Multiscale Visualization of Pathogenic Structural Variants

Armita Safa (armt.safa@gmail.com)
Janet Li (janli@bcgsc.ca)
Neera Patadia (neera.patadia@gmail.com)

Introduction

The advent of genome sequencing in the biomedical sciences has provided researchers with massive, high-dimensional datasets that can be used for a variety of purposes such as understanding genetic variation in the human population and elucidating the genomic cause of diseases (The DNA Universe, 2020). Genomic data can be considered as having a multiscale structure. This large, multiscale structure can make it difficult to interpret and understand genomic information.

Visualization tools can be used to better understand the overall structure of genomic information, as well as to gain insights into potential relationships within genomic data. A major area of interest in genomics is finding genetic variants within the genome. Genetic variants are considered to be any change in the sequence of nucleotides that make up a given DNA sequence in comparison to a reference sequence. Genetic variants can range in size, from single nucleotide variants (SNVs), to structural variants (SVs) which are any variants larger than 50 base pairs (bp) (Tattini et al., 2015).

Structural variants can take on a variety of forms including deletions, insertions, duplications, inversions and translocations. These structural variants can result in a range of functional consequences, and often contribute to the occurrence of diseases. Structural variants that cause disease are considered to be “pathogenic”. Pathogenicity of a variant can fall on a spectrum from being highly pathogenic or likely pathogenic to neutral or benign (Biesecker et al., 2018).

Biological data are often stored and shared in large, publicly available databases. Data from these databases can be downloaded as a text file and used for bioinformatic analysis (Landrum et al., 2014). The National Consortium of Biological Information (NCBI) provides clinically relevant structural variants and their corresponding pathogenicity annotations in its ClinVar database. In this work, we aim to use a curated set of ClinVar structural variants that have been annotated with pathogenicity classifications to develop a tool that visualizes a user’s structural variants in relation to ClinVar’s reviewed SVs. Data from ClinVar will be used to develop filtering mechanisms to allow for the visualization of variants of differing levels of pathogenicity. To demonstrate the utility of our visualization tool, we will use the Human HG002 dataset, which is a set of variants pertaining to a single individual (Zook et al., 2020). The SVs from HG002 will be queried against the set of ClinVar SVs to identify the pathogenicity of the SVs from HG002. These results will be displayed on a global view of the genome, as well as on individual chromosomes, providing different levels of detail in the multiscale data. Furthermore, we will provide details about associated disease information for individual SVs, based on the ClinVar annotations.

Related Work

There are a variety of tools that have recently been introduced for the visualization of SVs. This section will discuss the implementation and utility of these approaches, as well as their benefits and limitations.

Linear Genome Browser

Linear genome browsers were one of the first classes of tools used to visualize the human genome. The UCSC Genome Browser was initially developed during the Human Genome Project and allowed for the visualization of the DNA sequences of all 23 chromosomes. Linear genome browsers typically display the nucleotide sequence of interest below a reference sequence. The nucleotides that comprise the DNA sequence of interest and the reference genome are displayed in a horizontal view. Furthermore, custom views of the genome or "tracks" can be added to linear genome browsers in order to visualize different aspects of the genome such as genomic variants (Karolchik et al., 2009).

The Integrative Genomics Viewer (IGV) tool can be considered as a type of linear genome browser, which allows for the visualization of diverse genomic data types. The viewer consists of a series of rectangular panels. The top panel shows the region being investigated on a chromosome in a horizontal view. Data being visualized through IGV can also include annotations in regards to phenotype, experimental label or clinical label. These annotations can be visualized in the two leftmost columns, with the annotated categories listed vertically (Robinson et al., 2011).

While linear genome browsers have a wide range of utility in the visualization of genomic data, one caveat arises when considering the fact that they are based on visualizing short-read sequencing data. Short reads are not ideal for identifying structural variants, so linear genome browsers have not been optimized to visualize structural variant data (Yokoyama & Kasahara, 2020).

Ribbon

The Ribbon visualization tool provides a similar view to linear genome browser visualization tools but is designed to be compatible with long read sequencing data. Horizontally at the top of the visualization is a representation of the reference genome segmented into chromosomes. Users can select a chromosomal section to see the relevant sequence alignments of interest lined up vertically, as well as structural variants such as translocations (Nattestad et al., 2021). This visualization can be considered as an improvement over the IGV visualization tool due to its support of visualizing long-read sequencing data.

