Physician Focused Genetic Visualizations

Spencer King[*] Lev Morgan[†] Joshua Shapiro[‡]

CSE 557A Advanced Visualization
Washington University in St. Louis

Abstract

Year by year the cost of whole genome sequencing is dropping, increasing its viability in personalized medicine. The majority of existing visualization systems for genetic data are focused on researchers as opposed to physicians who may not have an in-depth genetics background. Genome data is complex and can be difficult to reason about without the appropriate background. While it's reasonable to expect any given physician to have some understanding of genetics, they likely won't have the depth of knowledge to explore the data set unaided. We propose and implement a visual analytics system that makes genome data more approachable to the average physician.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Visualization application domains—Visual analytics; Human-centered computing—Visualization—Visualization techniques—Visualization application domains—Scientific Visualization; Human-centered computing—Visualization—Visualization design and evaluation methods

1 Introduction

Whole genome sequencing began with the Human Genome Project in 1990 and the first full human genome was sequenced in 2003. According to the National Human Genome Research Institute the cost of genome sequencing was in the neighborhood of $100,000,000 in 2001 [7]. Over the past 20 years, the costs have fallen dramatically. Current estimates put the cost at roughly $1,000 per genome. This is a massive reduction in cost and is making precision or personalized medicine a more attractive research and business venture.

Precision medicine is a treatment strategy where physicians treat patients based on the specifics of their individual genetics. This potentially allows for more targeted and effective treatments, leading to better patient outcomes. For example, there is ongoing research in cancer treatment plans to determine the viability of personalized, genetic-based treatments [2].

At the moment, precision medicine is largely still in the domain of researchers. However, as the field continues to advance and the costs of whole genome sequencing continue to fall, precision medicine will likely be something for physicians to consider when treating their patients. The key issue is that the genome is quite complex and not approachable to someone without a background in human genetics. While all physicians will have some understanding of genetics, we believe it is unlikely that most will have the requisite depth of knowledge to fully explore and draw conclusions from a whole genome dataset. To address this concern, we performed a review of existing systems and implemented a prototype system of our own.

2 Related Work

This section presents some existing work in the field. The three main categories we have come across are Electronic Health Record integrations, tools for plant genetics, and tools focused on researchers.

2.1 Electronic Health Record Integration

Electronic Health Record systems (EHRs) are very widely used in hospitals and medical practices in the United States. It is reasonable to expect that much of the existing work in the field would be contained within these systems. Lau-Min et al. have detailed some of their work at Penn Medicine as they collaborated with Epic and Ambry Genetics to incorporate genomic data into their EHR [4].

Their data is coming from unstructured PDF files as opposed to the raw outputs from a genome sequencing tool or processing pipeline. In order to automatically extract useful information, Penn Medicine needed to enforce common procedures for generating these PDF files. These PDF files are the results from various genetic testing services and are not complete records of a patient's genome. In our opinion, this system is a good first step but misses out on advances in genomic research due to not interfacing directly with the patient's raw data.

2.2 Plant Genetics

There is a growing body of work focused on visualizing genomic data for crop scientists. While plants are obviously different from humans, many of the underlying file formats and concepts are the same. König et al. present BRIDGE, a system they designed to assist with exploring genetic diversity amongst plants [3]. They demonstrate their system visualizing a large database for barley plants.

BRIDGE is a web-based tool that allows for filtering on a wide variety of parameters such as phenotype, country of origin, and others. The system reads and writes commonly used genomic file formats and allows researchers with the appropriate background to dive deep into the data. While some of the features are specific to plant biology, this is the only publicly available system we have seen addressing similar problems as our system.

2.3 Research Tools

There are a wide variety of research tools available to geneticists. These provide a great deal of valuable information, however they are largely not approachable to individuals without a strong background in genetics. We do not provide a review of these tools here as we feel this is outside the scope of our system, however we did want to note there is a wealth of existing work for geneticists.

