snpXplorer documentation

v1.0 ~ June 17, 2020

Niccolo' Tesi

n.tesi@amsterdamumc.nl

Table of contents

1. **snpXplorer** in a nutshell

snpXplorer is a web application written in R and based on the package *shiny* that allows (*i*) visualization and superimposition of summary statistics from human genetic association studies (*e.g* GWAS and/or sequencing-based association studies) and (*ii*) functional annotation and pathway enrichment analysis of SNP-sets.

The application offers an *exploration* section and a *functional annotation* section. The *exploration* section consists of 3 main plotting panels that show association statistics, structural variations and gene expression per tissue. The user can choose association data of interest among the available summary statistics, or load own association statistics. **snpXplorer** allows the superimposition of association statistics from multiple studies to explore association trends across traits. **snpXplorer** supports both GRCh37 and GRCh38 versions of the reference genome. The user can browse the genome by input a specific gene or variant, or can manually scroll the genome. Multiple visualization parameters including the size of the window, association significance and Linkage disequilibrium patterns can be dynamically added. The basic visualization consists of genomic position on x-axis and association significance (in -log10 scale) on y-axis. Alternatively, association statistics can be displayed as p-value densities. **snpXplorer** reports the top 10 association showed in the region as well as all significant expression-quantitative-trait-loci (eQTL) in blood of variants displayed in the region. Additionally, structural variants (*e.g* insertions, deletions, inversions, etc) as well as gene-expression of the displayed at tissue level are shown.

The *functional annotation* section allows the user to perform variant-gene mapping and gene-set overlap analysis to investigate biological pathways enriched in the input variants. The variant-gene mapping procedure allows each gene to associate with one or more genes, and the gene-set overlap analysis is implemented with sampling techniques to avoid enrichment bias due to multiple genes mapping to the same variant. The user can input the desired set of variants, along with an email address: **snpXplorer** will perform the computations in background and send the results by email.

Additional information, including stand-alone installation and quick user-guide are available on the github page.

2. Background: what do you need to know

Genetic association studies

Genome-wide association studies (GWAS) and sequencing-based association studies are extensively used to study the genetic factors underlying a large variety of human phenotypes. Briefly, in association studies the frequency of genetic variants across the genome is associated with a phenotype, which can either be binary (*e.g* cases and controls), or linear (*e.g* age at death or metabolite/gene/protein abundance). As a result, summary statistics consisting of effect-size and p-value, can be generated for each variant-phenotype association. The effect-sizes quantifies the strength and the direction of the association, while p-value represents the statistical significance of that variant-phenotype association. The power to detect significant associations thus depends on the effect-size, the sample size and the significance threshold to be obtained.(Hong and Park, 2012)

Multiple testing correction in genetic studies

Due to the extremely high number of genetic variants in the genome (few tens of millions of variants), and since each variant is tested independently, multiple testing correction dramatically affect association statistics. The threshold normally adopted for genome-wide significance is $p<5x10^{-8}$, which is the Bonferroni corrected p-value assuming 1 million of tests. To reach such low p-value, large number of samples are required, which is not always feasible. For this reason, it can be highly informative to visualize the degree of association of a genomic region across different studies and traits, especially when the sample size is not sufficiently large.

Linkage disequilibrium

Linkage disequilibrium (LD) is formally defined as the non-random association of variants at different genomic loci. In other words, linkage disequilibrium can be seen as a measure of co-occurrence of variants at different genomic position. The strength of linkage disequilibrium between variants is quantified with different measures, including R^2, which ranges 0 (for independent variants) to 1 (variants in complete linkage). Variants that are in linkage disequilibrium are inherited together from the parents, and constitute the so-called haplotypes.

Interpretation of genetic associations

The interpretation of genetic associations is often complex: most of the variants in the genome are non-coding variants, for which the functional consequences are unclear. LD patterns, structural variations and chromatin states add other layers of complexity. Therefore, each variant should be investigated independently in the context of its genomic region, exploiting diverse annotation sources, and allowing uncertainty in the annotation databases.

3. Exploration section

The *exploration* section represents the main interface of **snpXplorer**. For simplicity, the different sub-sections are discussed separately.

