Genomic Data and Biological Pathway Visualizations

Mahika Dubey, Hadiseh Gooran, Veronica Rivera
University of California, Santa Cruz

Abstract

Most of the science fields work with big data. For example, in the field of biology, researchers work with the huge amount of data with different values and features which are not easy to read and interact, while they need to correctly understand the data. To address this issue, computer engineers and designers try to simplify the understanding of the big data by data visualization tools. In this project we tried to simplify the understanding of the biological pathways for biology and cancer researchers of the Daniel Kim's and Angela Brook's labs at UCSC, by developing the eVip and RasVis, which are interactive web-based visualization tools. eVip data visualization allows the exploration of genomic data in interactive and intuitive way, to show the categories of the different mutations and replications, and how different over expressions affects different genes. And the RasVis tool shows simple network based visualizations for isolated specific pathway data.

Overview

We have had discussion with grad students of Daniel Kim and Angela Brooks lab. They had big biology genomic data sets which was huge and difficult to read and they both wanted to have an interactive data visualisation of those data sets, to be able to have a better understanding of those data. After the discussion we decided to address both challenges in one project but after exploring the data and talk more with PIs we have found that the data sets and pathways are totally different and we started working on two different projects, eVip for Brooks and RasVis for Kim lab. Goal of both projects was to help these labs' researchers to have better understanding of these huge amount of data and help them to have a dynamic, interactive and intuitive data visualization.

Past Project Goals

We worked with Daniel Kim's and Angela Brook's lab at UCSC to create a visualization that helps them visualize genomics data. To obtain a better understanding of the types of visualization tasks that are important in their research, we had a free-form discussion with Angela and with two of Daniel Kim's graduate students. In this section we summarize the work that is done in both of these labs, as well as some of the most important takeaways from both meetings.

Daniel Kim's lab studies noncoding RNA in stem cells and cancer. They use genomic technologies to characterize noncoding RNA in single cells and they generate a lot of sequencing data in their experiments. The Kim lab also tries to understand different pathways

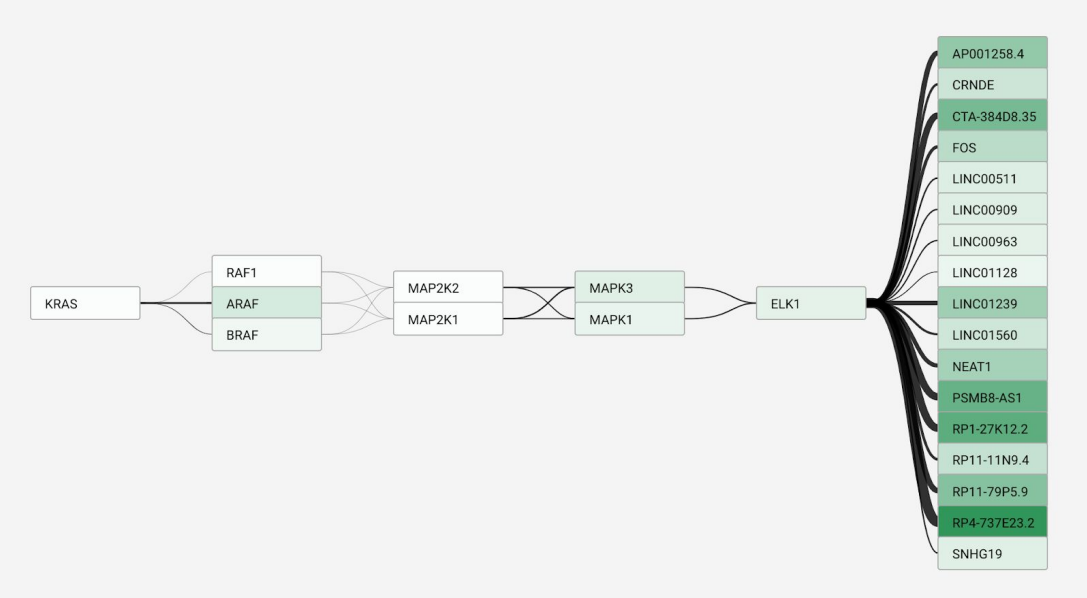
between proteins, RNA sequences and how these may contribute to cancer. Angela Brooks's lab focuses on studying somatic mutations that cause changes to the transcriptome. The work in her lab tries to understand how alternative splicing is regulated, and the members of this lab also develop computational techniques to analyze genome and transcriptome sequencing data, similar to the data generated in the Kim lab.

During our meeting with two graduate students in the Kim lab, we learned that they work heavily with kRas, a gene that is a driver mutation that drives other cells to become cancerous. They mentioned that both for their kRas data and other sequencing datasets that their lab generates, they would like a visualization that helps them analyze pathways between genes, giving them the ability to mine the data and look at interactions between pathways. They have an existing visualization called RasVis, developed by one of the undergraduate students from the Creative Coding lab using D3 to build on top of an existing design made in Adobe Illustrator. However, RasVis is currently not dynamic and contains hardcoded positional data for the network representation of a pathway. Together with the two students, we came up with a set of ideas for visualizations that may be helpful to them and other biology and bioinformatics research groups. Long term, they said that they would like to have a tool that can take genes from data input, find relevant pathways programmatically, and visualize them. Short term, they would be interested in any pathway visualization that helps visualize a signalling cascade. They would like to be able to automatically generate a network based on user provided data input, rather than using hard coded data. They mentioned that they would also benefit from a tool that parses data into a common format for visualizing, and that provides some interaction leading to external hyperlinks and metadata. Another idea that was brought up was being able to take a group of genes from data input, cross reference with sequencing data, and visualize the hierarchies.

During a separate meeting we met with Angela Brooks to learn about what types of visualization tools would be useful in her work. Much of the research conducted at the Brooks Lab involves identifying gene mutations that actually affect the progression of cancer cells. During our meeting with Angela, we discussed some of the shortcomings she sees with existing common visualizations in the field. The greatest issue she noted was the non-intuitive nature of the visualizations her lab is currently developing. When it comes to displaying biological data, there are huge databases to provide lots of contextual information, when in many cases, a simple diagram is desired, isolating a very small reaction or pathway that is being reasoned about. When visualizations cover broader outcomes, there is a lot of explanation required before they can be understood. One specific example is the large network visualization of pathways that is often used in bioinformatics papers. After our conversation, we received a dataset from one of Angela's graduate students who is working on an existing visualization for pathways. The dataset contains gene expression data for all samples in the project, and the pathway information comes from the Hallmark gene set from the Molecular Signatures Database.

Our goal in this project was initially to enhance RasVis (pictured below) and create an automated pipeline that could generate network visualizations for pathway data from both the Kim Lab and the Brooks Lab. We were to facilitate adding data to the visualization by having it

programatically read in the data file and generate the resulting pathway for the kRas gene. This would allow the visualization to be used for data exploration, rather than just for including in a



publication, since the visualization would have the ability to change depending on the data being used.

Literature Review

In this section we present a comprehensive summary of relevant literature we used in learning about biological pathway visualization and existing visualizations for genomics datasets. The articles cited are presented in chronological to facilitate understanding how they relate. In addition to presenting a summary of each article, we also provide comments our takeaways and opinions.

Title: Visualizing genomes: techniques and challenges
Authors: Cydney B Nielsen , Michael Cantor, Inna Dubchak , David Gordon & Ting Wang
Year: 2010

Summary and how it relates to visualization:
This paper provides a guide to genomic data visualization tools that facilitate analysis tasks by enabling researchers to explore, interpret and manipulate their data, and in some cases perform “on-the-fly” computations. It has been mentioned that the emergence of extensive sequence data resources opened new interfaces with computer science, fuelling fields like bioinformatics. One challenge is choosing an appropriate visual representation, beside this challenge, some types of primary data are unavailable owing to their prohibitive online storage requirements, and enabling real-time interaction with large scale datasets is nontrivial. They know “rapidly evolving” as one of the important consideration in the field of genomics. According to this research visualizing genomic has different categories of visualization. From our conversation with Kim’s Lab we need to visualize the sequencing data. In this paper they came up with that recent innovations in sequencing technology have been accompanied by a

growth in new assembly and alignment programs to cope with the shorter read lengths and larger numbers of reads but no standards have been reached. Another challenge about browsing the genome, is how a researcher can navigate this sequence to find regions of interest. The most considerable challenges in this types of visualization is the data type, data volume and data representation.

