GenomeSeqView: Visualization app for Normalized Exon Coverage Data

Pranav Manjunath, Mohammed Kanchwala, Adwait Sathe, Ph.D., Chao Xing, Ph.D.
McDermott Center Bioinformatics Lab, UT Southwestern Medical Center, Dallas, Texas

Abstract

Whole genome and whole exome sequencing have been adopted widely to find disease causing variants. One of the most important quality control steps is to check whether the genes of interest have enough read coverage for further processing. But since this is high dimensional data, there is a lack of visualization tools that show this information in a dynamic and speedy way. Our objective in this project was to develop such a comparative visualization tool which can create intuitive plots on the go.

Introduction

Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) [1] are used to obtain the nucleotide sequence of an individual's genome and exome (protein-coding portion of genome) respectively.

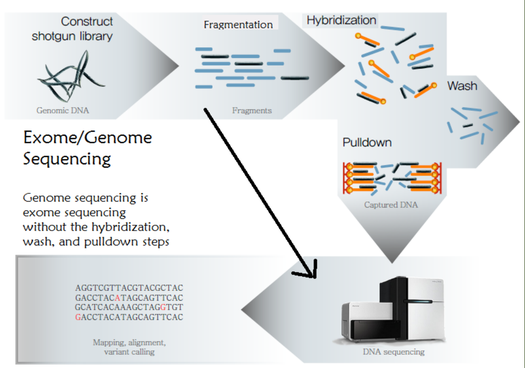


Figure 1: Flowchart showing steps in WGS/WES

Materials and Methods

Linux bash scripts were used to pre-process the sample files before analysis. The main tool used for the actual visualization is R Studio. Two particular R packages are important in the visualization process: the data.table package and the Shiny package.

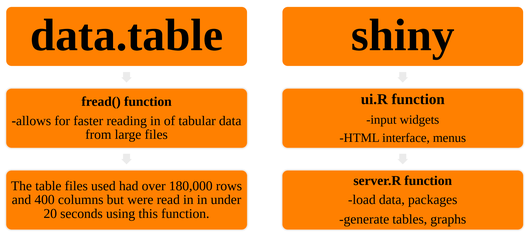


Figure 2: Flowchart showing organization of data.table and shiny packages in R and how they were utilized

Results

This is where the user uploads the data table, like output from featureCounts [3]

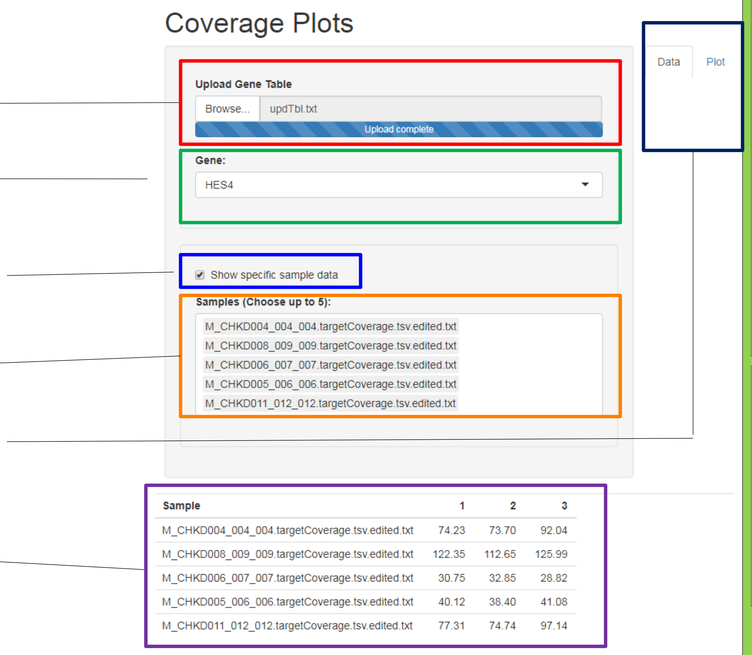
This menu appears only after the table is successfully uploaded and read in. This is where the user selects their gene of interest

This checkbox is selected if the user wants to see data for the selected gene from specific samples only

This menu is where the user selects the specific samples

This panel is where the user switches between seeing tabular data and the actual plots

This is the specific sample tabular data that is being displayed. If no specific samples were selected this table would instead contain data from all the samples for the selected gene



Sample	1	2	3
M_CHKD004_004_004.targetCoverage.tsv.edited.txt	74.23	73.70	92.04
M_CHKD008_009_009.targetCoverage.tsv.edited.txt	122.35	112.65	125.99
M_CHKD006_007_007.targetCoverage.tsv.edited.txt	30.75	32.85	28.82
M_CHKD005_006_006.targetCoverage.tsv.edited.txt	40.12	38.40	41.08
M_CHKD011_012_012.targetCoverage.tsv.edited.txt	77.31	74.74	97.14

Figure 4: A view of the HTML interface of the shiny app that shows a menu where the user selects a gene and up to 5 specific data samples, displaying a tabular version of the data from these samples corresponding to the selected gene