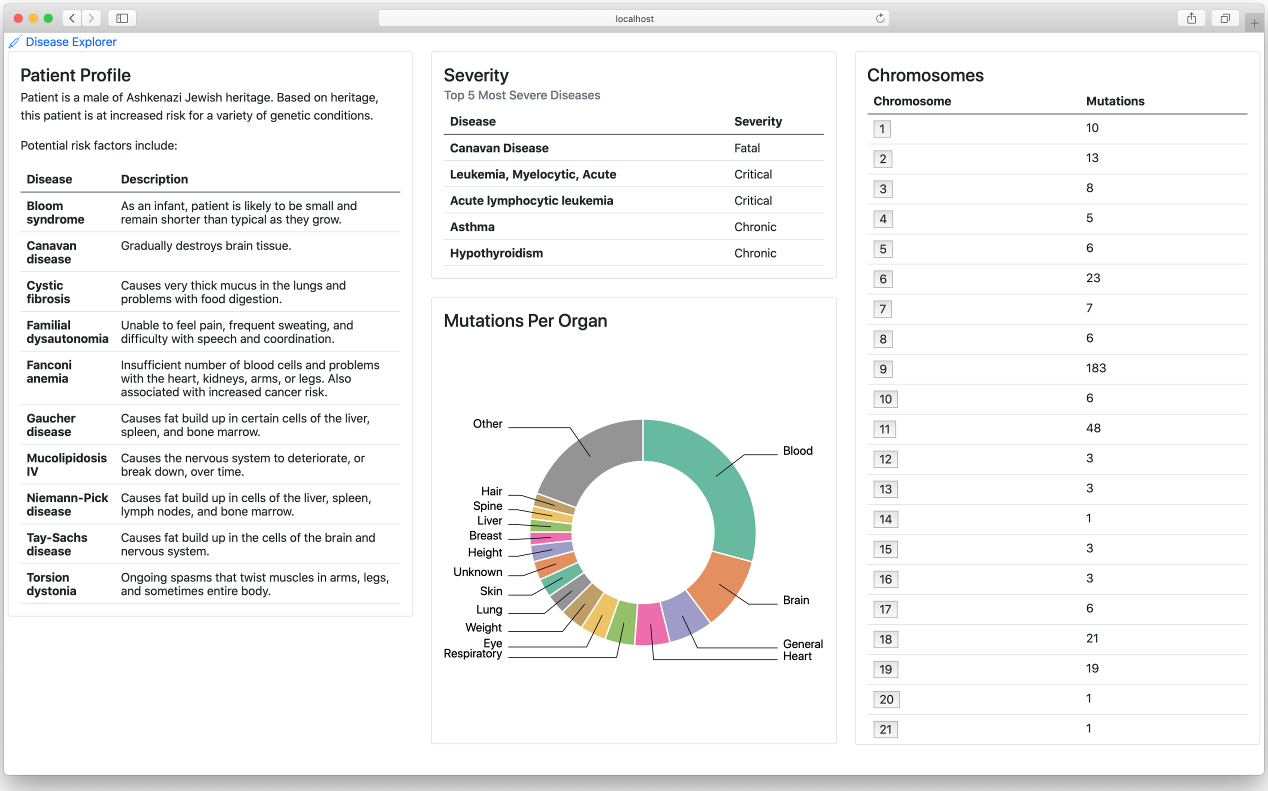
3 System Overview

This section presents the data description, requirement analysis, and overview of the proposed system.

3.1 Data Description

Whole genome data can be time consuming to collect and there are potential privacy issues to navigate when using a patient's data for prototyping a new system. Fortunately, we were able to rely on the Genome in a Bottle Consortium [8] for real, high-quality

[*] e-mail: spencer.king@wustl.edu
[†] e-mail: levmorgan@wustl.edu
[‡] e-mail: josh.shapiro@wustl.edu



Patient Profile

Patient is a male of Ashkenazi Jewish heritage. Based on heritage, this patient is at increased risk for a variety of genetic conditions.

