Computational Genomics Project 4

Patryk Prusak

Warsaw University of Technology

June 14, 2024

Abstract. The report describes the successful development of an algorithm for SV-eQTL analysis on provided RNA-Seq data and SV data. The designed program has been executed for chromosome 19 by performing linear regression on processed data resulting in 223 discovered SV-eQTL that can be considered significant. Regardless of the results, there is room for improvement, the correctness of the constructed analysis should be measured, and the algorithm could undergo a certain amount of fine-tuning to improve running times.

Contents

1 Problem Description

The aim of this project is to perform SV-eQTL analysis on RNA-Seq data and a list of population SVs. The deliverables should consist of code, a report, and a file with a list of discovered SV-eQTL along with the significance (in the form of p-value) of the performed regression.

Quantitative trait locus (QTL) are defined as areas of the genome connected to certain traits such as height or lung capacity. If the trait is an expression of a gene then we are faced with an eQTL - Expression quantitative trait loci [1]. To detect SV-eQTL one can investigate the relationship between the gene's structural variants and the gene's expression levels.

2 Application Instruction

The project has been implemented in the form of a Jupyter Notebook in Python. To reproduce the results one needs to install the required packages, supply the required datasets, and execute the cells one by one. As a result a file with discovered SV-eQTL will be produced.

3 Procedure

The SV-eQTL detection is realized on RNA data from the gEUVADIS consortium (GD660.GeneQuantRPKM) [2] and Structural Variants data from 1000 Genomes - Phase 3 [3]. A number of preprocessing steps first have been conducted to ensure the data is in the correct format.

1. Ensuring columns of interest are in a proper format (i.e. column indicating chromosome type initial is a mix of ints and strings).

2. Filtering the data by a specific chromosome (19). This has been done solely to reduce computation time.

3. Identyfing common columns between the two datasets.

4. Preparing to handle different structural variants separately:

   (a) Splitting the values in the "ALT" column in SV data by ",".

   (b) For each SV collecting relevant columns containing information regarding specific SV.

   (c) Mapping the columns representing alleles in the following format: 0|0 -> 0, x|0 -> 1, 0|x -> 1, x|x -> 2, where x stands for an index corresponding to SV in "ALT" column in the original split array.

5. For each gene, finding 3 closest RNA data points and creating a new data frame from the collected information.

With such dataset one can perform linear regression on each gene and chosen RNA-Seq data. One needs to fit the best line to data following the pattern visible in figure 1, note that values of ’x’ can range from 0 to 2. The Ordinary Least Squares statistical regression is conducted for each row of the created dataset and a result containing information (gene position, gene id, structural variant, rna id, rna position, p-value) about the procedure is collected and saved to "regressionResults.csv" file.

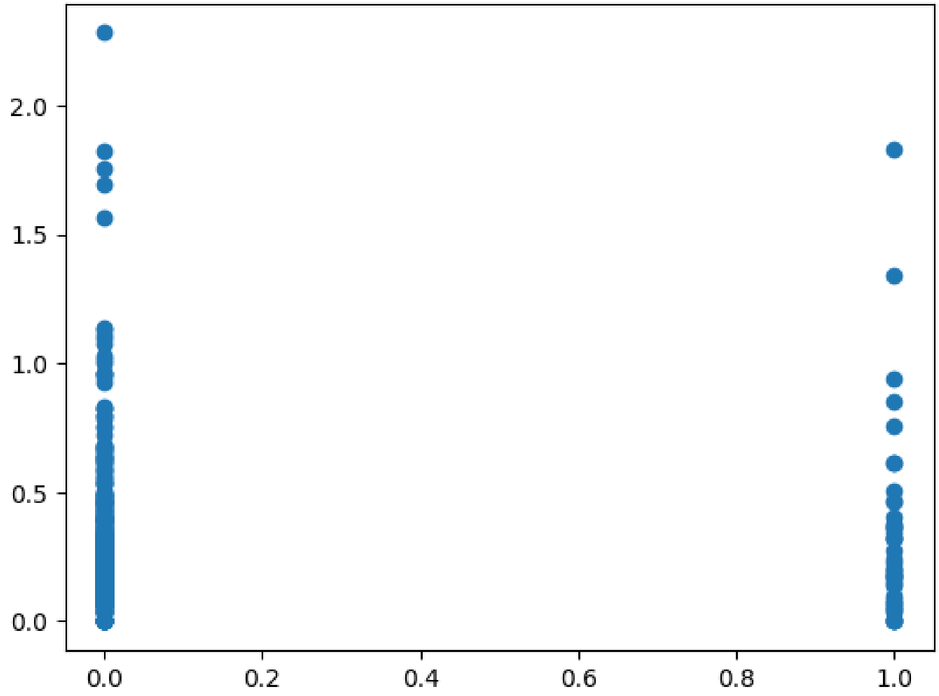

Figure 1: Data prepared for linear regression

4 Results

The aforementioned procedure resulted in 223 discovered SV-eQTL (filtered by chromosome 19) ranging from copy number variants to insertions, this is better represented by figure 2. The SV-eQTL analysis detected the CN0 variant most frequently.

