Visualisation of Interactions between Single Nucleotide Polymorphisms and Structural Variants

Tom Jacobs[1]

Supervisor(s): Marcel Reinders[1]**, Niccoló Tesi**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Tom Jacobs
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Niccolò Tesi, Andy Zaidman

Abstract

in bio-informatics visualisations are often used to relay the results of genome-wide association studies (GWAS), which can be used to get a better inside into the genetics of diseases. Over the years many websites have been developed, which can create visualisations for a variety of GWAS, with one of the most famous being locuszoom. Recent advancements in technology have allowed for collection of data which wasn't previously possible and which no visualisation tool currently supports. For the first time it is now possible to collect data on the interactions between two different types of genetic mutations: Single Nucleotide Polymorphisms, which consist of a single change in one nucleotide (one letter in a string of DNA) and Structural Variants, which consist of a change in a series of nucleotides. This paper builds upon an already existing visualisation tool snpxplorer and the goal is to design a visualisation, which captures thesse interactions and to add a new, easily integratable section to snpxplorer containing this new visualisation.

1 Introduction

Genetic factors influence the susceptibility to many diseases. One of the main approaches to study the genetics of diseases is by conducting genome-wide association studies (GWAS), which can find the association between single nucleotide polymorphisms (SNP) and diseases [7]. SNPs are the simplest and most common genetic variants and they consist of just a single change in one nucleotide. This means that in a string of DNA just a single letter different, while other letters around the SNP are unaffected by the mutation. In the past years thousands of GWAS have been conducted and thens of thousands associations with traits have been found, however often understanding the functional biological implications SNPs and how they affect the disease is complex and requires the integration of multiple levels of genomic data. Structural variants (SV) are a more complex type of genetic variation and consist of a series of nucleotides being changed in a number of different ways. Think about a string of DNA being copied, deleted, duplicated or rearranged [3]. Recent technological advancements now allow us to study interactions between SNPs and SVs to see whether maybe such interactions can also explain genetic diseases [2]. Before that can be done there first needs to be a better understanding of the causes behind the correlation between these SNPs and SVs. An important part to understanding this is being able to visualise this correlation effectively. Frontrunner in genetic visualisation locuszoom, currently doesn't offer a way to visualise these interactions. The aim of this project is to develop an interactive visualisation tool to show these SNP-SV interactions. This project will build upon the already existing website snpxplorer [6], which helps researches by visualising sections of the genome and any SNPs, SVs and disease-related traits within that section. With the new addition it will now also be possible to see correlations between SNPs and SVs.

2 Methods

Most inspiration for this project is drawn from snpxplorer [6], but that is just one of many websites you could go to to visualise genomic data. Another widely used one is locuszoom. The task on how to clearly show connections between multiple different mutations in the genome, however is a unique one and has never been done before, there are however guidelines and conventions that most visualisation tools of genomic data follow. Because DNA is 2D, it is very intuitive to lay out a region of DNA horizontally across the screen and highlight the relevant parts [5]. In the very common case when we want to compare multiple different characteristics of DNA, multiple different tracks can simply be used. A layout with multiple tracks and lines showing the relations between the different tracks can be seen in Figure 1. A special version of such a plot is a parallel coordinates plot, which can contain any number of tracks and links each track with the previous one and the next one. Our case deviates from a parallel coordinates plot, because the most important information (namely the strength of the correlation between the SV and SNP) will not be layed out along the main axis. This axis is strictly to show the genomic position. Therefore the final plot that will be used is a variation on a plot called a Miami plot (Fig. 2). A standard Miami plot has 2 tracks, with one shared axis, usually the x-axis.

