snpXplorer documentation

v2.0 ~ April 19, 2021

Niccolo' Tesi

n.tesi@amsterdamumc.nl

Table of contents

1. snpXplorer in a nutshell 3
2. Background: what do you need to know 4
3. Exploration section 5
    a. Input data 5
    b. Genome version 5
    c. Browsing options 5
    d. Visualization options 6
    e. Visualization panels 7
4. Annotation section 10
    a. Input data 10
    b. Variant-gene mapping 10
    c. Previous associations of the variants and gene 11
    d. Gene-set overlap analysis 11
5. Stand-alone version 13
6. Citation 14
7. Legend and abbreviations 15
8. References 16

1. ***snpXplorer*** in a nutshell

snpXplorer is a web-server application written in R and based on the package *shiny* that allows (*i*) the rapif exploration of any region in the genome with customizable genomic features, (*ii*) the superimposition of summary statistics from multiple genetic association studies, and (*iii*) the functional annotation and pathway enrichment analysis of SNP sets in an easy-to-use user interface.

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 of the selected genomic region. The user can choose association data of interest among the available summary statistics, and/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.

The first (and main) visualization panel shows the association statistics of the input data in the region of interest: genomic positions are shown on the x-axis and association significance (in $-log10$ scale) is reported on the y-axis. Both the x-axis and the y-axis can be interactively adjusted to extend or contract the genomic window to be displayed. Within this panel there are two ways to visualise the data. By default, each variant-association is represented as a dot, with dot-sizes optionally reflecting p-values. Alternatively, associations can be shown as *p*-value profiles.

The second panel shows structural variations (SV) in the region of interest. Structural variations are represented as segments: the size of the segment codes for the maximum difference in allele sizes of the SVs as observed in the selected studies. Depending on the different studies, structural variations are annotated as insertions, deletions, inversions, copy number alterations, duplications, mini-, micro- and macro-satellites, and mobile element insertions (Alu elements, LINE1 elements, and SVAs).

The third panel shows tissue-specific RNA-expression of the genes in the selected genomic window, from the GTEx consortium. The expression of any gene within the genomic region of interest across 54 human tissues is scaled and reported as a heatmap. Hierarchical clustering is applied on both the genes and the tissues, and the relative dendrograms are reported on the sides of the heatmap.

The side panel allows the user to interact with the exploration section. In order to guide the user through all the available inputs and options, help messages automatically appear upon hovering over items. The side panel reports (*i*) the top 10 variants with highest significance (together with the trait they belong to, in case multiple studies were selected), and (*ii*) the top eQTLs associations (by default, eQTLs in blood are shown, and this can be optionally changed), and cross-references including GeneCards, GWAS-catalog, and LD-hub. Finally, download buttons allow to download a high-quality image of the different visualisation panels as

well as the tables reporting the top SNP and eQTL associations, the SVs in the selected genomic window, and the LD table.

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 the desired run options (input type, reference genome, gene-sets for enrichment and tissue to consider to eQTL analysis) and an email address: ***snpXplorer*** will perform the computations in background and send the results by email.

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). Typically, the association statistics are calculated, for each variant, with regression models, and consist of effect-sizes and *p*-value. The effect-sizes indicate the direction and strength of the association, while *p*-value represents the statistical significance of the 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.(The International HapMap Consortium, 2007; Hong and Park, 2012) 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 complete 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.

