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. We will show a global view of the genome, as well as individual chromosomes, providing different levels of detail in the multiscale data. Furthermore, we will provide details about associated disease information for individual SVs.

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 often 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. Our visualization tool will allow a bioinformatician to prioritize variants by querying their dataset against validated, clinically relevant SVs in the ClinVar database. Variants that are present in ClinVar will be highlighted and annotated with metadata related to disease association, molecular function and pathogenicity.

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. Genomic variants will be annotated with clinical metadata as a means of prioritizing them during analysis. This annotation step is a “produce” goal as it generates new data from the input. Finally, the researcher may also be interested in recording the SVs identified in this analysis for further work.

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 with what attributes.

Once a bioinformatician identifies a set of targets, they may either want to identify a set of clinically relevant SVs 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 patient’s genome, for example “how many insertions are present on chromosome 21?”

Data and Data Abstraction

Structural variants are always identified as an alternate allele in relation to a reference genome (reference allele). The variants are defined by their positions along the reference, making the data inherently spatial. A single variant can be considered as an item within a tabular dataset or as a grid of positions along the reference. Our input dataset will be a set of sequence-resolved benchmark SV calls for the human individual HG002 (Zook et al., 2020). These SVs were called against the human reference genome build GRCh37 (Church et al., 2011). There are 46,024 structural variants (>= 50bp) in total, and 54 attributes associated with each SV. We will not need most of the attributes and will drop these from the dataset. The main attributes of interest are described in Table 1.

Table 1. Main attributes of interest in HG002 benchmark structural variant dataset.

Variable	Description	Type	Possible values

CHROM	Chromosome along which the SV occurs	Categorical	25 (22 autosomes, 2 sex chromosomes, mitochondria)
POS	Starting position of SV along chromosome	Spatial	1 - length of chromosome
END	End position of SV along chromosome	Spatial	1 - length of chromosome
SVTYPE	Type of SV	Categorical	Deletion, contraction, insertion, duplication, inversion
SVLEN	Difference of length between reference and alternate allele	Continuous	50bp - infinite (we will likely impose a cutoff)

The curated ClinVar database is a tabular dataset consisting of 175,870 SVs annotated against GRCh37. The dataset consists of 37 attributes that describe an SV's location and type, its phenotypes, academic review status, etc. The majority of these attributes are descriptive, while Chromosome, PositionVCF and Type will be used to match the SV to the input data. Missing, irrelevant and incomplete items/attributes will be removed from the dataset. The attributes that will be used in our tool are described in Table 2.

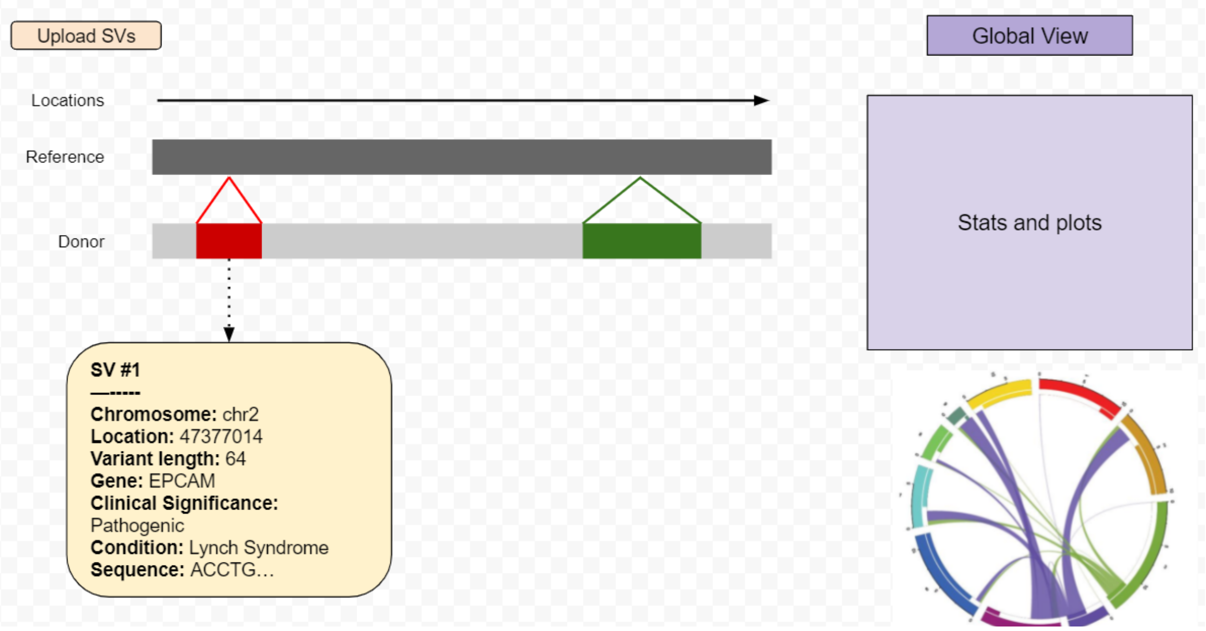
Table 2. Main attributes of interest in ClinVar dataset.