Potential risk factors include:

Disease	Description
Bloom syndrome	As an infant, patient is likely to be small and remain shorter than typical as they grow.
Canavan disease	Gradually destroys brain tissue.
Cystic fibrosis	Causes very thick mucus in the lungs and problems with food digestion.
Familial dysautonomia	Unable to feel pain, frequent sweating, and difficulty with speech and coordination.
Fanconi anemia	Insufficient number of blood cells and problems with the heart, kidneys, arms, or legs. Also associated with increased cancer risk.
Gaucher disease	Causes fat build up in certain cells of the liver, spleen, and bone marrow.
Mucolipidosis IV	Causes the nervous system to deteriorate, or break down, over time.
Niemann-Pick disease	Causes fat build up in cells of the liver, spleen, lymph nodes, and bone marrow.
Tay-Sachs disease	Causes fat build up in the cells of the brain and nervous system.
Torsion dystonia	Ongoing spasms that twist muscles in arms, legs, and sometimes entire body.

Severity

Top 5 Most Severe Diseases

Disease	Severity
Canavan Disease	Fatal
Leukemia, Myelocytic, Acute	Critical
Acute lymphocytic leukemia	Critical
Asthma	Chronic
Hypothyroidism	Chronic

Mutations Per Organ

Chromosomes

Chromosome	Mutations
1	10
2	13
3	8
4	5
5	6
6	23
7	7
8	6
9	183
10	6
11	48
12	3
13	3
14	1
15	3
16	3
17	6
18	21
19	19
20	1
21	1

Figure 1: Interface of the Overview screen of our proposed system.

genomic data. This consortium is hosted by the National Institute of Standards and Technology with the express purpose of aiding in innovation in human genome technologies and facilitating their use in clinical practice. We downloaded their Variant Call File (VCF) for an Ashkenazi Jewish family and used this as the base of our data set. Unlike existing systems, this data is in a standard format that is widely used in research-focused genomic applications.

This VCF file tells us in what ways this family's genomes differ from a human reference genome (e.g. variants), but does not provide any information about specific genes or diseases. The Variant Effect Predictor (VEP) database [5] provides information about specific variants such as the gene name, the frequency of variant, the type of the variant, and many other pieces of information. We used the Hail genomic analysis framework [1] to annotate our VCF with VEP.

Next, in order to get disease information, we used the DisGeNET database [6]. DisGeNET provides a variety of information, but for our system we were most interested in pairing gene names with disease names. DisGeNET does not provide descriptions of diseases and we are not aware of any databases that map DisGeNET's disease names to descriptions. Additionally, we are not aware of any database providing severity scores for these diseases. We manually added descriptions and severity scores using a variety of publicly accessible pages, such as WebMD and others.

Using all of this information we filtered our data. We removed variants without known genes and performed quality filtering to ensure all remaining variants are real variants. Additionally, we only kept damaging variants. A damaging variant is one that has a functional impact, e.g. a disease implication. These steps reduced our data from ~ 3 million variants to ~ 400 variants.

3.2 Requirement Analysis

Genomic data is complex and is often processed and explored using specialized tools aimed at researchers. In order for this data to be useful to a physician when treating a patient the data needs to be made more approachable. If physicians are all expected to have in-depth genetics backgrounds, both in the science and the computing technology, this would significantly limit the applicability of genomic data in a clinical setting. Our goal is to develop a system that can make this data approachable to all physicians, not just those with a specialized background in genetics.

The major milestones of our design and development process are the following.

Literature review. Before beginning our design and development process we needed to determine what, if anything, had already been done in this domain. This gave us an idea of the level of complexity required for our system to advance the state of the field.