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. The number of available summary statistics will constantly grow and we will try to keep the most updated and interesting studies in place. As of February 2021, *snpXplorer* includes genome-wide summary statistics of 23 human traits classified in 5 disease categories: neurological traits (Alzheimer's disease, by-proxy Alzheimer's disease, autism, depression and ventricular volume),(Alzheimer Disease Genetics Consortium (ADGC), *et al.*, 2019; Jansen *et al.*, 2019; Matoba *et al.*, 2020; MDD Working Group of the Psychiatric Genomics Consortium *et al.*, 2020; Vojinovic *et al.*, 2018) cardiovascular traits (coronary artery disease, systolic blood pressure, body-mass index and diabetes),(CARDIoGRAMplusC4D Consortium *et al.*, 2013; Forgetta *et al.*, 2020; Yengo *et al.*, 2018; the Million Veteran Program *et al.*, 2018) immune-related traits (severe COVID infections, Lupus erythematosus, inflammation biomarkers and asthma),(The GenOMICC Investigators *et al.*, 2020; Wang *et al.*, 2021; FinnGen *et al.*, 2021; Han *et al.*, 2020) cancer-related traits (breast, lung, prostate cancers, myeloproliferative neoplasms and Lymphocytic leukaemia), and physiological traits (parental longevity, height, education, bone-density and vitamin D intake).(Timmers *et al.*, 2019; Yengo *et al.*, 2018; Manousaki *et al.*, 2020; Regeneron Genetics Center *et al.*, 2020) These summary statistics underwent a process of harmonization: we use the same reference genome (GRCh37, hg19) for all SNP positions, and in case a study was aligned to the GRCh38 (hg38), we translate the coordinates using the liftOver tool.(Kent *et al.*, 2002) In addition, we only store chromosome, position and p-value information for each SNP-association.

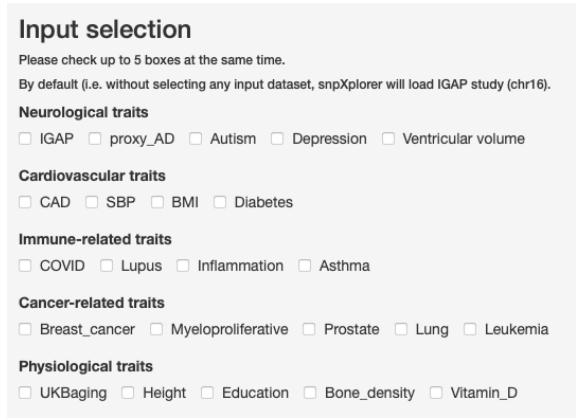

Figure 1: Input selection in **snpXplorer**.

Alternatively, the user can also upload own association statistics. ***snpXplorer*** recognizes multiple formats provided that at least a header for the columns is present. In principle, ***snpXplorer*** only needs genomic coordinates (*i.e.* chromosome and position) and *p*-value of association. The size of the uploaded file should not exceed 600Mb. ***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 visualise 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. We provide example files that the user can try out for the exploration section. These files can be found in the Help section of ***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.(Hinrichs, 2006) All annotations (genes to be displayed, recombination rates, structural variations) are automatically updated to the selected genome version.

Browsing option

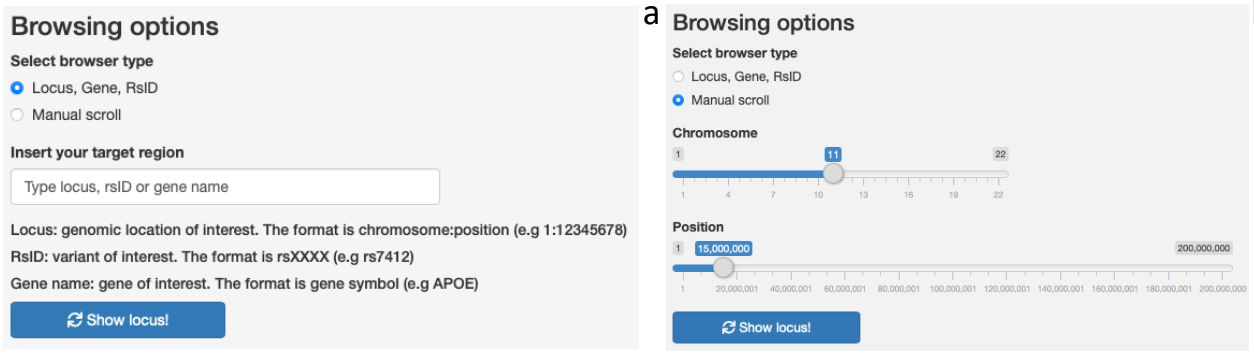

Figure 2: Genome options and Browsing options. Panel a refers to the "Locus, Gene, RsID" as browsing option, while panel b refers to the manual scroll option.