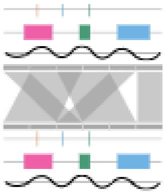


Figure 1: A visualisation with multiple horizontal tracks [5]

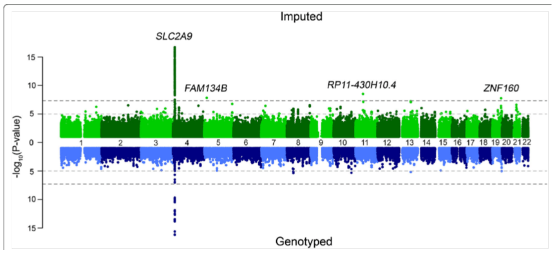


Figure 2: An example of a miami plot [4]

Then there are 2 y-axes to show two different types of data. This is used in genomic data visualisation to show and compare the p-values of multiple different studies or measurements. Our plot deviates from a standard Miami plot in the sense that it has three tracks. The top one shows the p-value of the correlation between the SNP and SV, the middle one shows the SVs location along the genome as well as the type and the bottom one shows the p-value of the association between the SNP and a chosen trait.

Data Selection

All the steps the user needs to take to get the data they want in the plot are layed out in Figure 3. First the user has the possibility to select a dataset corresponding with a specifc GWAS. This is then the dataset used to show certain traits. It is also possible to select a dataset which combines results from GWAS or has multiple traits for a certain SNP, for example the entire GWAS catalog dataset, reporting all associations of SNPs through GWAS. After that the user selects which part of the genome they would like to view. They can do this by choosing either a specific SNP, SV, trait or region. Choosing a SNP will show that specific SNP and no other SNPs, the p-value of the correlation between that SNP and one or multiple SVs (if one such correlation exists) and the p-value of the association with the trait form the first selected dataset. Choosing an SV will show that SV and all correlated SNPs within a given window (standard set to 25000) and all traits associated with those SNPs. Choosing a trait will show all associated SNPs and all SVs that are correlated with those SNPs. Choosing a region will show all SVs inside that region and again all SNPs with the given window and all traits associated with those SNPs. Since the dataset contains lots of SNP-SV correlations and most of them have fairly large p-values and are therefore not interesting to show in the plot, only those correlations with a p-value smaller than 5e-5 will be shown. Also the Bonferroni-correction [1] is applied to further decrease the amount of datapoints.

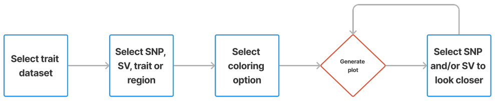

Select trait dataset
Select SNP, SV, trait or region
Select coloring option
Generate plot
Select SNP and/or SV to look closer

Figure 3: The steps a user goes through to see the data

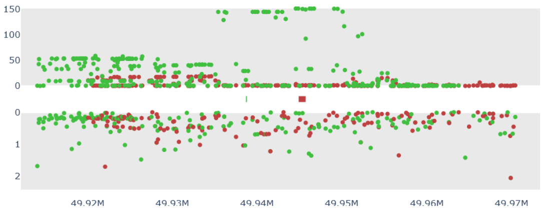


Figure 4: A plot showing 2 different SVs and their corresponding SNPs on the top and the traits on the bottom. The color option here is se to SV.

In order to create a further understanding of what might cause these correlations, other than just the genomic position, the user will be able to select from a variety of options regarding what extra data to show. The user can customize to color the graph based on (1) the type of SV, (2) the type of SNP, (3) the SV the SNPs and traits are correlated with. Handy in case you want to show multiple SVs in a single plot, (4) the Trait the SNP and SV are associated with. Handy to see which SV potentially has a high association with a trait and (5) whether the correlation between the SV and SNP is positive or negative.