Takeaways:
They discussed graphical methods designed for the analysis of denovo sequencing assemblies and read alignments, genome browsing, and comparative genomics, highlighting the strengths and limitations of these approaches and the challenges ahead. Different challenges are in visualizing genomics, such as existing large number of samples to compare, large number of data types, genomic features are sparse, there are many genome, capture variation on a graf, and one more very important challenge that we should add to those, is that there is a computational analysis on one side, and the human interaction on the other side. It should be as simple as user's understanding to be useful enough. It has been mentioned in this paper that today browsers have become a standard tool for exploring genomes, facilitating analysis of genome-anchored data, and providing a common platform for investigators to share, store and publish scientific discoveries, which in our research we are also trying to develop a web based interactive visualization.

Title: X-Chromosome epigenetic reprogramming in pluripotent stem cells via noncoding genes
Authors: Daniel H. Kim, Yesu Jeon, Montserrat C. Anguera, Jeannie T. Lee
Publication & Year: Seminars in Cell and Developmental Biology, 2011

Summary:
This paper is a discussion paper on how studying X-chromosome status can provide actionable indication of the epigenetic (relative to or arising from nongenetic influences on gene expression) quality of pluripotent (not yet mature) stem cells. The experiments described involved mice, however, there is a tight linkage found that is likely to exist in humans as well. This paper presents an interesting research question – given the confirmed relationship of X-chromosome state and pluripotent step cells in mice, how can we further understand how and why the X-chromosome status can affect stem cell development in humans? Further research in the area would be of utmost importance to progress in stem cell therapy and regenerative medicine, which depend on access to pluripotent cells.

Relation to Visualization:
This paper's content does not fall directly in line with the work being done by the Kim Lab for which we are to create a visualization. However, as a prominently cited publication from Dr. Kim, we felt it was important to survey and take note of the visualizations used. Interestingly, there were no visualizations used within the paper to enhance understanding of the presented hypothesis. Some simple visualizations were referenced, which were attached at the end of the paper. Some of these visualizations are below, included with their original captions:

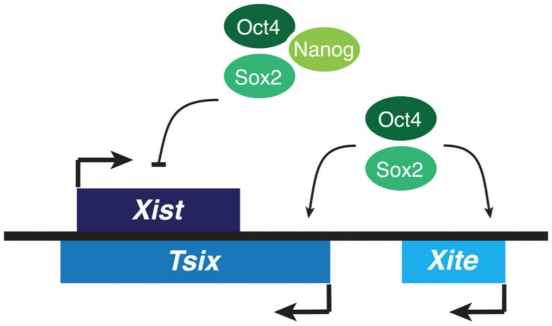


Fig. 1.
Xist regulation by the core pluripotency factors. Oct4 and Sox2 bind the noncoding *Tsix* and *Xite* loci, upregulating the expression of *Tsix*. Xist levels are also controlled by direct binding of Oct4, Sox2, and Nanog to *Xist* intron 1.

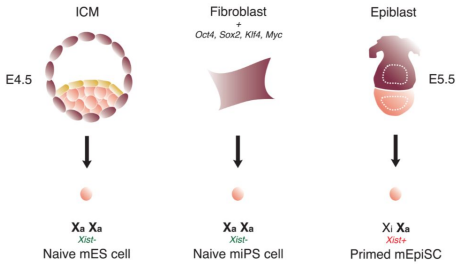


Fig. 2.
X-chromosome state in mouse pluripotent stem cells. Naïve mES and miPS cells represent the ground state of pluripotency, as evidenced by the presence of two active X-chromosomes and the absence of Xist RNA. Primed mEpiSCs have already undergone X-chromosome inactivation and represent a developmentally more advanced state.

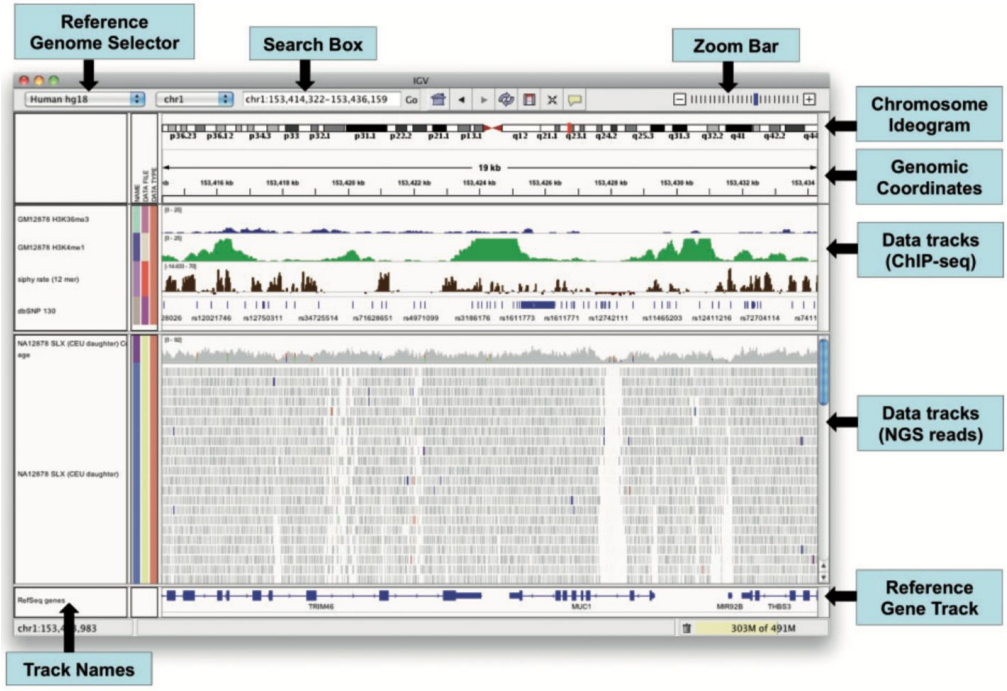
Takeaways:
The visualizations used in this paper were not very complex, as they mostly were aimed at summarizing simple hypotheses and relationships. However, we noticed that some of the actual markers used were either not helpful, or just not intuitive. In Figure 1 (attached above) we see a very simple visualization likely made in a word processor, that contained some bad alignment and unhelpful color schemes. While the information presented may be scientifically useful and well explained, the aesthetic choices for this visualization were not well organized. In Figure 2, shading is used in some of the graphic elements, does not add any value to the image, and instead creates an unusual artifact in places where outlines are blurred. This similar shading is repeated in other visualizations attached to the paper. Given the age of this publication, and the presence of the RasVis project, we know that visualizations are important to the Kim Lab, but it was interesting to explore some of the older works to see a progression in the style.

Title: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration
Authors: Helga Thorvaldsdottir, James T. Robinson, Jill P. Mesirov
Publication & Year: Briefings in Bioinformatics, 2012

Summary and Features of the Visualization Tool:
The large volume of data produced by genomics research poses a challenge in finding the best visualizations for the most effective analysis. The Integrative Genomics Viewer (IGV) is a free desktop application that can be used by all genomics researchers to analyze their own data and interact with publicly available datasets, allowing users to explore all levels of details in an easy and quick way.

This software was created as a direct response to the sheer amount and variety of data produced by current next-generation sequencing (NGS – a parallel processing method of analyzing large amounts of DNA) and array-based profiling methods in genomic research. While it is easy to automate much of the data processing, the integral role of the researcher remains: to analyze results and form hypothesis and conclusions. IGV serves to facilitate this task, giving

new and experienced users an intuitive platform with which to undertake analysis tasks in real time due to the novel way in which data is managed by the application. IGV uses data tiling, or the precomputing of data at various resolutions, while pushing the task of final rendering to runtime, to efficiently handle large amounts (up to terabytes and gigabytes) of data at a single time.



Another unique feature of IGV is the flexibility in file formats accepted – the software can handle any data that can be mapped into standard genomic coordinates. This greatly reduces the need to preprocess large datasets into very specific arrangements, and allows instant upload and immediate visualization of results. There are multiple viewing resolutions available, but the main IGV application window display is shown above, with labels showing the various features available to the user. State of the application can also be saved, letting users return to a previous configuration without needing to load all the data again.

At the time of publication there were some planned future goals for the IGV project. The continual expansion in the amount of data available for analysis means that loading data and managing memory is a constant challenge, especially since the purpose is to allow access to both overall aggregation and display, as well as details upon further interaction. Both applications require useful visualizations such that researchers can immediately draw insights. Additionally, novel formats of visualization will be necessary to highlight specific data involving pathways and other genomic data that can be isolated for further examination, including network views or other diagram formats. Lastly, the group intends to integrate some data driven methods of searching and exploration within the application.