Variable	Description	Type	Possible values
Chromosome	Chromosomal location	Categorical	25 (22 autosomes, 2 sex chromosomes, mitochondria)
PositionVCF	Starting position of SV along chromosome	Spatial	1 - length of chromosome
Type	Type of SV	Categorical	11 (copy number gain, copy number loss, deletion, duplication, fusion, insertion, inversion, microsatellite, tandem duplication, translocation, variation)
Name	ClinVar preferred named for the record	Categorical	Unique for each item
GeneID	Associated GeneID from NCBI's Gene	Categorical	8,094 unique genes reported

	database. Reported if there is a single gene, otherwise reported as -1		
ClinicalSignificance	Character, comma-separated list of calculated clinical significance	Categorical/ordinal (ordered by pathogenicity)	5 (benign, likely benign, pathogenic, likely pathogenic, uncertain significance)
PhenotypeList	Character, list of associated phenotype names	Categorical	7,190 unique associated phenotypes

Solution

The goal of this project is to create a visualization tool that shows which structural variants in an input dataset are present in the ClinVar database. For our initial implementation, we will match the HG002 SVs with the ClinVar SVs and directly visualize this dataset. We hope to make this tool generalizable, so a user can upload their own data and view the final results. A sketch of the final product is presented in Figure 1.



SV #1
——-----
Chromosome: chr2
Location: 47377014
Variant length: 64
Gene: EPCAM
Clinical Significance: Pathogenic
Condition: Lynch Syndrome
Sequence: ACCTG…

Figure 1. Sketch of the final visualization panel

Proposed User Scenario

A user has the structural variants from an individual. They wish to visualize these variants in comparison to the human reference genome as well as obtain more information about the significance of the variants. The user uploads a VCF (samtools n.d.) file containing the structural variants. The file must contain the required fields in VCF format. Our tool will perform a search in the

ClinVar database, looking for structural variants that match those uploaded by the user. A matching structural variant is scored by proximity of the location in ClinVar to the location in the uploaded file and the similarity of the contents of the two variants. Best matches are retrieved and used to further annotate the structural variants with additional information such as clinical significance and molecular consequence. This data is then shown on a sample individual's genome in comparison to the reference genome for better illustration. The user can inspect each variant individually by moving to that location. Statistics can also be presented in subplots.

Implementation

We first need to clean the data obtained from ClinVar and perform initial analyses. Next, to implement our proposed visualization, we need two main components: 1. the scripts for communication with the ClinVar database, data augmentation and filtering, and 2. the dashboard for user interaction and showing the final results. We intend to use the Python programming language for the scripts. Accessing the ClinVar database will be done using clinVar's application programming interface. After the data is gathered and processed, we use the D3.js (Bostock, Ogievetsky, and Heer 2011) framework to create the visualization.

Potential Issues and Solutions

- The user must specify which reference genome was used to retrieve the locations of the structural variants in the VCF file. The platform can provide multiple reference genomes or use liftOver("Genome Browser User's Guide" n.d.) files to map the locations to a specific genome reference genome build based on data available in the ClinVar database.
- ClinVar might not contain information for some of the novel structural variants uploaded by the user. For these cases, the user can add their own annotations to the VCF file and those will be shown instead of the ClinVar data.
- ClinVar has an extensive database. Querying this database can be computationally expensive as we also have to check the content similarity between the query and hits in the database. Hence, the searching phase might cause scalability problems. We have to make proper use of searching strategies and efficient use of databases to reduce the search overhead as much as possible.

Milestones

Our project deadlines and major milestones are outlined in Table 3. Before starting our visualization, we need to identify which HG002 SV match variants in ClinVar. The dataset will be cleaned to remove unnecessary attributes and items. The data cleaning and initial analysis does not require our entire team, but Armita's expertise will be helpful here, along with one other member. All group members will be involved in all other milestones.

Table 3. Project deadlines and milestones

Task	Deadline	Estimated Time (hours per person)	Description/Assignments

Project Pitch	Sep 29	2	-
Pre-proposal Meeting	Oct 13	-	All
Proposal	Oct 21	4	All
Data cleaning and initial analysis	Oct 27	1-2	Armita, one other person
Match HG002 variants to ClinVar dataset and finalize inputs	Nov 3	3	All
Start implementing UI	Nov 5	10	All
Written Update	Nov 16	4-5	All
Peer Project Review	Nov 17	2	All
Post-Update Meeting	Nov 24	-	All
Make any necessary changes to plans	Nov 26	1-2	All
Finish code	Dec 13	?	All
Final Presentation	Dec 15	4	All
Final Report	Dec 17	6	All

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). D³ 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 / Editoral Board, Andreas D. Baxevanis ... [et Al.], CHAPTER, Unit1.4. https://doi.org/10.1002/0471250953.bi0104s28

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

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