Finally, when the plot is shown and any particular SNP-SV interactions might stand out, the user has

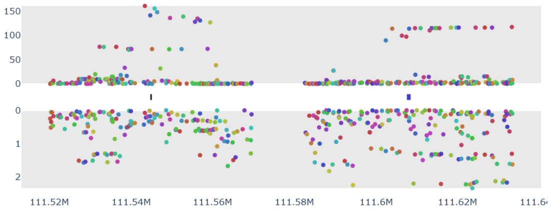


Figure 5: A plot with the color set to SNP (coloring each individual SNP)

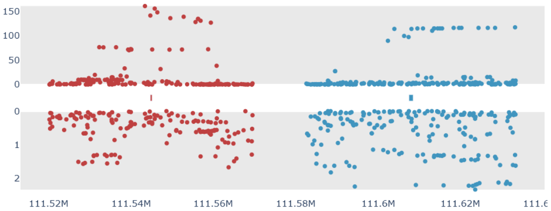


Figure 6: A plot with the color set to SV (coloring each individual SV)

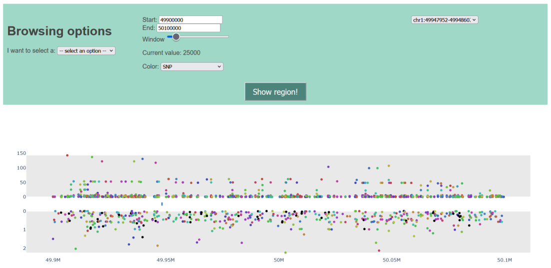
the option to select a specific SNP or SV to then only show that specific mutation.

The plots will be made using Plotly, which is graphing package for Python. Plotly was chosen because of its easy insertion into html for the web application and its intuitive interactability. Python was chosen to easily integrate this section into snpxplorer, which is also written in Python.

3 Results

3.1 The graphs

The main plot consists of three parts. The SNPs, the SVs and the traits as seen in Figure 4. On the top the p-value for the SNP-SV correlation can be shown and on the bottom the SNP-trait association. In the middle there is space for the SVs.


Browsing options

I want to select a: -- select an option --

Start: 49900000

End: 50100000

Window

Current value: 25000

Color: SNP

chr1:49947952-4994860?

Figure 7: The full UI

The main factor contributing to high load times is the start up of the app. The app will preload all data, which takes 44 seconds. However, in parallel with this project, this problem is also being researched, so this is not the primary concern. After the data has been loaded in, the data selection steps take a negligible amount of time. Finally the plot needs to be generated. The time this takes depends on the amount of data selected previously. Selecting one or multiple SNPs or SVs results in an amount of data so small that the load time of the plot is once again negligible. The real test is when choosing the 'region' option to load your data. Then all the SVs within the selected range will be plotted. Selecting a region of 1 million base pairs, with the window set to 25000 (this means that only SNPs within 25000 base pairs of the SV will be plotted) takes roughly 20 seconds. However, it often isn't necessary to take such a large region. In most cases 200,000 base pairs is more than enough, which takes roughly 3 seconds to load.

The plots are interactive up to a certain extend. Once the plot has been generated it is possible to zoom in to a section to look at a smaller range and scroll left and right, however the plots won't automatically update their data when scrolling further than the initially specified range. It is however not advised to load in very large regions and then zoom into the plot, because plotly will have to reload the plot every time a change happens caused by zooming or scrolling. This means that if you load a region of 1 million, every time you need to scroll left or right, you will have to wait 20 seconds. When hovering over a dot representing a SNP, the exact position, p-value and SV are shown. This is necessary in order for the user to be able to select this interaction in case they want to have a closer look at it. When hovering over an SV, the entire id is shown, which contains information about (1) the chromosome, (2) the position and (3) the type of type of variant.

3.2 Example

To illustrate how one might use the tool, the following case is provided. Say you want to know whether the SV with id chr1:111607529-111608515_JOIN has any SNPs associated with it that are also associated with Alzheimer. First you select the Alzheimer dataset, then you choose that you want to look at an

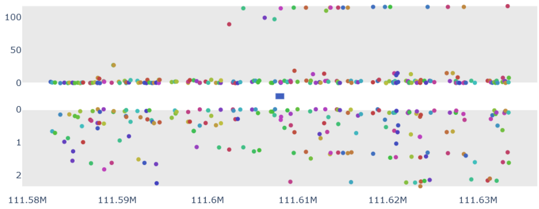


Figure 8: SV: chr1:111607529-111608515_JOIN

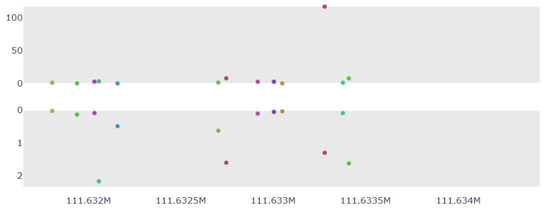


Figure 9: SV: chr1:111607529-111608515_JOIN zoomed in on a smaller region. The magenta dot on the right has a high association with the structural variant, but nothing special can be seen in the p-value with the trait.