MoMI-G: A Graph Based Genome Browser

MoMI-G is a web based genome graph browser that contains multiple panels that can be used to visualize different aspects of genomic structural variants. The panels contain three main views. The first view is a circos plot which provides a chromosomal level overview of the structural variants. Within the circos plot, the structural variants are represented by curved line segments on different regions within the chromosomes. The second view is a table, which contains metadata on

each annotated structural variant such as the type of structural variant (insertion, deletion, translocation, duplication, inversion), the chromosome the variant occurs on and the start and end position of the SV. Finally, the browser also contains a linear genome browser view which visualizes structural variant positions in relation to a reference genome (Yokoyama et al., 2019).

Task Abstraction

Clinical researchers can obtain several thousand or even millions of structural variant calls for a single sample. Identifying the medically relevant SVs within a set is crucial for determining the cause of disease and gaining a better understanding of the role of these genomic aberrations in human health. The biological relevance of a variant is often inferred by manually querying a database of known variants (e.g. ClinVar) for matches. The presence or absence of a variant in a database can be used as a metric for prioritizing variants for further analysis. The extra data available in the database is also used to annotate these variants.

At the analysis level, a clinical bioinformatician both consumes and produces new information from a large SV dataset. By analyzing SVs, the researcher can generate a new hypothesis, or verify or disconfirm an existing hypothesis about the variant's role in disease. The process of annotating variants with clinical and phenotypic data produces new information. Once a bioinformatician obtains a set of structural variants, they may either want to select a set of clinically relevant SVs from the dataset or compare several potentially relevant SVs to one another. They may also be interested in summarizing the entire dataset to get a global view of the genome or a region of it, for example "how many insertions are present on chromosome 21?" or "how many pathogenic variants are present in my dataset?" At the search level, a bioinformatician must browse through a set of SV calls to identify variants of interest. The locations of these SVs are unknown and the exact identity is unknown as well. The user will likely be browsing for SVs with specific clinical attributes, such as SVs labelled as "pathogenic" or "likely pathogenic". There may be cases where a user is simply exploring their dataset, as well, to see if any of their SVs are present in ClinVar, and what diseases are associated with it.

Data and Data Abstraction

Structural variants are identified in relation to a reference genome. Variants are defined by their start position along a certain chromosome in the reference, making them ordered and continuous. The human reference genome contains 25 distinct chromosomes, so the chromosome is a categorical "bin" within the reference. A variant is a single item within a tabular SV dataset. Our main input dataset will be a set of structural variants identified for the human individual HG002 (Zook et al., 2020) that have a match in the ClinVar dataset. The ClinVar dataset consists of 150,782 variants in total. The original HG002 dataset contained 46,024 variants, and 2,664 of these SVs have a match in ClinVar. These 2,664 variants will be presented in our proposed visualization, along with a summary of the 150,782 variants in ClinVar (Figure 1). A summary of the attributes in our dataset is presented in Table 1.

Table 1. Attributes in variant dataset.

Variable name	Description	Type	Possible values
Allele ID	Identifier/key for the ClinVar variant that the HG002 variant was matched to	Categorical	2,664 values: unique for each ClinVar variant
Chromosome	Chromosome that the variant is located on	Categorical	25 possible values: 22 autosomes + 2 sex chromosomes + mitochondria
Pos	Start position of HG002 SV along chromosome	Continuous	1 - length of chromosome
Type	Type of SV	Categorical	9 possible values: complex, deletion, duplication, duplication, insertion, inversion, microsatellite, tandem duplication, translocation, type, variation, copy number gain, copy number loss, fusion
ClinicalSignificance	Pathogenicity or clinical relevance for a variant	Ordered	4 possible values: Uncertain significance, benign, likely pathogenic, pathogenic
PhenotypeList	Phenotypes (diseases) associated with a variant	Categorical	11,583 possible values: Up to 5 phenotypes are given listed for a single variant; if more than 5 are associated, the number of phenotypes is given instead
Similarity	The similarity of the HG002 variant to the corresponding ClinVar variant. Given as a percentage.	Continuous	0-100

Solution

The goal of this project is to create a visualization tool that shows structural variants in the HG002 SV dataset that match variants in the ClinVar database, along with relevant clinical and metadata about the matches. A sketch of our proposed visualization is presented in Figure 1. For our initial implementation, we will use a custom dataset obtained by manually querying the HG002 SVs in the ClinVar dataset. We hope to make this tool generalizable so that a user can query their own data in the ClinVar dataset and view the final results, however, implementation of this feature will be time-dependent.

