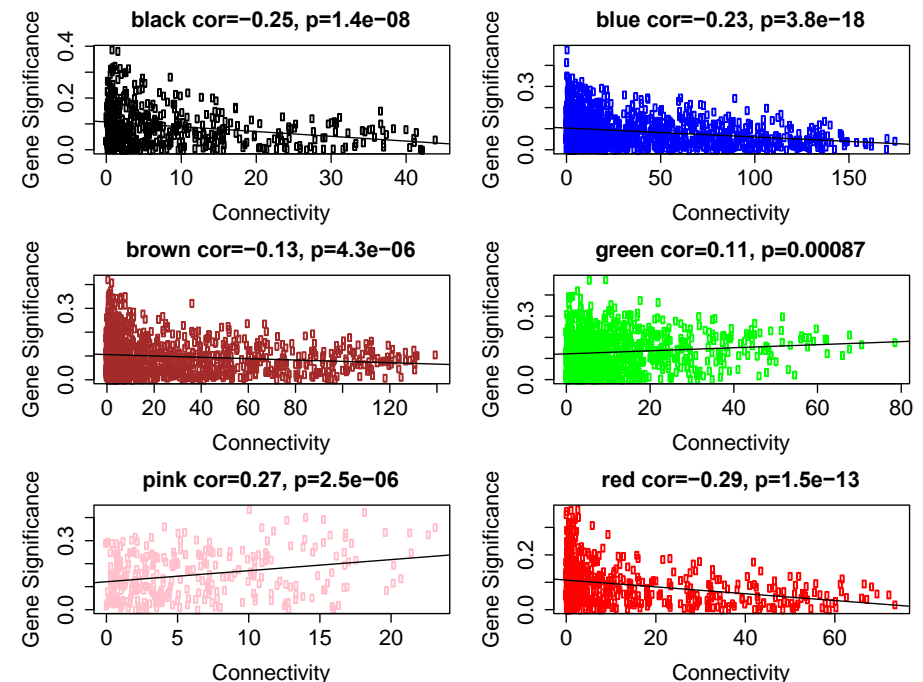


Scatterplot with cut-offs at dotted line to select for highly relevant genes (gene labels censored as xx).

Made using [ggplot2](#) package in R.

Gene significance vs connectivity plots to determine relevance of different modules identified.

Made using the [WGCNA](#) package in R.



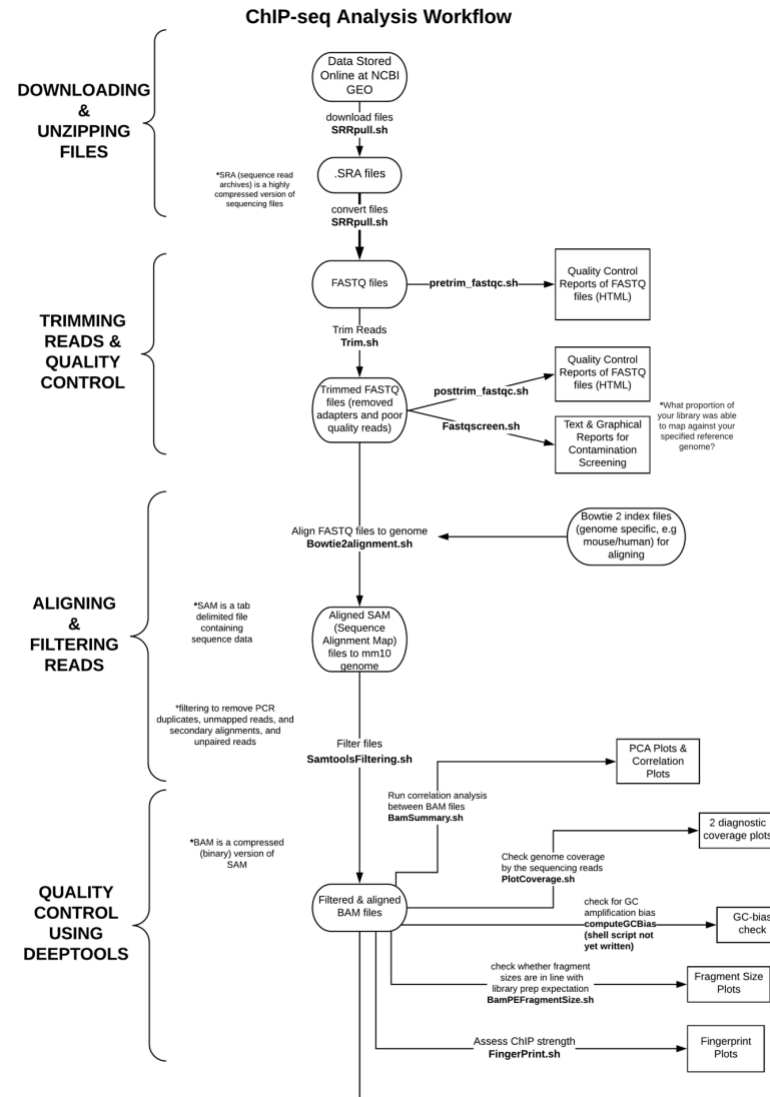
GitHub repository with more details about the analysis and background of the project can be found at:

https://github.com/justinjao/Gene_Expression_Project

ChIP-seq Workflow Diagram

Because of the pandemic, our lab had to temporarily shutdown, and so a lot of members started exploring more computational projects.