SV and type chr1:116 into the search bar, by then this one is the only SV remaining, so you select it. For color you select SNP, because it will make it the easiest to differentiate different SNPs using this setting and for the window you set 25000. The plot in Figure 8 will then show up. You can then zoom in on the SNPs that have particularly high p-values to check whether the p-value of their association with the trait is also high (Fig. 9).

4 Responsible Research

4.1 Code and Data

The code is available on https://github.com/TomJacobsGit/research_project

For this project two different datasets have been used. The most important is the dataset containing the interactions between the SNPs and SVs. More information about how this dataset was generated can be found in appendix A. The other dataset used is a dataset containing the p-values between SNPs and traits for the disease Alzheimer. This dataset can easily be swapped out for any other dataset resulting from a GWAS.

4.2 Ethics

This paper introduces visualisation that will be used by other researchers in the future and other papers might even draw conclusions based on these visualisations. Therefore it is important to make sure that the visualisations are accurate and non-misleading. The similarity to other visualisations in this field helps in this regard. It is important to note however that, because there are two different y-axes, which both occupy the same physical space in the visualisation their steps van differ in size wildly as is often the case with these plots.

5 Discussion

This tool will be a great addition to the other tools out there. It is difficult to really compare it to tools which have the exact same purpose, because there are none, but we can compare it to the already existing snpxplorer and locuszoom. It closely follows the UI and user flow of snpxplorer

6 Conclusions and Future Work

With this project the basis is laid for visualising SNP-SV interactions. However there are several additions that can still be made: (1) Adding Linkage-Disequilibrium (LD) to the coloring options.LD is a way to express the correlation between SNPs. Just like we are currently looking into the correlation between SNPs and SVs it has been known for a while now that SNPs are also correlated amongst themselves. Showing this in the plot will be helpful to see whether SVs belong to these groups that SNPs often form or are maybe correlated with lots of different SNPs. (2) Adding a fourth part to the plot showing genes. Adding genes will allow researches to look for more interactions.

References

[1] Hervé Abdi. The bonferonni and Šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3, 01 2007.

[2] Miriam Ciani, Luisa Benussi, Cristian Bonvicini, and Roberta Ghidoni. Genome wide association study and next generation sequencing: A glimmer of light toward new possible horizons in frontotemporal dementia research. *Frontiers in Neuroscience*, 13, 2019.

[3] Marc-André Lemay and Sidiki Malle. *A Practical Guide to Using Structural Variants for Genome-Wide Association Studies*, pages 161–172. Springer US, New York, NY, 2022.

[4] Reka Nagy, Thibaud Boutin, Jonathan Marten, Jennifer Huffman, Shona Kerr, Archie Campbell, Louise Evenden, Jude Gibson, Carmen Amador, David Howard, Pau Navarro, Andrew Morris, Ian Deary, Lynne Hocking, Sandosh Padmanabhan, Blair Smith, Peter Joshi, James Wilson, Nicholas Hastie, and Caroline Hayward. Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 generation scotland participants. *Genome Medicine*, 9, 03 2017.

[5] S. Nusrat, T. Harbig, and N. Gehlenborg. Tasks, techniques, and tools for genomic data visualization. *Computer Graphics Forum*, 38(3):781–805, 2019.

[6] Niccolo Tesi, Sven van der Lee, Marc Hulsman, Henne Holstege, and Marcel J T Reinders. snpXplorer: a web application to explore human SNP-associations and annotate SNP-sets. *Nucleic Acids Research*, 49(W1):W603–W612, 05 2021.

