A novel way of visualising eQTLs relative to SNP-SV pairs using Gosling.js

Sonny Ruff[1]
Supervisor(s): Marcel Reinders[1], **Niccolo 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: Sonny Ruff
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Niccolo Tesi, Andy Zaidman

1 Abstract

Genome-wide association studies (GWAS) are commonly used to identify genetic variants associated with human traits by comparing genetic differences between diseased and healthy individuals. One way to gain insights into the biological consequences of these variants is to use quantitative trait locus (QTL) analysis. These connect SNP with variations in gene expression levels among individuals. QTL studies are mostly done on single nucleotide changes, but as SVs are bigger and have greater impact on traits, SV-QTL connections are of great interest. Using Gosling.js, a tool was developed to easily display the links and significance between SNPs, associated eQTLs, and SVs. The main purpose of this tool is to provide clear visualizations, while also offering options for further exploration of the chromosome. The existing search functionalities from snpXplorer have been integrated and enhanced. Users can define a variable window size before querying, allowing for flexible data examination. Additionally, the tool supports requests for data across multiple tissues. For improved performance and usability, the option to select which data tracks are shown were added.

Keywords: SNP, SV, QTL, tool, visualisation

As this paper will discuss technical topics outside the field of computer science, here is a brief explanation of the most important reoccurring terms.

- **GWAS:** A **Genome-Wide Association Study** scans the genomes of many individuals to find genetic variants associated with specific diseases or traits.
- **BP: Base pairs** are the unit of measurement in genetics that represents a pair of complementary nucleotides on opposite strands of a DNA molecule.
- **Locus:** A specific, fixed position on a chromosome where a particular gene or genetic marker is located.
- **SNP:** A **Single Nucleotide Polymorphism** is a single base-pair variation in the DNA sequence among individuals.
- **SV:** A **Structural Variation** is a large-scale alterations of DNA of >50 bp in size, and can include insertions, deletions, duplications, inversions and translocations of DNA segments.
- **eQTL:** An **expression Quantitative Trait Locus** is a genomic locus that correlates with variations in gene expression levels among individuals.

2 Introduction

Genome-wide association studies (GWAS) are a common approach to study the genetic factors associated with human traits. These studies compare the frequency of genetic variants throughout the genome between a group of diseased individuals, and healthy control group, in order to find genetic modifiers of the disease of interest. However, when such genetic modifiers are significantly identified, understanding the functional biological consequences of these genetic variants is difficult. One way to gain information about this is to use quantitative trait locus (QTL) analysis. These analyses connect genetic variants with tissue/cell type-specific cellular functions in different biological stages. QTLs come in many forms, ranging from the length of repeating elements in a different part of DNA, gene expression in specific tissues, to more complex phenotypical traits like human height. An example of an eQTL can be seen in figure 1. Here the genes in an area of 10Mbp (megabase pairs) around a SNP are shown with their expression in certain tissues as the intensity of the heatmap. This also highlights a limitation of this tool. Although P-values associated with whole genes can be viewed, the values associated to individual SNPs can't be viewed. If observing those values at that level of detail were an option, it would be possible to then see whether any of them coincide with any SNPs related to SVs. Although there exist tools for searching and analysing QTLs, and SVs and SNP-SVs separately, very little research has been conducted towards producing a proper way of visualising the relation between the two, which is of great interest for genetics researchers in order to understand SNP-effects and generate new hypotheses. Currently, QTL studies are mostly done on single nucleotide changes, but as SVs are bigger and have greater impact on traits, SV-QTL connections are of great interest.

It is suggested that by using a visualisation library specifically made to display various genomic elements such as genes, marks and transcription units, the question whether a better way of visualising the connections between SNP-SV pairs and eQTLs can be made, can be answered. This involved gaining an understanding of what these connections mean and how visualization must aid researchers in exploring these connections. The aim of this research was to create a tool that can show the same SNP-SV and SNP-eQTL correlations as other tools, in a single plot, while being easy to use and offering improvements over drawing conclusions from currently available tools by hand. Furthermore, it should minimize the amount of data displayed at once, to avoid clutter. Clarity is essential, ensuring all graphs and data points are appropriately labeled, with the ability to request and display more detailed information easily, without significant layout changes.

