

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 to reducing the number of incidences by 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.

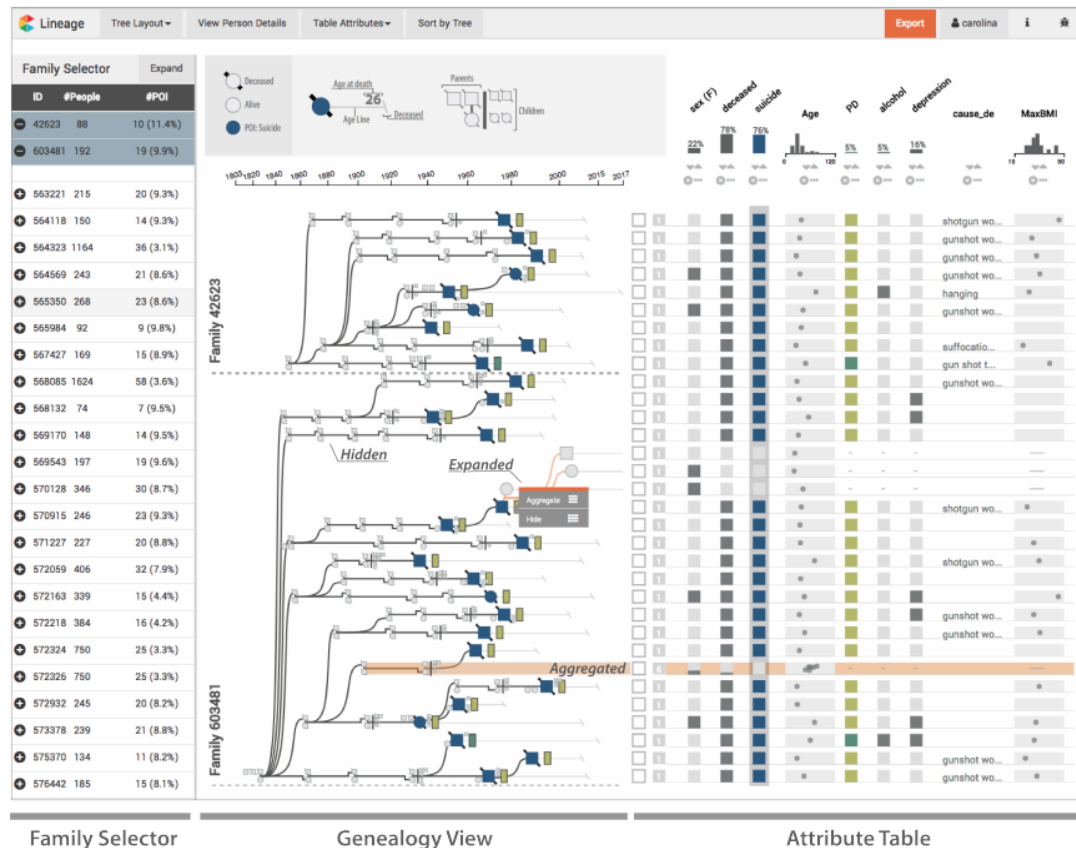
This project's main motivation will explore the seven families with several patients (n=198)'s genetic information associated with the polyps. We will try to understand the genetic information to compared familial clinical multivariate details to visualize all exposure risks. Thus, we know that colon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

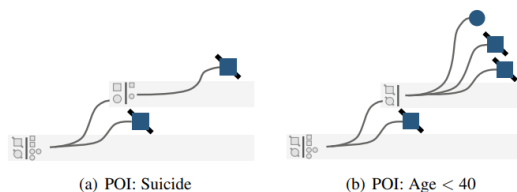
RELATED WORK

- (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."

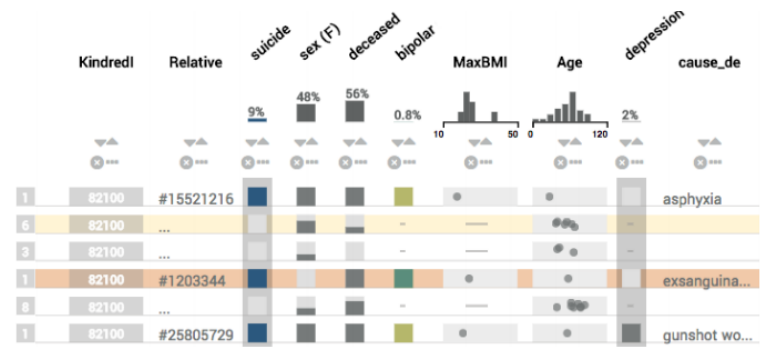
This paper describes the combination of hereditary and all environmental risk factors. The paper used the three layouts (i.e., Family selector, Genealogy view, and Attribute table). These layouts gave an easy interaction of the association of the family selector through the attributes. We tried to follow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