SVs in the HG002 dataset and ClinVar cannot be matched simply by looking for identical start positions along a chromosome because these positions may vary slightly. For example, there may be a 200 bp deletion starting at position 120 of chromosome 1 in the HG002 dataset, while it may be located at position 123 of chromosome 1 in the ClinVar dataset. While the start position is slightly different, since the two variants are close in proximity and size, they should be considered the same. To compare a query variant from HG002 to a target (in ClinVar), the query will be scored based on its similarity, proximity and length. This score will be used in our solution to filter variants, and is also a valuable derived attribute that will be encoded.

Our proposed solution is to implement a multi-view visualization representing different levels of detail within the HG200 dataset. We will provide a global view of the reference genome with a Circos plot (Krzywinski et al., 2009). Since the HG200 dataset is quite large, variants will be sorted and filtered based on their similarity score and/or pathogenicity, so that only the top 1,000 most similar, pathogenic variants are shown. This will provide an overview of relevant SVs of clinical significance in the dataset. The hue channel will be used to represent the pathogenicity of the variant, using a diverging colour scale, and the saturation channel will be used to represent the similarity score of the variant. Intrachromosomal variants will be encoded by a line mark in the radial track of the Circos plot, and interchromosomal variants (e.g. a translocation between chromosome 2 and 4) will be encoded by a line linking the two positions.

In addition to the Circos overview, the solution will provide a linked linear view of a region within the genome to allow the user to navigate through the genome, directly below the Circos plot. This region will be selected by a brushing action, where a user clicks and drags a region within the Circos plot. The brushed region within the Circos plot will also be highlighted to verify what region is being shown in the linear view. All variants present in the brushed region will be shown in the linear view, and the “brushable” region size will have an upper limit of 1 million base pairs to prevent the number of items being shown from growing too large. Intrachromosomal variants will also be encoded with a line mark in this linear view, while interchromosomal variants will be represented with arrows denoting the direction of the variant (i.e. translocation from the region will be encoded as an arrow pointing away or upwards, while a translocation to the region will be encoded as an arrow pointing towards or downwards).

Users will be able to identify details of a single variant by hovering over the variant in either the Circos or linear view, which will cause a pop-up to be shown with details of its ClinVar ID, specific location, the phenotypes/diseases that it is associated with, and its similarity score to its associated

ClinVar variant. Variants in the linked Circos and linear views can also be selected by clicking on the individual mark. The selected variants will be listed in a table on the bottom right of the visualization.

Stacked bar charts will be used to present summaries statistics of the ClinVar dataset and initial HG002 dataset (before matching variants). This will include visualizing the number of variants that fall into each pathogenicity category for a given chromosome for both the ClinVar and HG200 datasets. This will allow users to quickly examine and identify differences between the two datasets. To keep consistent with the Circos plot visualization, the hue channel will be used to represent pathogenicity and will be implemented using the same colour scheme.