Tools like QTLBase [2] and GTEx Portal are examples of very sophisticated tools that already exist. QTLBase has functionality to show associations between genes and tissues, while GTEx Portal emphasizes associations between genes' specific loci and tissues. However, neither of these tools provides correlations to SVs. The problem arises when there is a need for a broader overview of the different connections between various elements such as SNP-SV and SNP-eQTL correlations. With the existing tools this requires a significant amount of work, involving manually searching for suitable datasets, further filtering and merging to extract these larger associations. The goal of this research is to create a tool that streamlines the steps of searching for relevant SNP-SV and SNP-eQTL data, integrating them into a unified dataset, and providing visualizations wider relationships. This new tool will be added to snpXplorer [1], a publicly available web-server for displaying GWAS associations.

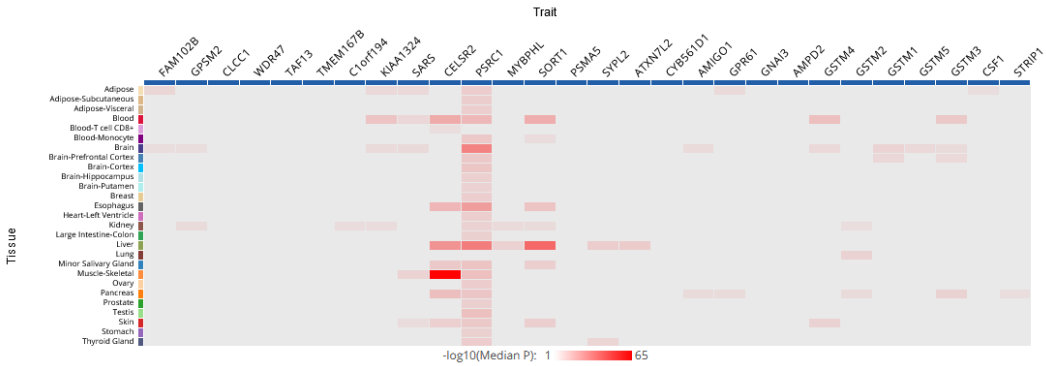

Figure 1: Heatmap showing an overview of eQTL for rs12740374 (± 10Mbp region) as shown on QTLbase [2]. On the x-axis are genes within the given range and on the y-axis a preset set of tissues. The hue of the cells indicate the -log10 value of the median P-value of the individual significant SNPs associated with the gene.

Materials & Methodology

2.0.1 Exploration of existing tools and relative implementation

The first step was finding existing methods of visualising SV, SNP-SV and QTLs. Of these tools the shortcomings were determined and what types of visualisation and user interface would improve it.

The libraries that were considered were Plotly, numpy and seaborn, because of how popular and widespread they are. They are very useful, high-level libraries, best suited for displaying relatively simple datasets. Two libraries that were specifically designed for visualising genomic data were also considered. The first was IGV.js, as it was used by some of the tools online. However, due to lack of documentation another library was considered. Gosling.js has a very thorough tutorial, documentation and list of examples to work with and therefore this one was chosen.

2.0.2 Data

Two datasets were being used: first a SNP-SV association dataset, and second an eQTL dataset. The SNP-SV dataset reports associations between SNPs and SVs based on 214 individuals. For each position of a SNP it contains information of the reference and alternate base and associated data like imputation quality, p-value and of course the SV. In the eQTL data the position entries also had reference and alternate base, but now also contained an array of tissues with associated p-values relating to expression strength. For it to be usable however, it required some transformation. However, due to it being 360.000 entries with many containing multiple associated tissues, some problems arose that will be discussed in the discussion section. An external dataset used for the gene track contains annotations for all genes, utilizing the genome assembly GRCh38. This dataset provides comprehensive information about gene locations, names and transcription units.

2.0.3 Implementation

The created visualisation is an extension on snpXplorer, which features basic visuals created using Plotly and matplotlib. For querying the SNP and eQTL data, pandas was used. Tabix, a specialized tool for retrieving genomic data from large csv files, was considered, but due to technical issues couldn't be used.

The SNP-SV and SNP-eQTL datasets are always first tried to be filtered, because of their immense size. Transformations are very slow, therefore, again preferably on as small a dataset as possible. Sometimes filtering isn't possible first, so the simplest possible transformation is done to extract the value to filter on.

