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.

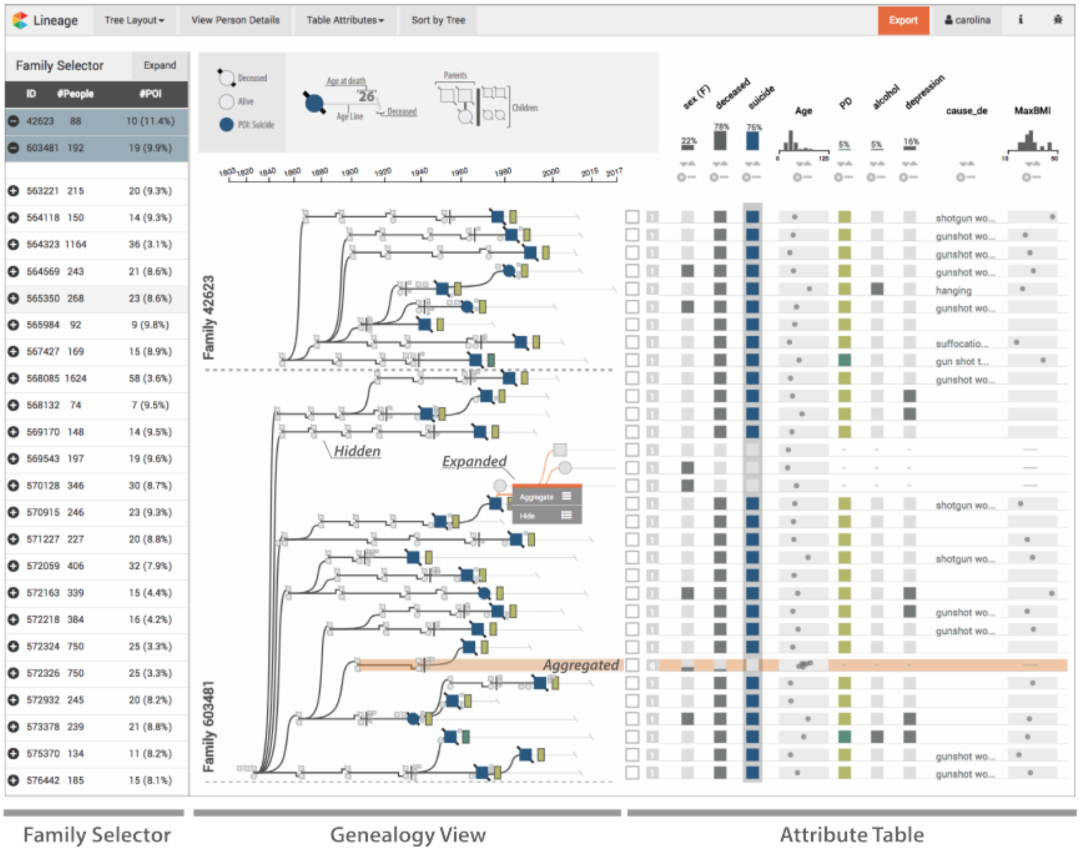
This project's main motivation will explore the seven families with several patients (n=198)'s genetic information associated with the polyps. We will try to understand the genetic information to compared familial clinical multivariate details to visualize all exposure risks. Thus, we know that colon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

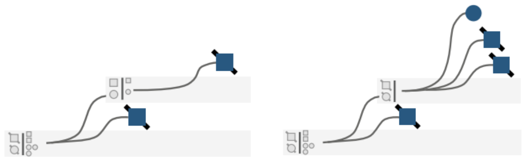
(1) The most inspiration on the technical design for the multivariate risk factors came from the paper called " Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs by Nobre et al."

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.



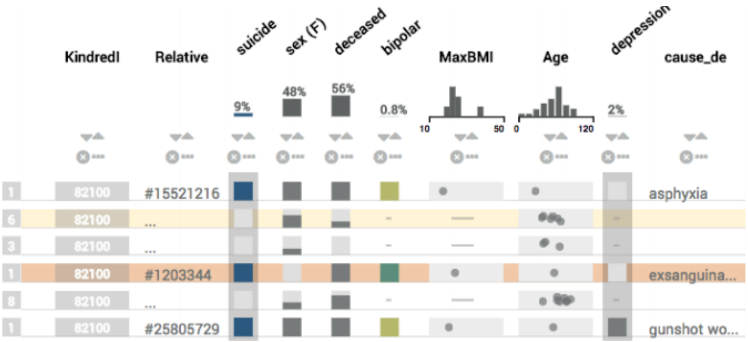
ID	#People	#POI
42623	88	10 (11.4%)
603481	192	19 (9.9%)
563221	215	20 (9.3%)
564118	150	14 (9.3%)
564323	1164	36 (3.1%)
564569	243	21 (8.6%)
565350	268	23 (8.6%)
565984	92	9 (9.8%)
567427	169	15 (8.9%)
568085	1624	58 (3.6%)
568132	74	7 (9.5%)
569170	148	14 (9.5%)
569543	197	19 (9.6%)
570128	346	30 (8.7%)
570915	246	23 (9.3%)
571227	227	20 (8.8%)
572059	406	32 (7.9%)
572163	339	15 (4.4%)
572218	384	16 (4.2%)
572324	750	25 (3.3%)
572326	750	25 (3.3%)
572932	245	20 (8.2%)
573378	239	21 (8.8%)
575370	134	11 (8.2%)
576442	185	15 (8.1%)