Input data

The first thing to do in **snpXplorer** is to choose the desired input data. Here, the user can either select one of the available summary statistics or upload own association statistics (Figure 1). The number of available summary statistics will constantly grow and we will try to keep the most updated and interesting studies in place. Alternatively, the user can also select own association statistics. There is not a fixed format for input data: in principle, **snpXplorer** only needs genomic coordinates (*i.e* chromosome and position) and p-value of association. All other information, if present in the input data, will NOT be used. As for the format, **snpXplorer** can read both tab-separated file and comma-separated files: a column header is mandatory: this will be used to infer the content of each column. Most headers are recognized. In addition, **snpXplorer** automatically recognizes PLINK (v1.9+ and 2.0+) association outputs, which makes it easier for the user to directly load association data from PLINK into **snpXplorer**.(Purcell *et al.*, 2007) One of the main features of **snpXplorer** is that it allows to visualize association statistics from multiple studies on top of each other, permitting comparisons. To do so, just select multiple input data, and let **snpXplorer** do the rest.

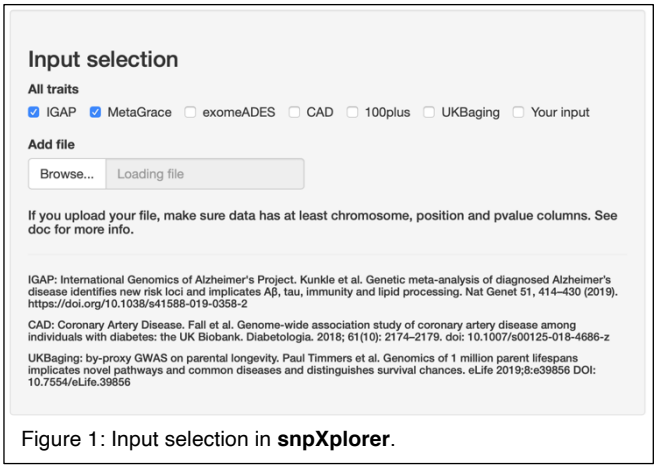

Figure 1: Input selection in **snpXplorer**.

Genome version

After choosing the input data, the user needs to specify the reference genome of interest (Figure 2). This is **particularly** important if the user is using own association data, so make sure that the association statistics and the selected genome version are the same. By default, GRCh37 (hg19) is used, as most of the summary statistics available are with respect to this genome version. However, any input data can be visualized in both genome versions: to do so, **snpXplorer** implements lift-over tool to translate genomic coordinates from one version to another. All annotations (genes to be displayed, recombination rates, structural variations) are automatically updated to the selected genome version.

Browsing option

In order to explore association statistics, the user needs to specify *how* to browse the genome. There are two main options: first, by genomic position, variant identifier or gene symbol (*Locus, Gene, RsID*, Figure 2a). The desired target region can be inserted in the text box. The format of the target region is described under the text box, with examples (see Figure 2a). The second browsing option is *Manual Scroll*. If this option is selected, two sliding bars appears in place of the text box. The first sliding bar codes for chromosome

number, while the second codes for genomic positions. In this way, the user can dynamically and freely scroll the genome in any chromosome and at any position (see Figure 2b). **In order to display associations data,**