Takeaways:

This paper directly dealt with visualization as a challenge in the genomics and bioinformatics community. Data generation in today's research labs is moving at a much faster rate than analysis, and the tools we use need to help close that widening gap and give researchers faster methods of reaching conclusions about their experiments. One of our biggest challenges, from discussions with graduate students at the Kim Lab, is dealing with different formats of data. IGV handles this problem relatively simply, by taking any file format that presents data that can be plotted in genomic coordinates. Following a similar model, we can aim to create a practical module that can pre-process any given data to match some specific labels, that we can then feed into our dynamic visualization generator. The next related challenge is that of isolating a single pathway or a single relationship within a large system to visualize for forming or explaining a hypothesis. This paper was published in 2012, and one of the highlighted future goals was to explore network views of relationships, specifically pathway data, which aligns exactly with the needs of both the Kim Lab and the Brooks Lab. Focusing on this area is therefore a good direction for us in looking for novel ways to develop visualization in a field that requires intuitive models quickly.

Title: Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser
Authors: Melissa S. Cline, Brian Craft, Theresa Swatloski, Mary Goldman, Singer Ma, David Haussler & Jingchun Zhu
Year: 2013

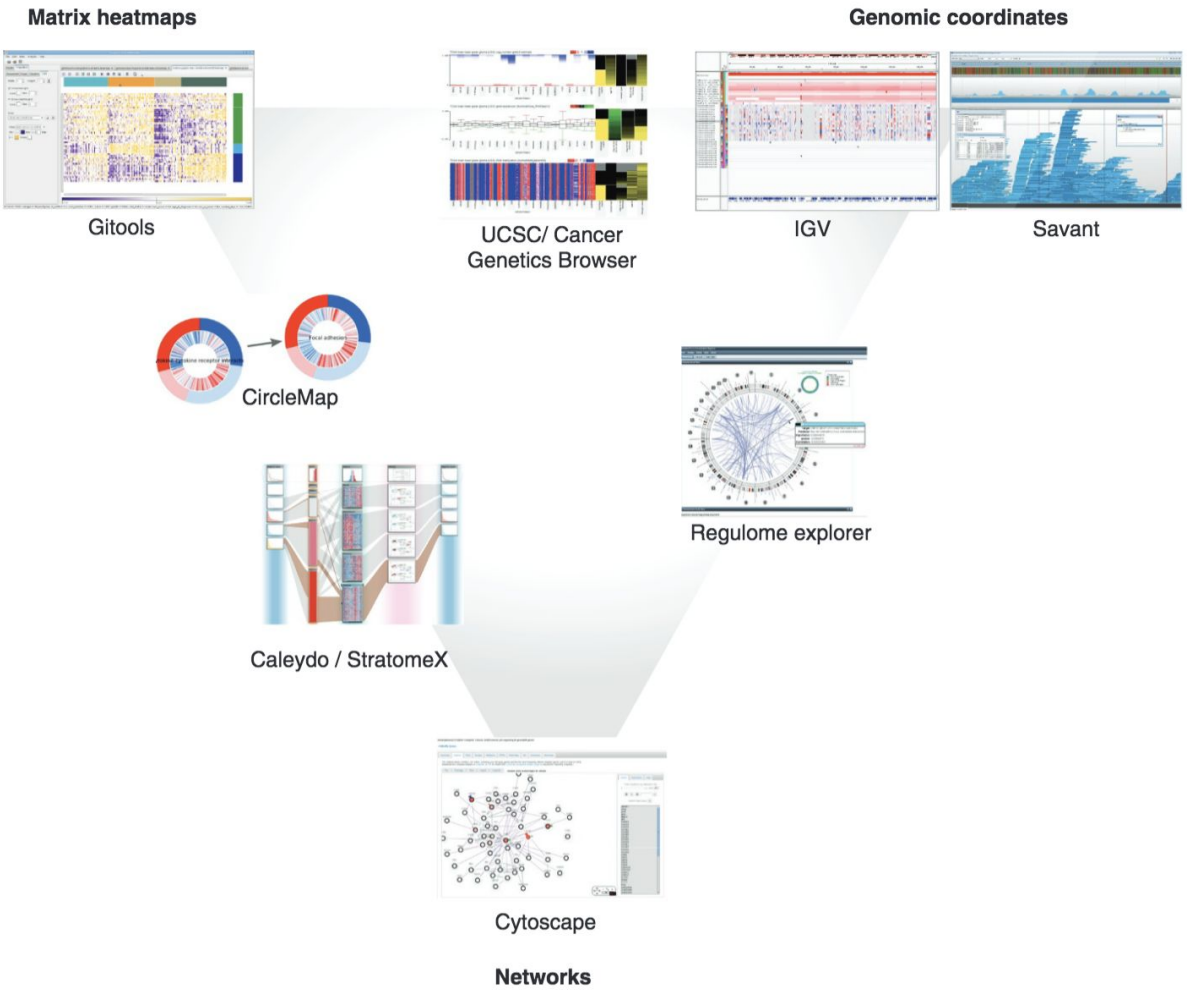
Summary and how it relates to visualization:
This is a web-based interactive visualization and exploration of cancer genomic data. There are different choices for cancer genome visualization tools. Genome-based, gene-based, pathway-based. The genome-based visualization only allows user to see one or two genomic regions at once, but they are not as effective for exploring possible connections between alterations in multiple different genomic regions. The UCSC Cancer Genomics Browser allows prob- based visualization in its gene-based viewing mode. It has heatmap-based methods which do not generally indicate how genes interact in a pathway. This visualization is effective for cancer analysis. This browser provides direct access to and visualization of data at specific genes or genomics.

This visualization has different methods including Genomic data and views, which displays genomic, clinical and annotation data in multiple view, such as heatmap, proportion, and boxplot. Annotated data, which each genomic dataset has a number of annotation field associated with it. Each annotation column is color-coded by its contents. Controlling the display, which user can zoom into the display in both horizontally and vertically. By clicking and dragging the mouse horizontally across the data, user can zoom into selected gene or genome regions. By doing it vertically user can zoom into into selected samples. Also user can resize the display of any dataset. Subgroup and online statics, which allow users to group samples. Genomic signature, which is algebraic expression over a set of genes. User can also make more complex signatures. Annotation upload and download, which user can upload their own sample as custom data. Kalpan-Meier plots, which users can choose either recurrence free survival or overall survival. Bookmarks, which provide users with hyperlinks to save the state of the browser and share analysis insights with other.

Takeaways:
What I have found useful in this paper is that it dealt with interactive data visualization in this subject. It discussed different methods of genomic visualization. Most of the biological data visualization are static. The existed data visualization for Kim's Lab is also static and it was for the publication. These are highly useful categories as sample that we can explore them for our interactive data visualization to make a well-designed data visualization with user-friendly and simple interactions.

Title: Visualizing multidimensional cancer genomics data
Authors: Michael P Schroeder , Abel Gonzalez-Perez and Nuria Lopez-Bigas
Year: 2013

Summary and how it relates to visualization:
In this paper they have been mentioned various data visualization tools have been developed in recent years to support genomic studies. They revisit the most common ways in which these data are visualized, and present selected tools that allow researchers to visualize multidimensional oncogenomics datasets effectively. They gathered different tools and resources for visualizing multidimensional cancer genomics data in a table in detail, which have different types of visualization including: Matrix Heatmap, Genomic Coordinates, and Networks. This table also help us to explore different existed multidimensional cancer genomics visualizations:



In this research they used four case studies. The description of the case studies focused on their biological interpretation. One case studies in this research have been Visual exploration of cancer drivers which discuss the Heatmap of oncogenomic alterations ordered by mutual exclusivity plotted with Gitools, another case study with the same data with the same color represented as a network of functional interactions between the genes, extracted from the cBio Cancer Genomics Portal, Heatmap of pathway expression levels plotted with Gitools, CNA and expression data for the EGFR gene region of glioblastoma samples as shown by IGV, and Adaptations of Circos plots of three breast tumors with three very different alteration landscapes. The next case study is Visualizing cause-effect relationships between different types of alterations, Visualizing cancer patient stratifications, and Visualizing global alteration profile patterns. For the interfacing the data visualization tool, because of the landscape of the multidimensional data seems fragmented, they have tried different tool types including web tool, desktop application, and command line application. In the subject of cancer genomic data the amount of the data it is possible to generate for an oncogenomics project continues to increase, requiring visualization tools that very efficiently load and process large amounts of data. The complexity of oncogenomics data and the multitude of questions to be addressed ensure that a static plot is often insufficient for data visualization. The user needs to explore the data interactively in order to address a wide range of questions. Several tools listed in