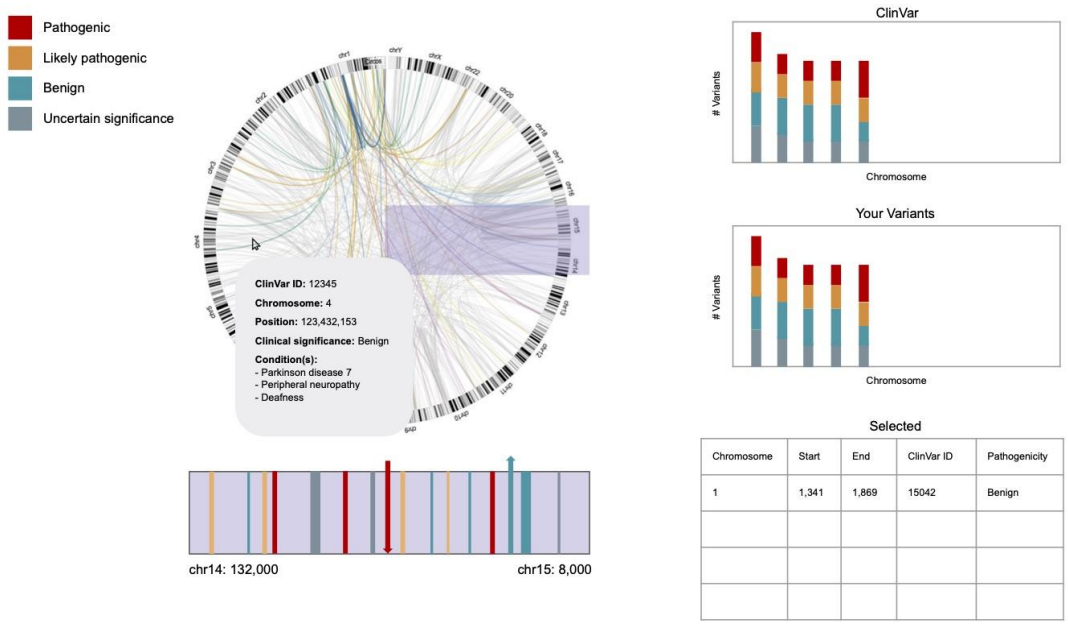


Figure 1. Sketch of the proposed visualization panel. The Circos plot and linear view are linked through brushing, shown by the purple region. When the user hovers over a variant in the Circos plot (or linear view), a popup will appear with more details. Two stacked bar charts on the right show the total number of variants in the whole ClinVar dataset per chromosome (top) and the number of ClinVar matches in the input data per chromosome (bottom). Variants selected by clicking on the Circos plot or linear view are presented in a table on the bottom right.

Implementation

We will clean the ClinVar and HG002 datasets with Python scripts and command line tools. A Python script will also be used to match and score the variants to the ClinVar variants. We will implement our proposed solution as a web application using the React library. The Circos and linear genome views will be created with the Gosling.js grammar and visualization package (L'Yi et al., 2021). The summary bar charts will be implemented with D3.js (Bostock, Ogievetsky, and Heer 2011). All interactions and other visualization elements will be handled with React.

Milestones

Our project deadlines and major milestones are outlined in Table 3. Before starting our visualization, we need to clean the large input datasets and identify which HG002 SV match variants in ClinVar. The data cleaning and initial analysis does not require our entire team, but Armita's expertise will be helpful here, since she works with structural variants in her research. All group members will be involved in all other milestones.

Janet will set up the React app, set up the main structure of the view and set up the Gosling specifications. She will also work on the stacked bar charts. Neera and Janet will be responsible for implementing the Circos plot, and Armita will be responsible for the linked linear view. We plan to have our data finalized and begin implementing the UI before the peer project review and post-update meeting. We will meet after the post-update meeting and discuss the feedback received, taking into account any suggestions and making changes to our proposed solution as necessary. In the final weeks of the term, we will finalize the visualization, with Janet implementing the interactive components of the visualization once the main plots are complete. If we have time after the proposed solution is complete, we will set up a backend to allow users to upload their own variants and perform matching with the ClinVar database.

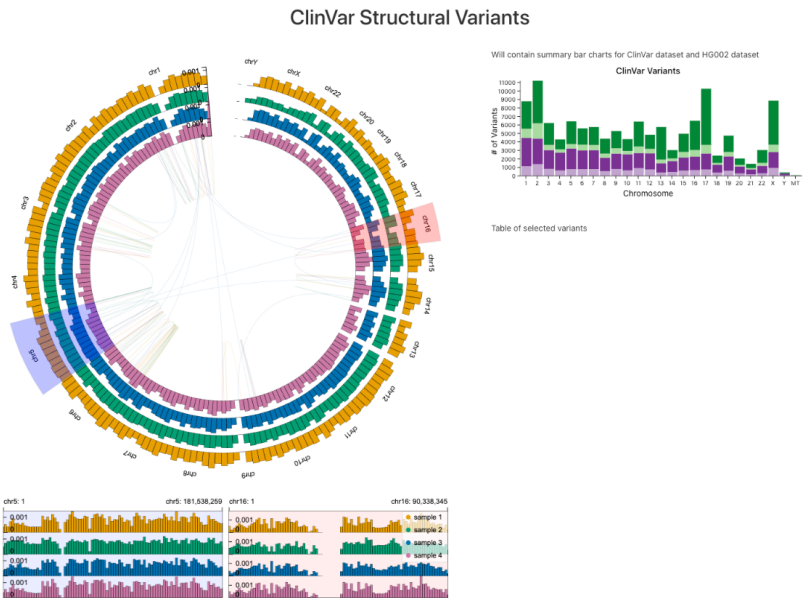
Table 3. Project deadlines and milestones