Results

The web page consists of two sections: an input form and the Gosling.js plot. Initially, the plot isn't visible. However, once the form is filled in and the submit button is pressed, the plot appears. If a large number of tissues is selected, rendering the plot can take several minutes. The input form contains a list for selecting tissues, a field to input target, a slider to select a window size ranging from 1000bp to 250000bp, and a field to enter a P-value. The target can be specified in 4 different ways, entering a locus (e.g. 1:1000000), specifying a range (e.g. 1:1000000-1200000), a gene (e.g. APOE), or a variant (e.g. rs7412).

At the top of the main plot is a horizontal diagram of the chromosome that's currently being viewed. This serves as a simple navigation, to display where on the currently selected chromosome the view is located. To avoid confusion about this, it can't be scrolled or zoomed. It contains a track displaying the G-bands of the current chromosome and indicates the left side of the plot below it, with a vertical black bar.

Underneath this is the main plot. Gosling.js plots can consists of multiple subplots called tracks. All these tracks are connected along a common x-axis, representing the position on the chromosome. Users can drag and zoom the window to change the range of loci displayed, allowing them to view neighboring values or zoom in on detailed regions for closer examination. In the created tool, there are 7 horizontal tracks, from top to bottom, contain the following information: genes, SNP-gene links, SNP-eQTL P-values, the marker track, SNP-SV P-values, SNP-SV links, and SVs.

The **gene** track shows horizontal lines drawn from the transcription starting point to the end point, with vertical markings for all exons. If a gene is located on the + strand, it is colored blue and shown on the top half of the track. If it is on the - strand, it is red and on the bottom half.

The **SNP-gene links** track and **SNP-eQTL P-values** track are both generated, grouped together, for each tissue that is queried. The SNP-gene links track shows top-to-bottom links between between the SNPs locus and associated gene starting locus on the x-axis. The SNP-eQTL P-values track shows the P-values of SNPs associated with gene expression in the selected tissue. When the user hovers over any of the points in the plot, a tooltip is shown with detailed information about that point. It contains the exact locus, exact P-value, associated gene name, reference base and alternate base. The values are displayed as points, as it is possible for multiple genes to be associated with the same SNP. These can be differentiated by P-value and color. The points are colored in groups, corresponding to the gene they are associated with, but due to limitations of Gosling.js, these change based on the number of groups visible in the window.

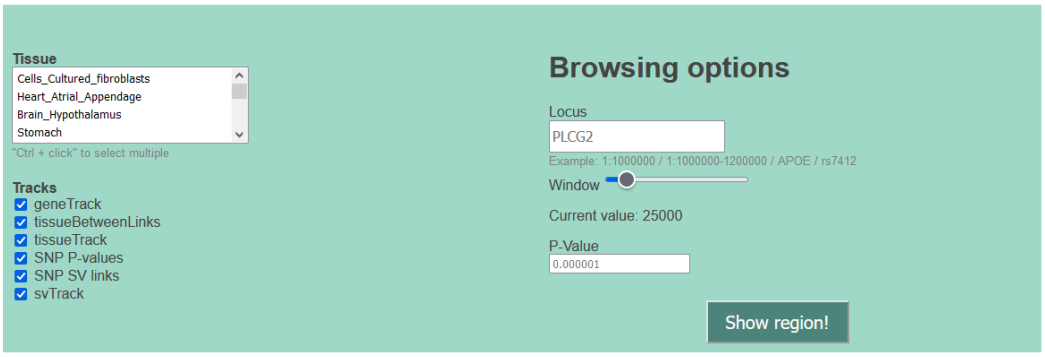

Figure 2: Main input form, showing the different input options.

The **marker track** in the middle only shows a dotted vertical line to indicate one of four things, based on the search query entered in the "Locus" field of the "Browsing options" form. When a locus or a variant is entered, it is simply located at that locus. If a range or a gene are entered, it indicates the start of the window that was entered, or the start locus of the gene.

The last three tracks are related to SVs. In structure they are very similar to the tracks related to QTLs, but flipped upside-down, such that the SNP P-values of both lie on top of each other, for easy comparison.

The first of them is the **SNP-SV P-values**, which shows the P-values of SNPs related to SVs, with the same option to hover over the points to display a tooltip with detailed information.