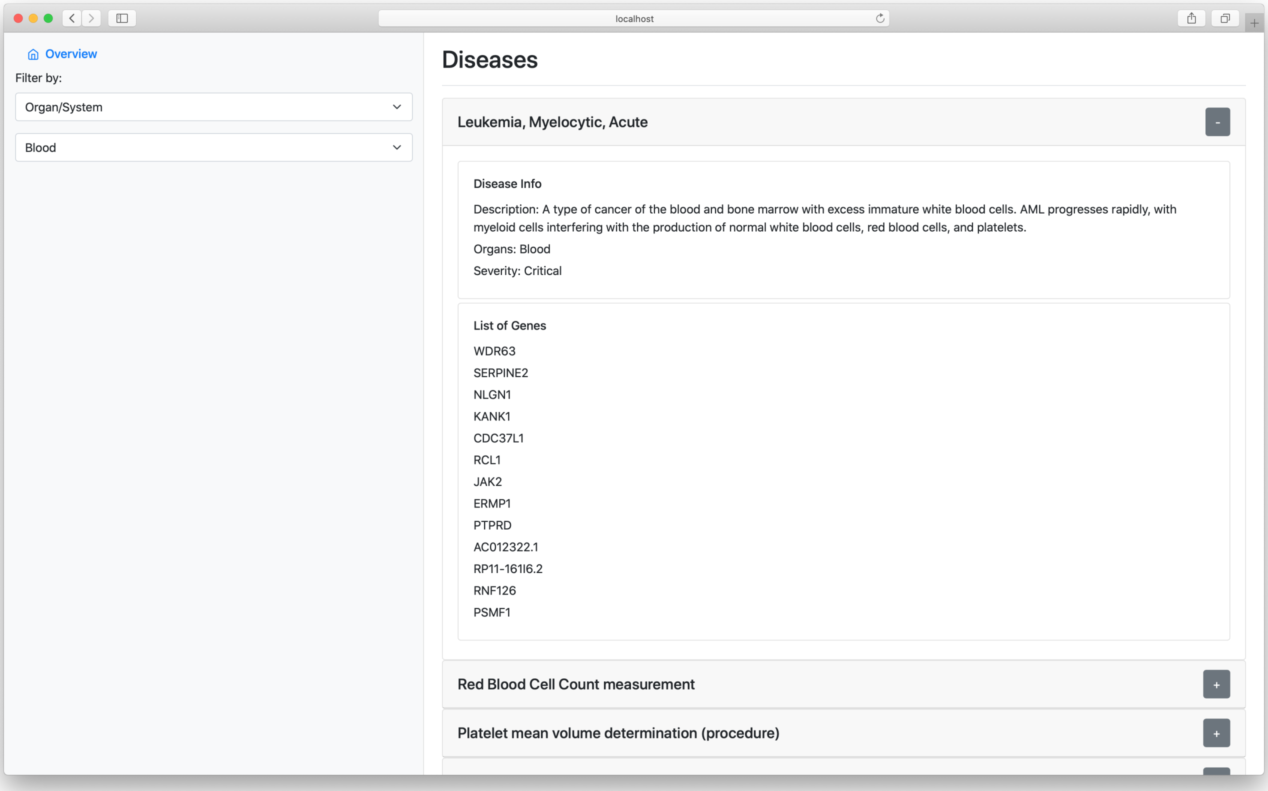
Collect data. We needed to collect and process real patient data before moving forward with design and development. Data availability would impact the extent of what we could include in our system.

Design mock-ups. We sketched mock-ups of our system before beginning development. We wanted to have an overall vision of our system to work towards during development.

Develop the prototype. The final step of our process was to put together the information we gathered to build our system.

3.3 Design Principles

When building our system we considered the following design principles.



Filter by:

Organ/System

Blood

Diseases

Leukemia, Myelocytic, Acute

Disease Info

Description: A type of cancer of the blood and bone marrow with excess immature white blood cells. AML progresses rapidly, with myeloid cells interfering with the production of normal white blood cells, red blood cells, and platelets.

Organs: Blood

Severity: Critical

List of Genes

WDR63
SERPINE2
NLGN1
KANK1
CDC37L1
RCL1
JAK2
ERMP1
PTPRD
AC012322.1
RP11-161I6.2
RNF126
PSMF1

Red Blood Cell Count measurement

Platelet mean volume determination (procedure)

Figure 2: Interface of the Disease Explorer screen of our proposed system.