Task	**Deadline**	**Estimated Time (hours per person)**	**Description & Assignments**	**Status**
Project Pitch	Sep 29	2	-	Complete
Pre-proposal Meeting	Oct 13	-	All	Complete
Proposal	Oct 21	4	All	Complete
Obtain data, initial analysis for proposal	Oct 21	2	Janet	Complete
Data cleaning	Oct 27	2	Neera	Complete
Match HG002 variants to ClinVar dataset and finalize inputs	Nov 3	4	Armita	Complete
UI implementation (initial; pre-"post-update" meeting)	Nov 5	2	React app setup: Janet	Complete
	Nov 16	10	Bar charts: Janet	In progress
	Nov 16	12	Circos plot: Neera & Janet	In progress
	Nov 16	12	Linked linear view: Armita	In progress
Written Update	Nov 16	6	All	Complete
Peer Project Review	Nov 17	2	All	To be

				completed
Post-Update Meeting	Nov 24	-	All	To be completed
Make any necessary changes to plans	Nov 26	2	All	To be completed
Finish implementing visualization	Dec 13	10	Interactive features: Janet	To be completed
	Dec 13	30	Other features: same assignments as in “UI implementation” milestone	To be completed
Final Presentation	Dec 15	6	All	To be completed
Final Report	Dec 17	8	All	To be completed

Progress

So far, Janet has identified and downloaded the required datasets. Neera performed the preliminary data cleaning and Armita wrote and ran the script to match the HG002 and ClinVar variants. Janet set up the main structure of the React app and Gosling plot components, and has begun implementing the stacked bar chart component. The current Circos plot and linear view are placeholders with test code, but the general structure of the app is complete.



Bibliography

Biesecker, L. G., Nussbaum, R. L., & Rehm, H. L. (2018). Distinguishing Variant Pathogenicity From Genetic Diagnosis: How to Know Whether a Variant Causes a Condition. JAMA, 320(18), 1929–1930. https://doi.org/10.1001/jama.2018.14900

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 Data-Driven Documents. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2301–2309. https://doi.org/10.1109/TVCG.2011.185

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., … Hubbard, T. (2011). Modernizing Reference Genome Assemblies. PLoS Biology, 9(7), e1001091. https://doi.org/10.1371/journal.pbio.1001091

Genome Browser User's Guide. (n.d.). Retrieved October 21, 2021, from https://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#Liftover

Karolchik, D., Hinrichs, A. S., & Kent, W. J. (2009). The UCSC Genome Browser. *Current Protocols in Bioinformatics*, *28*(1), 1.4.1-1.4.26. https://doi.org/10.1002/0471250953.bi0104s28

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research, 42(Database issue), D980–D985. https://doi.org/10.1093/nar/gkt1113

L'Yi, S., Wang, Q., Lekschas, F., & Gehlenborg, N. (2021). Gosling: A Grammar-based Toolkit for Scalable and Interactive Genomics Data Visualization. OSF Preprints. https://doi.org/10.31219/osf.io/6evmb

Nattestad, M., Aboukhalil, R., Chin, C.-S., & Schatz, M. C. (2021). Ribbon: Intuitive visualization for complex genomic variation. Bioinformatics, 37(3), 413–415. https://doi.org/10.1093/bioinformatics/btaa680

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative Genomics Viewer. Nature Biotechnology, 29(1), 24–26. https://doi.org/10.1038/nbt.1754

SAM/BAM and related specifications. (2021). [TeX]. samtools. https://github.com/samtools/hts-specs (Original work published 2012)

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Frontiers in Bioengineering and Biotechnology, 3, 92. https://doi.org/10.3389/fbioe.2015.00092

The DNA Universe. (2020). A Journey Through The History Of DNA Sequencing. The DNA Universe BLOG. https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing

Yokoyama, T. T., & Kasahara, M. (2020). Visualization tools for human structural variations identified by whole-genome sequencing. Journal of Human Genetics, 65(1), 49–60. https://doi.org/10.1038/s10038-019-0687-0

Yokoyama, T. T., Sakamoto, Y., Seki, M., Suzuki, Y., & Kasahara, M. (2019). MoMI-G: Modular multi-scale integrated genome graph browser. BMC Bioinformatics, 20(1), 548. https://doi.org/10.1186/s12859-019-3145-2

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., Sherry, S., Koren, S., Phillippy, A. M., Boutros, P. C., Sahraeian, S. M. E., Huang, V., Rouette, A., Alexander, N., Mason, C. E., Hajirasouliha, I., Ricketts, C., Lee, J., Tearle, R., … Salit, M. (2020). A robust benchmark for detection of germline large deletions and insertions. Nature Biotechnology, 38(11), 1347–1355. https://doi.org/10.1038/s41587-020-0538-8