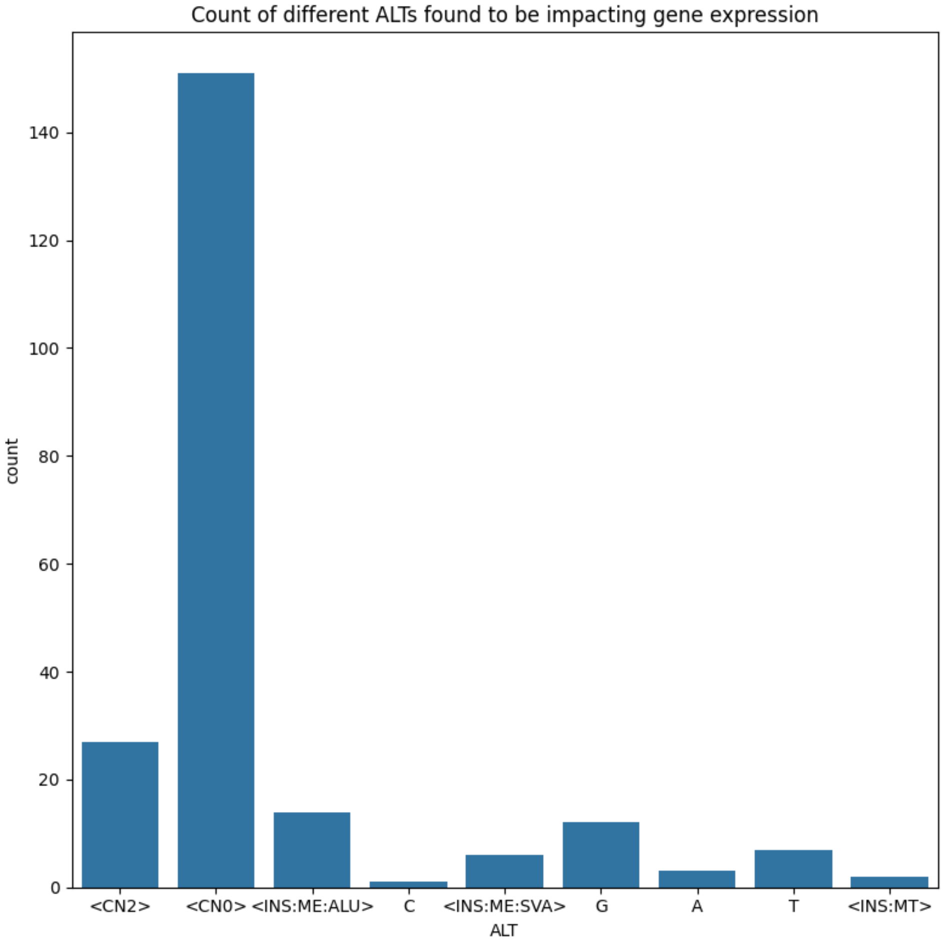

Figure 2: Count of different ALTs found to be impacting gene expression

5 Conclusions

An algorithm capable of SV-eQTL analysis has been successfully designed and implemented. Moreover, Provided datasets have been preprocessed and prepared for linear regression. Most importantly, linear regression has been conducted successfully resulting in 223 discovered SV-eQTL for chromosome 19. However, the prepared solution lacks performance comparison with established and robust methods to ensure the solution's correctness and the designed algorithm could be improved in terms of efficiency to ensure the analysis could be comfortably conducted for all chromosomes.

List of Figures

References

[1] Nica, A. C., Dermitzakis, E. T. "Expression quantitative trait loci: present and future. Philosophical transactions of the Royal Society of London." In: *Biological sciences* (2013), 368(1620). URL: https://doi.org/10.1098/rstb.2012.0362.

[2] Emmanouil Dermitzakis, Natalja Kurbatova Tuuli Lappalainen. "RNA-sequencing of 465 lymphoblastoid cell lines from the 1000 Genomes." In: *BioStudies, E-GEUV-1.* (2012). URL: https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-GEUV-1.

[3] *1000 Genomes - SV from Phase 3.* URL: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/.