Uncertainty. Representing uncertainty is a common challenge in many visual analytics systems. This is a particularly important problem in medicine. In the context of genomic data, having a specific gene variant associated with a disease does not necessarily mean that individual will develop the disease. In many cases, this only means they are at an increased risk for the disease. As far as we are aware, there is not a publicly available database that clearly provides these evaluations. In light of this, our system does not attempt to represent uncertainty.

User behavior. Adapting to user behavior is an active area of research in visual analytics. Our system's primary goal is to distill the key insights from the data and present them to a physician. Our expectation is that user behavior will play a larger role downstream of our system when it comes to determining next steps for a patient.

Use common visual metaphors. In order to make the data approachable we use common visual metaphors such as tables, donut charts, and plain text descriptions. We want to avoid overloading the user with more complex visualizations common in research oriented tools. Additionally, we use a traditional left-to-right and top-to-bottom layout with which we expect the vast majority of users to be familiar.

3.4 System Overview

We designed our system with the intent of making complex genomic data accessible to all physicians. To that end, we developed two main views. The first screen, shown in Fig. 1, is the Overview screen. This provides a high-level overview of a patient's data and attempts to highlight the key aspects for the physician. The second screen, shown in Fig. 2, is the Disease Explorer. After observing key insights on the Overview screen, the physician can switch to the Disease Explorer to take a more in-depth look at the patient's data. These screens and an example workflow will be detailed in later sections.

4 Visual Design

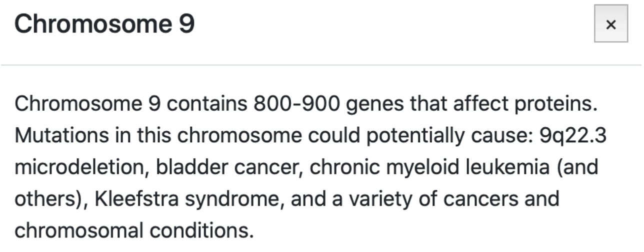
As previously noted, our system is primarily focused on making key insights known up front in order to better guide a physician's exploration of the data. The two major sections of our system, the Overview and Disease Explorer, are presented below along with an example workflow.

4.1 Overview

The Overview screen is shown in Fig. 1. This screen provides a broad overview of a patient's data. The primary goal of this screen is to provide a physician with the key insights about the patient at a glance.

Patient Profile. The Patient Profile provides high-level information about this patient, such as their sex and heritage. Based on this information, physicians are provided a table of diseases for which the patient is at an increased risk. For the patient presented in Fig. 1, it will be important to remember Canavan disease and Fanconi anemia for later.

Severity. This shows the top 5 diseases ranked by severity scores. The diseases listed here are ones for which the patient has genetic variants, but may not be guaranteed to exhibit in practice. The interesting items to note here are Canavan disease and the two forms of leukemia. From the Patient Profile a physician would already know this patient has an increased risk of Canavan disease. Additionally, Fanconi anemia is a blood disorder often associated with an increase cancer risk, which is potentially indicative of the



Chromosome 9

Chromosome 9 contains 800-900 genes that affect proteins. Mutations in this chromosome could potentially cause: 9q22.3 microdeletion, bladder cancer, chronic myeloid leukemia (and others), Kleefstra syndrome, and a variety of cancers and chromosomal conditions.

Figure 3: Interface of the Chromosome modal.

two types of leukemia. Just by looking at two sections, a physician would already have developed important insights about this patient.

Mutations Per Organ. This section shows the percentage of mutations in a variety of organs or systems. In the donut chart, a physician can clearly see that this patient has a significant number of mutations in genes relating to the blood. In this case it is over 25%. This observation further validates knowledge from the previous sections concerning Fanconi anemia and leukemia, both of which are blood disorders.

Chromosomes. This section provides counts of the mutations in each chromosome. By clicking on a chromosome, a physician can read more detailed information about what mutations in that chromosome may indicate. An example is shown in Fig. 3.

4.2 Disease Explorer

The Disease Explorer screen is shown in Fig. 2. On this screen physicians are able to explore the patient's data based on the insights they gathered on the Overview screen. On the left, they have the option to filter by disease, organ/system, severity, gene, and chromosome. On the right, they are presented with a view of all diseases meeting the filtering criteria. Each disease displays a description, affected organ or system, severity score, and a list of impacted genes for this specific patient.