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



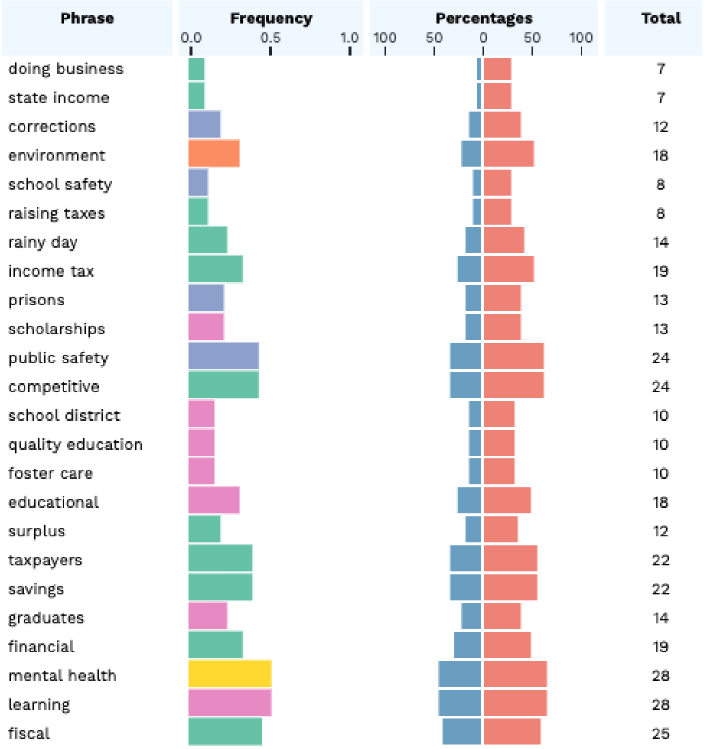
(a) POI: Suicide (b) POI: Age < 40

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

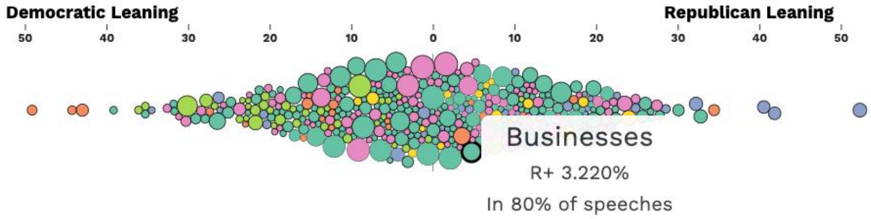


Nevertheless, all this table and family selector idea was inspired by our data visualization.

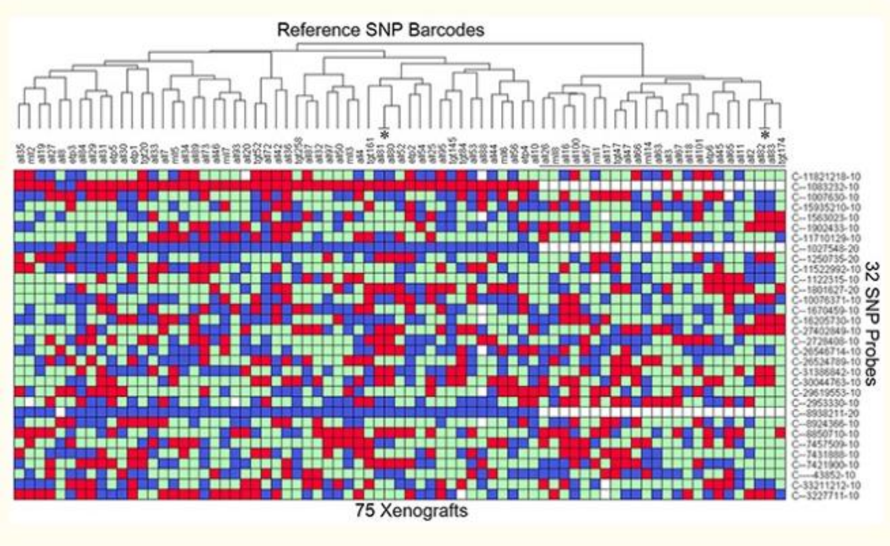
(2) Another inspiration source was from homework six from the class. For our visualization, we wanted to compare the family ID (= KinderID) and the number of the risk factors in the ascending/descending method of sorting. Thus, we can interact with the genetic information that families have the highest 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.



(3) 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)

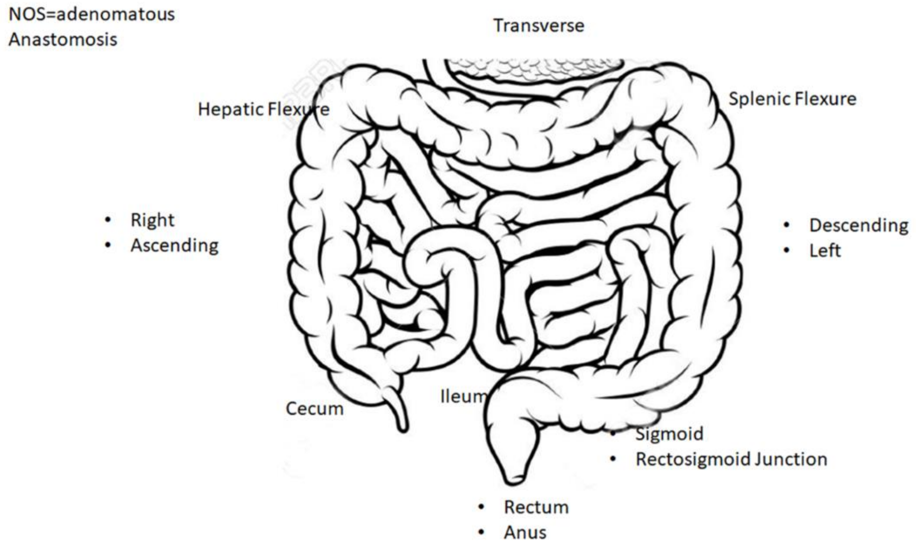
Smoking: Cigarette smoking is an established risk factor for colorectal cancer (PMID: 20587792, PMID:19088354)

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.



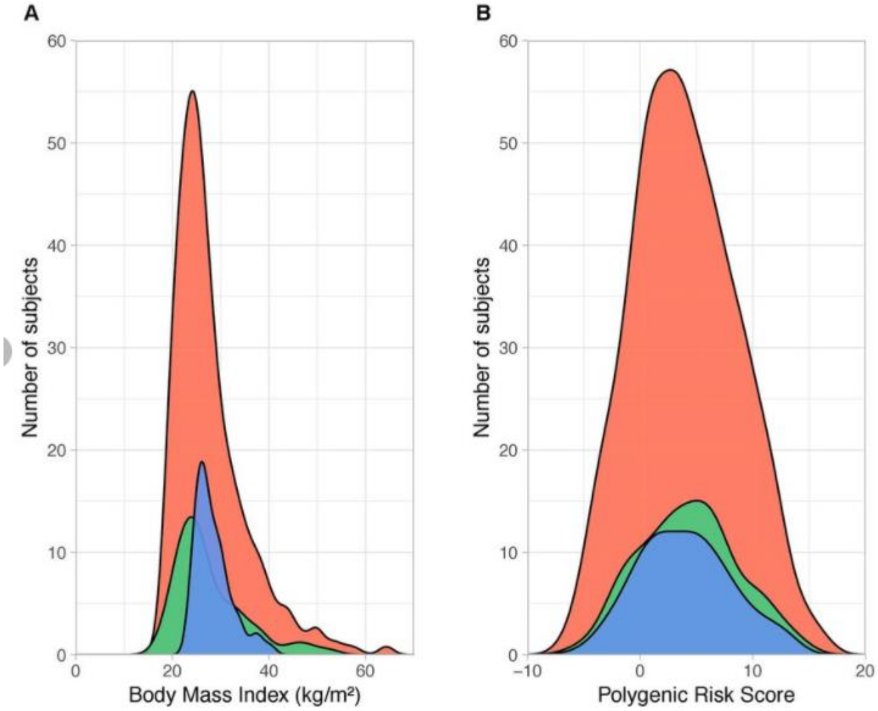
Polyps type: There are 19 types of polyps.
We will combine all the three different unknowns (i.e., No biopsy taken, Lost polyp) into one unknown. Otherwise, we will keep the same for the rest. Total, we reduced the type to 9.