[7] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1, 2021.

A Genration of dataset containing SNP-SV interactions

This appendix was fully written by Niccolò Tesi and not edited by me.

Population of the study We included N=214 individuals: of these, 93 are patients diagnosed with Alzheimer's disease from the Amsterdam University Medical Center,1 while the remaining N=121 are cognitively healthy centenarians from the 100-plus Study cohort,2 who are individuals of at least 100 years of age with good cognitive and physical abilities. Additional information regarding the samples included and the cohort is available elsewhere.3

PacBio HiFi long-read whole-genome sequencing and data processing All samples were sequenced using a Pacific Biosciences (PacBio) Sequel-IIe System with 2 hours of pre-extension time and 30 hours of collection time. After sequencing, raw reads were collected and analyzed through an in-house pipeline (freely available at https://github.com/holstegelab/snakemake_pipeline and previously described in depth.4 Briefly, this pipeline generates high quality HiFi reads for each sample and aligns data to the reference genome (GRCh38, patch release 14). When available, individual BAM-files from the same sample were merged into a single BAM-file. SNP were called using deepvariant,5 and compared with SNP array data available for all samples.

Structural variant calling, repeat annotation, and fine-mapping analysis Structural variants (SVs) were identified by using both reference-based and de novo assembly approaches. First, candidate SVs were identified using sniffles2,6 specifying –minsupport 2 and –mapq 20. We then retained only SVs that were observed in at least 11 genomes (2.5More specifically, TR annotations of GRCh38 were downloaded from the UCSC Genome Browser by retaining 'Simple_repeat', 'Low_complexity', and 'Satellite' annotations, and merging all annotations that were within 100 bp. SVs were classified as TRs if there was a 50Furthermore, each SV was classified as a TE by aligning the corresponding sequence of each SV with HMMER (v3.3.2, default parameters) to the repeat database, DFAM, using the human markov model (v3.7). All SVs with significant alignments as reported by HMMER were classified as TEs.

QTL analysis of SNPs and SVs We explored the potential of using a quantitative-trait-locus (QTL) based approach to estimate the relationship between SNPs and SVs. Compared to LD analysis, QTL analysis models linearly the relationship between SNPs and quantitative traits, e.g gene expression levels, blood pressure, or height. Given the higher variability of SVs compared to SNPs and the higher chance of multiple alleles being present for a SV, here, we considered the size of SVs as a quantitative trait. Thus, we tested the association between SV size and SNP genotype of all SNPs within 500kb up/downstream the SV location. For the association, we used linear regression models while adjusting for population stratification (PC 1-5). QTL analysis was performed with plink.

References

1. van der Flier, W. M. Scheltens, P. Amster-

dam Dementia Cohort: Performing Research to Optimize Care. Journal of Alzheimer's Disease 62, 1091–1111 (2018).

2. Holstege, H. et al. The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description. European Journal of Epidemiology (2018) doi:10.1007/s10654-018-0451-3.

3. Tesi, N. et al. Cognitively Healthy Centenarians Are Genetically Protected against Alzheimer's Disease Specifically in Immune and Endo-Lysosomal Systems. http://medrxiv.org/lookup/doi/10.1101/2023.05.16.23290049 (2023) doi:10.1101/2023.05.16.23290049.

4. Salazar, A. et al. An AluYb8 Retrotransposon Characterises a Risk Haplotype of TMEM106B Associated in Neurodegeneration. http://medrxiv.org/lookup/doi/10.1101/2023.07.16.23292721 (2023) doi:10.1101/2023.07.16.23292721.

5. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36, 983–987 (2018).

6. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol (2024) doi:10.1038/s41587-023-02024-y.

7. Tesi, N. et al. Characterising tandem repeat complexities across long-read sequencing platforms with TREAT. Preprint at https://doi.org/10.1101/2024.03.15.585288 (2024).