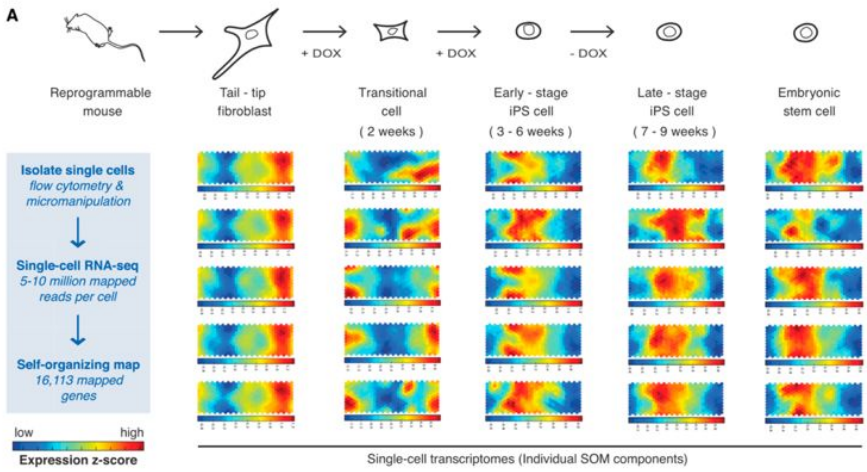
this paper including IGV, Gitools and Caleydo make us of interactive visualization techniques to make this possible. Other web frameworks with various visualization and some optional analysis possibilities are being developed, including the cBio Cancer Genomics Portal, IntOGen and Regulome Explorer. Open source and plug in architecture facilitates quick adoption of these new platforms. The important efforts have been made in recent years to create visualizing multidimensional cancer genomics data in visualization tools that can explore multidimensional genomic datasets. Further efforts are needed to develop those resources and to create new tools to meet the changing needs of the field.

Takeaways:
It was very useful to have a small collection of genomic related visualization in this paper, and be able to not only check the different visualization tools and methods, but also be able to compare them. It gave us a view of the 2D visualization in this subject The interesting point of this paper is in all of these visualizations they have used web tool, desktop application, and command line application which are two dimensional. And the question on this paper could be what is the reason that most of the existed data visualization in multidimensional genomic data are 2D. Is that because of the complexity of the data? How can we approach to enhance the existing visualizations of biological pathways to represent large pathway relationships in 3D platforms.

Title: Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming
Authors: Daniel H. Kim, Georgi K. Marinov, Shirley Pepke, Zakary S. Singer, Peng He, Brian Williams, Gary P. Schroth, Michael B. Elowitz, Barbara J. Wold.
Year:2015

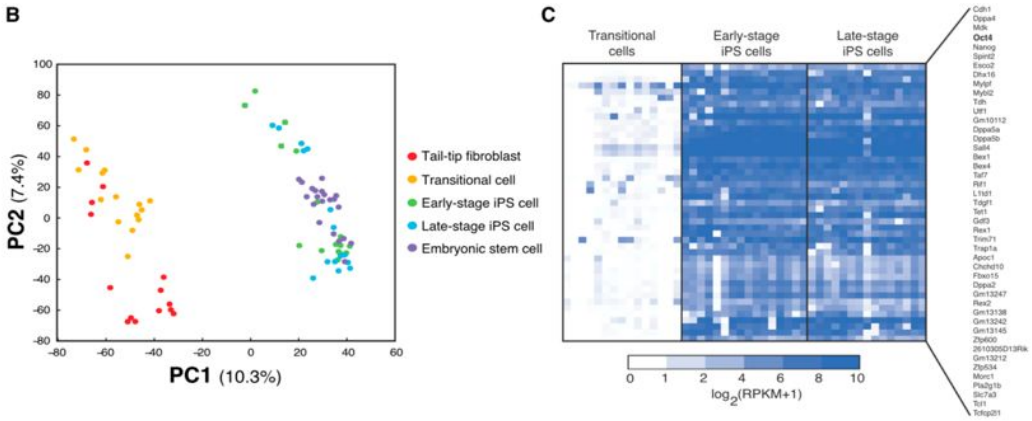
Summary and how it relates to visualization:
This is a paper that is co-authored by Daniel Kim at UCSC and is an example of the kind of work done in his lab. While the paper focuses more heavily on the biology and scientific experiments that are typical of work done in his lab, the paper also presents some static visualizations. These visualizations are characteristic of bioinformatics research and provide examples of the types of data and kinds of visualizations used by biologists and cancer researchers. For example, throughout the paper, the authors frequently use self-organizing maps (SOMs) to structure transcriptome data at the single-cell level.

The paper focuses on Long noncoding RNAs (lncRNAs), which are useful in epigenetic regulation. In this paper the authors use single-cell RNA sequencing to characterize the expression patterns of over 16,000 genes. Single-cell RNA sequencing can help gain transcriptome-level understanding of how individual cells are reprogrammed, which is used to study the plasticity of the somatic cell state. In one experiment, the authors performed single-cell RNA sequencing on cells from the “reprogrammable” mouse. They performed some wet-lab experiments with cells from the mice’s tails and generated data about the expression patterns of over 16,000 protein-coding and noncoding genes, including lncRNAs. This high-dimensional data was analyzed using a SOM, as shown in the figure below:

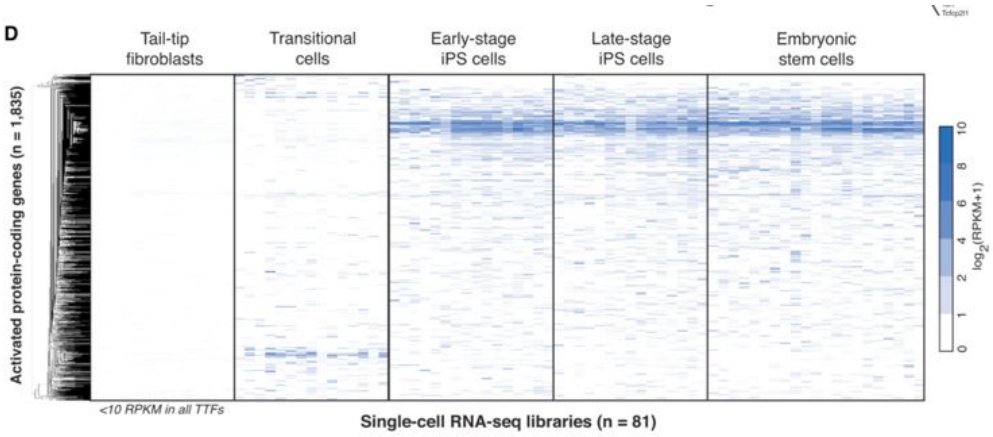


This visualization was used to provide a way to display similarity relationships via a heat map. Spatial proximity represents similarities in expression patterns. Each single-cell transcriptome can be visualized as a component of the SOM, so in this example there are five representative single-cell components shown for each of the five cell types.

The paper also highlights visualizations that arise out of other experiments common in biology labs. For example, this paper mentions that in order to examine how the individual transcriptomes from different points in time were related, the authors conducted principal component analysis (PCA). This led to the visualizations depicted below:



Another method common in this area is hierarchical clustering of genes. An example of how that could be represented in a visualization is included below. In this example, the group includes 1,835 protein-coding genes.



Takeaways:
This paper is one example of how many visualizations in biology and bioinformatics are static. They are primarily used for publications, rather than for data exploration. Also, this paper shows the variety in both the types of data gathered by scientists like Daniel Kim and Angela Brooks. Not only is the data extremely varied, but also the visualizations differ in how they represent the data. There are not many single visualization systems that have the ability to represent multiple datasets. It seems as though each dataset has its own type of visualization, such as a heat map or a scatterplot. In addition, these visualizations are not dynamic, so they do not have the ability to be easily altered when new data is inputted. As we see in other papers and as is supported by our conversation with two of Daniel Kim's graduate students, these themes and others surrounding visualization of complex data come up regularly in biology and cancer research.

Title: A Taxonomy of Visualization Tasks for the Analysis of Biological Pathway Data
Authors: Paul Murray, Fintan McGee, Angus G. Forbes
Year: 2016

Summary and how it relates to visualization:
This paper presents a taxonomy for visualizing biological pathway data based on interviews with expert biologists. Biological pathway data represents chains of interactions that occur in a biological process. In Daniel Kim's lab, the two graduate students we talked to explained that they study the kRas protein and how it drives mutations in other cells, creating a cascade that leads to cancer. This is an example of something that can be modeled using a biological pathway.