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?

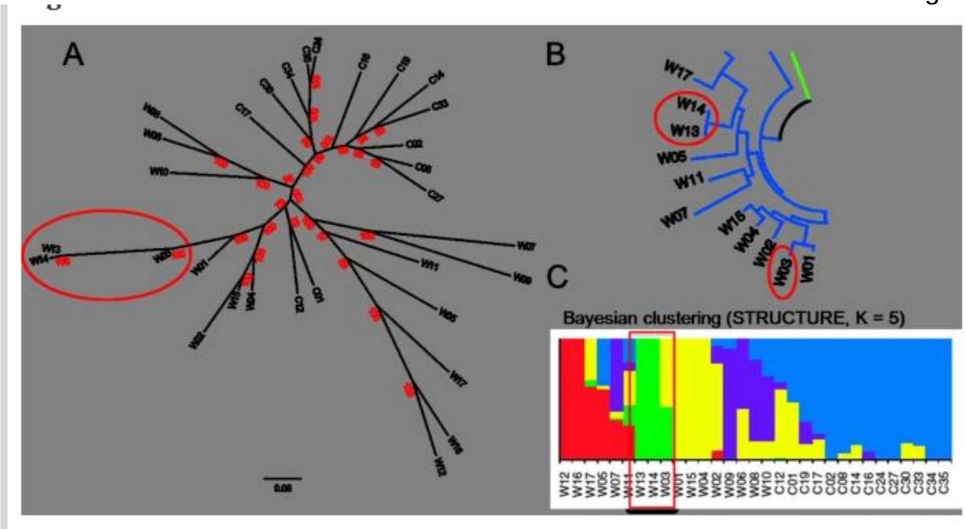
To initially look at our data, we tried to visualize genetic information to gain each family's insight. To see the below image:



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

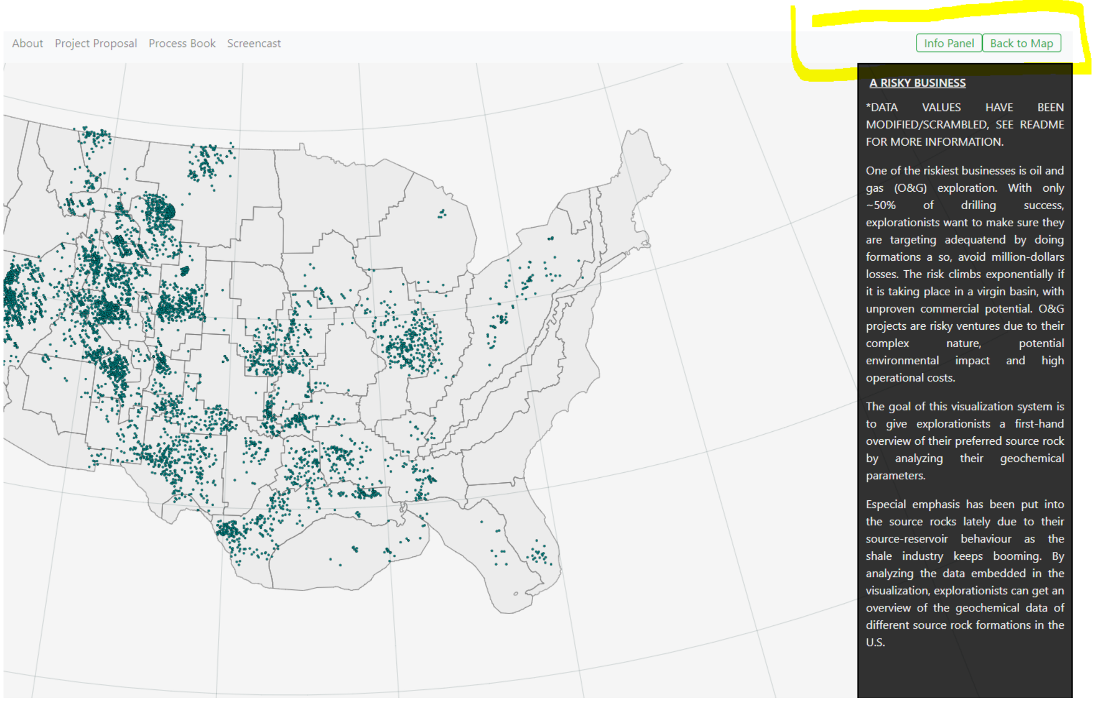
DESIGN EVOLUTION

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


A
B
C
Bayesian clustering (STRUCTURE, K = 5)
W12 W16 W17 W05 W07 W11 W13 W14 W03 W01 W15 W04 W02 W09 W06 W08 W10 C12 C01 C19 C17 C02 C08 C14 C16 C24 C27 C30 C33 C34 C35

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)



A RISKY BUSINESS

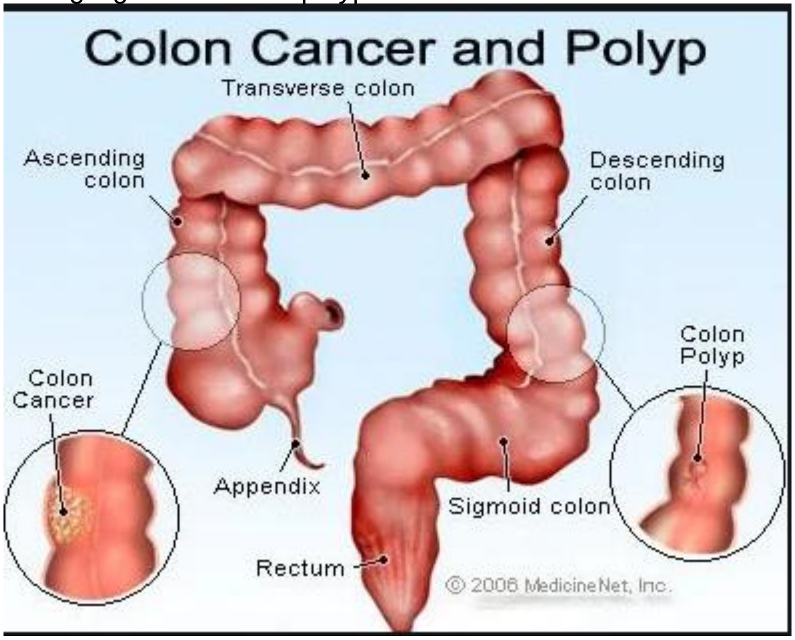
*DATA VALUES HAVE BEEN MODIFIED/SCRAMBLED, SEE README FOR MORE INFORMATION.