In Fig. 2, the physician is investigating the large number of mutations found in the blood. As would be expected, leukemia is featured prominently. Also visible in the figure is a ”Red Blood Cell Count measurement” which would potentially be indicative of the Fanconi anemia previously observed on the Overview screen.

5 Future Work

We are satisfied with our system thus far, however we have a few points we would like to address in future work.

Uncertainty. We would like to devise a method of representing uncertainty. Uncertainty will play a significant role in a physician's decision-making process regarding which tests or treatments to order.

Additional disease data. We believe the Disease Explorer would benefit from additional disease data. At the moment it is rather barebones. Information such as common treatment plans or population level statistics concerning long-term outcomes would make the system much more helpful. This may also be one avenue for addressing issues with uncertainty.

Family history. Taking family history into consideration could help make the Overview screen more accurate, especially for patients without a unique or well-studied heritage. There are existing methods in the genetics literature for determining inheritance of specific genes, however this would significantly increase the complexity of our data processing pipeline. In the interest of building a working prototype in a reasonable time frame, we elected to postpone this functionality.

EHR integration. We believe that EHR integration is going to be vital to adoption. These systems are widely used in the United States and other developed nations. If our tool does not exist within a physician's current workflow, it is less likely to be adopted. Additionally, EHR systems hold a wealth of information about patients and diseases which may help to address the previous points.

User testing and expert opinions. We would like to have actual physicians use our system and provide feedback. While developing and using our system we feel that we have learned valuable information about the sample patient. However, none of us are physicians and may not approach the data in the same way. It will be important to have actual physicians confirm our system is intuitive and helpful.

6 Conclusion

This work presented a prototype of a system geared towards physicians for exploring a patient's genomic data. Our goal was to make these complex data sets approachable, without an explicit need for a background in human genetics. To the best of our knowledge, there are no existing, publicly available systems accomplishing this task. Existing systems are either focused heavily on researchers or are proprietary. In the case of proprietary systems, it is possible that similar work has already been completed. However, this is not something we were able to verify. We believe that work in this field will become highly important in the coming years as personalized medicine becomes more viable. We hope our work will be continued in the future, either by ourselves or others, and that it will make a difference in the future of medicine.

Acknowledgments

This work was produced as a part of Washington University in St. Louis' CSE 557A Advanced Visualization course. The authors wish to thank Dr. Alvitta Ottley and Shayan Monadjemi for their instruction and advice.

References

[1] Hail. Hail - scalable genomic data analysis. Hail - GitHub, March 2021.

[2] N. C. Institute. Precision medicine in cancer treatment. National Cancer Institute - Educational Resources, October 2017.

[3] P. König, S. Beier, M. Basterrechea, D. Schüler, D. Arend, M. Mascher, N. Stein, U. Scholz, and L. Matthias. Bridge – a visual analytics web tool for barley genebank genomics. *Frontiers in Plant Science*, 11:701, June 2020. doi: 10.3389/fpls.2020.00701

[4] K. S. Lau-Min, S. B. Asher, J. Chen, S. M. Domcheck, M. Feldman, S. Joffe, J. Landgraf, V. Speare, L. A. Varughese, S. Tuteja, C. VanZandbergen, M. D. Ritchie, and K. L. Nathanson. Real-world integration of genomic data into the electronic health record: the pennchart genomics initiative. *Genetics in Medicine*, 23:603–605, Apr. 2021. doi: 10.1038/s41436-020-01056-y

[5] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17(122), June 2016. doi: 10.1186/s13059-016-0974-4

[6] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48:D845–D855, Jan. 2020. doi: 10.1093/nar/gkz1021

[7] K. A. Wetterstrand. The cost of sequencing a human genome. National Human Genome Research Institute - Educational Resources, December 2020.

[8] J. M. Zook, D. Catoe, and M. Salit. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, 3(160025), June 2016. doi: 10.1038/sdata.2016.25