Visualization of Utah population based genetic and clinical features in colorectal cancer

Process Book

By HyoJoon Park & Seyoun Byun

Overview and Motivation

Colorectal cancer is the third most common cancer diagnosed yearly, in both men and women, in the United States and the second leading cause of cancer-related deaths when men and women are combined. Colorectal cancer treatment is beneficial when reducing the number of incidences when removing colon polyps. Also, early diagnosis and survival are better with detecting the polyps in the colon.

Interestingly, past studies reported that the asserted pedigree and twin studies indicate that 20-30% of colon cancer cases arise in the inherited susceptibility. Also, 3-5% of colon cancer occurs in inherited syndrome. However, the risk of adenomatous polyps in men and women or different ages concerning the family history of colon cancer cases is not studied well.

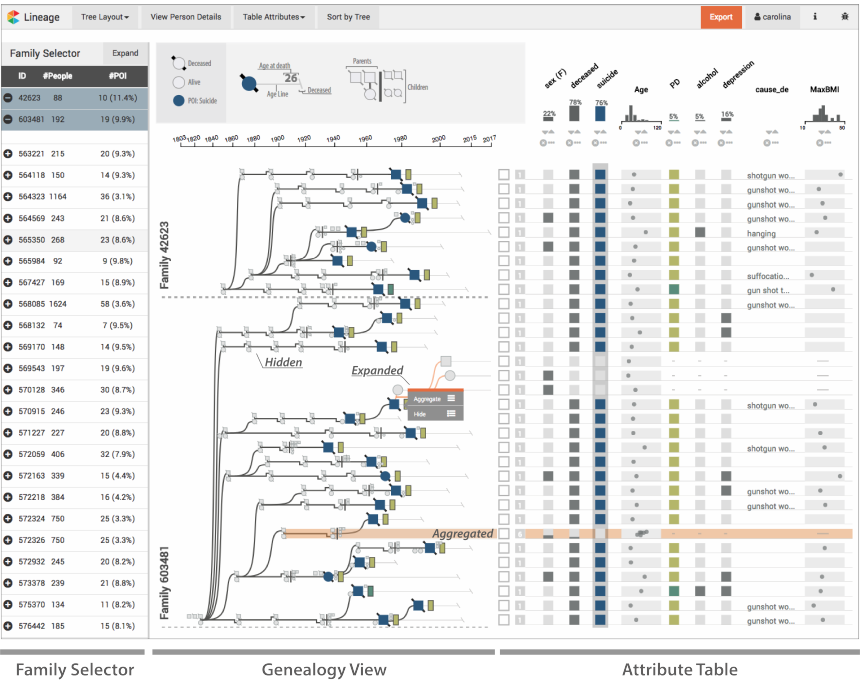
Association studies reported genetic variants and exposure risk factors, including BMI, smoking, exercise, alcohol consumption, NASID, and hormone menopause.

This project's main motivation will explore the seven families with several patients (n=198)'s genetic information associated with the polyps. We will try to understand the genetic information to compared familial clinical multivariate details to visualize all exposure risks. Thus, we know that colon cancer is associated with a genetic inheritance based on multi factors.

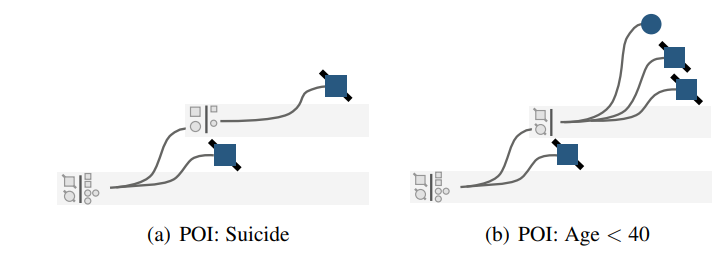
Lastly, we will try to understand the polyp size with the relationship with the family's inheritance. Therefore, in this visualization, we evaluated the polyp size with the location and the clinical factors in huge extended families with a strong family history representing the familial high-risk colorectal cancer classification.

Related Work

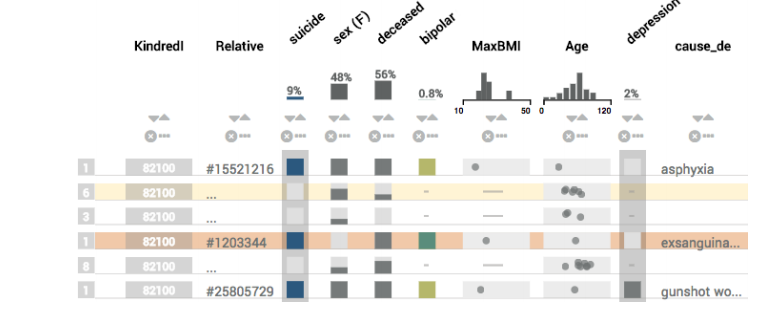
1. The most inspiration on the technical design for the multivariate risk factors came from the paper called " Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs by Nobre et al."

This paper describes the combination of hereditary and all environmental risk factors. The paper used the three layouts (i.e., Family selector, Genealogy view, and Attribute table). These layouts gave an easy interaction of the association of the family selector through the attributes. We tried to follow a similar format of our attributes of risk factors.

However, this visualization mainly focused on genealogies and survival information (ex. death, suicide, alive, etc.).



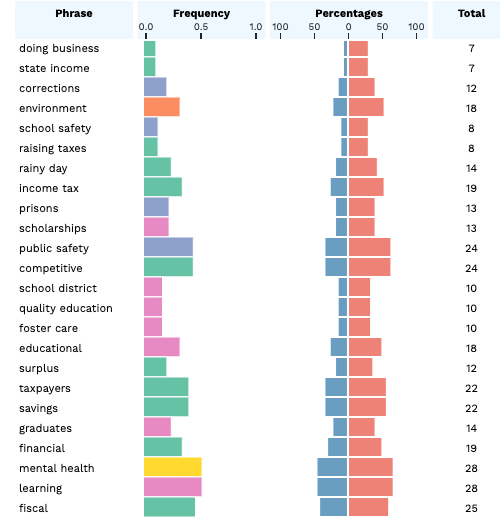
On the table, it was well organized with the risk factors, and the highlighting all column seems very useful.