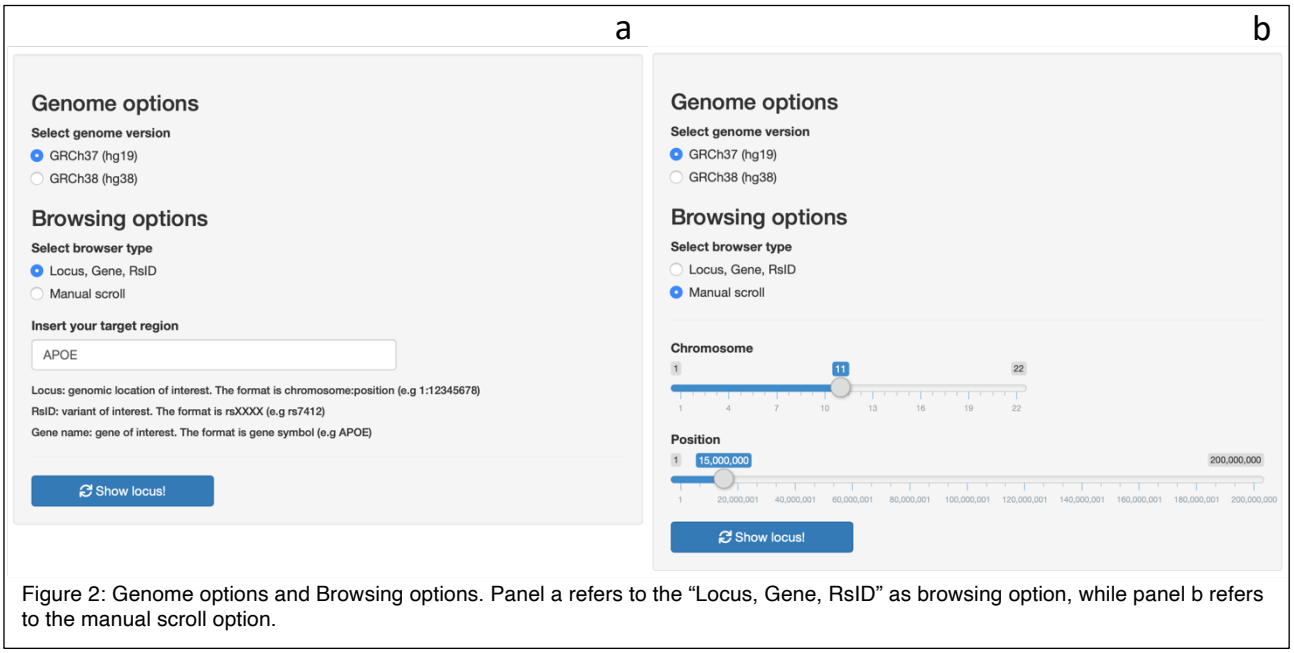

Figure 2: Genome options and Browsing options. Panel a refers to the "Locus, Gene, RsID" as browsing option, while panel b refers to the manual scroll option.

the use must click the *Show locus* button. This action is required every time the user change the input data, the genome version, the browsing option or any visualization parameter.

Visualization options

Once the user has clicked the *Show locus* button, **snpXplorer** will run and will display the requested information. Normally, the procedure is relatively fast, but connection and server load can influence **snpXplorer** reactivity. On the left side of the **snpXplorer** page, the visualization options are reported (see Figure 3). These options include:

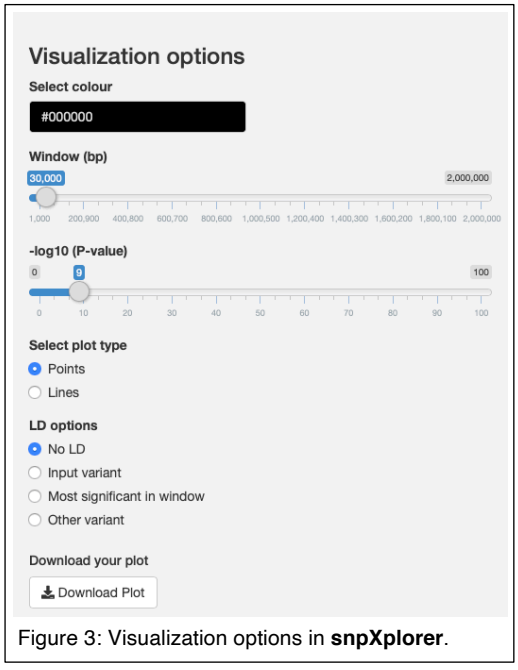

Figure 3: Visualization options in **snpXplorer**.