Node-link diagrams are often the first choice used to represent biological pathway data in existing visualizations. An example of this is type of representation is Cytoscape, an open source bioinformatics platform for visualizing molecular interaction networks (https://cytoscape.org). However, there is evidence that other visualization techniques, such as matrix visualizations, may be more optimal than node-link diagrams in certain cases. This biological pathway data is stored in many formats such as BioPAX and KEGG.

Before presenting the taxonomy, this paper highlights some of the reasons why it is important to study how to create effective biological pathway visualizations by explaining that there are several reasons why this data is difficult to visualize. Data may contain thousands of points that are connected by many different relationship types, such as feedback loops and cascades. Most visualizations used in bioinformatics papers are static visualizations. Yet, the complexity and size of biological datasets often make these visualizations appear cluttered and difficult to interpret. Therefore, it is extremely important for people working in the visualization community to be able to understand the analysis tasks important in bioinformatics work, so that they can better help biologists and cancer researchers disseminate the results of their experiments.

The interviews conducted to generate the taxonomy for biological pathway visualization were free-form and used to understand the research process of the researchers interviewed, as well as the tasks they perform when analyzing data and the structure of their datasets. The resulting taxonomy assigns analysis tasks to three categories named Attribute, Relation and Modification. Below is an image from the paper showing the categories and subcategories of the taxonomy:

Table 2 A summary of the biological pathway visualization task taxonomy

Category	Example task
Attribute tasks	
(A1) Multivariate	Find all up-regulated genes in a biological pathway. Integrate results of a laboratory experiment into existing protein-protein interaction networks.
(A2) Comparison	Compare a biological pathway to a pathway with the same functionality in a reference species.
(A3) Provenance	Determine which studies provides the evidence for a link between two genes.
(A4) Uncertainty	Understand which pathway components have the strongest empirical evidence relationships.
Relationship tasks	
(R1) Attributes	Find all translocations of entities in a given biological pathway.
(R2) Direction	Find the products or output of a biochemical reaction.
(R3) Grouping	Expand a module entity to include all child-entities in the visualization.
(R4) Causality	Find all genes downstream of the currently selected entity, which may be affected by a change in regulation.
(R5) Feedback	Identify potential feedback loops in gene regulation.
Modification tasks	
(M1) Annotate	Update out-of date-information in a pathway data set, or create a personalized pathway relevant to a specialized research topic.
(M2) Curate	Identify errors and update historical data.

In describing the different components of this taxonomy, the authors of this paper mention that it is important for biology researchers to be able to visualize complex data while viewing a pathway and being able to view extra additional external resources related to a specific data point. Other important components of future visualizations should include ways to relate similar pathways or compare a single pathway in different states. Existing approaches of comparative visualizations include juxtaposition, superposition and explicit encoding of differences using color. Additionally, some of the biologists interviewed in this paper also highlighted the important of being able to look at the history of original sources related to their data and being able to understand and visually represent uncertainty.

Takeaways:

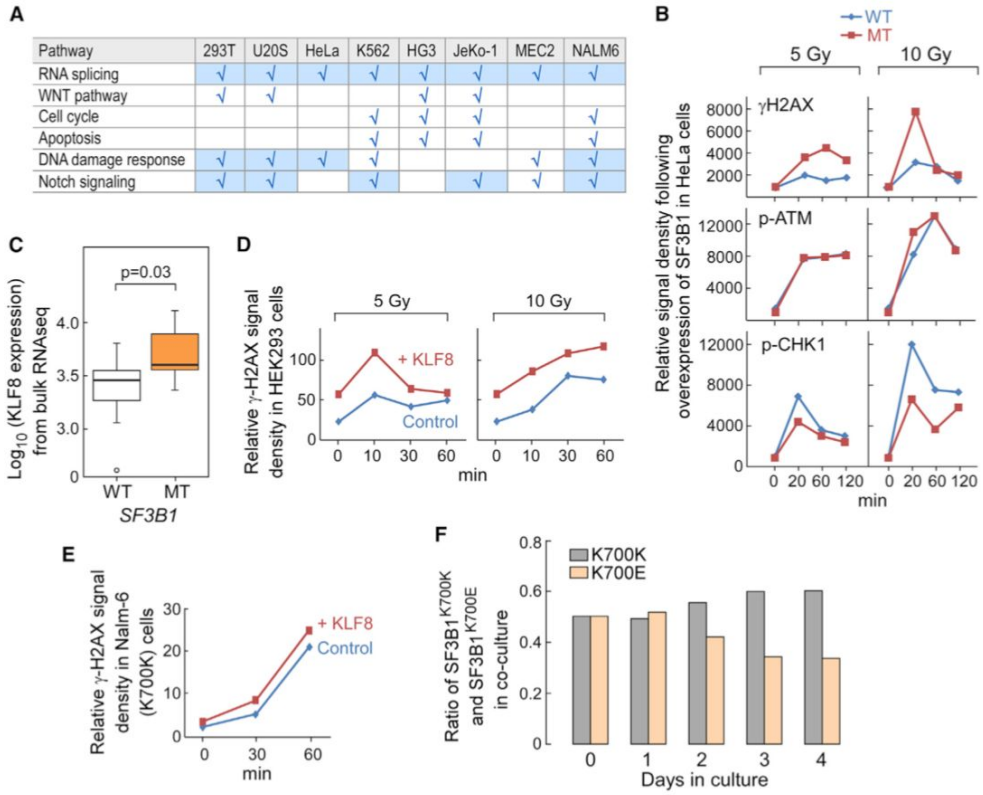
One of the things we found interesting about this paper is that many of the suggestions and themes presented surround biological pathway visualizations also came up in our meeting with the graduate students in the Kim lab. This tells us that even though more work in visualizing biological pathways has been done since this paper was published, there is still a lot more work to be done, and that our project could be built on in the future to develop a highly useful data exploration tool for this type of work. Additionally, this paper lays the foundation for some of the challenges that motivate the Xena platform paper. The Xena platform was designed and developed to address some of the problems biologists and cancer researchers have while visualizing and exploring complex datasets, many of the same problems presented in this paper.

Title: Transcriptomic Characterization of SF3B1 Mutation Reveals its Pleiotropic Effects in Chronic Lymphocytic Leukemia
Authors: Lili Wang, Angela N. Brooks, Jean Fan, Youzhong Wan, Rutendo Gambe, Shuqiang Li, Sarah Hergert, Shanye Yin, Samuel S. Freeman, Joshua Z. Levin, Lin Fan, Michael Seiler, Silvia Buonamici, Peter G. Smith, Kevin F. Chau, Carrie L. Cibulskis, Wandi Zhang, Laura Z. Rassenti, Emanuela M. Ghia, Thomas J. Kipps, Stacey Fernandes, Donald B. Bloch, Dylan Kotliar, Dan A. Landau, Sachet A. Shukla, Jon C. Aster, Robin Reed, David S. DeLuca, Jennifer R. Brown, Donna Neuberg, Gad Getz, Kenneth J. Livak, Matthew M. Meyerson, Peter V. Kharchenko, Catherine J. Wu
Publication & Year: Cancer Cell, 2016

Summary:
Mutations in the gene SF3B1 are known to affect the progression of Chronic Lymphocytic Leukemia (CLL) cells in humans. This paper identifies specific pathways affected by the mutation to isolate why and how they affect these cells. The study yielded the conclusion that the mutations of SF3B1 is not the main cause of CLL progression, but they do aggressively accelerate CLL cell development in the case when there already exist some other cancer-driving alterations to DNA. The experimental procedure included analyzing blood samples taken consensually from human donors and patients, and data analysis was done alongside usage of some large databases and software libraries including JuncBase, DESeq2, and SCDE. JuncBase is used for identification of alternative splicing events (a process used during gene expression to code multiple proteins with a single gene) from RNA sequencing data. DESeq2 and SCDE are R packages that aid in gene expression analysis.

Relation to Visualization:
Most of the visualizations used in this paper were static images displaying some sort of statistical analysis done on various splice sites of the CLL cell or gene expression data in SF3B1 mutations. The diagrams were highly technical and require a significant amount of explanation to understand or read them. This paper was co-authored by Angela Brooks, and is relevant to the current research being done at the Brooks Lab. Statistical plots generally have a consistent layout, and many comparisons are done with plots with identical axes, indicating a potential use for 3D and dynamic visualizations that could accelerate evaluation. A specific visualization idea discussed with Angela Brooks was transitioning to a network visualization of pathway data, or some graphical icon usage instead of a simple and somewhat non-intuitive chart like one

used in the paper (A), and attached below. There are also some other examples of repetitive and technical diagrams for which new styles of visualization could be explored.