One of the riskiest businesses is oil and gas (O&G) exploration. With only ~50% of drilling success, explorationists want to make sure they are targeting adequatend by doing formations a so, avoid million-dollars losses. The risk climbs exponentially if it is taking place in a virgin basin, with unproven commercial potential. O&G projects are risky ventures due to their complex nature, potential environmental impact and high operational costs.

The goal of this visualization system is to give explorationists a first-hand overview of their preferred source rock by analyzing their geochemical parameters.

Especial emphasis has been put into the source rocks lately due to their source-reservoir behaviour as the shale industry keeps booming. By analyzing the data embedded in the visualization, explorationists can get an overview of the geochemical data of different source rock formations in the U.S.

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

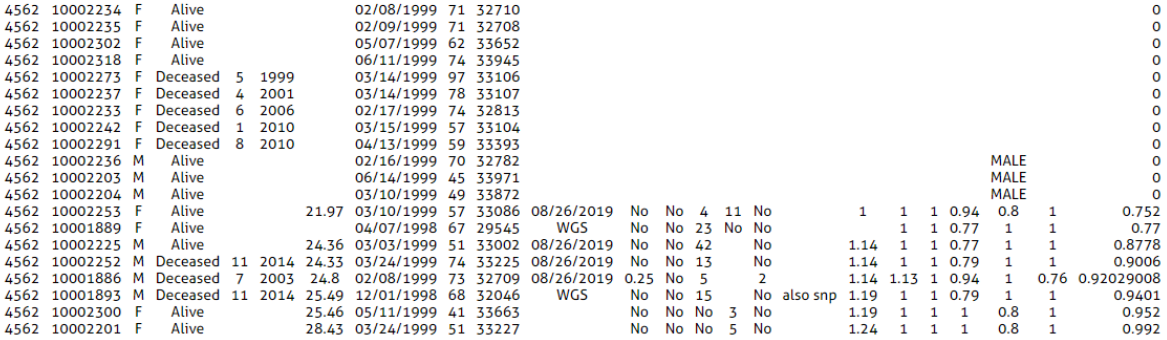
Entry Nov 7th: Setting up the Click button for the different screens and basic density plot for the genetic

- Polyps
- Project Proposal
- YouTube Tutorial
- Contact

Description

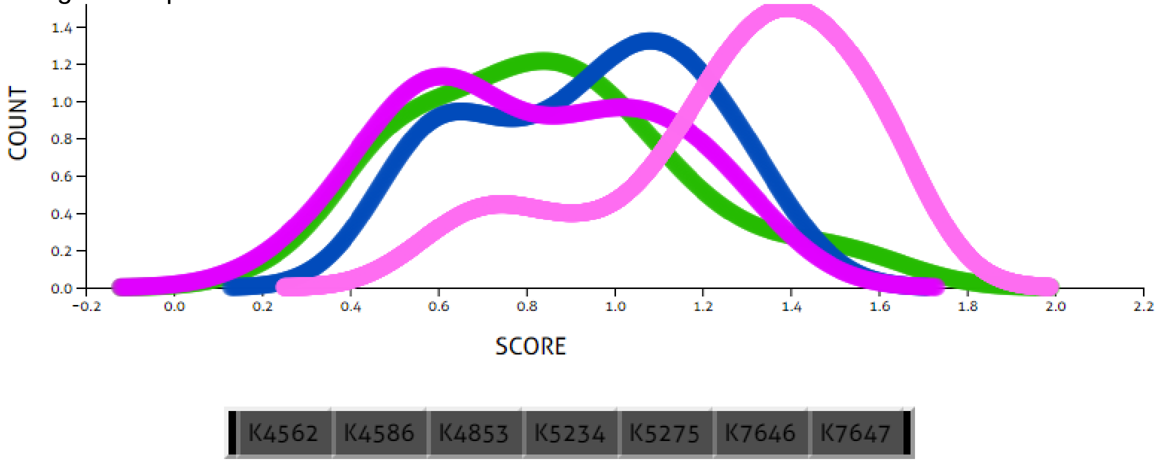
Utah Population based genetic and clinical feature in colorectal cancer