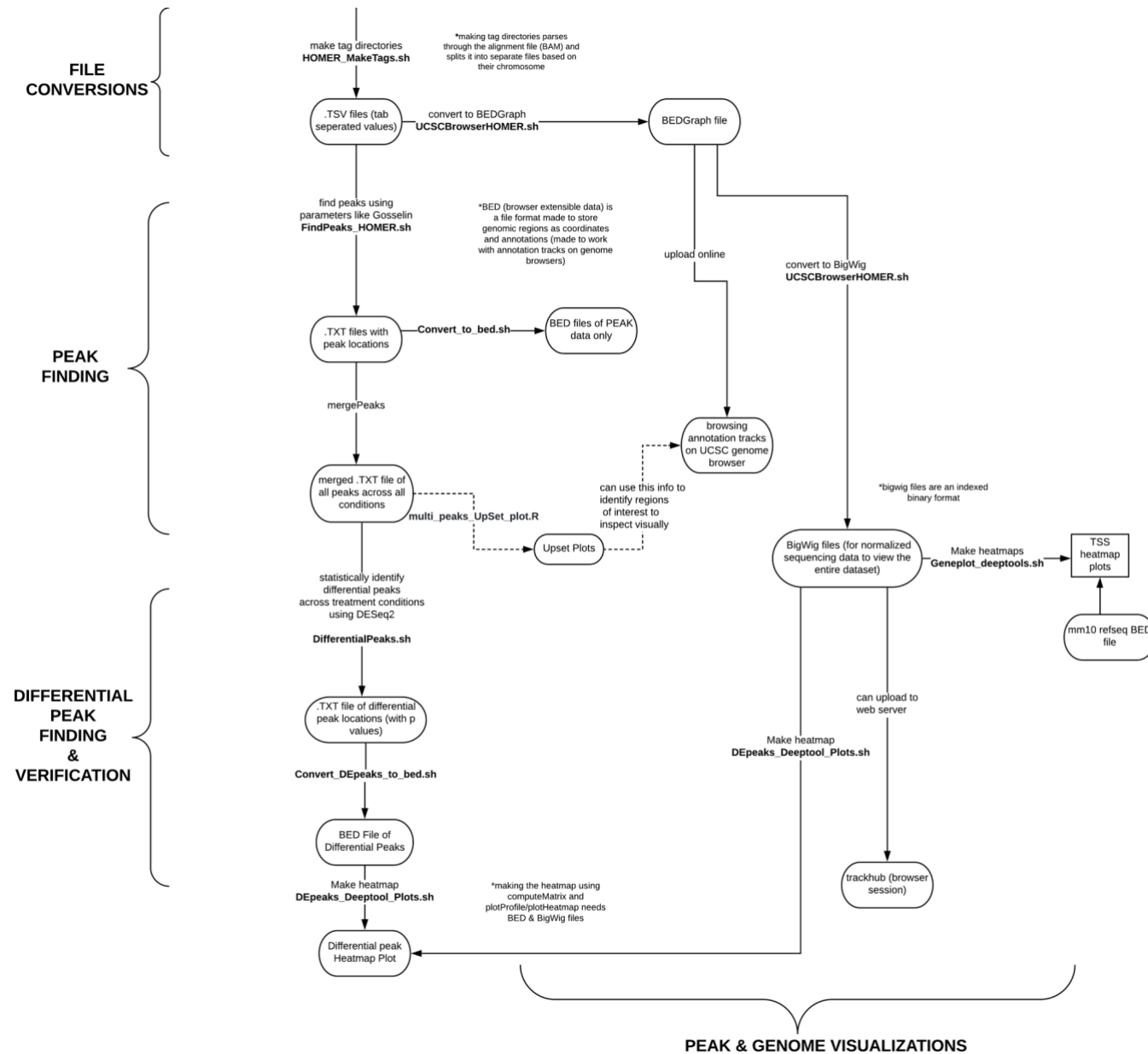
One such analysis is called Chromatin-Immunoprecipitation sequencing analysis (ChIP-seq). To help members of my lab learn the data analysis pipeline (such as what scripts to run, and what each step of the analysis was doing), I made this flowchart.



GitHub repository with more details about the analysis and background of the project can be found at:

<https://github.com/ciernalab/Alder-ChIPseq-Tutorial>

Part 2 of the
flowchart:

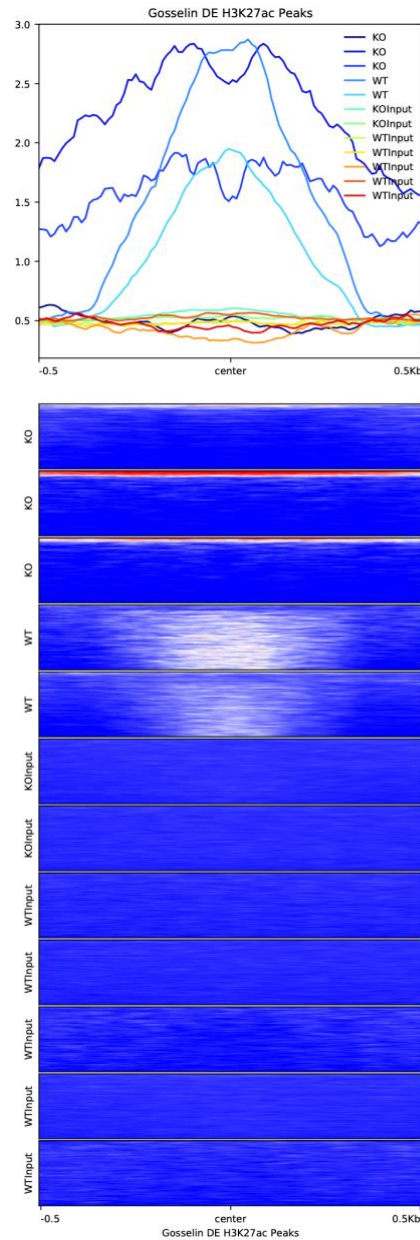


GitHub repository with more details about the analysis and background of the project can be found at:

<https://github.com/ciernalab/Alder-ChIPseq-Tutorial>

Peak-finding (one of the figures that can be generated from ChIP-seq).

