Genetic Output Analysis Tool (GOAT)

Beatriz Kanzki*, Victor Dupuy, Cedric Urvoy, Fodil Belghait, Alain April**
École de Technologie Supérieure (ÉTS)
1100, rue Notre-Dame ouest,
Montréal, QC, Canada
beatriz.kanzki.1@ens.etsmtl.ca
victor.dupuy.1@ens.etsmtl.ca
cedric.urvoy.1@ens.etsmtl.ca
fodil.belghait.1@etsmtl.ca
alain.april@etsmtl.ca

François Harvey, François- Christophe Marois-Blanchet, Michael S. Phillips, Johanne Tremblay** and Pavel Hamet**
Centre de Recherche du Centre Hospitalier de l'Université de Montréal, CRCHUM
900, rue Saint-Denis,
Montréal, QC, Canada
francois.harvey.chum@ssss.gouv.qc.ca
françois-christophe.marois-blanchet.chum@ssss.gouv.qc.ca
pgxdoc@gmail.com
johanne.tremblay@umontreal.ca
pavel.hamet@umontreal.ca

* First Author, ** Co-Senior Authors & to whom correspondence should be sent.

***Abstract*—To identify complex associations in large patient cohorts, researchers use genome wide association studies (GWAS). This type of study involves a vast amount of clinical and genetic data. In order to analyze and visualize these complex datasets efficiently we have developed an open source Genetic Output Analysis Tool (GOAT) that facilitates the visualization and annotation of GWAS data. GOAT offers the ability to interactively search GWAS datasets via specific queries to identify significant associations between multiple SNPs and phenotypes. It was designed to be scalable and operate on top of "Big Data" technologies. It is programmed in python and can be connected directly to any database using an Apache open source server. This paper outlines some of GOAT's leading features and internal structure. Finally, we present future development plans for GOAT that will provide researchers with improved visualization functionality and the ability to mine GWAS data easily.**

Index Terms*— *Bioinformatics; GOAT; GWAS; Genomic region; Visualization tool, Software Engineering, Spark, Big Data.

I. Introduction

Genome Wide Association Studies (GWAS) are a widely utilized method applied to large genetic datasets to identify genetic variants associated with specific diseases or phenotypes and their associated complications [1]. Output from these types of studies involves the analysis of millions of single nucleotide polymorphisms (also called SNPs) which are examined concurrently and classified by their strength of association with the disease. GWAS results are initially visualized using Manhattan plots where the –log of the associated SNP p-values are plotted against physical map position of SNPs in relation to their chromosomal position within the genome [2]. It is instrumental that researchers have the ability to query all this complex data interactively across multiple phenotypes even though most of the associations do not reach a nominal significance threshold, and specially if the researcher's field involves complex diseases. Complex diseases involve SNPs responsible for the disease that will be in a non-coding area of the DNA, or will be in linkage desequilibrium (LD) with the most significant SNPs, so it's very important to capture all SNPs surrounding those areas [3]. Thus, there is a current need for improved visualization tools that can help geneticists to capture "ALL" relevant information and allow a better understanding of complex "OMICS" data at multiple levels facilitating the generation of novel pathogenetic hypotheses. Current tools to visualize these kind of datasets lack performance, efficiency and don't always provide publication ready graphs. Furthermore, the datasets used are usually stored in a relational database and, for the average genomic researcher, becomes challenging to effectively query very large amounts of data. These considerations constituted our impetus to design the Genetic Output Analysis Tool (GOAT).

This paper explains GOAT's key functionalities, and compares it to two well known tools: LocusZoom [4] and Integrative Genomics Viewer (IGV) [5]. Finally we outline plans for future improvements of GOAT.

II. GOAT 'S KEY FUNCTIONALITIES

The first version of GOAT was designed to give researchers the flexibility and independence to query multiple GWAS without requiring a bioinformatician expert's help. GOAT allows to query multiple GWAS datasets using either gene name or SNP rs ID.

The idea behind this interactive interface is to enable researcher's who don't have the appropriate programming skills or database knowledge to search for the most interesting results, nor the knowledge about bioinformatics file formats while at the same time, generate visual tools adapted to the genetic research domain to support their hypothesis. GOAT achieves this by selecting the most interesting SNPs and phenotypes, above a 0.001 threshold on the p-value, from its database and generates a result page containing an interactive Manhattan plot and a table result, where all significant SNPs are ordered from the most significant to the least significant (see Fig. 1). The table under the graph provides the user with the ability to filter, sort, select or mask specific columns, so that all the relevant information needed for analysis and interpretation can be displayed and extracted. The interactiveness of the Manhattan plot enables the researcher for

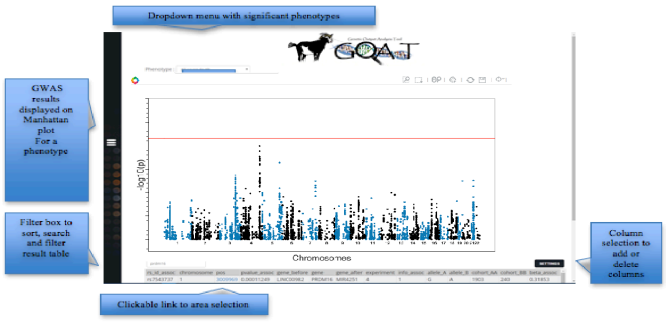

Figure 1. Result page generated after query by gene or rs ID with a selected threshold of 0.001.