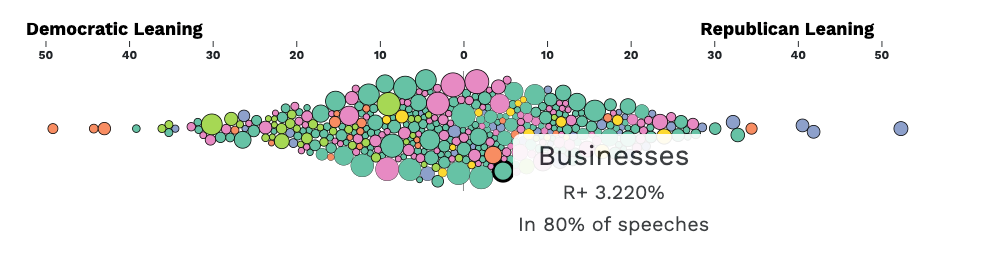
Nevertheless, all this table and family selector idea was inspired by our data visualization.

1. Another inspiration source was from homework six from the class. For our visualization, we wanted to compare the family ID (= KinderID) and the number of the risk factors in the ascending/descending method of sorting. Thus, we can interact with the genetic information that families have the highest or low value on the risk factors. Also, for the polyp’s information table, we will have the sorting method as well.

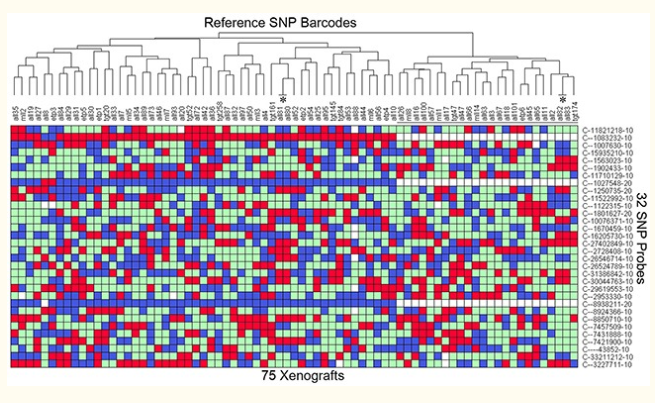
Visualizing the axis number seems a great idea because the reader may know the range of each feature based on each column's axis. We will do all the clinical risk factors for the box with the axis.



Hoovering all the information would be ideal as homework six too. Since if we draw the data as a box in each column, it will be hard to know the exact value. Therefore, it will be a great mouse over the hoovering method.



1. The first optional heatmap idea came from the paper "A single nucleotide polymorphism genotyping platform for the authentication of patient derived xenografts by El-Hoss et al.." Since this one had been a draw with the R package tool, we will consider adding it because this heatmap will give individual information of the SNP (genetic).



Questions

We build a visualization that may utilize interactive plots to help viewers explore relations between polyps (i.e., size, location) and other biological and family-related risk factors. The plots allow users to sort and filter data such that the relations can be viewed from different perspectives. That interaction would enable users to conclude how various factors affect the incidence of colon cancer.

At first, we were considering each family's average number of each feature. However, we realized that the average boxplot might have problems with the complex data - 1) missing values- We have much missing value due to the nature of the clinical information. 2) We are not able to see the individual level value.

Therefore, we decided that table visualization would best show the polyps data and clinical risk factors.

Other questions considered and that may be still floating around are:

* + "What are the best association with both clinical risk factors and polyp’s information (size and location)"
  + "What is the survival associated with the polyps and clinical risk factors."

The last two questions would help understand the risk factors and polyp’s information on each level even though the analysis of both risk factors in one slide may confuse the difference between clinical data and the polyp’s information. Survival association with the family sectors may have a conclusion. However, due to the massive data missing, it would be better with the basic research of exploring the multivariate first that would be the best for the users.

Data

The Utah population database (UPDB) was used to identify the seven families. Colorectal cancer cases in the families were contacted by the Utah Cancer Registry through mail requesting them, or their next of kin, permission to be contacted by the study. In total, there are seven large kindreds with multiple colorectal cancer cases (n=198) included in this study. The medical records were obtained on colorectal cancer cases. Published guidelines evaluated adenomatous and hamartomata’s polyposis syndromes.

Due to the IRB, the data is not appropriated to provided.

**KinderID:** Family Id

**Sample ID:** Sample Id

**Sex:** Male and Female

**BMI:** Body Mass Index is a complex phenotype that may interact with genetic variants to influence colorectal cancer risk. (PMID:32324875)

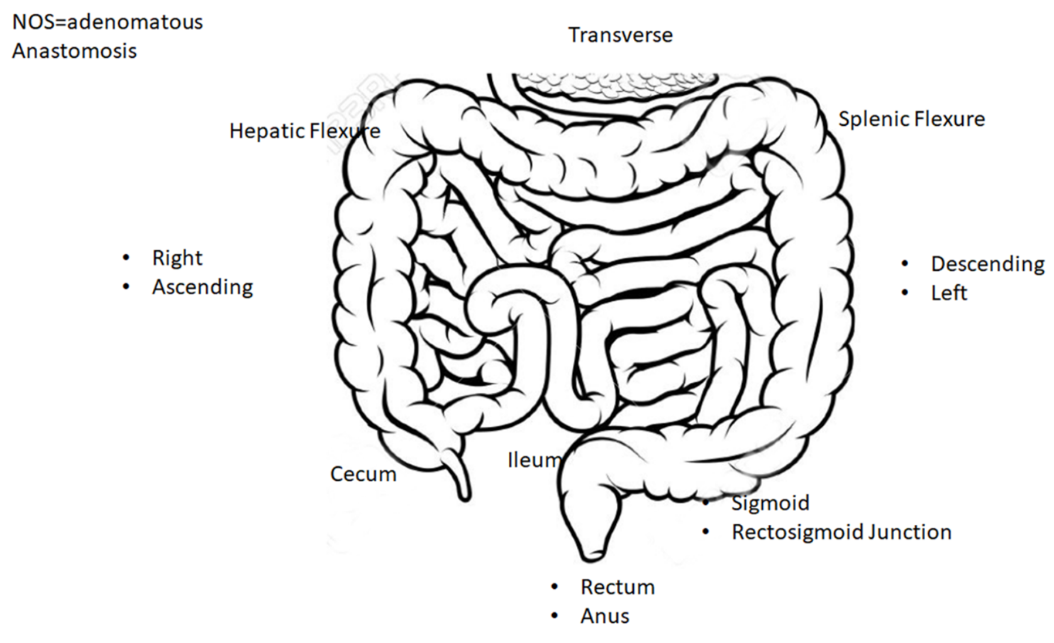
**Smoking:** Cigarette smoking is an established risk factor for colorectal cancer (PMID: 20587792, PMID:19088354)

**Exercise:** Exercise will decrease the mortality and risk of recurrence for colorectal cancer (PMID:31139306)

**NASIDs:** Non-steroidal anti-inflammatory drugs- Men who used aspirin were also more likely to use the NASIDs, and the Aspirin/NSAIDs would prevent colorectal cancer and cardiovascular disease (PMID:26940135)

**HRT:** Hormone Replacement Therapy- Epidemiologic studies evaluating hormone therapy use and colorectal cancer risk by the status of cell-cycle regulators are lacking (PMID: 22511578)

**Polyps site:** There are 15 types of the site. However, we will be divided into Cecum, Ascending, Transverse, Descending, Sigmoid, and NOS. (Due to same position but different name). Total 9 types were annotated in the picture.