Made using [HOMER](#) software.



GitHub repository with more details about the analysis and background of the project can be found at:

<https://github.com/ciernalab/Alder-ChIPseq-Tutorial>

BC Cancer Projects

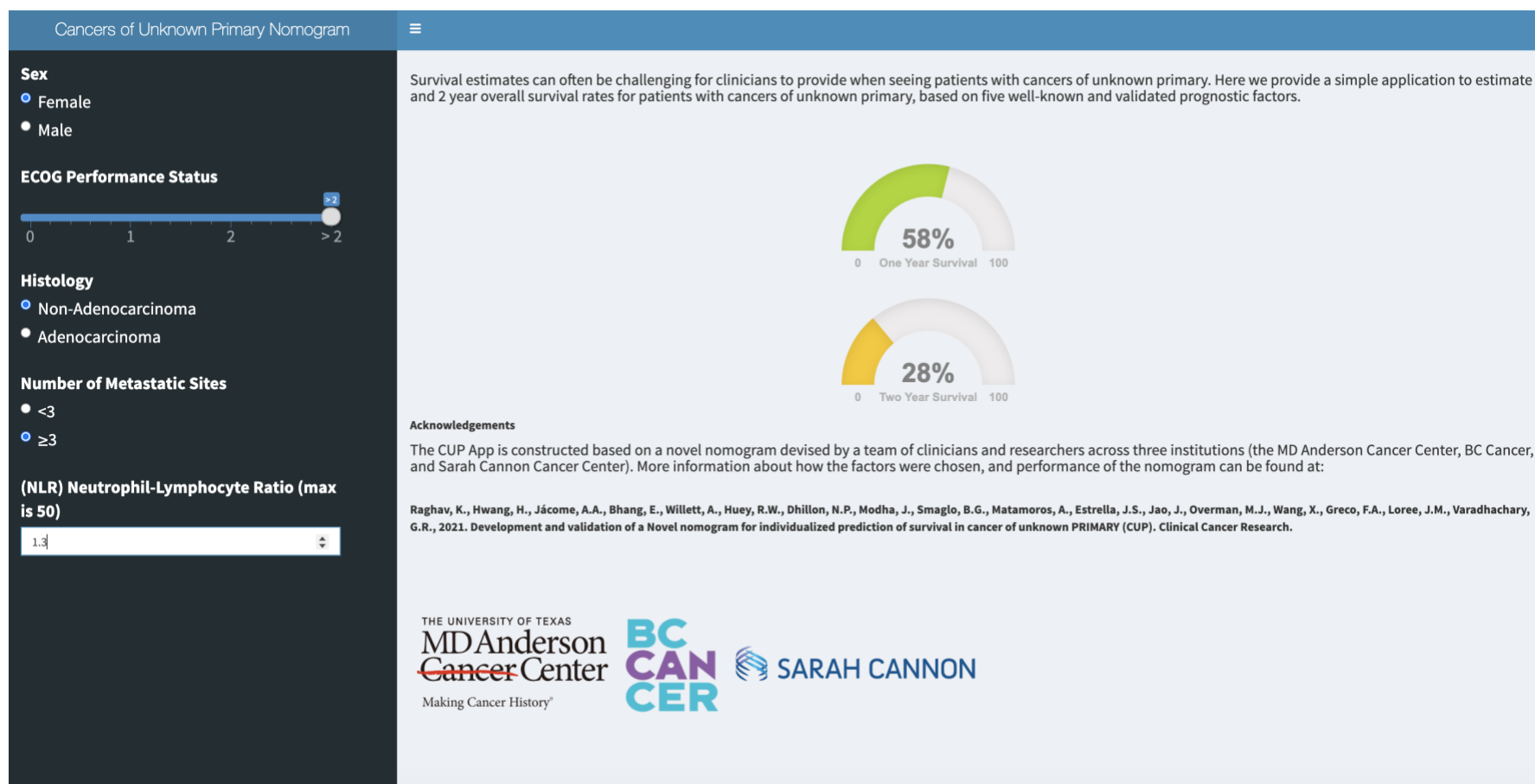
BC Cancer Agency

Period of Involvement: January 2019 – September 2019, July 2020 – September 2020

Cancer of Unknown Primary Nomogram

An interactive dashboard I built using the [Shiny](#) package in R, in the summer of 2020.

The tool was devised to allow real-time use in the clinic for assisting oncologists with assessing 1 and 2 year survival rate from cancers of unknown primary.