Takeaways:
Statistical plots are relatively standard in Bioinformatics, but being able to draw conclusions from them requires a certain amount of training. While this is expected within the field, it is worth considering new forms of visualization that can help people from outside the field to understand them quicker, or at least at a high level. From our discussion with Angela Brooks, there is an interest in exploring such models, especially those that provide the greatest amount of information on specific relationships without requiring so much explanation. While most of our ideas for creating new visualizations for statistical data analysis had to do with 3D and dynamic exploration, these ideas do not translate well to the space of paper and publications, in which static diagrams are somewhat required. It raises an interesting long term question on the nature of publications - how can 3D and dynamic models be represented in traditionally 2D literature? Practically, it makes sense that generation of new 2D diagram styles will be more immediately useful, but in a future sense it is worth exploring the benefit of 3D visualizations at least in the space of researchers analyzing their own data sets to gain insights.

Title: The UCSC Xena Platform for cancer genomics data visualization and interpretation
Authors: Mary Goldman, Brian Craft, Akhil Kamath, Angela Brooks, Jing Zhu, David Haussler

Year: 2018

Background and Goal of visualization:
UCSC Xena is a web-based visualization tool for multi-omic data as well as clinical and phenotypic annotations. It allows for visual integration as well as data exploration. Multi-omic refers to a process that combines "ome" data such as genome and transcriptome to analyze complex biological datasets (Wikipedia). The tool consists of a web-based Xena Browswer and multiple Xena Hubs that contain databases that connect to the Xena Browser simultaneously. Xena is used to interpret cancer genomics datasets, either the datasets that are provided by the visualization tool itself, or private datasets imported by individuals or labs.

Xena was designed to tackle several problems related to bioinformatics and genomics data, although it is targeted towards cancer researchers. The first challenge in this area is that the growing amount of genomics data also includes an increase in the different types of data, or data modalities, that are represented. Some of the most common types of data are those on somatic mutations, copy number and gene expression. Each modality provides unique information about the genome, so it is important to be able to interpret different kinds of data in order to have a more comprehensive understanding of a tumor's genomic events. Another difficulty in this space is that the data is highly distributed. There are very large data sets such as The Cancer Genome Atlas (TCGA) and Genomic Data Commons (GDC) that are used by researchers all over the world. Additionally, many researchers generate their own smaller data sets to go along with these larger datasets. Therefore, it becomes difficult to share data with people in other parts of the world and there is no way to easily connect the data.

How does the visualization work:
Xena consists of two main components: the web-based Xena Browser and the backend Xena Hubs. The Xena browser is what allows biologists and other users to explore genomics data, which is stored in the Xena hubs. Xena browser is a javascript application, primarily written using React. Xena Hub is written in Clojure and serves the data over HTTP. Xena hosts data from some of the most prominent cancer genomics databases including TCGA, International Cancer Genome Consortium (ICGC) and GDC in seven public Xena hubs that together host 1557 datasets from more than 50 different cancer types. However, there is also the option for users to use their own data by installing their own Xena hub and configuring the settings to make the data public or private. Xena was designed using a two-part system in order to facilitate using both public and private data together while keeping private data secure and only visible to the intended user. This also makes it easier to add more data because as more data is added, more Xena hubs are created but the overall system architecture remains the same.

Xena supports the following types of data: somatic and germline SNPs, INDELs, large structural variants, copy number variation, gene-, transcript-, exon-, protein-expression, DNA methylation, ATAC-seq peak signals, phenotype, clinical data, and sample annotations. It supports queries on thousands of samples and can return slices of genomic and clinical data within a few seconds, which is claimed to have been a challenge for many prior tools.

Features and reasoning:

The visualizations available in Xena include the Xena Visual Spreadsheet, survival analysis, scatter plots, bar graphs, statistical tests and genomic signatures. Some of these visualizations are also integrated with the UCSC Genome Browser (https://genome.cshlp.org/content/12/6/996.full.pdf+html). It is also formatted in a way that promotes collaboration, through pdf downloads of the visualizations themselves and sharable bookmarks used to highlight important and interesting visualizations to collaborators. Xena also provides links to examples of visualizations created using the tool.

The Xena Visual Spreadsheet is designed to facilitate viewing different types of data side by side. This allows for a more biologically complete understanding of a specific genomic event. This feature is a grid where each column is a slice of genomic or phenotypic data such as gene expression or age, and each row is a single entity, such as a cell line or tumor sample. The Xena Visual Spreadsheet shows genomic data in both gene-centric and coordinate-centric views. In gene-centric views, data is mapped to a gene or part of a gene and in coordinate-centric views, data such as copy number variation, simple mutations and DNA methylation is displayed.

Xena also provides a Kaplan-Meier analysis that displays a visualization showing survival data. It also provides bar charts, box plots and scatter plots that automatically compute chi-squared, t-test and ANOVA. Another possible view is the Transcript View, which enables users to compare transcript-level expressions between two different groups of samples for all transcripts of genes in that sample. Xena also has a text-based search, analogous to Microsoft Word's "find" capability. This allows users to highlight, filter and group samples to focus the visualization on the samples they are interested in exploring more in depth.

eVip Visualization

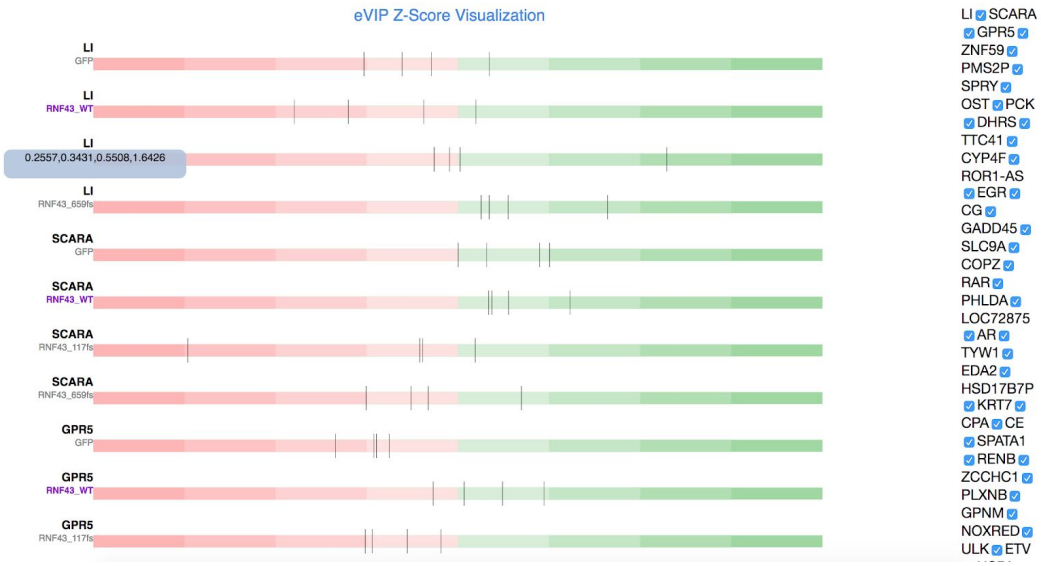
We worked with one of Angela Kim's graduate students, Alexis Thornton, to create a visualization for her eVip data. This data consists of gene expression data and is currently used by Alexis as input into an unspecified additional piece of software to generate scores that are used to create an existing visualization that represents the average difference in gene expression between all genes. The eVip data consists of several gene ensemble IDs that represent different genes, along with 4 replicates for each gene: RNF 43_659fs (longer variant), RNF 43_117fs (shorter variant), RNF 43_WT (wild type version of the gene) and GFP (a control). Each of the replicates for each gene contain a z-score, which is important to Alexis in comparing the different replicates of a specific gene. Even though this data is not immediately useful in creating Alexis's existing visualization, she said that being able to easily compare the z-scores may be useful in interpreting and understanding the results from her work in the existing visualization for gene expression. To help in understanding the type of data we were working with, we have included a picture of a segment of the data below.