**Polyps type:** There are 19 types of polyps.

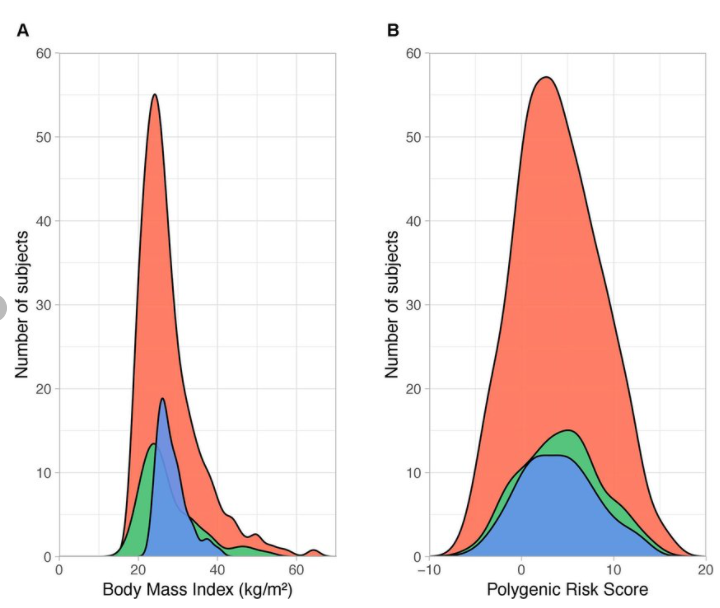
We will combine all the three different unknowns (i.e., No biopsy taken, Lost polyp) into one unknown. Otherwise, we will keep the same for the rest. Total, we reduced the type to 9.

**Polyp size:** There are seven different categories with size. We will use the shape to visualize the size.

Exploratory Data Analysis

What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

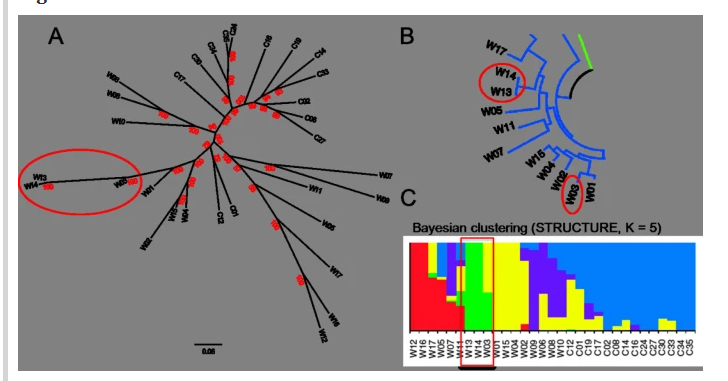
To initially look at our data, we tried to visualize genetic information to gain each family's insight. To see the below image:



First, we tried to look at how others do visualize graphs with the polygenic risk score. We got the insights how we put the x-axis and y-axis for our polygenic risk score data. We decided to interact with each family’s polygenic risk score; then, we can compare using seven different colors to resemble the families