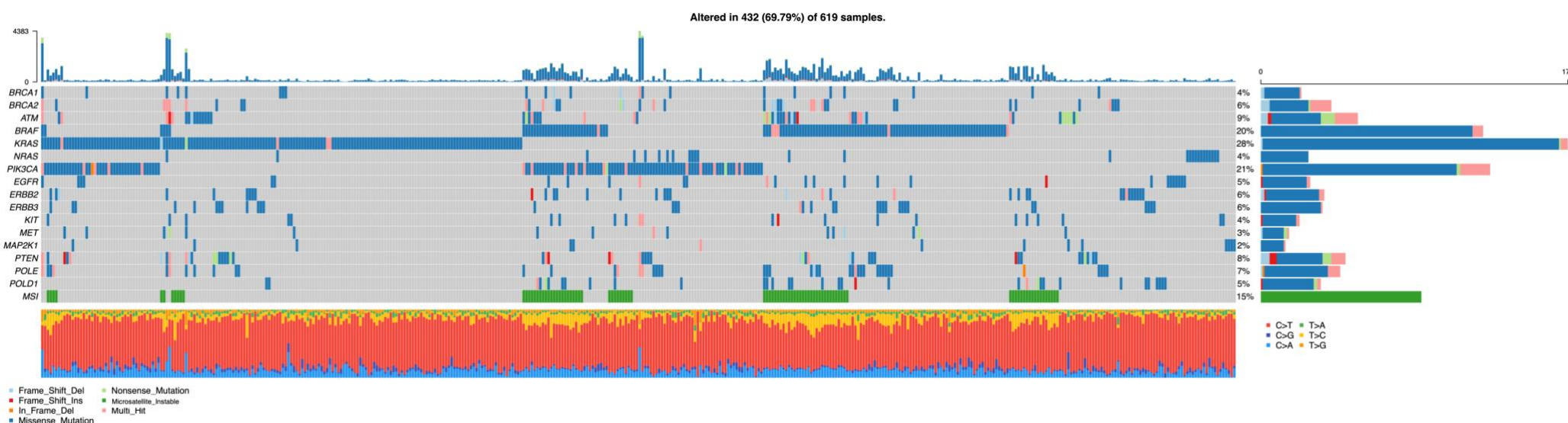
GitHub repository with the code can be found here: https://github.com/justinjao/CUP_Nomogram_App

Publication about how the nomogram was devised can be found at: <http://doi.org/10.1158/1078-0432.CCR-20-4117>

BRCA Analysis Visualizations

A project I worked on during my time as a clinical research assistant in 2018. I was looking at gene expression profiles of colorectal cancer patients with a certain type (BRCA1/2) of mutation. The project is currently on hiatus, but I still managed to explore a new tool for visualizing gene information, and made some neat graphics which I wanted to share!

An oncoplot, which shows the mutation profile of our cohort of patients, with annotations for the types of mutation, and what the mutation was, and how the gene was mutated.

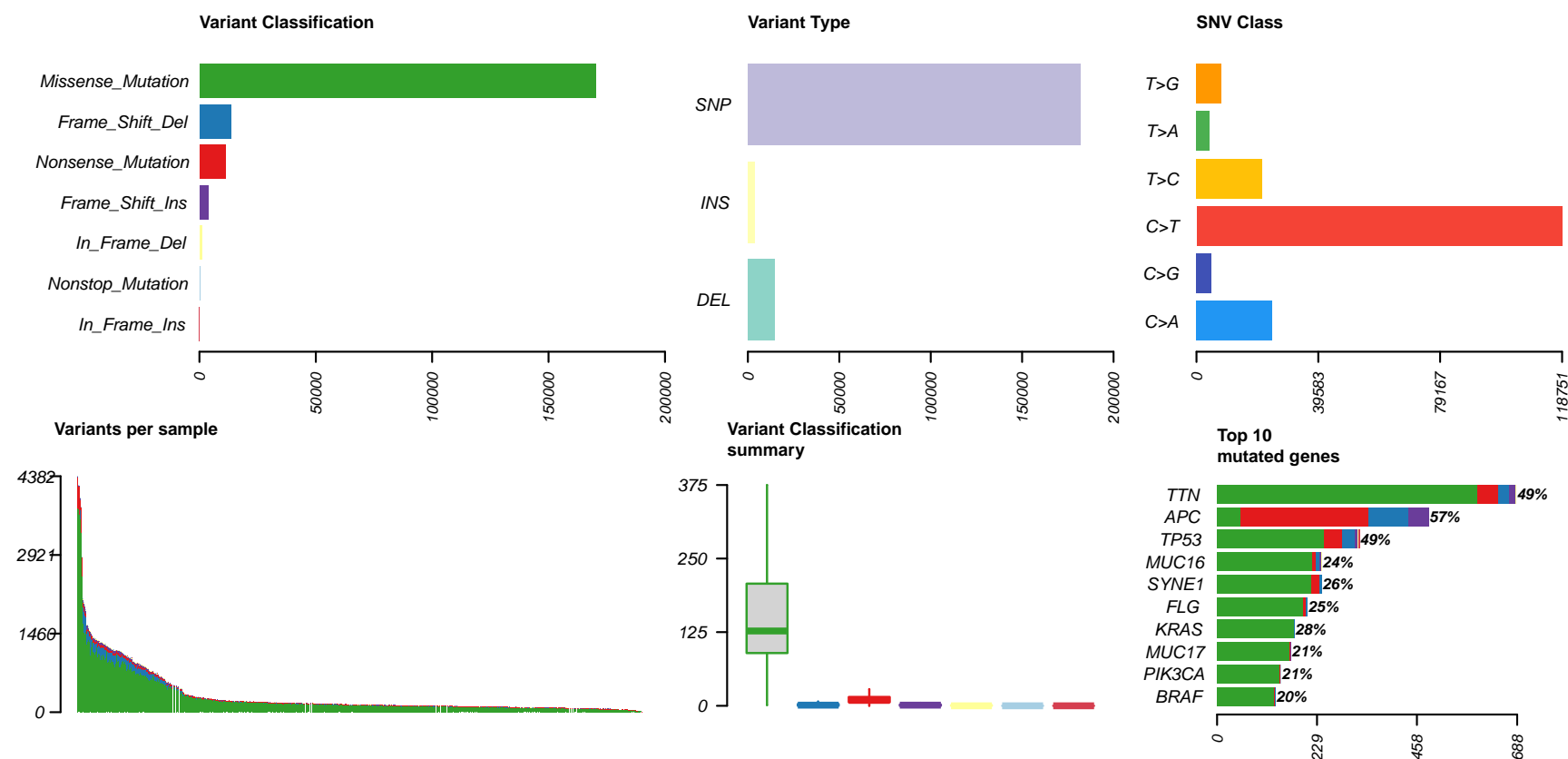


Made using the [maftools](#) package in R.