id	RNF43_WT_4	RNF43_WT_3	RNF43_WT_2	RNF43_WT_1	GFP_4	GFP_3	GFP_2	GFP_1	RNF43_659fs_4	RNF43_659fs_3	RNF43_659fs_2	RNF43_659fs_1	RNF43_117fs_4	RNF43_117fs_3	RNF43_117fs_2	RNF43_117fs_1
ENSG00000198242.13	0.3024	0.1094	-0.1280	0.9370	-1.0952	-0.1731	-0.4922	-0.1250	-0.0241	0.6749	0.5268	0.5824	1.8746	-2.2823	-1.4579	0.7703
ENSG00000134108.12	-0.6278	0.0309	-0.7000	0.2914	1.5018	1.3131	1.5731	1.1274	-1.3516	-0.8480	-1.4773	-0.7508	0.7395	-0.3860	0.1154	-0.5510
ENSG00000276644.4	0.1284	0.1917	-0.7127	-0.3272	0.8217	0.6441	0.6924	0.6964	-0.1575	0.6385	-0.2218	0.0816	-2.9864	-0.6416	-0.3711	1.5234
ENSG00000182141.9	0.6693	0.8966	0.7770	0.0328	-0.0337	1.1870	1.6086	-0.2424	-1.6091	0.1856	-0.6233	0.8666	-1.9700	-0.9853	-0.0768	-0.6828
ENSG00000167578.17	0.0239	0.0782	-0.4171	-0.3919	0.7666	1.1346	0.1677	1.2442	0.3945	-0.0669	0.2587	-2.8328	1.3020	-0.1215	-0.5883	-0.9522
ENSG00000236830.6	0.1872	-1.2277	-1.5023	-0.7278	2.2592	-0.8537	0.3215	0.4669	0.3723	-0.0548	0.2122	-1.7337	0.4447	0.5792	0.6017	0.6551
ENSG00000197557.6	-0.5834	0.2394	-0.6555	0.0845	0.7152	0.9320	0.6840	0.9811	-1.6818	-1.1047	-1.7513	-0.9075	-0.1384	1.1062	1.1052	0.9752
ENSG00000278616.1	-0.5913	-0.4719	-0.9343	-1.0241	1.6230	0.9415	1.1917	0.6486	0.7307	-0.1839	-0.3941	-0.0359	1.7004	-1.3294	-0.8607	-1.0102
ENSG00000146083.11	-1.1762	-0.3973	-0.5637	-1.2513	0.5952	0.4836	-0.8358	0.9851	-0.3585	-0.0222	0.1311	0.0518	2.9726	-0.4025	0.1661	-0.3780
ENSG00000070087.13	0.5583	0.4926	0.6907	0.9418	0.7055	0.6186	0.5121	0.7888	-0.1122	-0.3974	-0.0269	-0.4356	-3.1614	-0.9351	-0.2903	0.0504
ENSG00000204946.9	0.4682	0.1514	-0.0235	0.3067	-1.3004	0.0918	-1.5685	0.5789	-0.9662	-1.3218	-0.3165	-0.2194	2.2107	-0.2060	1.3524	0.7621
ENSG00000153561.12	-0.8895	-0.1961	0.1207	-1.3216	0.7698	0.3322	0.9056	-0.1234	0.4035	0.7861	0.0699	0.2696	-2.8987	1.0116	0.7660	-0.0056
ENSG00000179262.9	0.2951	0.0560	-0.5053	-1.3211	-0.2743	-1.4607	-0.0573	-0.9790	-0.0754	-0.4887	-0.0454	0.3629	2.8866	0.5242	0.8305	0.2520
ENSG00000104833.10	0.3988	-0.9136	0.3576	-0.9307	0.3046	-1.0677	-0.0185	0.3693	-0.4647	1.4283	-0.5650	-0.8339	2.8339	-0.3161	-0.4801	-0.1024

Design Process

For the visualization we used D3 and JavaScript. Our visualization runs on a Chrome webpage. We created a Python parser that takes in the eVip data (a txt file) and parses it into a JSON file which is then used to create the visualization. Initially when designing the visualization we wanted to develop a single network visualization for both the RasVis dataset from Daniel Kim's lab, as well as the eVip dataset from Angela Brooks's lab. However, during a second meeting with Alexis where we described our design ideas, she said that she did not see a good way of integrating her data with the RasVis data, and that a separate visualization tool would be more useful to her work.

Given this information, we set out to create a visualization based on a bullet chart. Our visualization design was inspired by Stephen Few's bullet chart design and the implementation of Clint Ivy, Jamie Love and Jason Davies in D3 [13]. We wanted an efficient way for Alexis to be able to see and compare the z-scores across the four gene replications. Our initial design idea consisted of a list of ensemble IDs that could be checked on or off in order to display and hide the data. We created a "bullet" line that represents each gene replicate, where each line contains tick marks that correspond to the z-scores for that replicate. We also highlight the WT replicate, since that is the non-mutation gene replicate among the four and the one against which the other three gene replicates are compared. Below is an image of our final visualization:



Feedback and Design Iterations

The figure above depicts our final visualization design. Prior to this design, we met with Alexis to obtain feedback on our working prototype, and some of her ideas were incorporated into the final product. Ideas that we were not able to modify in time for submission are left as future work and will be further discussed in the Future Work section. In this section we comment on the feedback that was used to modify our design.

In our first iteration prototype instead of listing the gene names in the right column we listed the the ensemble IDs that were provided in the original dataset. However, when we met with Alexis she mentioned that the users who would interact with the visualization would not know what the ensemble IDs mean, so it would be more helpful to use the gene names instead of the IDs. In addition, our first iteration prototype listed the four replicates of each gene in a different order than in the image above. Alexis suggested that we show the replicates in the depicted order, since that would facilitate her comparisons. She also provided feedback on the aesthetics of our design. Originally we had the color gradient for each line going from light read to dark red, followed by dark green to light green. However, Alexis said that she thought having the lighter colors closer to the middle of the line, the 0 mark, would be more intuitive to users.

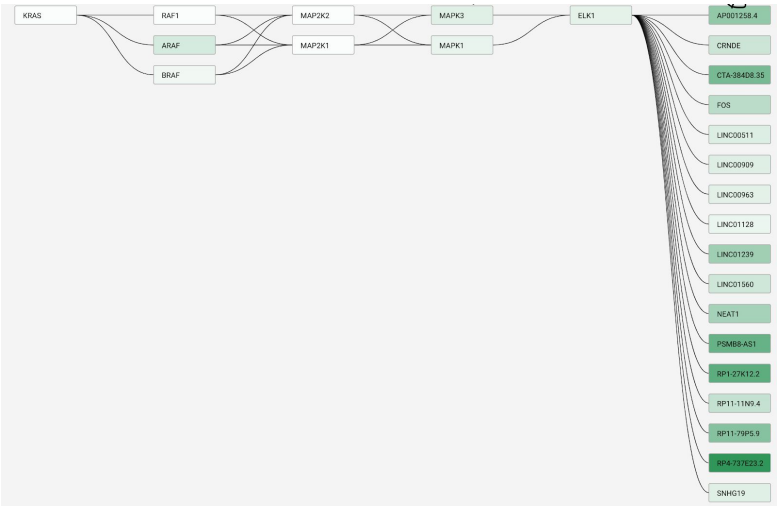
Our initial prototype allowed the user to select how many genes they wanted represented in the entire visualization by modifying a single value in our python script parser. However, Alexis told us that there are 81 genes that she is the most interested in, so we modified our visualization to only show the bullet lines for those 81 genes, in order to facilitate the use of our visualization for comparison purposes and eliminate the need for the user to modify the code.

Our final prototype also includes the ability for the user to hover over the gene and replicate name in order to view the z-scores corresponding to the lines marked on that particular line. The z-scores in the hover are listed in ascending order. We originally wanted to add the hover over each line marking. However, because a lot of the values are very close to each other it was

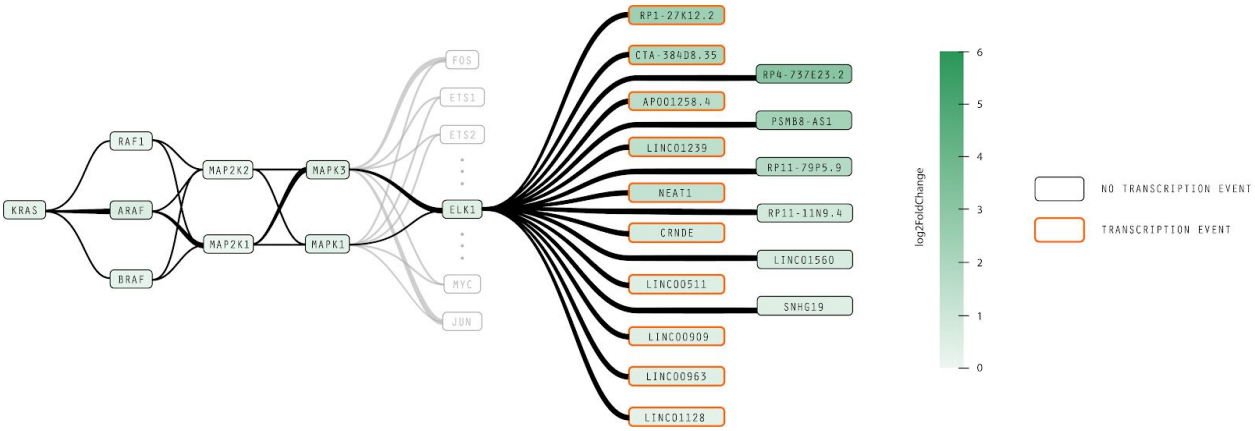
difficult to place the cursor directly over the thin tick mark, and making wider tick marks resulting in overlapping data that was no longer readable. In the future, we may change this by using a different scale on the bullet line that allows tick marks to be more spaced out.