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



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

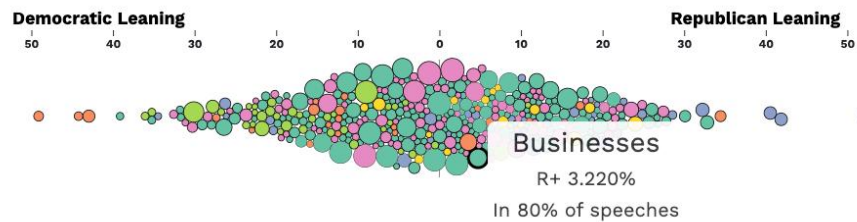


Nevertheless, all this table and family selector idea was inspired by our data visualization.

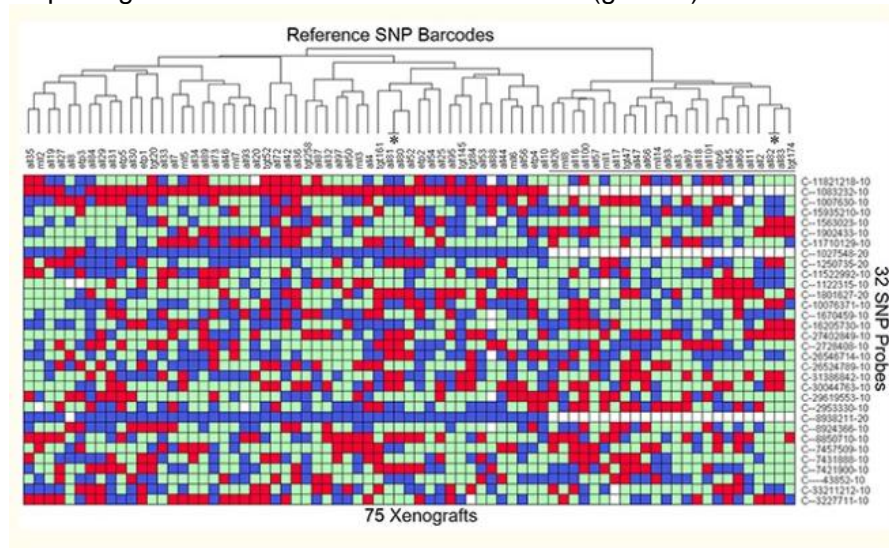
- (2) Another inspiration source was from homework six from the class. For our visualization, we wanted to compare the family ID (= KinderID) and the number of the risk factors in the ascending/descending method of sorting. Thus, we can interact with the genetic information that families have the highest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

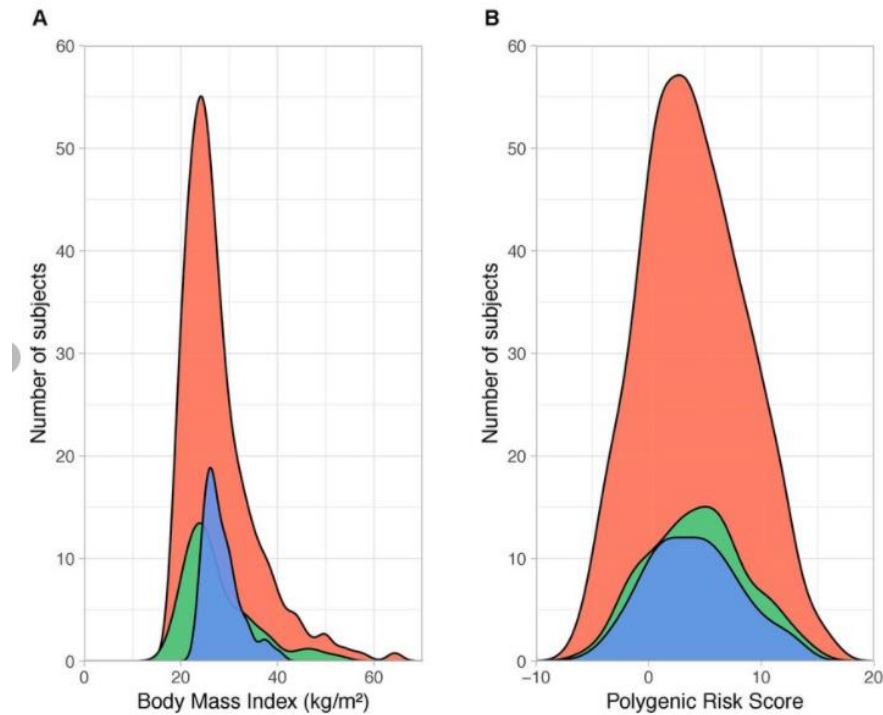
Polyps type: There are 19 types of polyps. We will combine all the three different unknowns (i.e., No biopsy taken, Lost polyp) into one unknown. Otherwise, we will keep the same for the rest.