Underneath it is the **SNP-SV links** track. Just like the SNP-eQTL P-values track, this one shows top-to-bottom links, but now between the SNPs and associated SV starting locus.

At the bottom is the **SVs** track. In it, SVs are displayed with a simple grey coloration from their start to their end.

The checkboxes makes it able for the user to select which of the 6 tracks, outside the marker track, they want to be displayed. Since displaying is tied to the querying of data, due to the way Gosling.js works, deselecting tracks that aren't necessary in a particular use case, could significantly increase loading time of the plot. It also decreases the amount of elements that have to be displayed, which could help if a user's computer struggles with the visualisation. The most simple example would be deselecting the link tracks. Although it isn't immediately visible anymore which SNPs correspond to which gene or SV, that data can still be viewed with help of the tooltips. Besides that, the coloring still shows what other SNPs share those traits.

Another potential use case could be searching for tissues that have significant SNPs in a certain area of interest. By only selecting the **SNP-eQTL P-values** track, a plot like in figure 4 can be generated. It can be seen that an area such as 2:111.000.000-114.000.000 has a high number of significant SNPs contained in it, which then opens up the possibility for a more detailed search in this area.

An example of a real application of the tool would be to view the correlated eQTL and SVs of variant rs6966331. It is already known that there exist a correlated eQTL in Whole_Blood and SVs. When querying the tool with a window size of 1000, a P-value of 10e-5, only selecting Whole_Blood as the eQTL tissue and selecting all track types, a

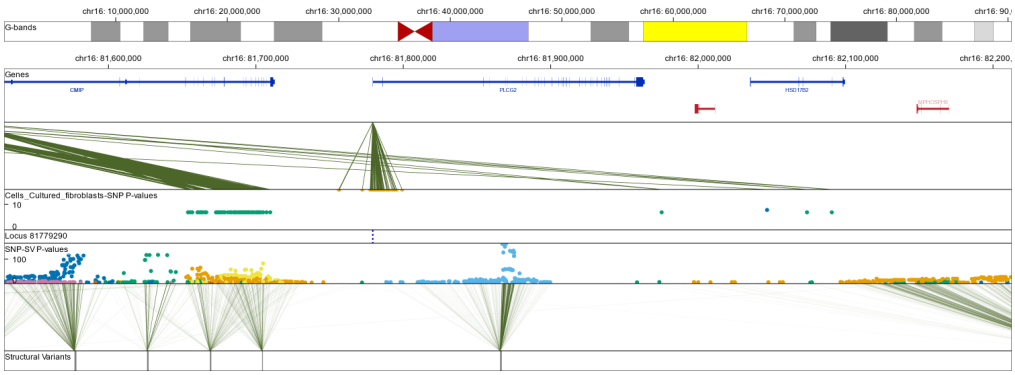

Figure 3: snpXplorer QTL visualisation example. The navigation track is visible at the top, showing the full range of chromosome 16. Beneath it, with a seperate x-axis, are the 7 main tracks available visible.

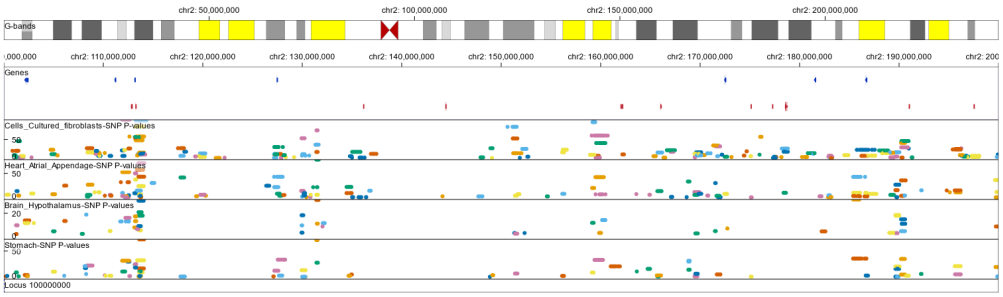

Figure 4: An example plot showing the eQTL P-values of 4 different tissues. Some tissues have groups of points, where others have none.

plot as seen in figure 5, will appear. Zooming out a little bit and adjusting the x offset, we can get an overview like in figure 6. Here it can be clearly seen that variant rs6966331 indeed has correlation with gene NME8 and two correlated SVs. When hovering over each respective element, detailed information about them can be viewed; Fig. 8, Fig. 9, Fig. 10. Zooming out even further, the full correlated gene and the location of the two SVs can be easily found by tracing the links, as seen in figure 7. It can be seen that there are two SVs correlated with SNP rs6966331, one close to the right and one relatively far away to the left. Additionally, 14 other neighboring SNPs that are correlated with both the same gene and these SVs can be easily identified, providing further insights into the genetic structure surrounding this region. These variants, that both have associations with eQTL and SVs, are very interesting, as SVs typically have a greater impact on genetic variation and phenotype than SNPs.