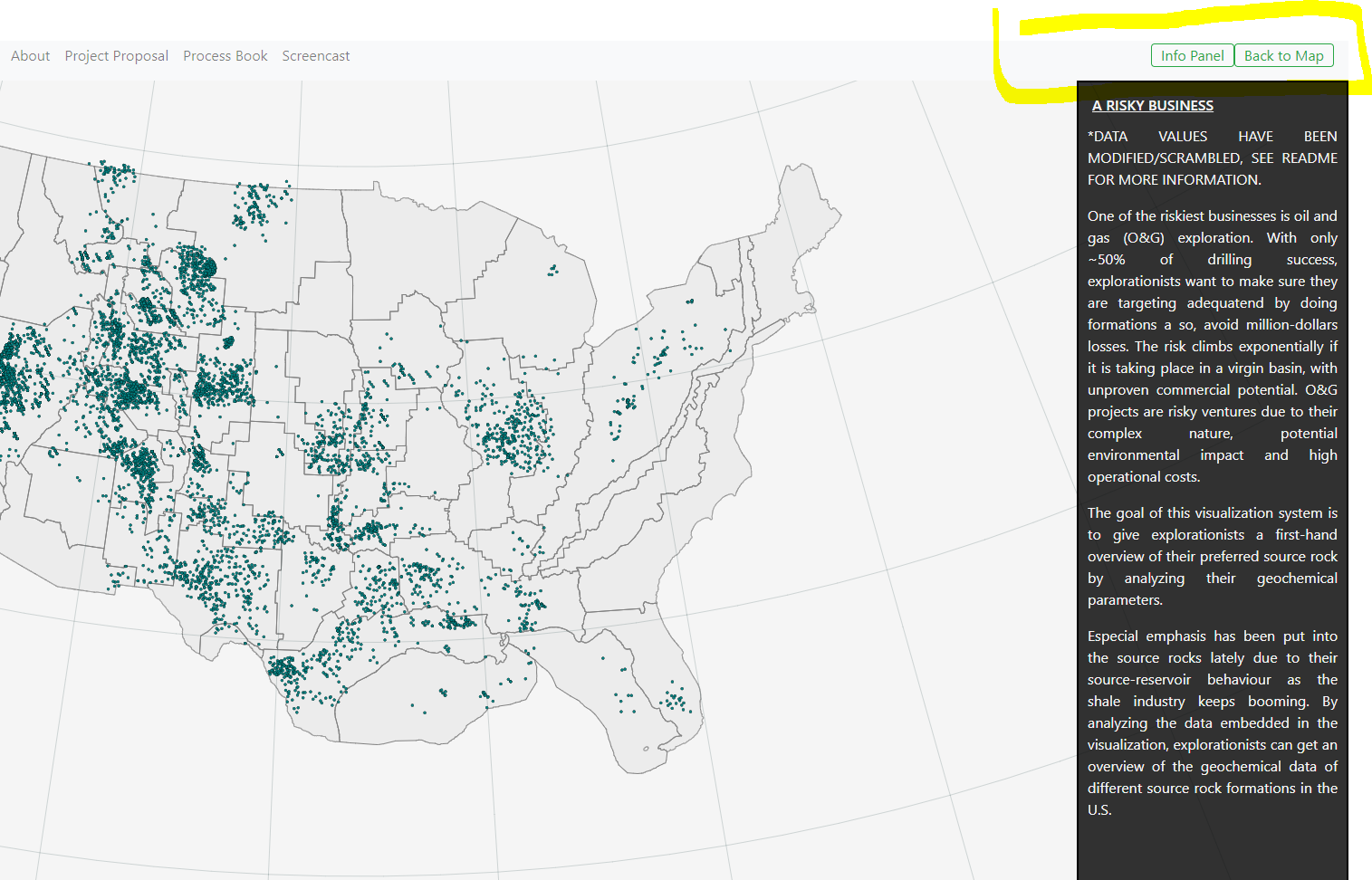
Design Evolution

We considered the different visualization of the SNP information-below image:

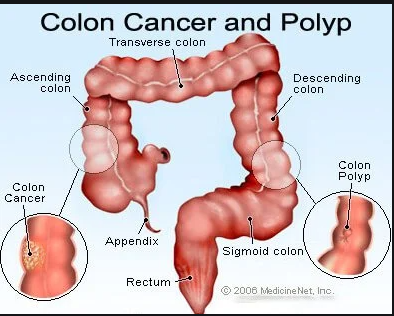


This is the tool paper "SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data by lee et al." We were considered the SNP phylogenetic tree of SNP. We did not decide to do this because this visualization needs more explanation because to understand this visualization, the reader needs more genetics. Heatmap would be more intuitive than this SNP tree.

We also tried to put our description in the right side of the first and second scree, however, due to big number of the heatmap and the table, we decided to the mouse hoovering the information. (The Idea below: Description panel)



We wanted to have information of the colon location. Understand the site information before we highlighted from the polyp site.

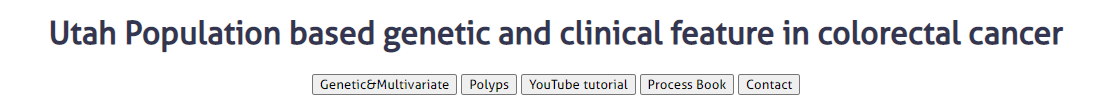
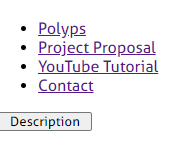


Implementation

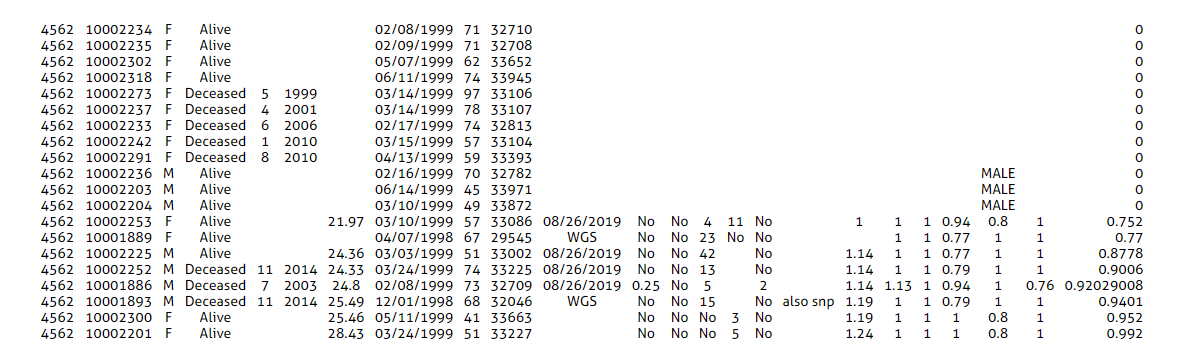
Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

**Entry Nov 3rd:** Implemented basic json files to load into JavaScript

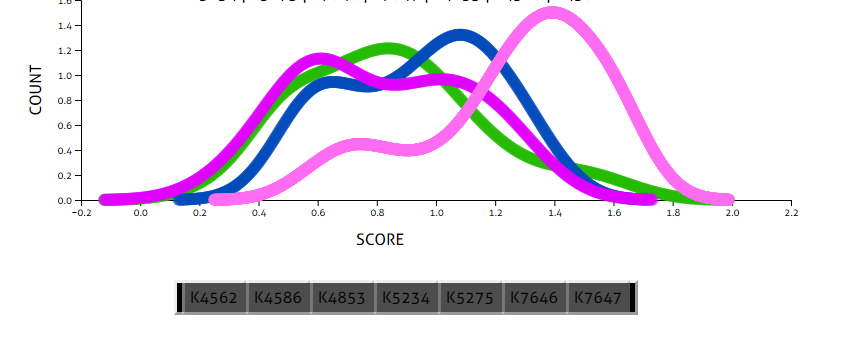
**Entry Nov 7th:** Setting up the Click button for the different screens and basic density plot for the genetic



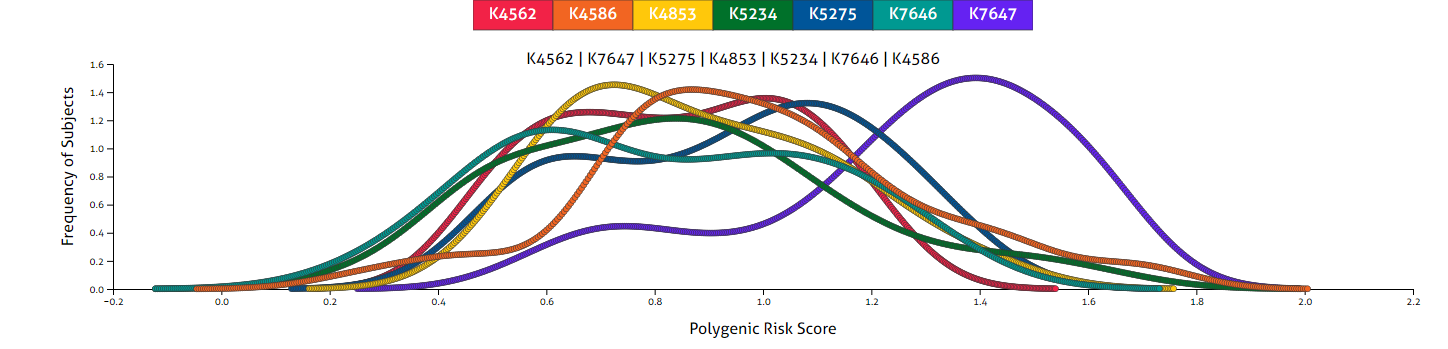
**Entry Nov 10th:** Trying to visualize the click icon again, table build up