Future Work
One of Alexis's suggestions that we were not able to implement but we thought was a good idea is to add spacing between every four lines to make it easier to distinguish between the replicates for each gene. In addition, we would have liked to implement a feature that makes it easier to compare different gene replicates when they are not placed next to each other in the visualization. Our current visualization generates the lines for each gene's replicates in the order in which they are listed in the JSON file. However, this makes it difficult to compare genes that are not next to each other, such as the first and last entries. In the future we could add a feature that allows the user to move the bullet lines around using the mouse so that they have control over the placement, or a feature that allows the user to select the genes they want to compare by clicking on the lines, and then viewing them in a pop-up comparison window.

RasVis Visualization

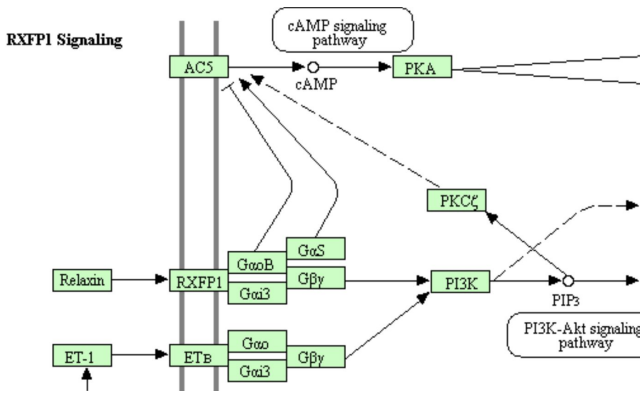


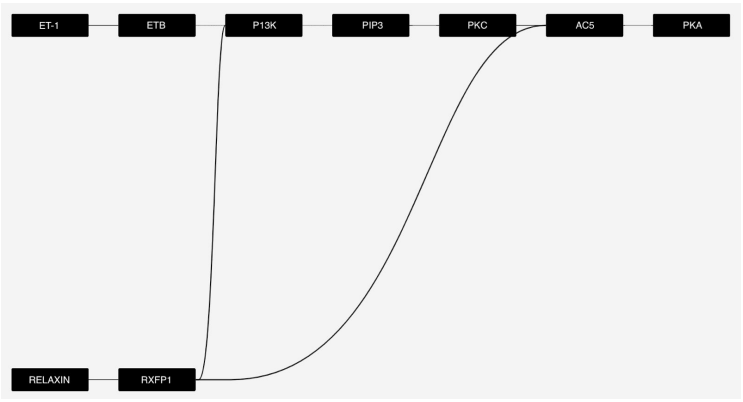
Design Process
The existing visualization for RasVis (pictured earlier and above in a non-centered format) has been created by an undergraduate designer Cassia Artenagara using Adobe Illustrator and D3 on manually created CSV data. This network-based visualization is a clean way to isolate a specific biological pathway and allow researchers to both understand their own data results and publish their findings and hypotheses. The existing visualizations, especially those created in Adobe Illustrator with additional design elements are primarily being used in publications and talks as informational diagrams. We aimed to enhance the visualization by testing on new datasets, and creating additional modules for automated data parsing and interactivity. Below is a different visualization with added design features.



Datasets

We received a new dataset for a different pathway from graduate students in the Kim Lab. The existing pathway visualization for the dataset (ET-1 and RELAXIN to PKA) was difficult to read and not very clear in aesthetic design. Using the D3 code for RasVis, we set up a simple network visualization for the pathway (colors and centering have not been implemented).





Future Work

A few features we are looking to improve upon include a more interactive module for editing network-based pathway visualizations. Given a visualization, the user should be able to click and drag to rearrange nodes, set certain links and nodes to different transparency settings, and add colored borders and other additional elements to the diagram. Additionally, the existing model is very dependent on manual data creation, and we would like to implement further functionality to allow quick generation from new datasets through the use of a data parser.

Conclusion

From our discussions with grad students in the Kim Lab and the Brooks lab we identified that our original idea of creating a single pathway visualization for the data across the two labs was not feasible. Instead, we mostly focused on creating a visualization for Alexis Thornton's (Brooks Lab) eVip data, since there is currently no existing visualization for this dataset. With our remaining time, we made some changes to the existing RasVis visualization, which will be continued in the Creative Coding lab in subsequent quarters.

Speculative Component

An interesting approach to enhancing existing network visualizations of biological pathways, such as RasVis, would be to represent large pathway relationships in 3D. Current bioinformatics papers represent these large datasets in flat diagrams that are often confusing to parse due to the large amount of represented pathways in spherical form. If however, we could take these same datasets and represent them in a 3D structure that could be interacted with on the web, the same data would become a lot more accessible to collaborators and other researchers seeking to gain insights. As these visualizations already have a network format, we could use a tabular representation of the data alongside A-Frame to create a 3D explorable network for pathway visualization. In the cases of massive datasets, having a VR interface could also be a useful next step in allowing exploration of various resolutions. However for the purposes of quick data exploration, a web-based interface seems to be the best solution.

References

[1]. Lili Wang et al. 2016. Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia. *Cancer Cell* 30, 5 (November 2016), 750–763. DOI:http://dx.doi.org/10.1016/j.ccell.2016.10.005

[2]. H. Thorvaldsdottir, J.T. Robinson, and J.P. Mesirov. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14, 2 (March 2012), 178–192. DOI:http://dx.doi.org/10.1093/bib/bbs017

[3]. Daniel H. Kim, Yesu Jeon, Montserrat C. Anguera, and Jeannie T. Lee. 2011. X-chromosome epigenetic reprogramming in pluripotent stem cells via noncoding genes. *Seminars in Cell & Developmental Biology* 22, 4 (June 2011), 336–342. DOI:http://dx.doi.org/10.1016/j.semcdb.2011.02.025

[4]. Melissa S. Cline et al.2013. Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser. *Scientific Reports*3, 1 (2013). DOI:http://dx.doi.org/10.1038/srep02652

[5]. Mary Goldman, Brian Craft, Akhil Kamath, Angela N. Brooks, Jingchun Zhu, and David Haussler. 2018. The UCSC Xena Platform for cancer genomics data visualization and interpretation. *bioRxiv*(2018). DOI:http://dx.doi.org/10.1101/326470

[6]. W.James Kent et al.The Human Genome Browser at UCSC. https://genome.cshlp.org/content/12/6/996.full.pdf html

[7]. Daniel H. Kim et al.2015. Single-Cell Transcriptome Analysis Reveals Dynamic Changes in lncRNA Expression during Reprogramming. *Cell Stem Cell*16, 1 (2015), 88–101. DOI:http://dx.doi.org/10.1016/j.stem.2014.11.005

[8]. Anon. 2018. Multiomics. (September 2018). https://en.wikipedia.org/wiki/Multiomics

[9]. Paul Murray, Fintan Mcgee, and Angus G. Forbes. 2017. A taxonomy of visualization tasks for the analysis of biological pathway data. *BMC Bioinformatics*18, S2 (2017). DOI:http://dx.doi.org/10.1186/s12859-016-1443-5

[10]. Cydney B. Nielsen, Michael Cantor, Inna Dubchak, David Gordon, and Ting Wang. 2010. Visualizing genomes: techniques and challenges. *Nature Methods*7, 3 (2010). DOI:http://dx.doi.org/10.1038/nmeth.1422

[11]. Keiichiro Ono. Cytoscape. https://cytoscape.org/

[12]. Michael P. Schroeder, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. 2013. Visualizing multidimensional cancer genomics data. *Genome Medicine*5, 1 (2013), 9. DOI:http://dx.doi.org/10.1186/gm413

[13]. Mike Bostock. Bullet Charts. https://bl.ocks.org/mbostock/4061961