Gakit Richard Yu - Data Viz - Project 3
NetId: gry861

Complement C4

Protein / Gene Copy Number Visualization

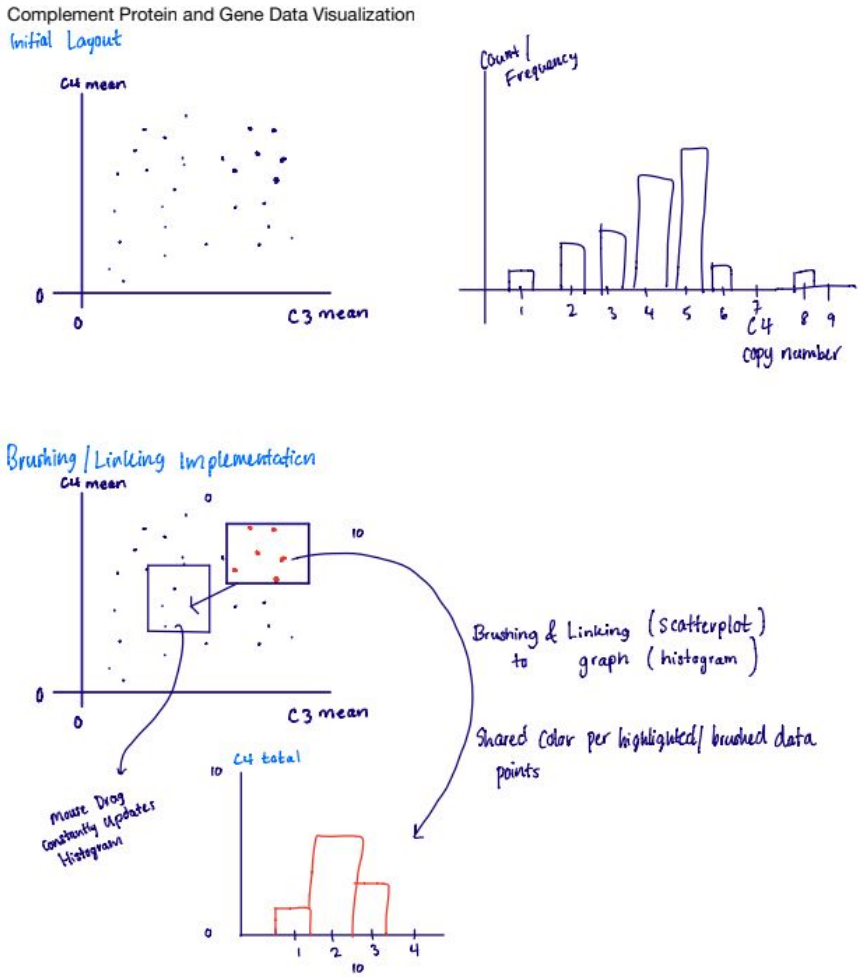
Description of the Data Domain

For Project Three I decided to work with an old data set provided to me by my father (Dr. Chack-Yung Yu of Nationwide Childrens' Hospital) that pertains to his research of the autoimmune disease: Systemic Lupus Erythematosus (SLE), shorthanded named Lupus. In his research he conducted blood samples of over 500 patients of the disease and investigated a correlative relationship with the disease and a genetic marker, specifically the copy number variation number of the C4 gene, which is involved in the autoimmune complement system.

This dataset consists of the results of those 500 blood samples, which were analyzed for both protein levels of the C4 protein, as well as the closely related C3 protein, and their DNA was sequenced to reveal the copy number variations of the C4 gene present. It is believed that abnormally low concentrations of the C4 and C3 protein levels may be associated with the disease's onset.

In this assignment I aimed for an interactive and easy way for a researcher to look at the relationship between C4 and C3 protein levels and associate them with the C4 total copy number, which is the number of copies a person has of the C4 gene (Copy number variation). Within this I aimed to generate a brushing and linking application using CSS, JS, and d3 to suit this goal.