**Entry Nov 14th**: We got a review of the proposal from TA and update the data source explanation. Working on the plot.

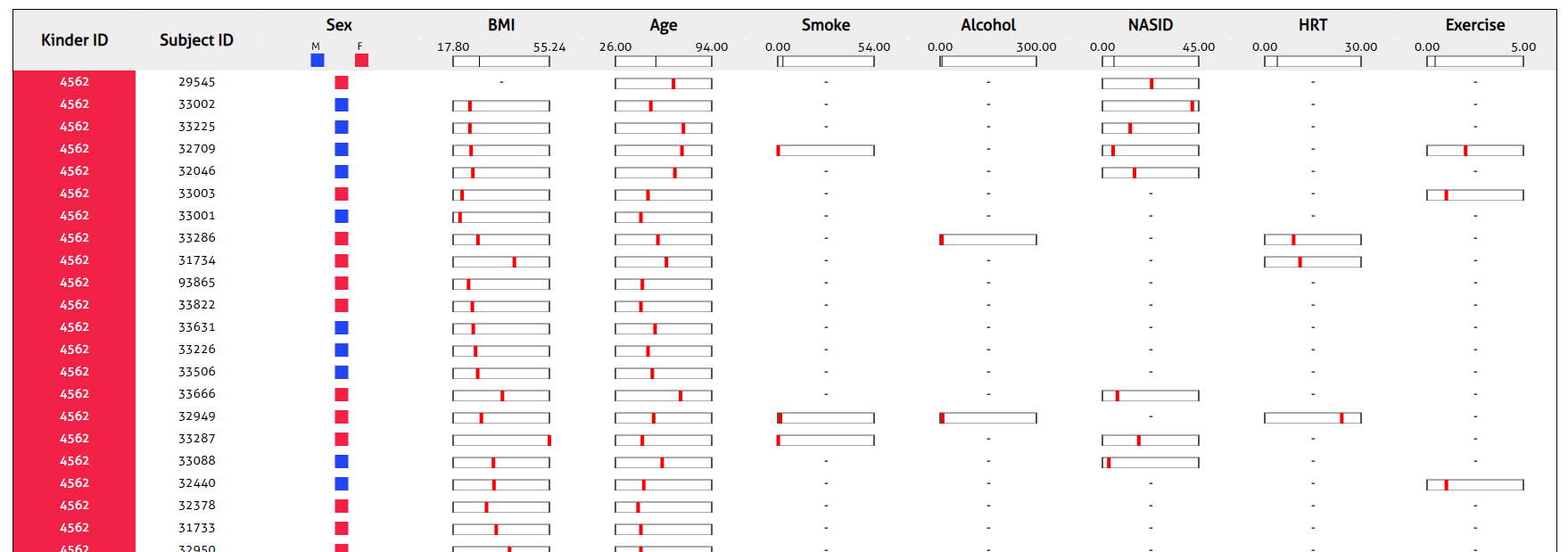


Nov 16th: Finished the density plot

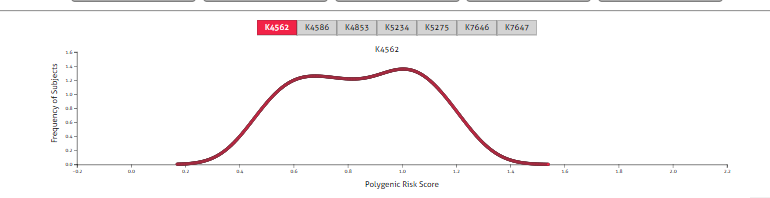


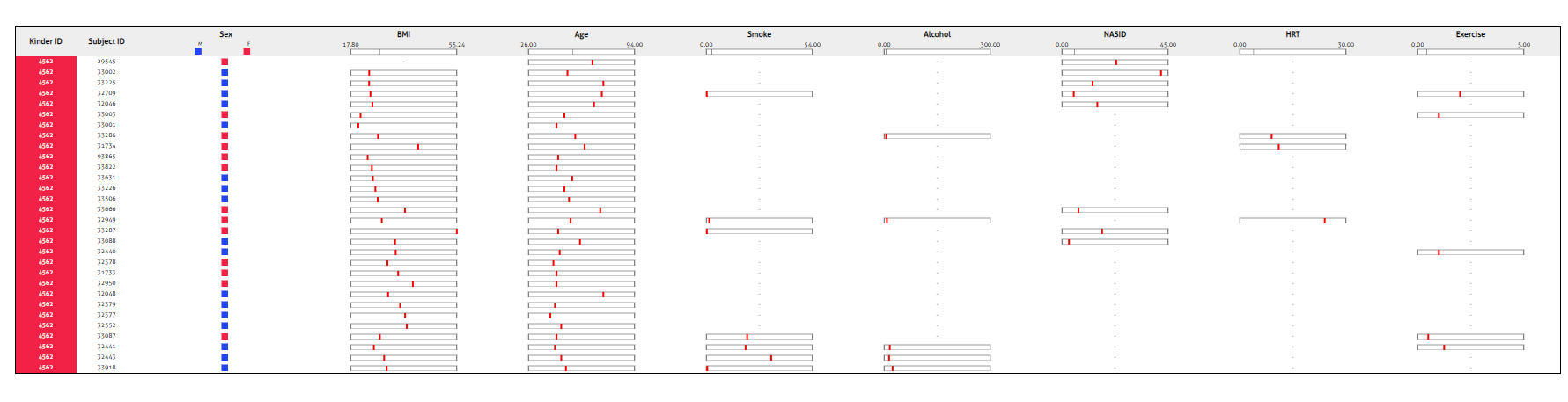
**Nov 17th:** Met with TA and got some advice for the heatmap and feedback from our project.