- Color (in case multiple input data are selected, multiple options for colors will appear);

- Window size (bp) which is indicative of the number of base-pairs upstream and downstream the desired browsing option. This parameter controls the width of the x-axis;

- -log10 (P-value) which controls for the significance level to be shown (height of the y-axis);

- Plot type which controls whether points or p-value densities should be plotted;

- LD options which adds LD information on the plot. This option is available only when a single study is plotted. The user can select the variant to calculate LD for the input variant (in case this was the browsing option), for the most significant variant in the region, or a variant of choice (in which case a text box is prompted and the user need to insert the genomic position of the desired variant);

- Download your plot which intuitively allows the user to download high-quality pdf of the visualized image;

On the side panel, below the aforementioned parameters, two tables shows (*i*) the top 10 most significant variants in the region along with their chromosomal location, input ID and association significance, and (*ii*) the top 15 most significant expression-quantitative-trait-loci (eQTL) of the variants in the plotted region. In the latter table genomic location (in the reference genome of choice), variant alleles, minor allele frequency, effect-size (normalized coefficient of the linear regression model used to calculate variant-gene associations), adjusted p-value and gene name are shown. Under both tables, a legend guides the user through the interpretation of the columns.

Visualization panels

The visualization consists of 3 separate panels reporting (*i*) SNP summary statistics (Figure 4), (*ii*) structural variants (Figure 5) and (*iii*) tissue-specific RNA-expression.

SNP summary statistics panel

The main visualization panel shows association summary statistics of the input data in the genomic region of interest (Figure 4). Here, genomic positions are plotted on the x-axis, and association significance (in -log10

