snpXplorer: a web application to explore SNP-associations and annotate SNP-sets

Niccolo Tesi [1,2,3], Sven van der Lee [1,2], Marc Hulsman [1,2,3], Henne Holstege [1,2,3] and Marcel J.T. Reinders [3*]

[1] Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2] Department of Clinical Genetics, Amsterdam UMC, Amsterdam, The Netherlands

[3] Delft Bioinformatics Lab, Delft Universtity of Technology, Delft, The Netherlands

* To whom correspondence should be addressed.

Running title: snpXplorer to visualize and annotate genetic association studies

Abstract

Genetic association studies are largely used to study the genetic basis of numerous traits. However, the interpretation of genetic associations is often complex and requires the integration of multiple sources of annotation. We developed *snpXplorer*, a web-server application for exploring SNP-association statistics across human traits and functionally annotate sets of SNPs. *snpXplorer* allows superimposition of association statistics from multiple studies, and displays regional information including SNP-associations, structural variations, recombination rates, eQTL, linkage disequilibrium patterns, genes and gene-expressions per tissue. *snpXplorer* can be used to compare levels of association across different phenotypes, which may help the interpretation of variant consequences. Interestingly, given a list of SNPs, *snpXplorer* performs variant-to-gene mapping and gene-set overlap analysis to identify molecular pathways that are enriched in the genes likely associated with the variants of interest.

Availability: *snpXplorer* is freely available at http://snpxplorer.eu.ngrok.io. The source code and application documentation are available at https://github.com/TesiNicco/SNPbrowser.

Contact: n.tesi@amsterdamumc.nl

Supplementary information: *snpXplorer* documentation is available at https://github.com/TesiNicco/SNPbrowser/blob/master/snpXplorer_documentation.pdf.

Keywords: visualization, GWAS, SNP-annotation, integration, gene-set analysis

Introduction

Genome-wide association studies (GWAS) and sequencing-based association studies are a powerful approach to investigate the genetic basis of complex traits and their heritability. Facilitated by the cost-effectiveness of both genotyping and sequencing methods and by established analysis guidelines, the number of genetic association studies has risen steeply in the last decade. As an example, as of October 2020, the GWAS Catalog, a database of genetic association studies, contains 4,761 publications and 213,519 variant-trait associations.[1] However, the interpretation of the functional effects of genetic variants is often challenging: most genetic variants in the genome rely in non-coding regions, which requires variants to be investigated independently combining diverse annotations sources. In addition, to better understand how genetic factors affect different traits, it can be informative to explore the extent of association of a genomic region on different phenotypes. To address these challenges, we have developed *snpXplorer*, a web-server application written in R that allows (*i*) exploration and superimposition of summary statistics from multiple genetic association studies, and (*ii*) functional annotation and pathway enrichment analysis of SNP-sets in a catchy and easy-to-use user interface.

snpXplorer

Exploration section

snpXplorer is a web-server application written in R and based on the R package *shiny* that offers an exploration section (*Figure 1A*-C) and a functional annotation section (*Figure 1D-G*).[2] The exploration section represents the main interface: first, an input data must be chosen, which can either be one of the available summary statistics datasets or the user can upload own association data. The number of available summary statistics will constantly grow and we keep the most updated studies in place. Alternatively, the user can select own association statistics: multiple formats can be recognized by *snpXplorer* including, among others, PLINK (v1.9+ and 2.0+) association outputs, which makes it easier for the user to directly load association data from PLINK.[3] Multiple datasets can be selected as inputs, and these will be displayed on top of each other with different colors (*Figure 1A-B*). To do so, the user just need to select multiple input data.

After selecting the input type, the user should set the preferred genome version (GRCh37 and GRCh38). By default, GRCh37 is used, however *snpXplorer* can translate genomic coordinates from one reference version to another using the *liftover* tool.[4] In order to browse the genome, the user can either input a specific genomic position, gene

name or variant identifier, or select the Manual scroll option, which allows to dynamically survey the genome in any chromosome and at any position.