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

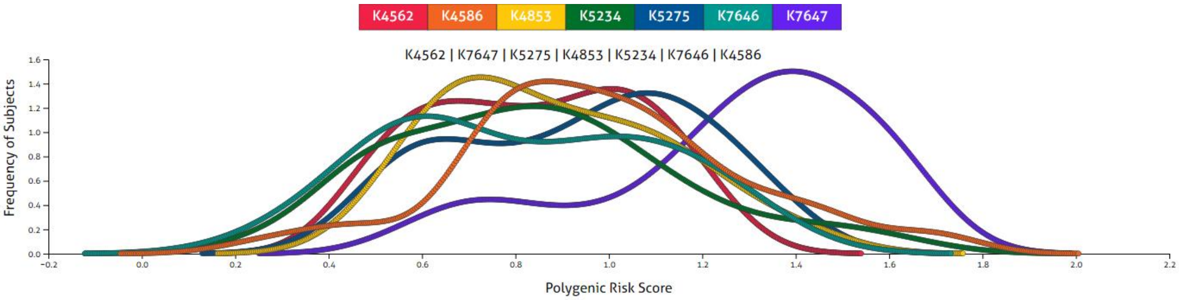


4562	10002234	F	Alive				02/08/1999	71	32710														0
4562	10002235	F	Alive				02/09/1999	71	32708														0
4562	10002302	F	Alive				05/07/1999	62	33652														0
4562	10002318	F	Alive				06/11/1999	74	33945														0
4562	10002273	F	Deceased	5	1999		03/14/1999	97	33106														0
4562	10002237	F	Deceased	4	2001		03/14/1999	78	33107														0
4562	10002233	F	Deceased	6	2006		02/17/1999	74	32813														0
4562	10002242	F	Deceased	1	2010		03/15/1999	57	33104														0
4562	10002291	F	Deceased	8	2010		04/13/1999	59	33393														0
4562	10002236	M	Alive				02/16/1999	70	32782												MALE		0
4562	10002203	M	Alive				06/14/1999	45	33971												MALE		0
4562	10002204	M	Alive				03/10/1999	49	33872												MALE		0
4562	10002253	F	Alive			21.97	03/10/1999	57	33086	08/26/2019	No	No	4	11	No		1	1	1	0.94	0.8	1	0.752
4562	10001889	F	Alive				04/07/1998	67	29545	WGS	No	No	23	No	No			1	1	0.77	1	1	0.77
4562	10002225	M	Alive			24.36	03/03/1999	51	33002	08/26/2019	No	No	42		No		1.14	1	1	0.77	1	1	0.8778
4562	10002252	M	Deceased	11	2014	24.33	03/24/1999	74	33225	08/26/2019	No	No	13		No		1.14	1	1	0.79	1	1	0.9006
4562	10001886	M	Deceased	7	2003	24.8	02/08/1999	73	32709	08/26/2019	0.25	No	5		2		1.14	1.13	1	0.94	1	0.76	0.92029008
4562	10001893	M	Deceased	11	2014	25.49	12/01/1998	68	32046	WGS	No	No	15		No	also snp	1.19	1	1	0.79	1	1	0.9401
4562	10002300	F	Alive			25.46	05/11/1999	41	33663		No	No	No	3	No		1.19	1	1	1	0.8	1	0.952
4562	10002201	F	Alive			28.43	03/24/1999	51	33227		No	No	No	5	No		1.24	1	1	1	0.8	1	0.992

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

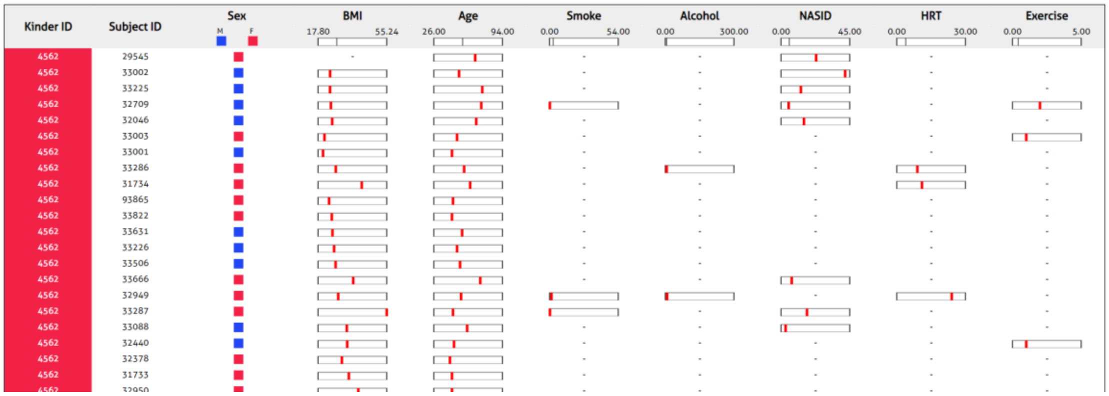


Nov 16th: Finished the density plot



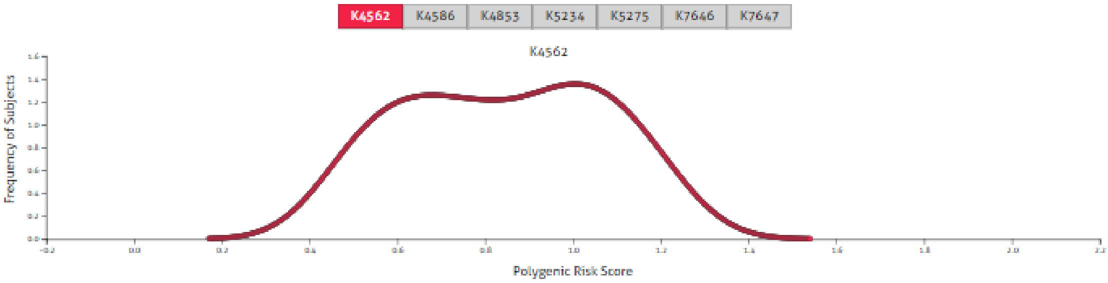
Nov 17th: Met with TA and got some advice for the heatmap and feedback from our project.

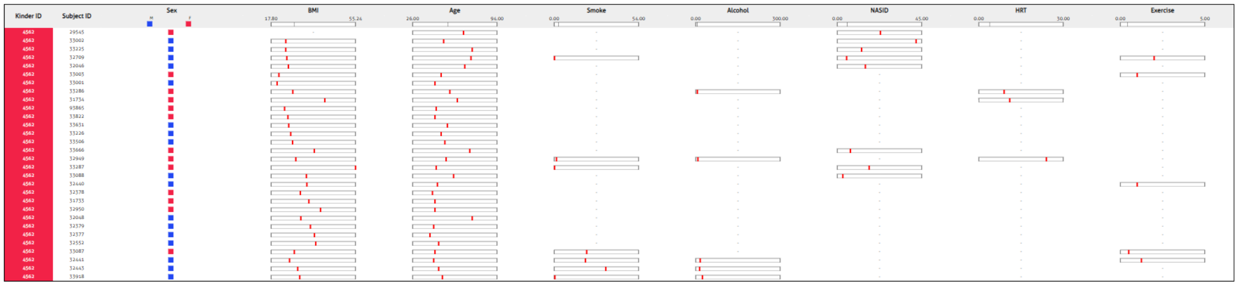
Nov 19th: Table for the 1st screen was added.



Kinder ID	Subject ID	Sex M 🟦 F 🟥	BMI 17.80 55.24	Age 26.00 94.00	Smoke 0.00 54.00	Alcohol 0.00 300.00	NASID 0.00 45.00	HRT 0.00 30.00	Exercise 0.00 5.00
4562	29545	🟥	-		-	-		-	-
4562	33002	🟦			-	-		-	-
4562	33225	🟦			-	-		-	-
4562	32709	🟦				-		-	
4562	32046	🟦			-	-		-	-
4562	33003	🟥			-	-	-	-	
4562	33001	🟦			-	-	-	-	-
4562	33286	🟥			-		-		-
4562	31734	🟥			-	-	-		-
4562	93865	🟥			-	-	-	-	-
4562	33822	🟥			-	-	-	-	-
4562	33631	🟦			-	-	-	-	-
4562	33226	🟦			-	-	-	-	-
4562	33506	🟦			-	-	-	-	-
4562	33666	🟥			-	-		-	-
4562	32949	🟥					-		-
4562	33287	🟥				-		-	-
4562	33088	🟦			-	-		-	-
4562	32440	🟦			-	-	-	-	
4562	32378	🟥			-	-	-	-	-
4562	31733	🟥			-	-	-	-	-
4562	32950	🟥			-	-	-	-	-