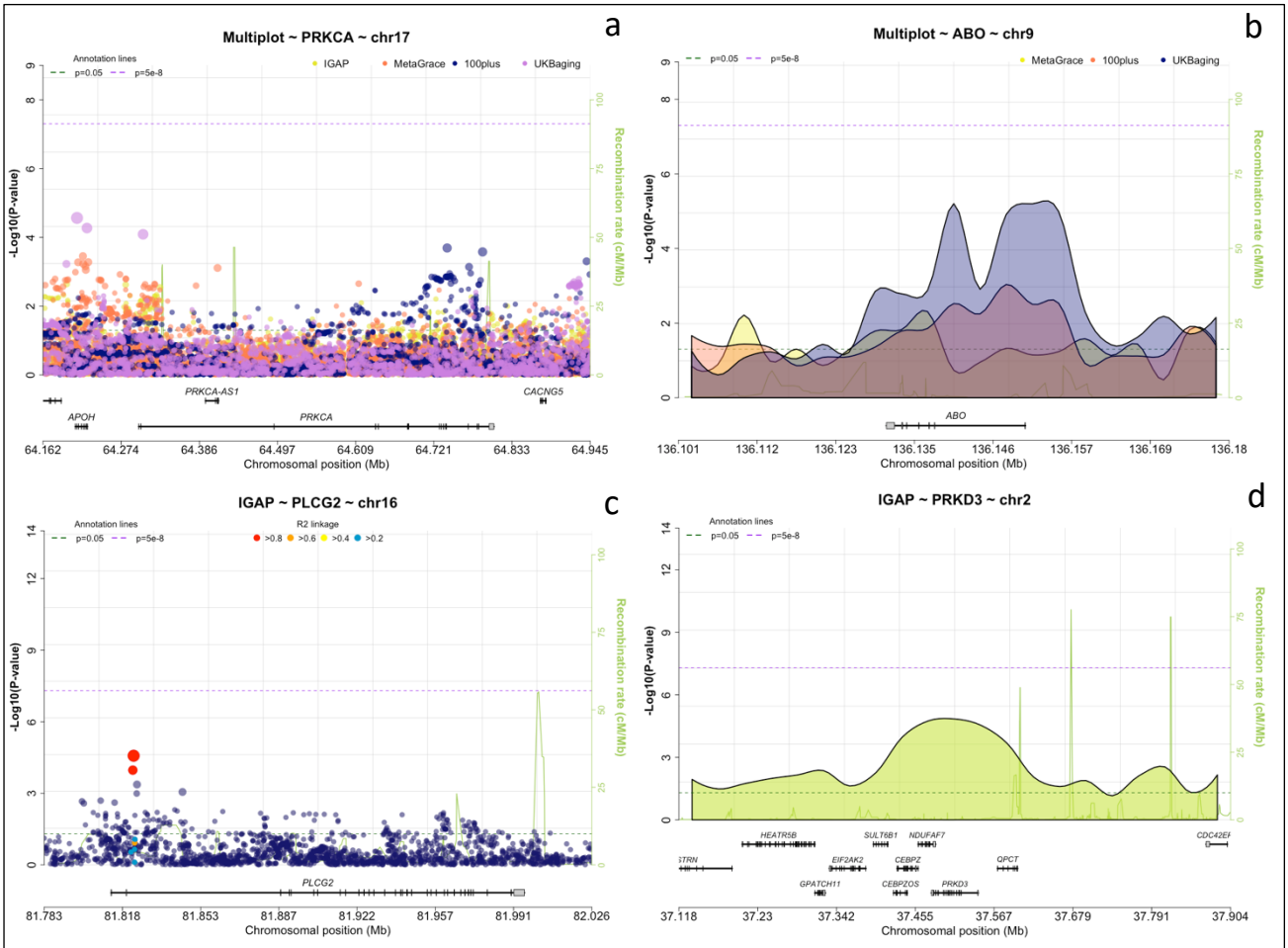

Figure 4: SNP association statistics. The figure shows examples of the SNP summary statistics representation. Four different studies are being plotted in the panel a using points as plot type. Different colors code for different studies. In panel b, three studies are plotted on top of each other using p-value density as plot type. Panel c show a single study and highlight the linkage disequilibrium patterns. Last, panel d shows a single study plotted as p-value density. In all panels, dashed lines refer to nominal significance of association ($p<0.05$) and genome-wide evidence of association ($p<5x10^{-8}$). Green lines represent recombination rates and are extracted from Hap Map II.

scale) is plotted on the y-axis. The user can dynamically extend or contract the genomic window to be displayed as well as tune significance level of association. Association statistics can be showed as points, with each dot referring to a variant-phenotype association (Figure 1a and 1c), or as p-value densities (Figure 1b and 1d). In the latter case, 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 positions as predictors. Regression parameters can be dynamically adjusted to make the regression line more or less fitted to the underlying data. Linkage disequilibrium patterns are optionally showed: these are calculated within the European samples of the 1000Genome project (N=504), and LD levels are showed in different colors (Figure 1c).(1000 Genomes Project Consortium *et al.*, 2015) Gene names are from RefSeq (version 98): in case of multiple gene models for a given gene, we used the one with the largest number of exons.(O'Leary *et al.*, 2016) In case the plotting region is large and it contains a larger number of genes to be displayed, these are automatically arranged on multiple lines. Recombination rates are from HapMap II and are dynamically normalized and adjusted to the region of interest.(The International HapMap Consortium, 2007)

Structural variants panel

The second main visualization panel shows any structural variant present in the region of interest.(Chaisson *et al.*, 2019) Structural variants can be seen also from GWAS association statistics as "holes" where no certain call could be done for any variant (see Figure 5 for an example). Structural variants are annotated to insertions, deletions, inversions, copy number alterations, duplications, Alu elements, LINE1 elements, and SVA. See Legend and abbreviation section for a description of each variant type. Structural variants are shows as ranging from the starting position to the ending position of the largest allele. Structural variations are plotted in the same genomic region as the main association plot, and are dynamically adjusted when the user tune the window to be plotted. This dataset can help in understanding the causal effects of certain variants, and are becoming more and more used in genetic research, especially with the advent of long-read sequencing technologies.