**Nov 19th:** Table for the 1st screen was added.

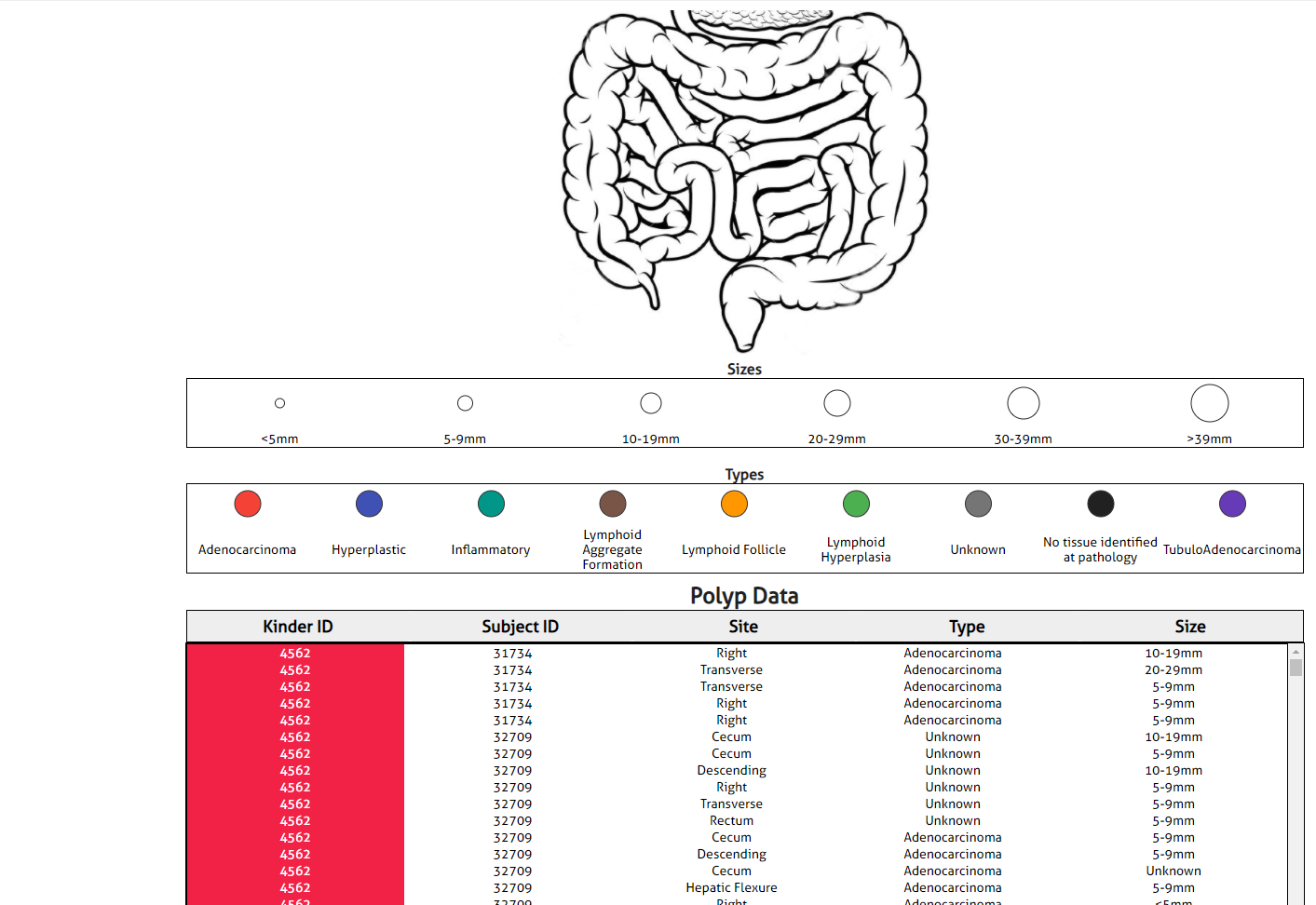


**Nov 22nd:** We made the interaction when we click the Kinder ID from the density plot, then it will show only same Kinder ID from the table