The visualization consists of 3 separate panels showing (*i*) the SNP summary statistics (Figure 1A-B), (*ii*) the structural variants (*Figure 1A*), and (*iii*) the tissue-specific RNA-expression (*Figure 1C*). The main visualization panel shows association statistics of the input data in the region of interest: genomic positions are plotted on the *x-axis* and association significance (in $-log_{10}$ scale) on the *y-axis*. The user can extend or contract the genomic window to be displayed and adjust significance levels of association. Linkage disequilibrium (LD) patterns are optionally shown for the most significant variant within the region, the input variant, or a different variant of choice: these are calculated within the European samples of the 1000Genome project (N=504).[5] There are two major visualization types. By default, each variant-association is represented as a dot (*Figure 1B*). Alternatively, associations can be shown as *p-value* densities: the selected region is divided in bins and a polynomial regression model is fitted to the data, using variant significance as dependent variable and genomic position as predictor (*Figure 1A*). Regression parameters including the number of sliding windows and the smoothing value can be dynamically adjusted. Gene names (from RefSeq v98) as well as recombination rates (from HapMap II) are dynamically adapted to the plotted region.[6,7]

The second panel shows structural variations in the region of interest (*Figure 1A*). These are extracted from a recent study concerning the discovery of structural variations at genome-wide level across sequencing platforms.[8] Structural variations are represented as segments: larger segments suggest bigger structural events, and we report the largest difference in allele length on top of each segment. Structural variations are annotated to insertions, deletions, inversions, copy number alterations, duplications, Alu elements, LINE1 elements, and SVAs.

The third panel shows tissue-specific RNA-expression of the genes displayed in the selected genomic window, from the GTEx consortium (*Figure 1C*).[9] The expression of any gene within the genomic region of interest across 54 human tissues is scaled at the level of each tissue, and reported as a heatmap.

Additionally, we report in the side panel the association statistics of the top 10 variants with highest significance as well as the top eQTLs associations. Finally, a download button allows to download the high-quality image of the entire exploration section.

Functional annotation section

The functional annotation section takes as input any list of variants (typically one variant per line) along with the input type and an email address: *snpXplorer* will run a functional annotation and enrichment analysis in the background and send the results by email.

The first step of the annotation section is the variant-gene mapping, that is, to link genetic variants to the most likely affected gene(s). We do so combining annotation from the *Combined Annotation Dependent Depletion* (CADD, v1.3), *expression-quantitative-trait-loci* (eQTL) in blood from the Genotype-Tissue-expression consortium (GTEx v8), and positional mapping up to $500kb$ from the reported variants (from RefSeq v98).[6,9,10] CADD annotation is used to inspect each variant's consequences: in case of coding variants (*e.g.* synonymous or missense variant), we confidently associate the variant with the corresponding gene. Alternatively, we first consider possible eQTL associations and in case these are not available, we included all genes at increasing distance d from the variant position (starting with $d \leq 50kb$, up to $d \leq 500kb$, increasing by $50kb$ until at least one match is found. This procedure allows multiple gene(s) to be associated with a single variant, depending on variant effect and position. We represent the final variant-gene mapping with a set of figures showing the sources of variant annotation (*i.e.* CADD, eQTL or genomic position), the number of genes associated with each variant, the distribution of mapped genes across chromosomes (*Figure 1D*), and a circular summary of variant annotation source, minor allele frequency and chromosomal distribution (*Figure 1E*).

Prior to enrichment analysis, we look whether the input variants and their associated genes were previously associated with any trait in the GWAS catalog, and plot the frequency of the top overlapping traits (*Figure 1F*). However, we realized that allowing multiple genes to associate with each variant could result in an enrichment bias, as neighboring genes are often functionally related. To control for this, we implement sampling techniques (1,000 iterations): at each iteration, we (*i*) sampled one gene from the pool of genes associated with each variant (thus allowing only 1:1 relationship between variants and genes), and (*ii*) looked whether the resulting genes were previously reported in the GWAS catalog.[1] Averaging by the number of iterations, we obtain an unbiased estimation of the overlap of the variants-associated genes with each trait in the GWAS catalog.