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?

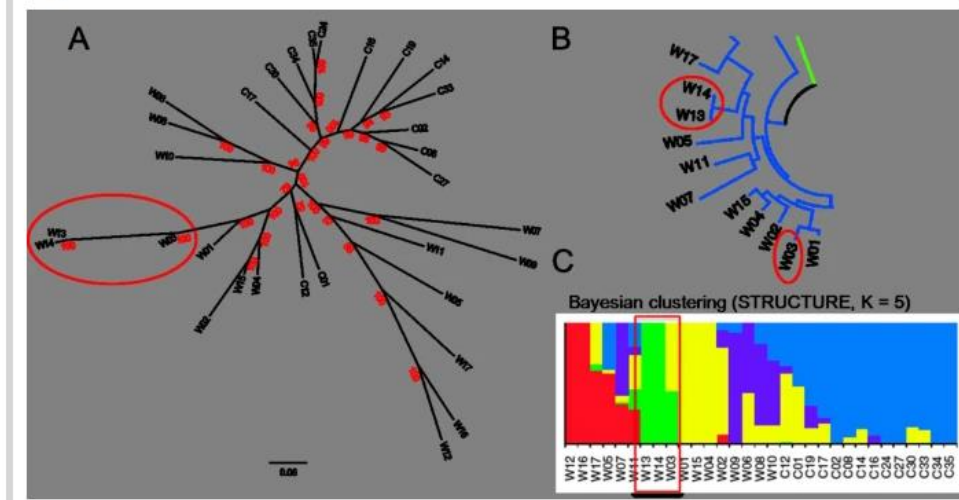
To initially look at our data, we tried to visualize genetic information to gain each family's insight. To see the below image:



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.

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

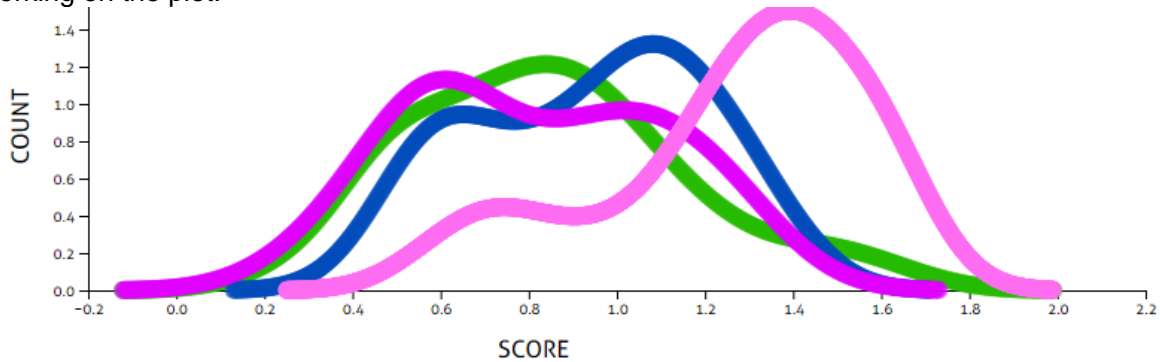
Utah Population based genetic and clinical feature in colorectal cancer

[Genetic&Multivariate](#)
[Polyps](#)
[YouTube tutorial](#)
[Process Book](#)
[Contact](#)

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			04/10/1999	57	33086	08/26/2019	No	No	4	11	No		1	1	0.94	0.8		1	0.752
4562	10001889	F	Alive			03/07/1998	67	29545	WGS	No	No	23	No	No		1	1	0.77	1	1		0.77
4562	10002225	M	Alive			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	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	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	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			05/11/1999	41	33663		No	No	No	3	No	1.19	1	1	1	0.8	1		0.952
4562	10002201	F	Alive			03/24/1999	51	33227		No	No	No	5	No	1.24	1	1	1	0.8	1		0.999

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



K4562	K4586	K4853	K5234	K5275	K7646	K7647
-------	-------	-------	-------	-------	-------	-------

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?