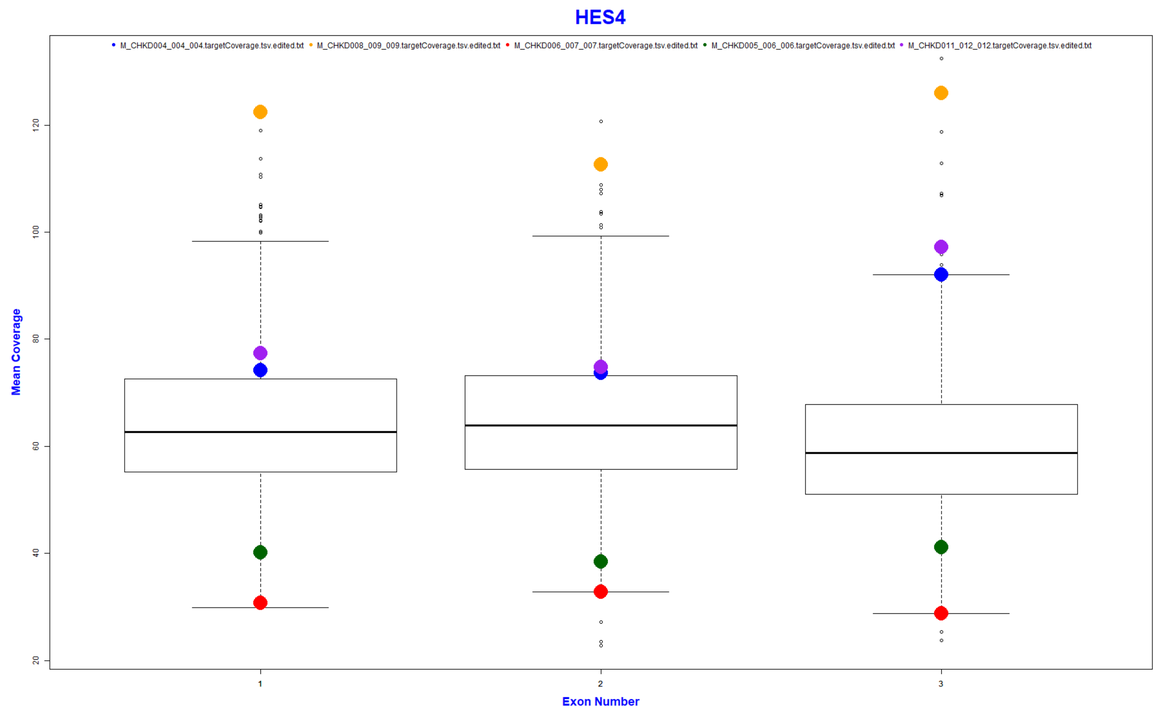


Figure 5: A boxplot with scatter overlays for the gene HES4 made by selecting the options shown in Figure 4

Discussion

The major advantage of this application tool is that a large amount of data can be visualized dynamically on the go by selecting the genes very quickly. In this project we have tested 400 samples with coverage data for 187,383 exons. We have tested up to 2,000 samples and plan to test for more.

This visualization tool helps convert the normalized read counts like RPKM (Reads Per Kilobase of transcript per Million mapped reads) and TPM (Transcripts per Million mapped reads) into a visual representation of the genomic regions. The regions which are enriched with reads have greater coverage depth. Here the plots can be produced for single or multiple genes with coverage shown per exon. The advantage of this is that Homozygous exon deletions can be identified and visualized easily [2].

Future Direction

In the future, making the input format more flexible would be beneficial because the app currently only accepts specifically formatted table files. We would also like to add some more biological analysis so that potential disease-causing exon deletions can be identified within the app itself without the need for outside analysis. One idea we have is to integrate the app with an existing R script called HMZDelFinder [4] that can find these deletions using RPKM data.

Literature cited

1. Biesecker, Leslie G, and Robert C Green. "Diagnostic Clinical Genome and Exome Sequencing | NEJM." New England Journal of Medicine, Oxford University Press, 19 June 2014, www.nejm.org/doi/full/10.1056/NEJMra1312543.
2. Gambin T, Akdemir ZC, Yuan B, et al. Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. Nucleic Acids Research. 2017;45(4):1633-1648. https://doi.org:10.1093/nar/gkw1237.
3. Yang Liao, Gordon K. Smyth, Wei Shi; featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics, Volume 30, Issue 7, 1 April 2014, Pages 923–930, https://doi.org/10.1093/bioinformatics/btt656
4. Gambin, T., Akdemir, Z. C., Yuan, B., Gu, S., Chiang, T., Carvalho, C. M. B., ... Lupski, J. R. (2017). Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. Nucleic Acids Research, 45(4), 1633–1648. http://doi.org/10.1093/nar/gkw1237

Acknowledgments

I would like to start off by thanking Dr. Joel Goodman, Ms. Lynn Tam, and Ms. Maria Sandlin for the opportunity to partake in the STARS program. Getting to work in a professional environment at such a young age has taught me so much and will prove to be invaluable the next time I get an internship or job opportunity. I would also like to thank Dr. Chao Xing for giving me the opportunity to intern in the McDermott Center of Bioinformatics and making me feel welcome from day one. Lastly, I would like to thank Mr. Mohammed Kanchwala and Dr. Adwait Sathe for their incredible mentorship, without which this project would certainly not have been possible.