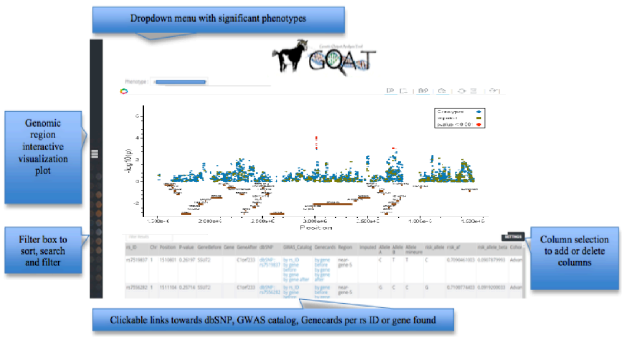

Figure 2. Area selection result page generated by selecting SNP on chromosome.

fast identification of interesting peaks providing information on rs number, gene name and p-values in a tag. GOAT also has the ability to interactively zoom into any genomic region. When a SNP is selected, from the general view containing the Manhattan plot, a new window appears providing the user with an interface named "Area Selection" which displays interactively the selected SNP within a +/- 1500 kbp interval. This clearly shows all surrounding genes in the region (see Fig. 2) and allows easy navigation features like: 3D views, zooming capabilities, mouse over display, and the ability to download the resulting graph, at any stage of the process, in a publication ready format. After comparison with LocusZoom and IGV a summary table highligting GOAT's features was produced (See Table 1). We compared performance, efficiency and the ability to produce publication ready plots. While these tools have different functionalities, very modern features were implemented in GOAT in order to create a whole new visualisation experience that, we hope, will satisfy even the most advanced genomic researcher.

Table 1. Summary of results after comparison of GOAT's performance, efficiency and the ability to produce publication ready plots to visualize genomic regions, with commonly used tools such as LocusZoom and IGV.

Plots comparison	LocusZoom	IGV	GOAT
Genomic interval	+/- 200 kbp	+/- 124 mbp depending on chr *	+/- 1.5 mbp
Zooming	No	Yes	Yes
3D view	No	No	Yes
SNP indentification	Yes	Yes	Yes
Interactive	No	Yes	Yes
Publication ready	Yes	No	Yes
Time to generate	~20 seconds	~ 1 second	~ 5 seconds
Prior data formatting	Yes	Yes	No

**chr: chromosome.*

III. GOAT'S SOFTWARE DESIGN

For the user interface (UI), technologies such as React[6], SASS[7], SMACS[8] were used in order to allow scalability and easyness of future maintenance of the UI. For the core functionalities, the recent Python framework named Django [9] was used. Other popular python scientific libraries such as Numpy, Pandas and Bokeh, from Continuum Analytics, to manipulate all tables and create interactive graphs were used. GOAT is composed only of open source software.

CONCLUSION

GOAT's key functionalities gives genomic researchers a user friendly user interface and gives them the ability to be autonomous during their GWAS discovery iterations. It also generates publication ready Manhattan graphs, as well as allow them to interactively data mine massive amount of genetic data. GOAT is released as an open source project on the github under GOAT_Genetic_Output_Analysis_Tool.git. Future releases of GOAT will offer features such as phenotype comparisons, Linkage desequilibrium, recombination rate of SNPs, do statistical analysis and use recent machine learning techniques. The performance and scalability of GOAT is planned to be improved by replacing the current MySQL back-end with an Apache Spark back-end [10].

REFERENCES

[1] Londin E, Yadav P, Surrey S, Kricka LJ, Fortina P. Use of linkage analysis, genome-wide association studies, and next-generation sequencing in the identification of disease-causing mutations. In Pharmacogenomics 2013 Jan 1 (pp. 127-146) Humana Press.

[2] Balding DJ. A tutorial on statistical methods for population association studies. Nat. Rev. Genet., 2006 Oct 1;7(10):781-91.

[3] Grossman E, Messerli FH. Hypertension and diabetes. Advances in cardiology. 2008;45:82.

[4] Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010 Sep 15;26(18):2336-7.

[5] Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2012 Apr 19:bbs017.

[6] Gackenheimer C. What Is React?. InIntroduction to React 2015 (pp. 1-20). Apress.

[7] Prabhu A. Introduction to Preprocessors. InBeginning CSS Preprocessors 2015 (pp. 1-12). Apress.

[8] Jain N. Review of different responsive CSS Front-End Frameworks. Journal of Global Research in Computer Science. 2015 Apr 2;5(11):5-10.

[9] Moore D, Budd R, Wright W. Professional Python Frameworks: Web 2.0 Programming with Django and Turbogears. John Wiley & Sons; 2008 Jan 22.

[10] Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. InProceedings of the 9th USENIX conference on Networked Systems Design and Implementation 2012 Apr 25 (pp. 2-2). USENIX Association.