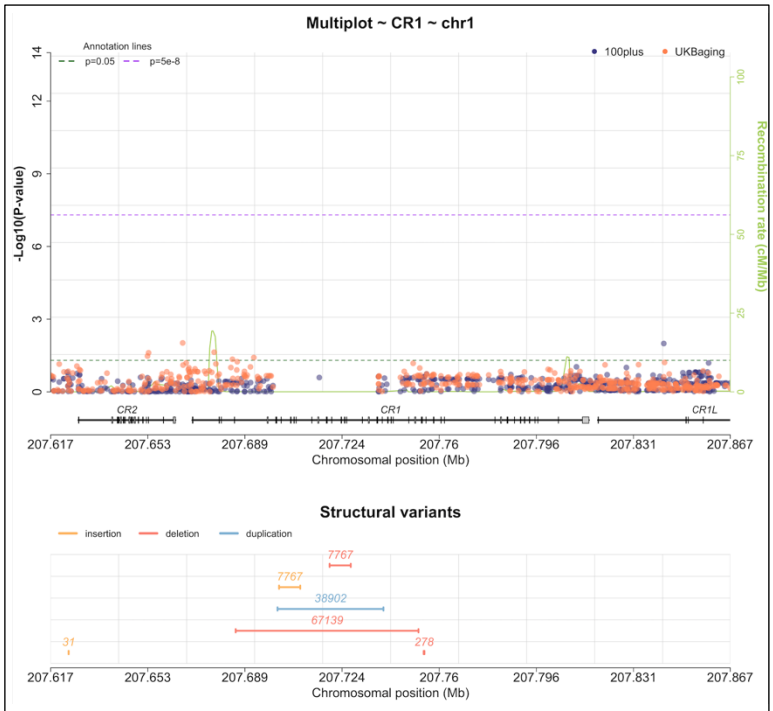

Figure 5: Structural variants within the region of interest.

Gene-expression per tissue

The last visualization panel consists of gene-expression per tissue from GTEx consortium: the expression of any gene within the genomic region of interest across 54 tissues in humans is plotted.(GTEx Consortium, 2013) Gene-expression is scaled at the level of the tissues.

4. Annotation section

The *functional annotation* section is integrated in the **AnnotateMe** package, and it is designed to perform variant-to-gene mapping and gene-set overlap analysis in order to explore the biological processes enriched in the set of input variants.

Input data

Input variants for **AnnotateMe** should be pasted in the relative text box. The format of the input variants is not strict: the user can input data in multiple formats (chromosome:position, rsid, etc) and should specify the input data type. After inputting data and the relative format type, the user should insert an email address. **This is particularly important as the results will be sent by email.** Once also the email address has been inserted, the user can update the page (using the *Show locus* button) and **AnnotateMe** analysis will start. **AnnotateMe** will run in background and will send results by email when completed. Usually, an analysis including 100 variants takes less than 5 minutes to perform.

Variant-gene mapping

The first step in **AnnotateMe** is the variant-gene mapping. The variant-gene mapping associates a set of variants with a set of genes. This step is one of the most delicate and important: since most of GWAS hits is in non-coding regions, understanding the functional consequences of each variant is not trivial. We linked each variant to its likely affected gene(s) combining annotation from Combined Annotation Dependent Depletion (CADD, v1.3), expression-quantitative-trait-loci in blood (eQTL) from GTEx consortium (v8) and positional mapping up to 500 kb from the reported variants (RefSeq version 98).(Rentzsch *et al.*, 2019; GTEx Consortium, 2013; O'Leary *et al.*, 2016) CADD annotation is used to inspect each variant's consequences: in case of coding variants (*e.g* synonymous or missense variant), we confidently associated the variant with the corresponding gene. We consider LD patterns in doing so, thus we check variant consequences of all variants in high LD with the input variants. Alternatively, we first considered possible eQTL associations and in case these are also not available, we included all genes at increasing distance *d* from the variant (starting with *d<=50* kb, up to *d*<=500 kb, increasing by *50* kb). This procedure allows multiple gene(s) to associate with a single variant, based on annotation uncertainty. We represent variant-mapping annotation with several plots showing the source of annotation of all variants (Figure 6a), the number of genes associated with each variant (Figure 6b), the