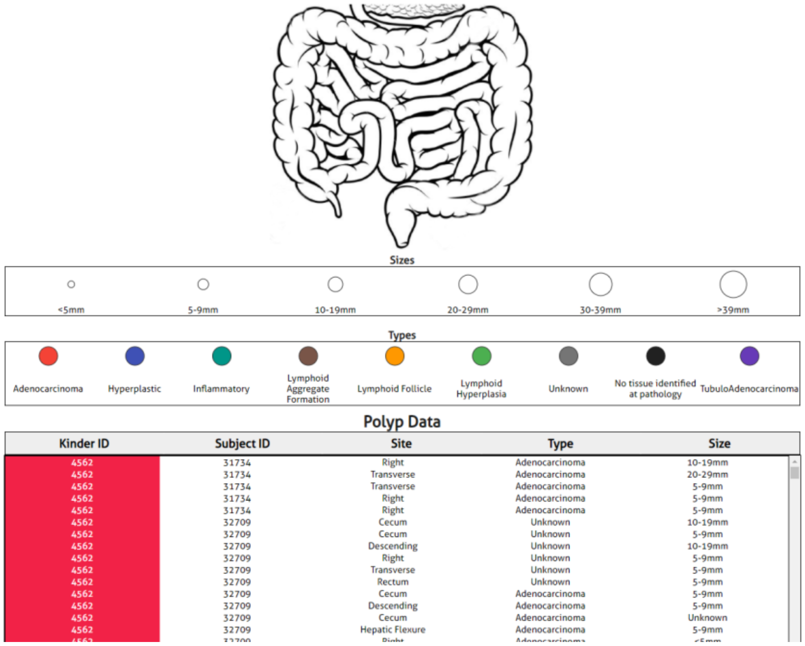
Nov 22nd: We made the interaction when we click the Kinder ID from the density plot, then it will show only same Kinder ID from the table





Kinder ID	Subject ID	Sex M 🟦 F 🟥	BMI 17.80 55.24	Age 26.00 94.00	Smoke 0.00 54.00	Alcohol 0.00 300.00	NASID 0.00 45.00	HRT 0.00 30.00	Exercise 0.00 5.00
4562	29545	🟥							
4562	33002	🟦							
4562	33225	🟦							
4562	32709	🟦							
4562	32046	🟦							
4562	33003	🟥							
4562	33001	🟦							
4562	33286	🟥							
4562	31734	🟥							
4562	93865	🟥							
4562	33822	🟥							
4562	33631	🟦							
4562	33226	🟦							
4562	33506	🟦							
4562	33666	🟥							
4562	32949	🟥							
4562	33287	🟥							
4562	33088	🟦							
4562	32440	🟦							
4562	32378	🟥							
4562	31733	🟥							
4562	32950	🟥							
4562	32048	🟦							
4562	32379	🟦							
4562	32377	🟦							
4562	32552	🟦							
4562	33087	🟥							
4562	32441	🟦							
4562	32443	🟦							
4562	33918	🟦							

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.

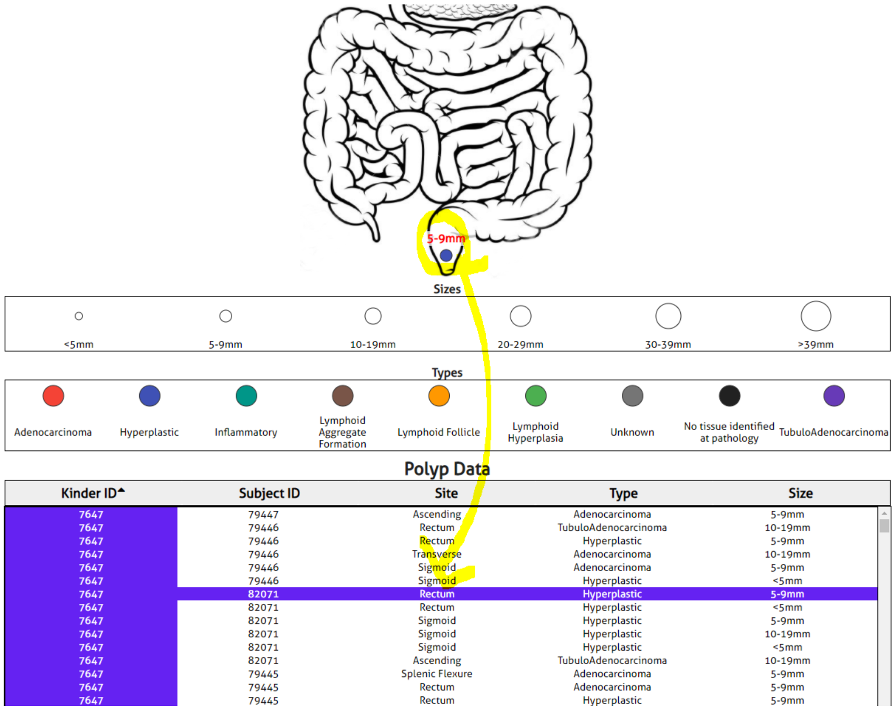


Polyp Data

Kinder ID	Subject ID	Site	Type	Size
4562	31734	Right	Adenocarcinoma	10-19mm
4562	31734	Transverse	Adenocarcinoma	20-29mm
4562	31734	Transverse	Adenocarcinoma	5-9mm
4562	31734	Right	Adenocarcinoma	5-9mm
4562	31734	Right	Adenocarcinoma	5-9mm
4562	32709	Cecum	Unknown	10-19mm
4562	32709	Cecum	Unknown	5-9mm
4562	32709	Descending	Unknown	10-19mm
4562	32709	Right	Unknown	5-9mm
4562	32709	Transverse	Unknown	5-9mm
4562	32709	Rectum	Unknown	5-9mm
4562	32709	Cecum	Adenocarcinoma	5-9mm
4562	32709	Descending	Adenocarcinoma	5-9mm
4562	32709	Cecum	Adenocarcinoma	Unknown
4562	32709	Hepatic Flexure	Adenocarcinoma	5-9mm
4562	32709	Right	Adenocarcinoma	<5mm