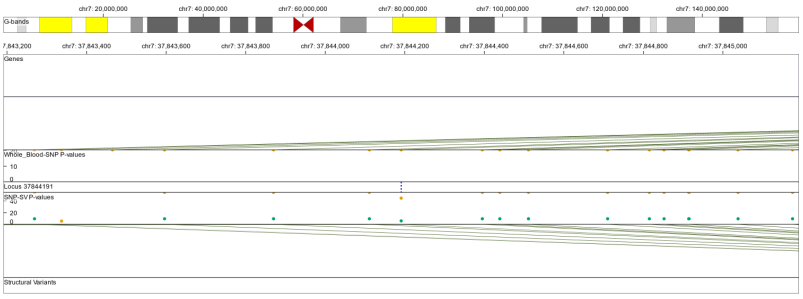

Figure 5: snpXplorer QTL visualisation plot for rs6966331, P-value = 10e-5, window = 1000. The locus of the variant is indicated by the blue dotted line in the middle track. 14 other SNPs that have both an eQTL in whole blood and an SV can be seen, 5 left of locus 37,844,191 and 9 right of it.

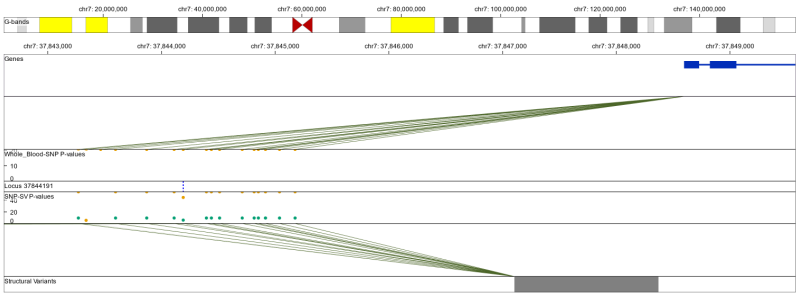

Figure 6: snpXplorer QTL visualisation plot for rs6966331, P-value = 10e-5, window = 1000. X-axis has been zoomed out to have the start of gene NME8 in view on the top right. At the bottom SV 7:37,847,104-37,848,371 is visible.

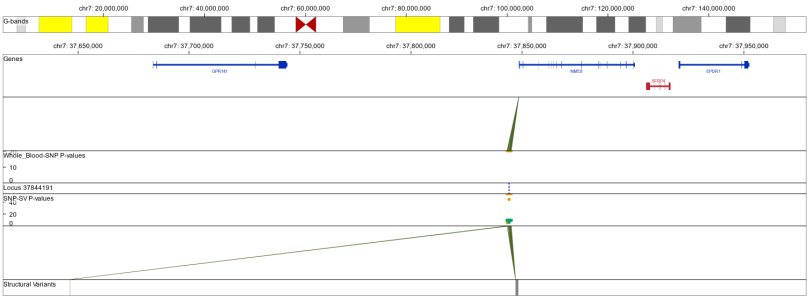

Figure 7: snpXplorer QTL visualisation plot for rs6966331, P-value = 10e-5, window = 1000. X-axis has been zoomed out to fully see gene NME8 at the top and both associated SVs at the bottom.

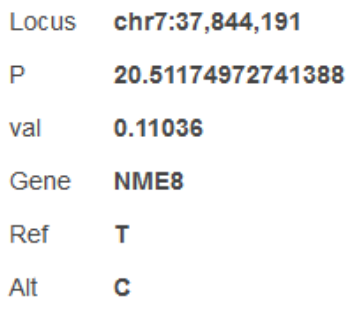

Figure 8: Tooltip showing detailed information for the correlated whole blood eQTL of variant rs6966331.