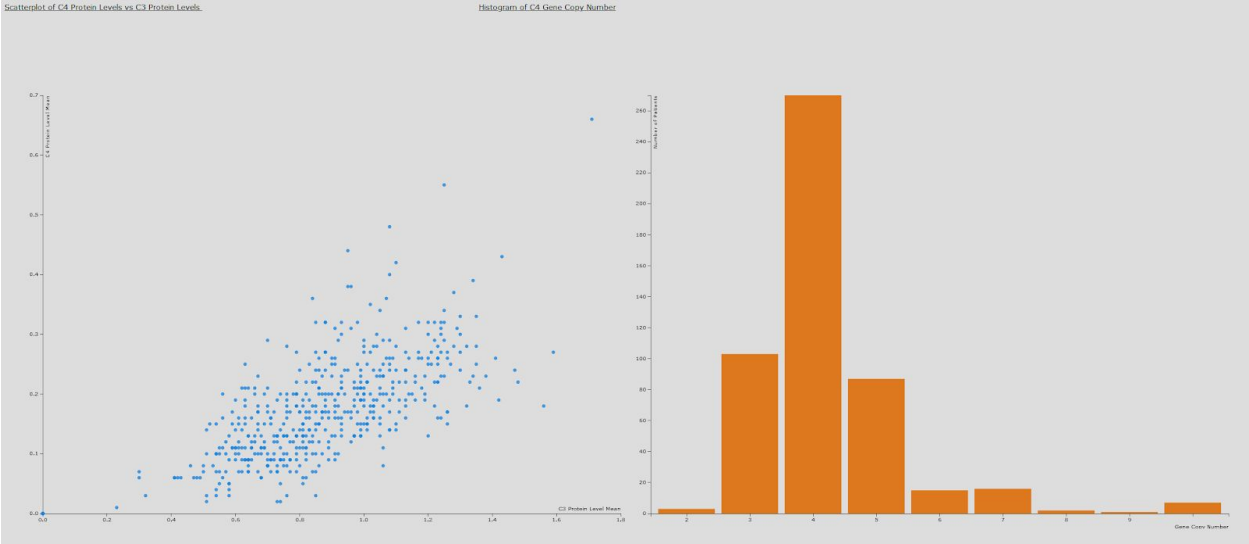
Storyboard



The above shows the storyboard for this application. I planned to have an overall scatterplot that plotted two quantitative variables, the protein concentration of C4 and C3. This graph would support brushing to select varying data points to be input to a linked histogram that enables easy visualization of copy number variations.

With this, a user can utilize d3's brush component to click and drag certain regions of the scatterplot and view an updated version of the copy number variation histogram on the fly.

Final Application Description



The final application essentially captures the goal of the storyboard, enabling users to use a brushing and linking function to constantly update a histogram of copy number variations.

The first visualization is a scatter plot with two quantitative variables corresponding to the C4 and C3 protein levels using a relative [0,2] measurement. This graph largely shows a linear relationship between the two variables. It enables the brushing and linking concept, allowing a user to select a region of data with the brush tool via click/drag to update the variables within the secondary visualization, the bar chart or histogram. The brush tool was assigned to color selected datapoints with the same coloration as the fill of the histogram to link the two together accordingly.

As previously mentioned, the bar chart is constantly updated using the brushing and linking concept and adjusts based on the brushed region of the scatterplot. While developing, I noticed that examining the frequency per copy number variation was difficult due to the y-axis scale being in the range of hundreds, so a hover/tab indicator on mouse hover was implemented to show the frequency numerically.

Alterations within the Final App and Storyboard

The final application largely followed the storyboard, with several adjustments to better suit the data such as the hover functionality on the histogram to view the frequency amount.

Development Process

A large portion of the development process was spent determining the dataset and filtering irrelevant fields.

This is the full variable list of the dataset:

```
Variable     Description
-------------------------------
patient No.        Identifier/Key
C4 mean
      Quantatitive variable attributing to the mean concentration level of the C4
            protein
C4 mean <10
      Boolean/Nominal variable used for grouping on whether C4 mean <= 0.10.
C4 range
      Quantatitive variable attributing to the variance/range present among C4
            protein levels.
C3 mean
      Quantatitive variable attributing to the mean concentration level of the C3
            protein.
C3 range
      Quantatitive variable attributing to the variance/range present among C4
            protein levels.
C3 mean <80
      Boolean/Nominal  variable used for grouping on whether C3 mean <= 0.80.
Low C3, C4<10; MEAN
      Boolean/Nominal  variable used for grouping on whether C4 mean <= 0.10 AND C3
            mean <= 0.80.
LOW C4, NOR C3
      Boolean/Nominal  variable used for grouping on whether C4 mean <= 0.10 but C3
            mean > 0.80.
C4T gene
      Quantatitive variable attributing to total copy number of C4 (T = A+B = S+L)
C4A gene
      Quantatitive variable attributing to copy number of C4A
C4B gene
      Quantatitive variable attributing to copy number of C4B
C4L gene
      Quantatitive variable attributing to copy number of C4S    (Short)
C4S gene
      Quantatitive variable attributing to copy number of C4L (Long)
```

For instance, the original dataset consisted of columns that had scientific significance but not data analysis importance such as a boolean value on whether C4 protein concentration was below a value of 0.10. The combination of this boolean value among others perhaps has some meaning to access disease onset or severity, but are variables that are more based on practical approach rather than inform numerical relationships.

In addition, various patients had incomplete data, perhaps missing C4 copy number or one of C3/C4 protein concentration levels that needed to be filtered.

Once the data was cleaned, implementation of the visualization component took the majority of the remaining time, as storyboarding and planning were not very time intensive. Over the course of a week, identification of a dataset, data wrangling/cleaning took approximately 4 hours; planning the visualization was approximately one hour, and implementing the final visualization took about 6-7 hours for a total of approximately 12 hours working on this assignment. Having not been experienced with d3 in the past, this development time could have likely been shortened given past experience as it implements simply the drawing of two graphs, one of which is linked to the other mono-directionally.