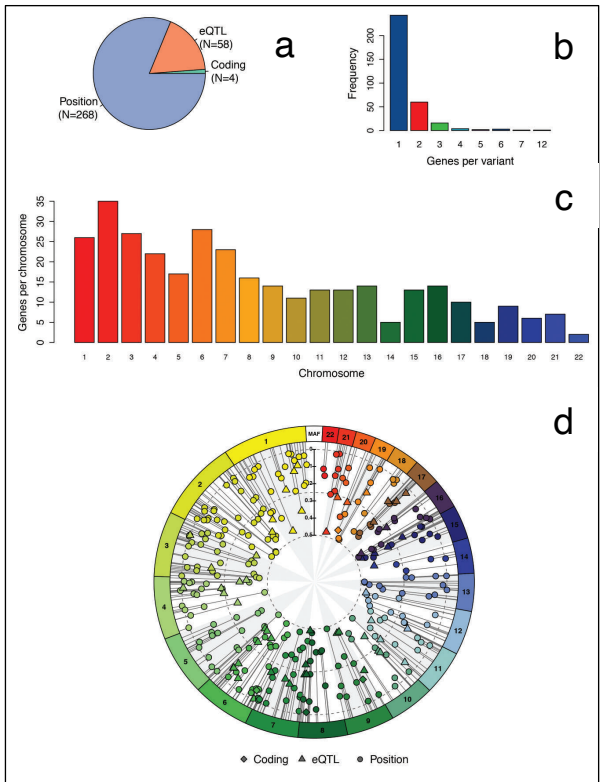

Figure 6: Variant-gene mapping from **AnnotateMe** tool.

distribution of the mapped genes across chromosomes (Figure 1c) and a circular summary visualization showing the source of variant mapping, variant frequency and chromosomal distribution (Figure 6d). All plots will be sent to the user by email.

Previous associations of the variants and gene

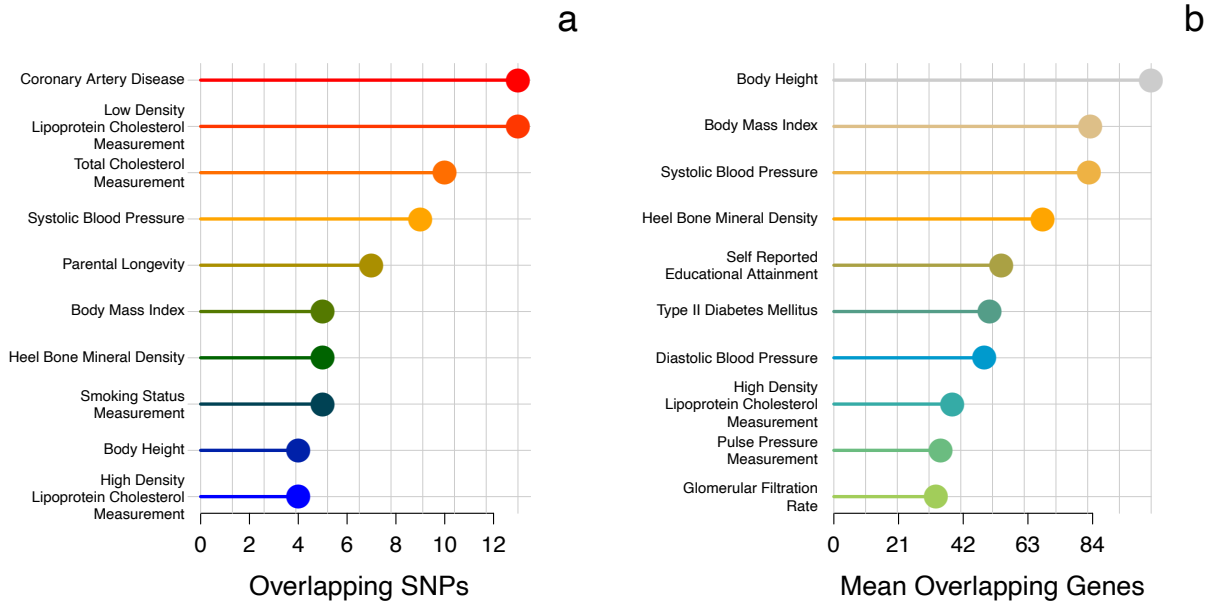

Figure 7: Overlap with GWAS catalog. The plot shows how we represent overlapping associations in the context of GWAS catalog. The overlap is done both at the level of variants and genes.