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



Sizes

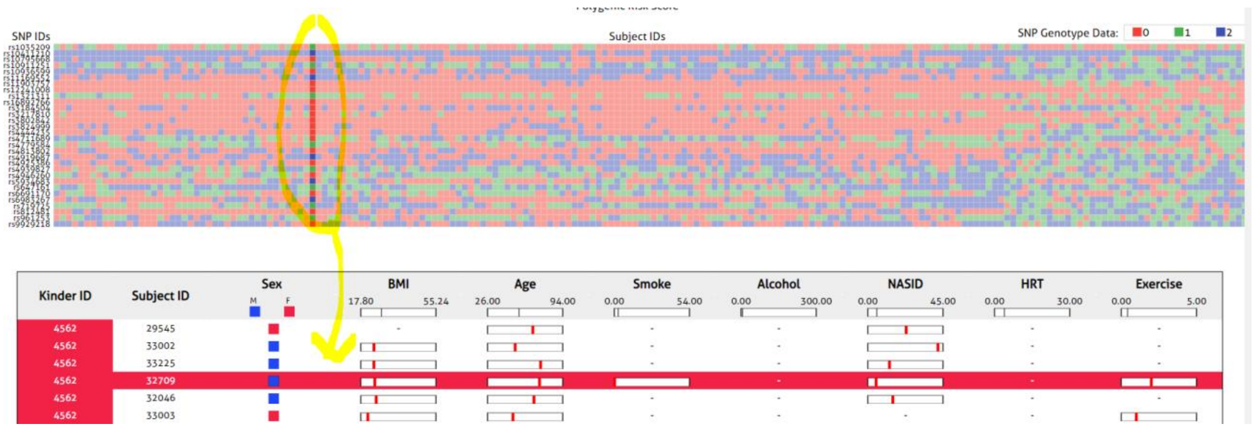
○	○	○	○	○	○
<5mm	5-9mm	10-19mm	20-29mm	30-39mm	>39mm

Types

●	●	●	●	●	●	●	●	●
Adenocarcinoma	Hyperplastic	Inflammatory	Lymphoid Aggregate Formation	Lymphoid Follicle	Lymphoid Hyperplasia	Unknown	No tissue identified at pathology	TubuloAdenocarcinoma

Polyp Data

Kinder ID⏶	Subject ID	Site	Type	Size
7647	79447	Ascending	Adenocarcinoma	5-9mm
7647	79446	Rectum	TubuloAdenocarcinoma	10-19mm
7647	79446	Rectum	Hyperplastic	5-9mm
7647	79446	Transverse	Adenocarcinoma	10-19mm
7647	79446	Sigmoid	Adenocarcinoma	5-9mm
7647	79446	Sigmoid	Hyperplastic	<5mm
7647	82071	Rectum	Hyperplastic	5-9mm
7647	82071	Rectum	Hyperplastic	<5mm
7647	82071	Sigmoid	Hyperplastic	5-9mm
7647	82071	Sigmoid	Hyperplastic	10-19mm
7647	82071	Sigmoid	Hyperplastic	<5mm
7647	82071	Ascending	TubuloAdenocarcinoma	10-19mm
7647	79445	Splenic Flexure	Adenocarcinoma	5-9mm
7647	79445	Rectum	Adenocarcinoma	5-9mm
7647	79445	Rectum	Hyperplastic	5-9mm



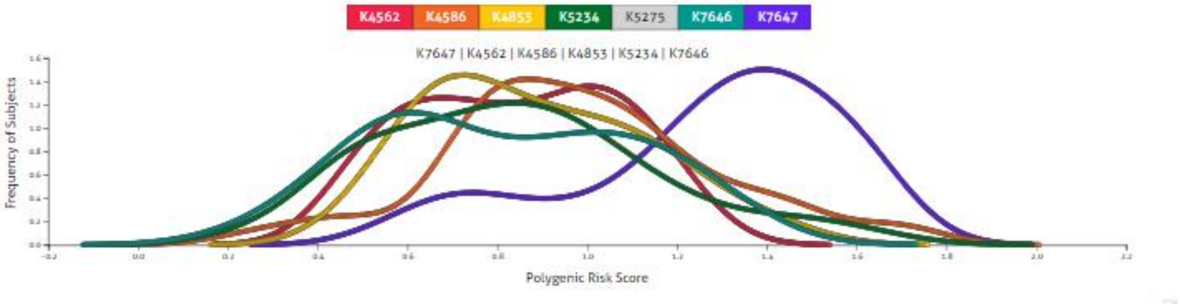
Kinder ID	Subject ID	Sex M 🟦 F 🟥	BMI 17.80 55.24	Age 26.00 94.00	Smoke 0.00 54.00	Alcohol 0.00 300.00	NASID 0.00 45.00	HRT 0.00 30.00	Exercise 0.00 5.00
4562	29545	🟥	-		-	-		-	-
4562	33002	🟦			-	-		-	-
4562	33225	🟦			-	-		-	-
4562	32709	🟦				-		-	
4562	32046	🟦			-	-		-	-
4562	33003	🟥			-	-	-	-	

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.

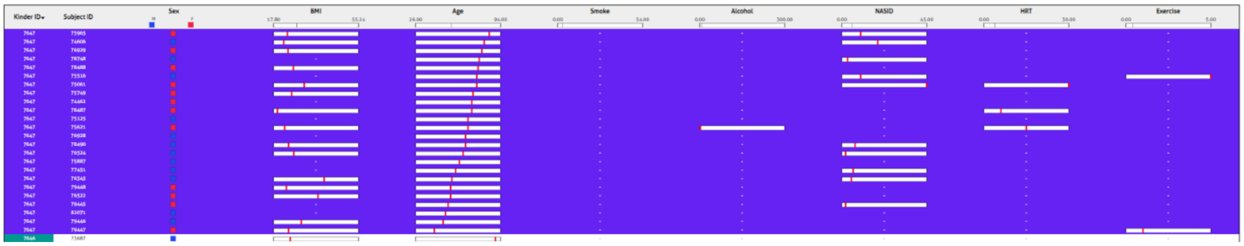
EVALUATION

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.

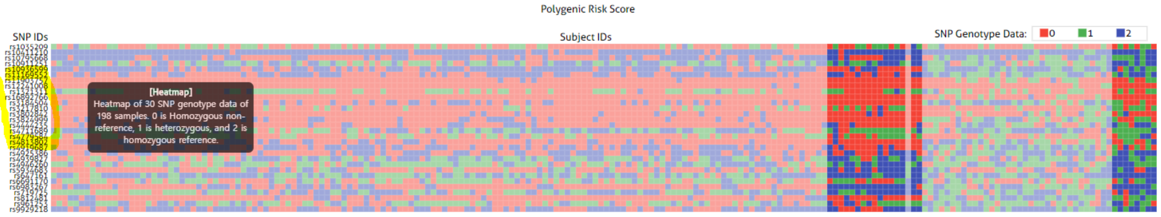


After understanding the general idea of the K7647 polygenic risk score, we were able to learn that clinical information of the samples in the K7647. This family does not have any or only few information of the smoking, alcohol consumption, HRT, and exercise. However, we understand that the subject’s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