Side note: I know this figure is busy and hard to interpret. If I were to actually prep it for usage, I would certainly make it more concise and easier to draw insights from.

No links available for this project.

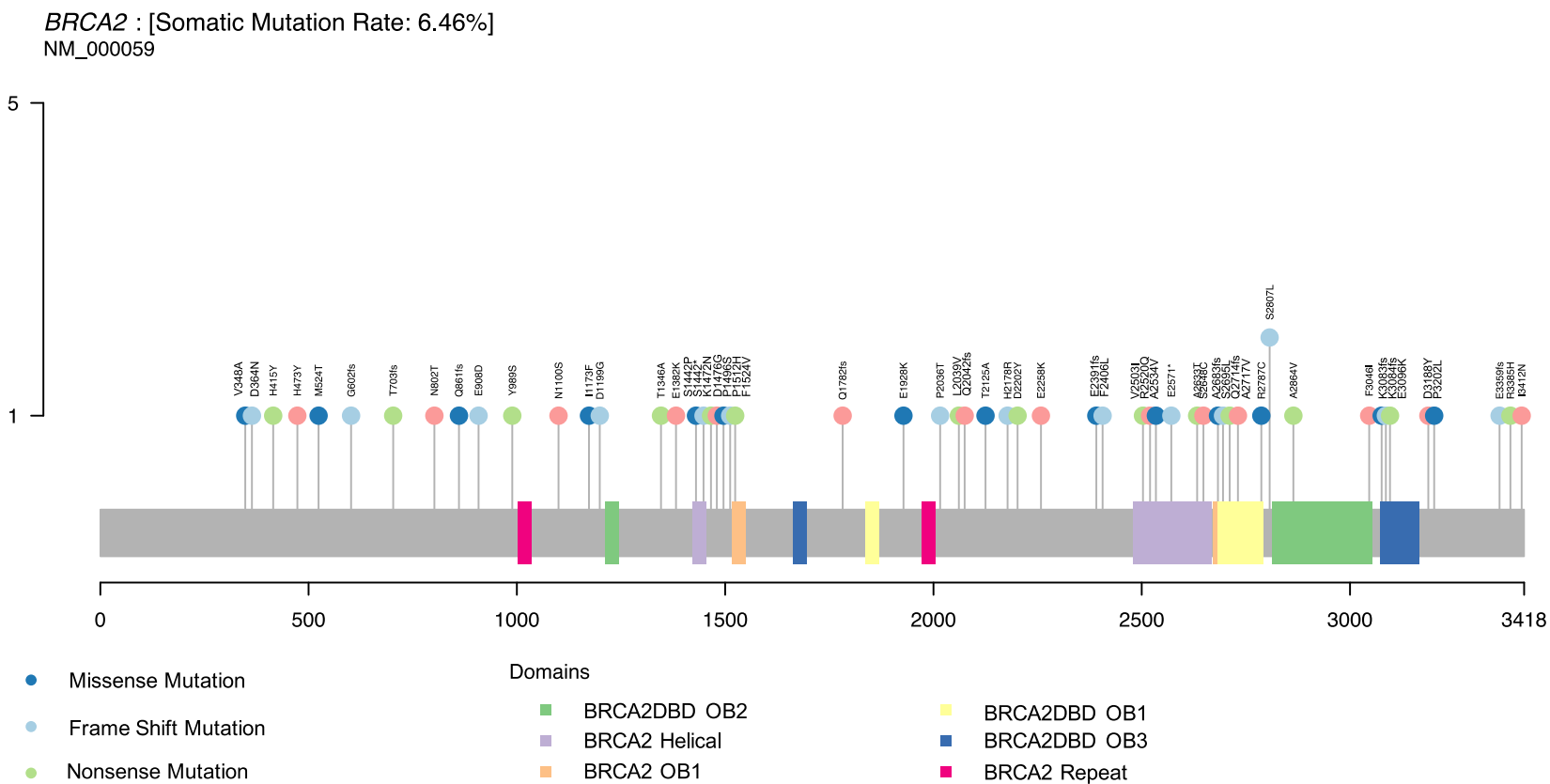
Some further summary visualizations of the data, exploring the mutation profiles of the patients.



Made with the [maftools](#) package in R.

No links available for this project.