Finally, we perform a gene-set enrichment analysis to find molecular pathways enriched in the genes associated with the input variants. Once again, to avoid enrichment bias due to multiple genes mapping to the same variant, we used sampling techniques: at each iteration, we (*i*) sample one gene from the pool of genes associated with each variant and (*ii*) perform gene-set enrichment analysis with the resulting list of genes. Gene-set overlap analysis is performed specifying the Gene Ontology Biological Processes as background, and the resulting *p-values*

are corrected for multiple testing with False Discovery Rate (FDR<5%). To reduce the complexity of the enriched biological processes, we use the REVIGO tool, which removes redundant terms based on a semantic similarity measure.[11] We choose *Lin* as semantic distance measure and allowed *small* similarity among terms in order to be clustered together.[12] Enrichment results are finally displayed in an embedded space via eigenvalue decomposition of the pairwise distance matrix (*Figure 1G*).

Once the analysis is completed, summary figures and intermediate files are sent to the user as a compressed folder, and are deleted from our server.

Technical details and source code

snpXplorer is freely available as a web-server at http://snpxplorer.eu.ngrok.io. However, the tool can also be installed on your local machine. In order to download *snpXplorer*, please clone the github repository from https://github.com/TesiNicco/SNPbrowser. Keep in mind that when cloning *snpXplorer* into your system, additional files should be downloaded, including all summary statistics. In order to run, *snpXplorer* needs to have R (v3.5+) and python (v3+) correctly installed in your system. *snpXplorer* uses the following R packages: "*shiny*", "*data.table*", "*stringr*", "*ggplot2*", "*liftOver*", "*colourpicker*", "*SNPlocs.Hsapiens.dfSNP144.GRCh37*", "*rvest*" and "*plotrix*", "*parallel*", "*lme4*", "*ggsci*", "*RColorBrewer*", "*gprofiler2*", "*GOSemSim*", "*GO.db*", "*org.Hs.eg.db*", "*pheatmap*", "*circlize*", "*devtools*", "*treemap*", "*basicPlotteR*", "*gwascat*", "*GenomicRanges*", "*rtracklayer*", "*Homo.sapiens*", "*BiocGenerics*". Additionally, the following python libraries should be correctly installed: "*re*", "*werkzeug*", "*robobrowser*" and "*sys*". For a typical functional annotation of a set comprising 100 SNPs, *snpXplorer* will take about 20 minutes.

Abbreviations

eQTL: expression-quantitative-trait-loci, represent the association of a genetic variant with a change in RNA expression of a transcript.

LD: linkage disequilibrium, defined as non-random association of variants at different genomic positions. **GRCh37/GRCh38:** Genome Reference Consortium human build 37/38 are the reference human genomes. **Gene:** genomic region that is transcribed into an RNA transcript.

RsID: unique variant identifier.

SNP: single nucleotide polymorphisms, *i.e.* the most basic type of genetic variants. Correspond to the substitution of a single nucleotide in the DNA.

Insertion: an insertion is the addition of one or more nucleotide into a DNA sequence.

Deletion: a deletion is a mutation in which a part of a chromosome or sequence of DNA is left out.

Duplication: defined as any duplication of a genomic region that contains a gene.

Inversion: a mutation that cause a genomic region to be inverted. The overall number of nucleotides does not change.

Alu element: a transposable elements, also known as jumping gene. Alu elements are rare sequences of DNA that can move (or transpose) themselves to new positions within the genome of a single cell.

Line1 element: Long Interspersed Nuclear Elements are a group of retrotransposons that are widely spread in the human genome (they constitute >20% of human genome).

SVA: SINE-VNTR-Alus are specific retrotransposons that are associated with disease in humans.

Acknowledgements

The authors declare no conflict of interests. This work was supported by Stichting Alzheimer Nederland (WE09.2014-03), Stichting Diorapthe, horstingstuit foundation, Memorabel (ZonMW projectnumber 733050814) and Stichting VUmc Fonds.