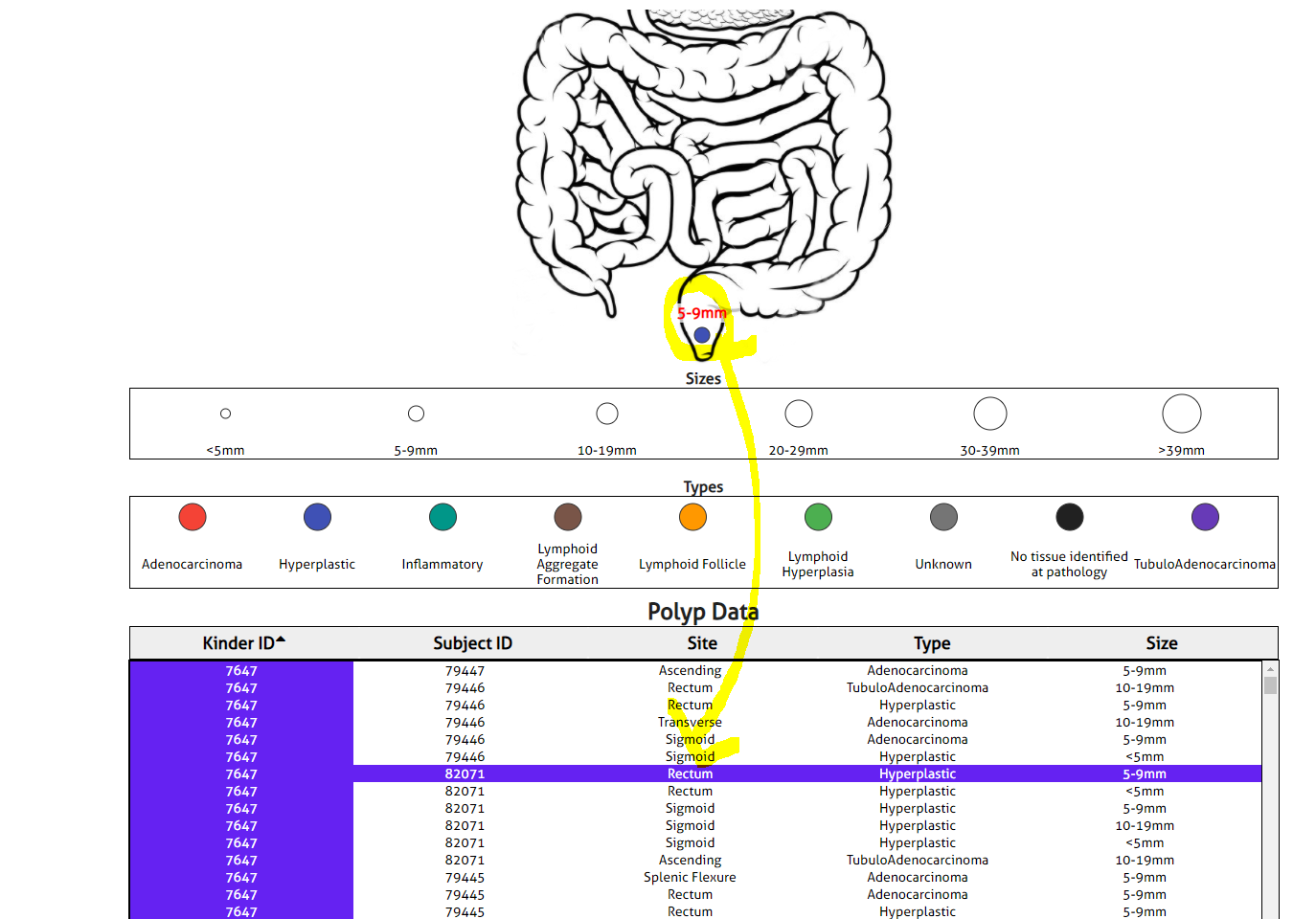


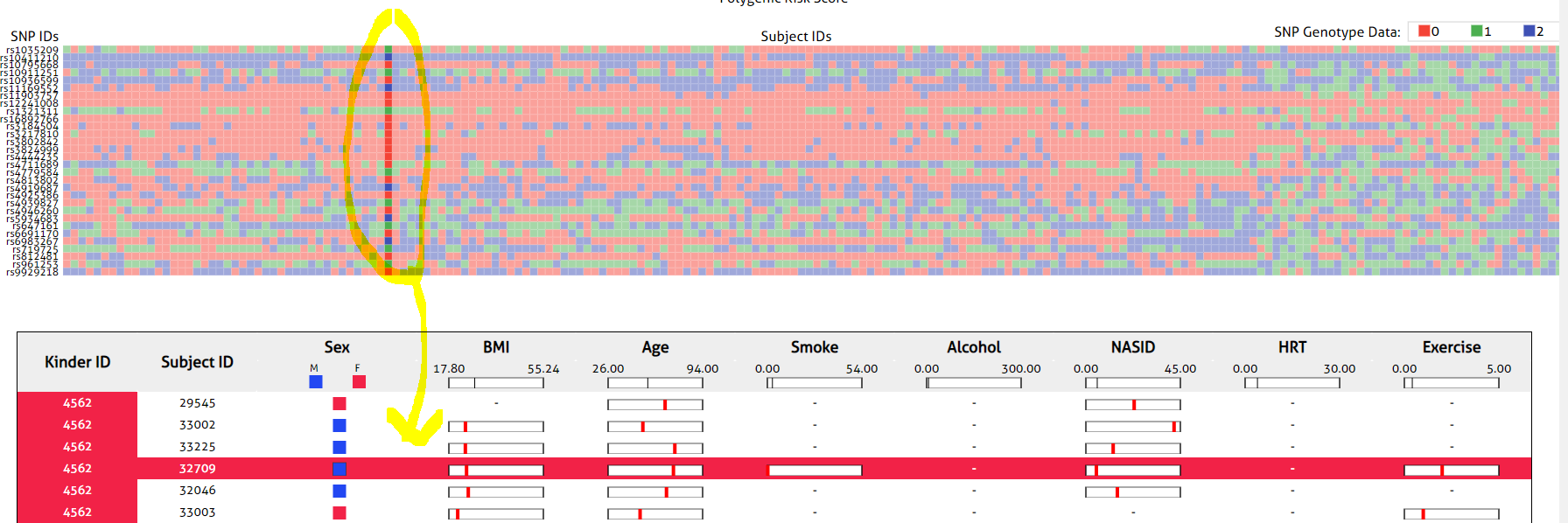
**Nov 23rd:** We started implementing screen 2. Found the colon photo and organized the number of sizes, number of types, and locations. First, it was a big circle for the color, but we removed the sizes and added the colors for the type—finished table for the polyp data.



**Nov 24th:** We did the interaction for the first screen and second screen. For the first screen, we made the table sorting by ascending/descending. For the second screen, when we click the row, the polytype can be in the picture.

**Nov 26th:** Heatmap created and set up everything where we organize in each screen. Heatmap was more significant than we initially expected. Therefore, we put the density plot on the top, then the heatmap, and then the first screen table. Same as the second screen, we Put the picture on the top and the size and type information box for the next, and the table is the bottom. We also calculated the average of each features (ex BMI, Age, Smoke, Alcohol, NASID, HRT and exercise). Then each



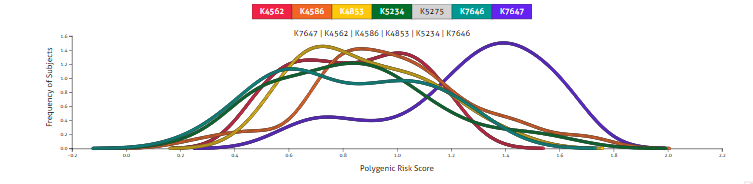


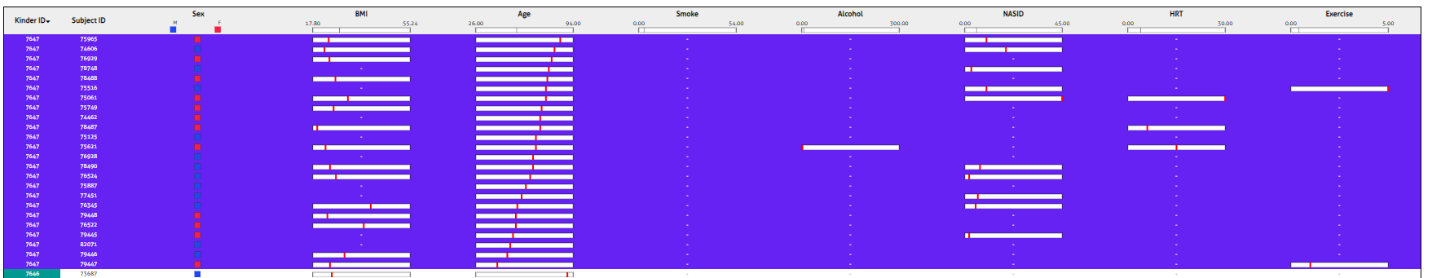
**Nov27th:** We realized that the Subject ID does not match with the heatmap, so we matched the subject ID. Put description as mouse hoovering. All the value became hoovering. We recorded the YouTube.