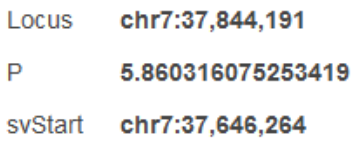

Figure 9: Tooltip for the first of the two correlated SVs of variant rs6966331, showing P-value and starting locus of the SV.

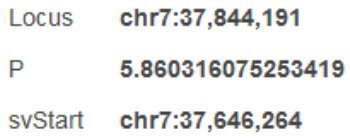

Figure 10: Tooltip for the second of the two correlated SVs of variant rs6966331, showing P-value and starting locus of the SV.

Discussion and Recommendation

A quickly apparent problem is the querying speed. The loading and filtering of datasets are handled by pandas. As this library is not optimized for these operations, simple functions like filtering become bottlenecks. As discussed in the implementation section, using Tabix or a similar database tool, such as the one created by group member Nick van Luijk, could significantly reduce querying time and improve overall usability.

Another point of improvement, that could reduce querying time and improve clarity, would be to have the groups of SNPs and their associated SV be colored the same. However, coloring marks in Gosling is not something that can be done by supplying the data with a color value. Currently, the colors are derived from the name of the associated gene, which means there is no further manual control over them. This feature was kept in, to show groups of SNPs with either a common associated gene, or common SV. The associated gene or SV

is however not colored corresponding to this. Improving this would open up the possibility for users to choose color schemes and choosing colors that are sufficiently far apart for easy differentiation, as this would improve usability for users with color vision deficiency (like myself).

Another future change would be adopting a different format for the data files. Having Gosling load CSV files is limiting. Using a format like BEDDB, as demonstrated in various Gosling examples, might improve performance and functionality. For instance, displaying exons of genes can be done with simple filter functions in Gosling, rather than requiring a complete transformation of the CSV data.

That said, there are several advantages to consider. With the current use of pandas and CSV data files, extending the current implementation should be relatively easy. Pandas is a commonly used library that is both powerful and easy to learn. The CSV format is also widely used and user-friendly, which eliminates the need for transforming data into a less common format.

Conclusion

The created tool succeeds in its original goal of helping researchers to integrate and overlap multiple genomic datasets such as SNP, SV and QTL, but not without its flaws.

The timespan of 2 months given to complete this research, greatly limited the extent to which tasks such as in-depth researching, learning the necessary programming languages and libraries, implementing the tool, and addressing bugs, could be completed. Nonetheless, a lot was achieved.

Integration with the existing snpXplorer codebase should be straightforward as the new endpoints for data retrieval can simply be incorporated next to the existing ones. No further new dependencies were introduced.

With this framework in place, new possibilities for further development open up. Integration of other datasets and the possibility for the user to upload their own data, could be beneficial. Another possibility is a generalization of the current used method of joining the two main datasets. This could open up features, like deciding in the webpage itself which datasets should be joined together to show connections between them. Further, Gosling.js offers dummy tracks, which can be used with other visualization libraries, facilitating integration with, for example, a lower-level visualization library like D3.js. This would allow for far more intricate and customized graphics. In conclusion, this project represents a start. With further creativity and development, the tool can be expanded to offer even greater capabilities, providing researchers with a robust resource for genomic data visualization.

Responsible Research

The tool was developed on Windows 10 and has not been tested on other operating systems. It was built using Python 3.12.0 and Flask 3.0.3. All code is available on GitHub at https://github.com/sonnyruff/snpXplorer-QTL. No other proprietary software has been used. All data sources are either publicly accessible or can be requested.

References

[1] N. Tesi, S. Van Der Lee, M. Hulsman, H. Holstege, and M.J.T. Reinders. SnpXplorer: A web application to explore human SNP-associations and annotate SNP-sets. *Nucleic Acids Research*, 49:W603–W612, 2021.

[2] Zhanye Zheng, Dandan Huang, Jianhua Wang, Ke Zhao, Yao Zhou, Zhenyang Guo, Sinan Zhai, Hang Xu, Hui Cui, Hongcheng Yao, Zhao Wang, Xianfu Yi, Shijie Zhang, Pak Chung Sham, and Mulin Jun Li. QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. *Nucleic Acids Research*, 48(D1):D983–D991, January 2020.