References

1. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005–D1012. doi:10.1093/nar/gky1120
2. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
3. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–575. doi:10.1086/519795
4. Kent WJ, Sugnet CW, Furey TS, et al. The Human Genome Browser at UCSC. *Genome Res*. 2002;12(6):996-1006. doi:10.1101/gr.229102
5. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74. doi:10.1038/nature15393
6. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733–745. doi:10.1093/nar/gkv1189
7. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007;449(7164):851-861. doi:10.1038/nature06258
8. Chaisson MJP, Sanders AD, Zhao X, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*. 2019;10(1):1784. doi:10.1038/s41467-018-08148-z
9. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580–585. doi:10.1038/ng.2653
10. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019;47(D1):D886–D894. doi:10.1093/nar/gky1016
11. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. Gibas C, ed. *PLoS ONE*. 2011;6(7):e21800. doi:10.1371/journal.pone.0021800
12. McInnes BT, Pedersen T. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *J Biomed Inform*. 2013;46(6):1116-1124. doi:10.1016/j.jbi.2013.08.008

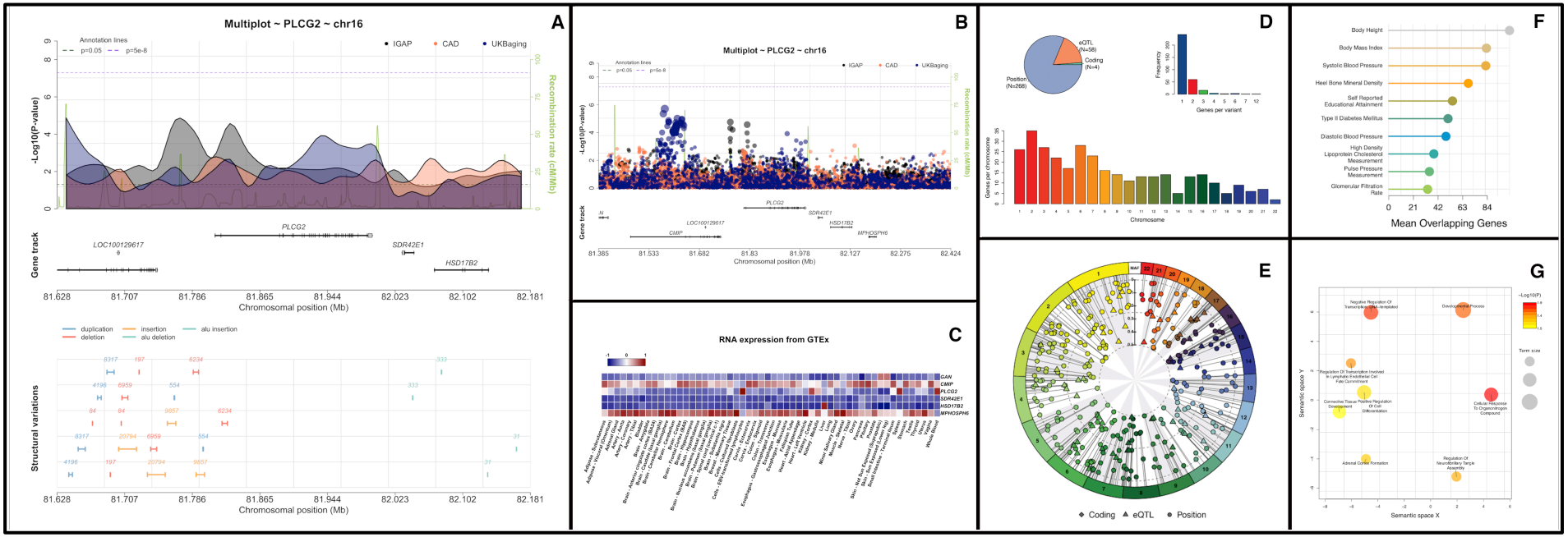

Figure 1: *snpXplorer* visualization (A-C) and functional annotation (D-G) plots.