Given the list of variants and genes associated with the input variants, we first look these variants and genes in the GWAS catalog, a database including, as of June 2020, 4493 publications and 179364 associations in total.(Buniello *et al.*, 2019) In this database, we sought for previous associations with any trait. Similarly, we also look whether the genes associated with these variants were previously reported to associate with any trait. 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 (1000 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. Averaging by the number of iterations, we obtain an unbiased estimation of the overlap of the PRS-associated genes with each trait in the GWAS catalog. We represent overlaps with GWAS catalog as barplots, where each bar is representative of a trait as found in the GWAS catalog: this is done both at the level of the single variants (Figure 7a) and at the level of the genes (Figure 7b).

Gene-set enrichment analysis

The final step in the functional annotation procedure is the gene-set overlap analysis, where the idea is to explore the biological processes enriched in the list of 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 overlap analysis with the resulting list of genes. Gene-set overlap analysis is performed with *GOSt* function as implemented in R package *gprofiler2*, with Biological Processes (GO:BP) as background, excluding electionic annotations and correcting p-values using FDR. Finally, we averaged p-values for each enriched term over the iterations (N=1000). To reduce the complexity of the resulting enriched biological processes, we exploited the web-server tool REVIGO.(Supek *et al.*, 2011) This tool summarizes enrichment results by removing redundant terms based on semantic similarity measure, and displays remaining terms in an embedded space via eigenvalue decomposition of the pairwise distance matrix (Figure 8). We chose *Lin* as semantic distance measure and allowed *small* similarity among terms in order to be clustered together.

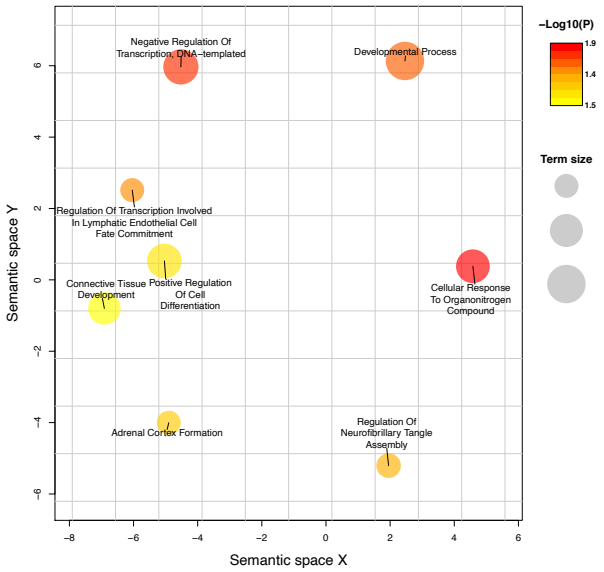

Figure 8: REVIGO output summarizes clusters of biological processes similar in terms of semantic similarity and displays remaining terms in an embedded space.

5. Stand-alone version

snpXplorer is freely available online as a web-server at XXX. However, the tool can also be installed on your local machine if the user needs more customized input data. In order to download **snpXplorer**, you should clone the github repository from https://github.com/TesiNicco/SNPbrowser. Please follow instructions on our updated github page for information on to do this. In principle, you just need R, few R packages to be correctly installed and python (v3). We have implemented a script to automatically check whether your system is ready-to-go or needs to have additional packages/libraries installed. Keep in mind that when cloning **snpXplorer** into your own system, additional annotation sources should be downloaded, including all summary statistics. For this reason, we recommend to contact us (n.tesi@amsterdamumc.nl) in case you would like additional summary statistics to be loaded into **snpXplorer** or you really want to have a local version in your system.

6. Citation

If you find the **snpXplorer** or **AnnotateMe** useful, don't forget to cite us. Our manuscript is currently under revision, and as soon as it will be publicly available, we will update this document with the correct citation. For the moment, you may want to use our preprint article on bioRxiv. In addition, **AnnotateMe** tool was used before in publications, and we report here the citations.

6. Legend and 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.

7. References

1000 Genomes Project Consortium *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

Chaisson,M.J.P. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, **10**, 1784.

GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Hong,E.P. and Park,J.W. (2012) Sample size and statistical power calculation in genetic association studies. *Genomics Inform.*, **10**, 117–122.

O'Leary,N.A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–745.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Rentzsch,P. *et al.* (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.

Supek,F. *et al.* (2011) REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE*, **6**, e21800.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.