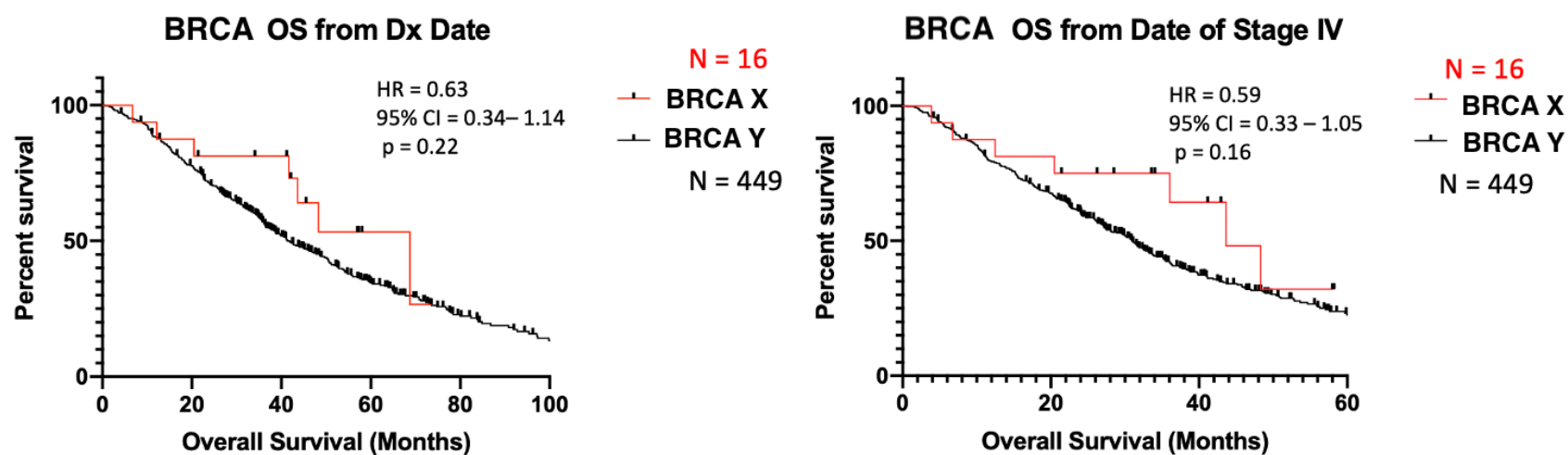
A lollipop plot (i.e. lollipop) plot, representing the mutations in the cohort and where they're located on the gene itself.



Made using the [maftools](#) package in R.

No links available for this project.

A survival plot (Kaplan Meier curve), visualizing the overall survival of the cohort based on their mutation status.



Made using [Prism GraphPad](#) software.

No links available for this project.

Genome Institute of Singapore Projects

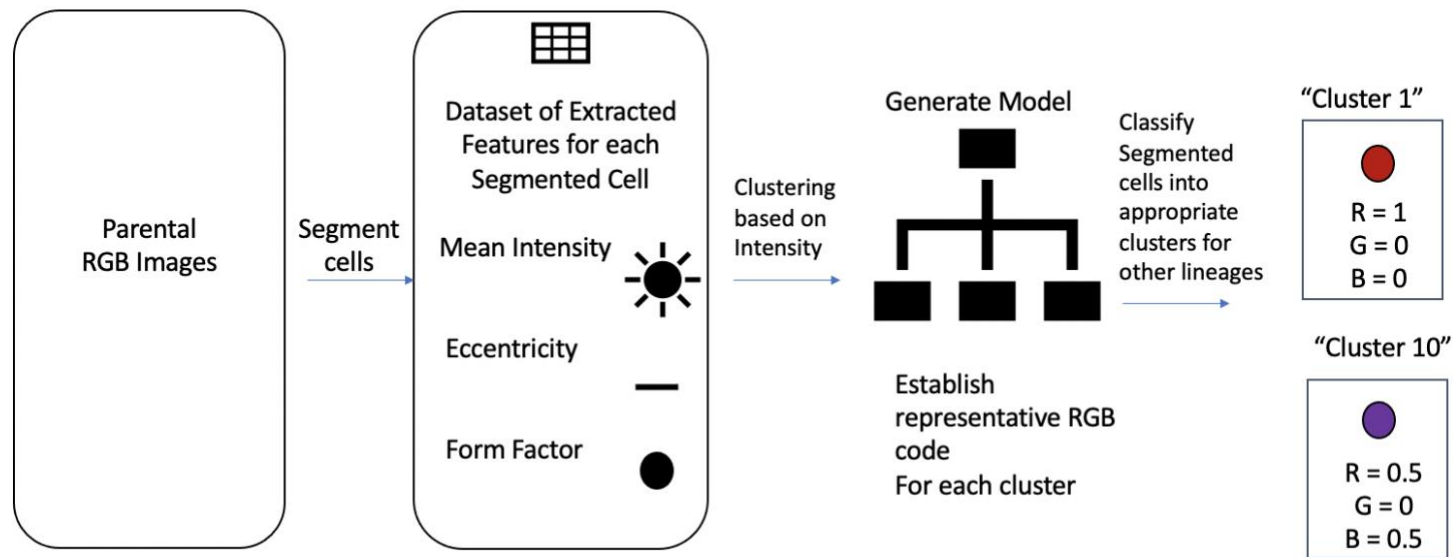
Genome Institute of Singapore (Centre for High Throughput Phenomics)

Period of Involvement: June 2018 – December 2018

RGB Analysis Project

The first real analysis project I worked on. We were tracking changes in population dynamics of colorectal cancer cells as they evolved, and I used image data (fluorescence intensity in red, green and blue light channels) to accomplish this. I first performed image segmentation to extract image features. Then, I used k-means clustering to group the cells based on their colour profile, and random forest classification to match the colour profile of cancer cells as they evolved, to determine how the cell populations changed.