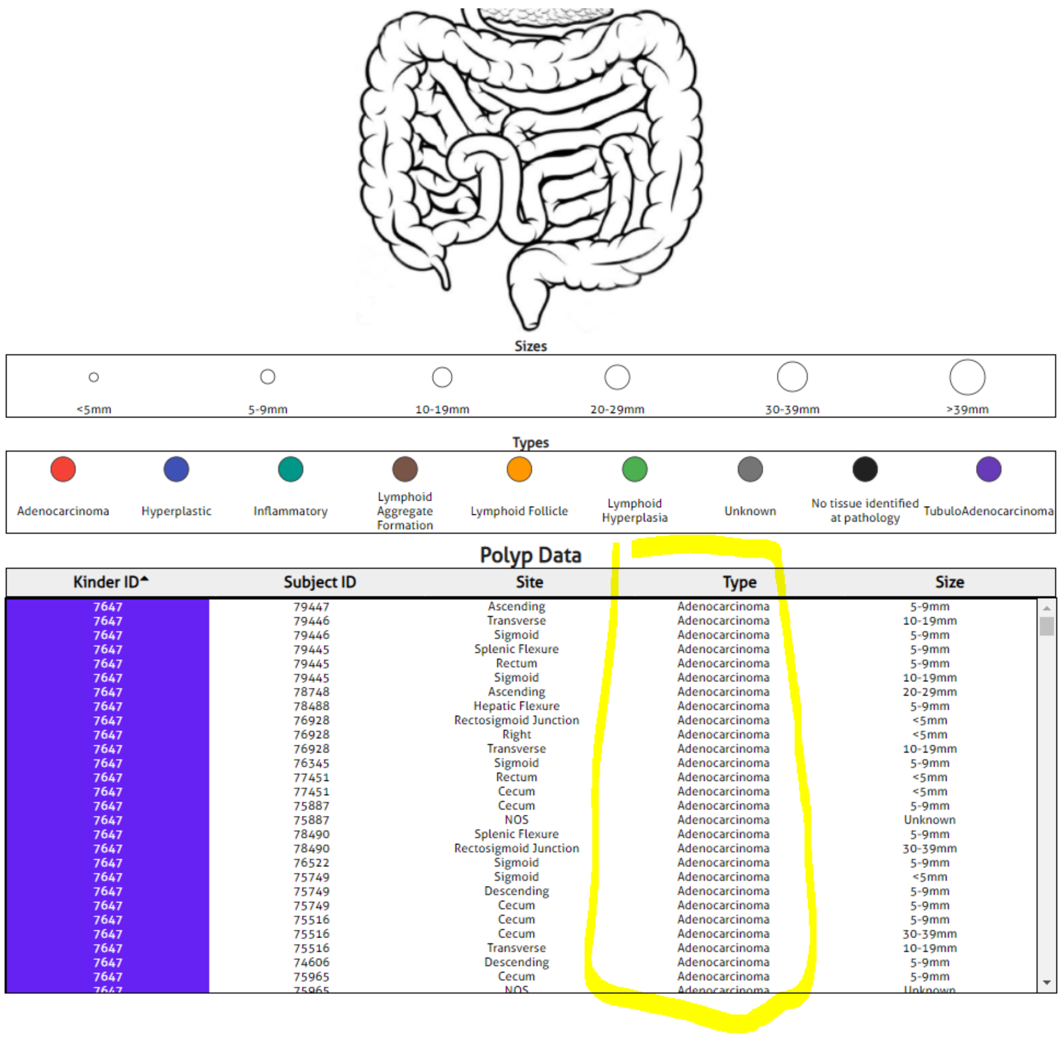
Kinder ID	Subject ID	Sex M 🟦 F 🟥	BMI 17.80 55.24	Age 26.00 94.00	Smoke 0.00 54.00	Alcohol 0.00 300.00	NASID 0.00 45.00	HRT 0.00 30.00	Exercise 0.00 5.00
7647	75965	🟥							
7647	74606	🟦							
7647	76939	🟥							
7647	78748	🟦							
7647	78488	🟥							
7647	75516	🟦							
7647	75061	🟥							
7647	75749	🟥							
7647	74462	🟥							
7647	78487	🟥							
7647	75125	🟦							
7647	75621	🟥							
7647	76928	🟦							
7647	78490	🟦							
7647	76524	🟦							
7647	75887	🟦							
7647	77451	🟦							
7647	76345	🟦							
7647	79448	🟥							
7647	76522	🟥							
7647	79445	🟥							
7647	82071	🟦							
7647	79446	🟦							
7647	79447	🟥							
7646	73687	🟦							

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.



[Heatmap]
Heatmap of 30 SNP genotype data of 198 samples. 0 is Homozygous non-reference, 1 is heterozygous, and 2 is homozygous reference.

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.



Polyp Data

Kinder ID▲	Subject ID	Site	Type	Size
7647	79447	Ascending	Adenocarcinoma	5-9mm
7647	79446	Transverse	Adenocarcinoma	10-19mm
7647	79446	Sigmoid	Adenocarcinoma	5-9mm
7647	79445	Splenic Flexure	Adenocarcinoma	5-9mm
7647	79445	Rectum	Adenocarcinoma	5-9mm
7647	79445	Sigmoid	Adenocarcinoma	10-19mm
7647	78748	Ascending	Adenocarcinoma	20-29mm
7647	78488	Hepatic Flexure	Adenocarcinoma	5-9mm
7647	76928	Rectosigmoid Junction	Adenocarcinoma	<5mm
7647	76928	Right	Adenocarcinoma	<5mm
7647	76928	Transverse	Adenocarcinoma	10-19mm
7647	76345	Sigmoid	Adenocarcinoma	5-9mm
7647	77451	Rectum	Adenocarcinoma	<5mm
7647	77451	Cecum	Adenocarcinoma	<5mm
7647	75887	Cecum	Adenocarcinoma	5-9mm
7647	75887	NOS	Adenocarcinoma	Unknown
7647	78490	Splenic Flexure	Adenocarcinoma	5-9mm
7647	78490	Rectosigmoid Junction	Adenocarcinoma	30-39mm
7647	76522	Sigmoid	Adenocarcinoma	5-9mm
7647	75749	Sigmoid	Adenocarcinoma	<5mm
7647	75749	Descending	Adenocarcinoma	5-9mm
7647	75749	Cecum	Adenocarcinoma	5-9mm
7647	75516	Cecum	Adenocarcinoma	5-9mm
7647	75516	Cecum	Adenocarcinoma	30-39mm
7647	75516	Transverse	Adenocarcinoma	10-19mm
7647	74606	Descending	Adenocarcinoma	5-9mm
7647	75965	Cecum	Adenocarcinoma	5-9mm
7647	75965	NOS	Adenocarcinoma	Unknown

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.