evaluation

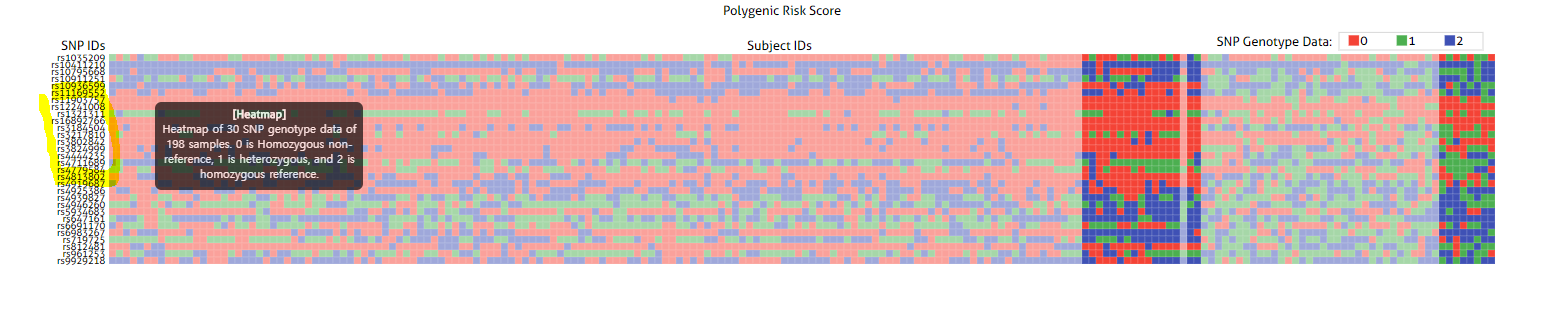
What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

We learned that the general understanding for the data given. First, the Kinder ID (K7647) have many samples tends to have higher polygenic risk score than others.

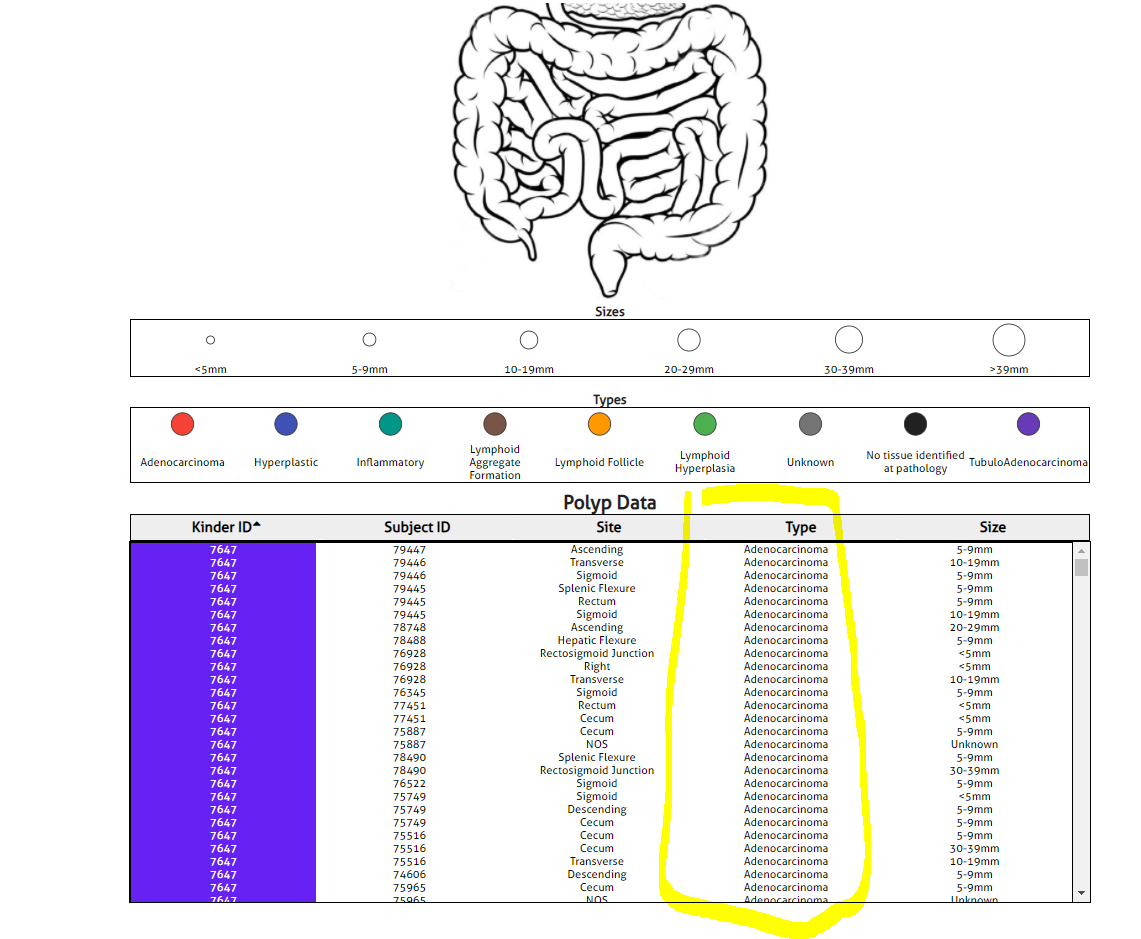


After understanding the general idea of the K7647 polygenic risk score, we were able to learn that clinical information of the samples in the K7647. This family does not have any or only few information of the smoking, alcohol consumption, HRT, and exercise. However, we understand that the subject’s age is older than the average also many subject’s BMI is lower than the average.

Also, we can understand the heatmap for the K7647 information too. Red rectangle (SNP genotype 0) represents the homozygous reference and we know which SNP marks have the most homozygous marks for the groups of K7647.



Finally, when we look at the screen 2, we can see that the K7647 polyp type s are all adenocarcinoma and very small compare to other families.



As our objects is explore relations between polyps, clinical multivariate and genetics, we were able to explore that all information.

If we want to improve this visualization, we will make boxplot for the clinical information and the polyps’ data. Then, this will visualize better than table.