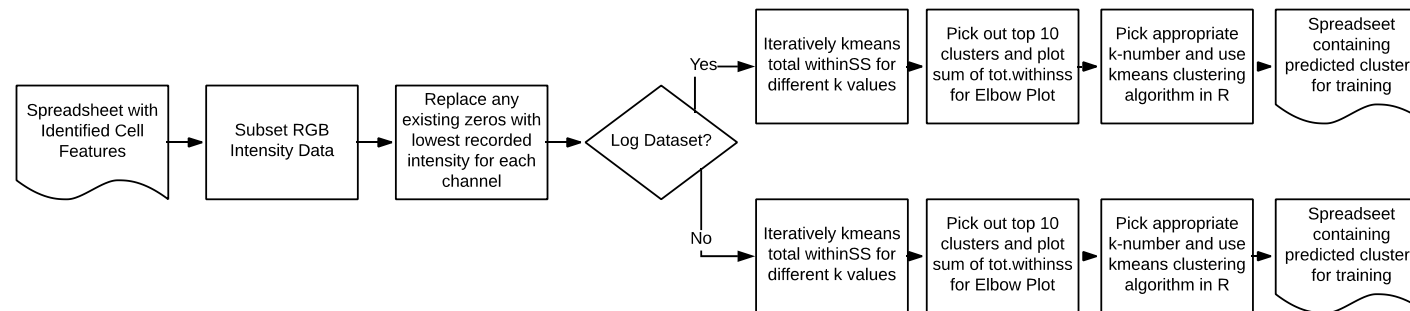
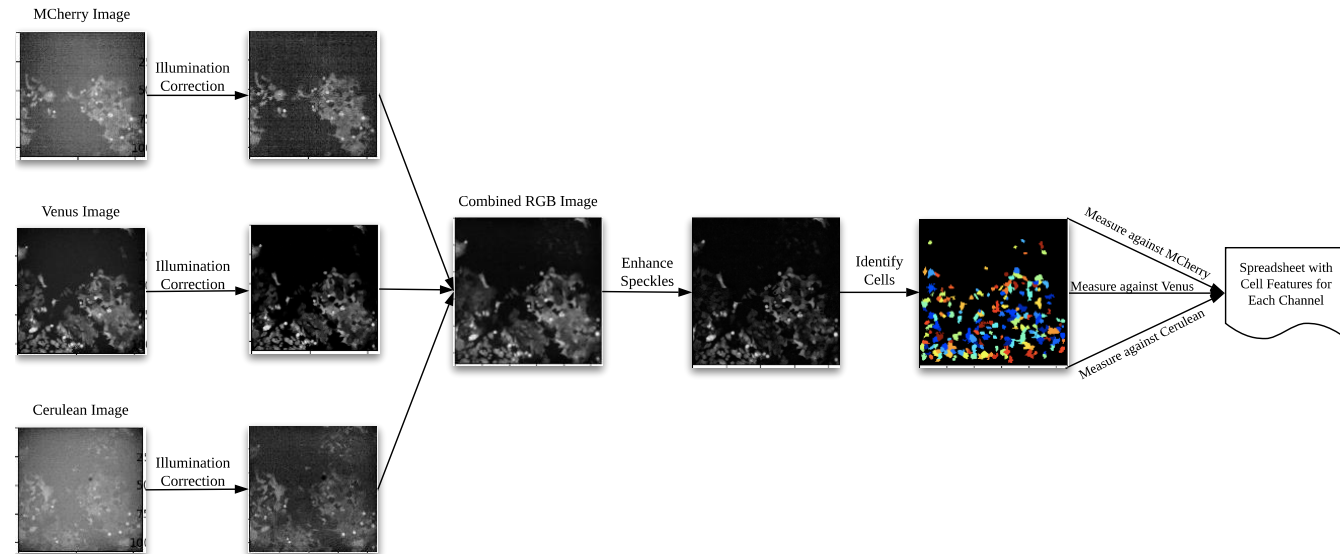
Workflow



No links available for this project.

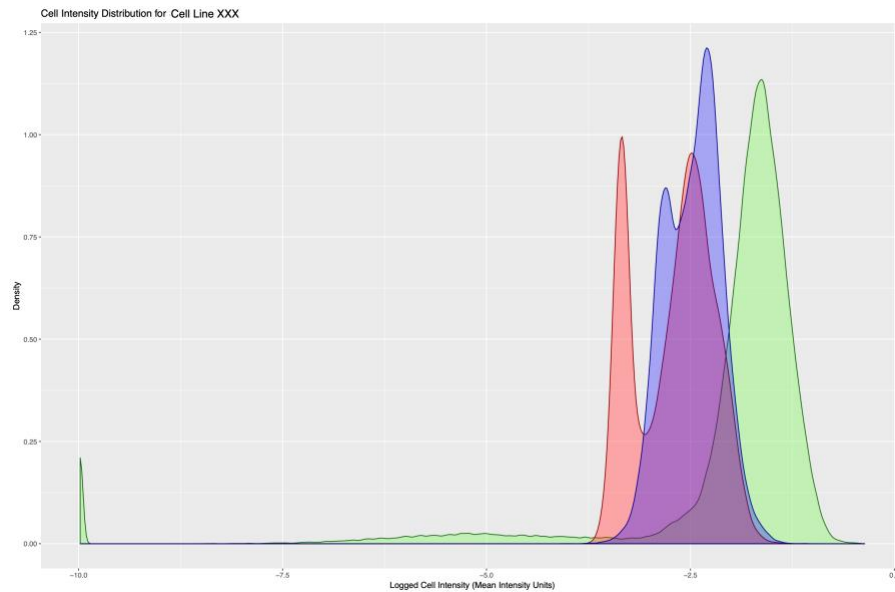
Image Segmentation Pipeline Workflow

Image segmentation was performed using [CellProfiler](#) software.



Data pre-processing steps

No links available for this project.

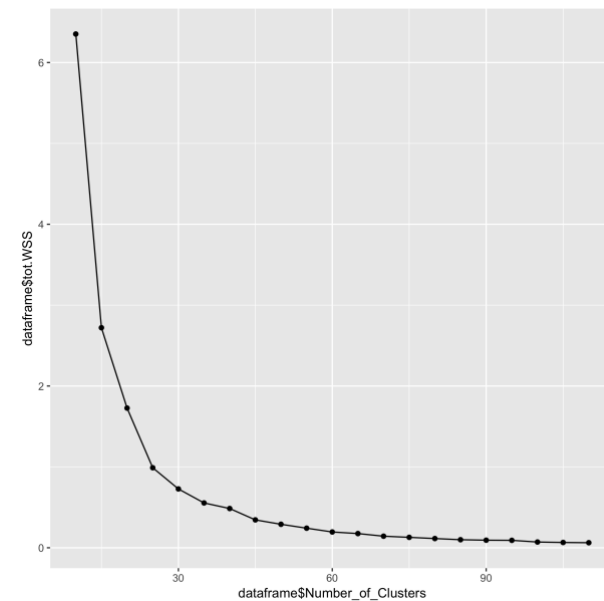


Plots to visualize initial colour distributions of cancer cells.

Made using [ggplot2](#) package in R.

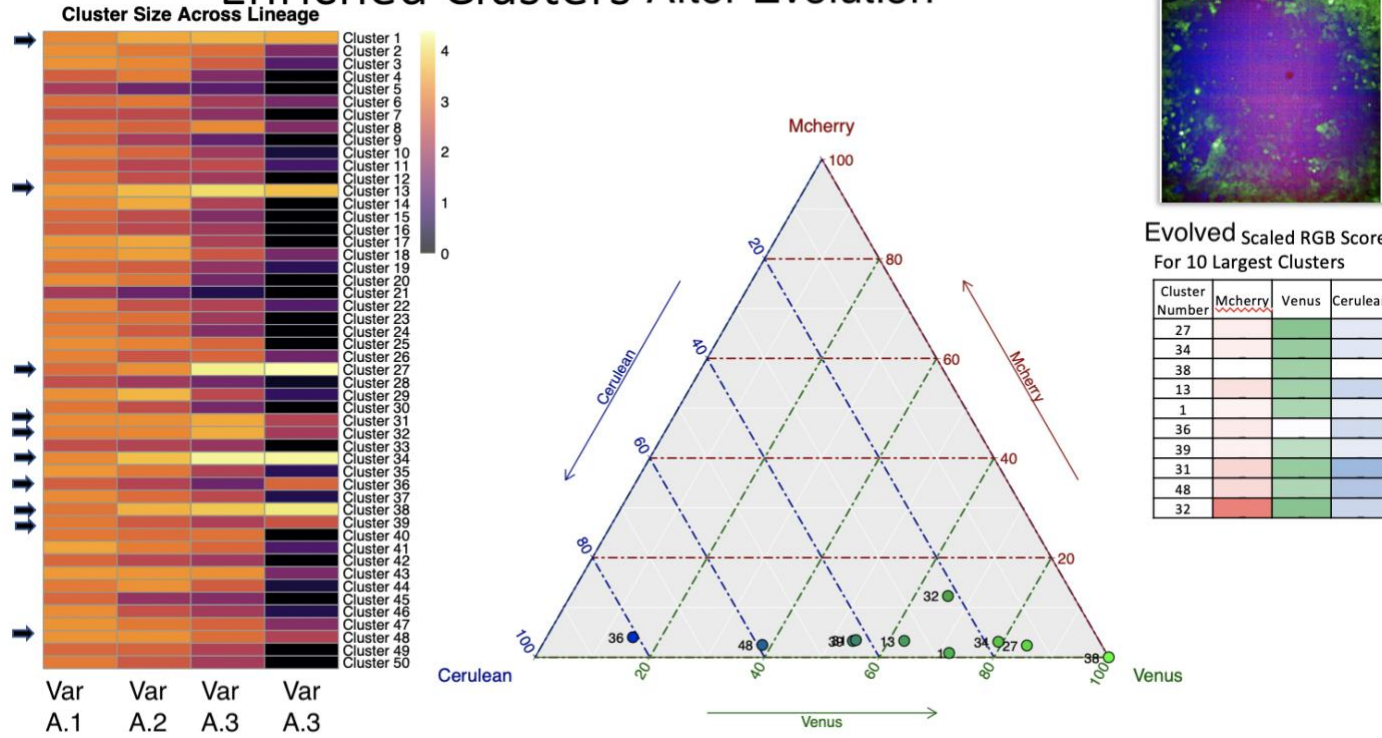
Elbow-plot used to determine appropriate number of clusters to use for k-means clustering.

Made using [ggplot2](#) package in R.

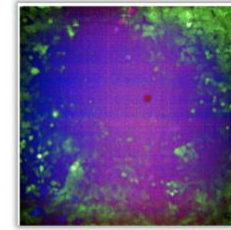


No links available for this project.

Cell Line X Enriched Clusters After Evolution



← Sample image



← table showing scaled colour distribution

Heatmap showing changes in clusters of cell populations as cancer cells evolve (left to right).

Made using [pheatmap](#) package in R.

A ternary plot used to visualize the colour distribution of the clusters generated.

Made using [ggtern](#) package in R.

No links available for this project.