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, 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:

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

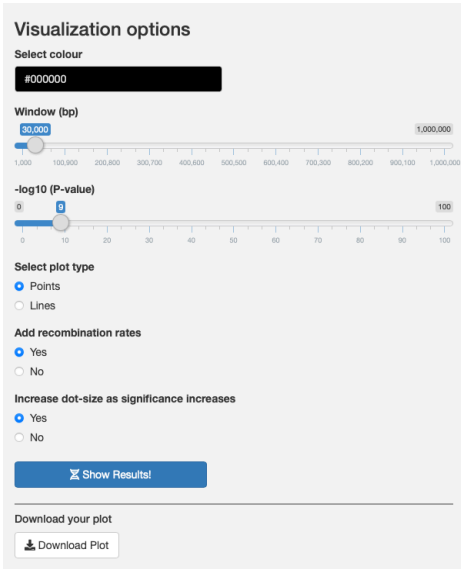

Figure 3: Visualisation options in **snpXplorer**.

- 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;

- Add recombination rates which shows/hide recombination rates;

- Increase dot-size as significance increases which increases dot size as a function of significance;

- Show Results! Button which is a copy of the other Show Results! button, and can be used similarly;

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

The side panel allows the user to interact with the exploration section. It also reports (i) the top 10 variants with highest significance (together with the trait they belong to, in case multiple studies were selected), and (ii) the top eQTLs associations, and cross-references including GeneCards, GWAS-catalog and LD-hub. Finally, download buttons allow to download a high-quality image of the different visualisation panels as well as the tables reporting the top SNP and eQTL associations, the SVs in the selected genomic window, and the LD table. Under both tables, a legend guides the user through the interpretation of the columns.

- LD options which adds LD information on the plot (Figure 4). 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). In addition, we provide the possibility to choose the population of the individuals used to calculate LD (available are all population of the 1000Genome project: European, American, African, South Asian and East Asian.(1000 Genomes Project Consortium *et al.*, 2015)

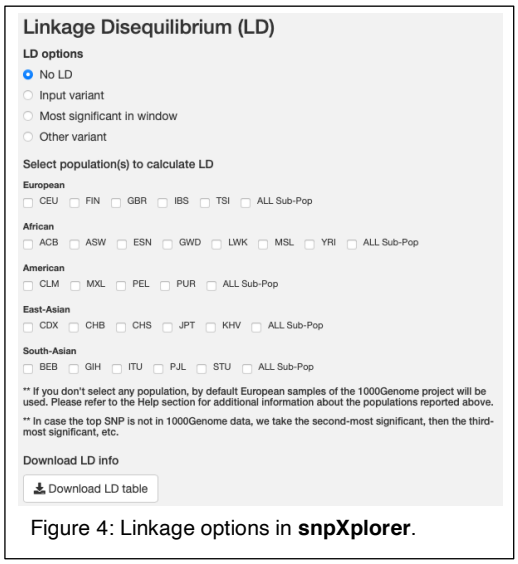

Figure 4: Linkage options in **snpXplorer**.

Visualization panels

The visualization consists of 3 separate panels reporting (*i*) SNP summary statistics (Figure 5), (*ii*) structural variants (Figure 6) and (*iii*) tissue-specific RNA-expression (Figure 7).

SNP summary statistics panel

The first (and main) 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 $-log10$ scale) is plotted on the y-axis.

Both the x-axis and the y-axis can be interactively adjusted to expand or contract the genomic window to be displayed. Within this panels there are two ways to visualize the data. By default, each variant-association is represented as a dot, with dot-sizes reflecting *p*-values (Figure 5a and 5c). Alternatively, associations can be shown as *p*-value profiles: to do so, (*i*) the selected region is divided in bins, (*ii*) a local maximum is found in

each bin based on association *p*-value, and (*iii*) a polynomial regression model is fitted to the data, using the *p*-value of all local maximum points as dependent variable and their genomic position as predictors. Regression parameters, including the number of bins and the smoothing value, can be adjusted.

Gene names (from RefSeq v98) as well as recombination rates (from HapMap II) are always adapted to the plotted region.(The International HapMap Consortium, 2007; O'Leary *et al.*, 2016) Linkage disequilibrium patterns are optionally shown for wither the most significant variant in the selected regions, the input variant, or a different variant of choice. Such linkages are calculated within the samples of the 1000Genome project, with the possibility to select preferred populations to calculate LD from.

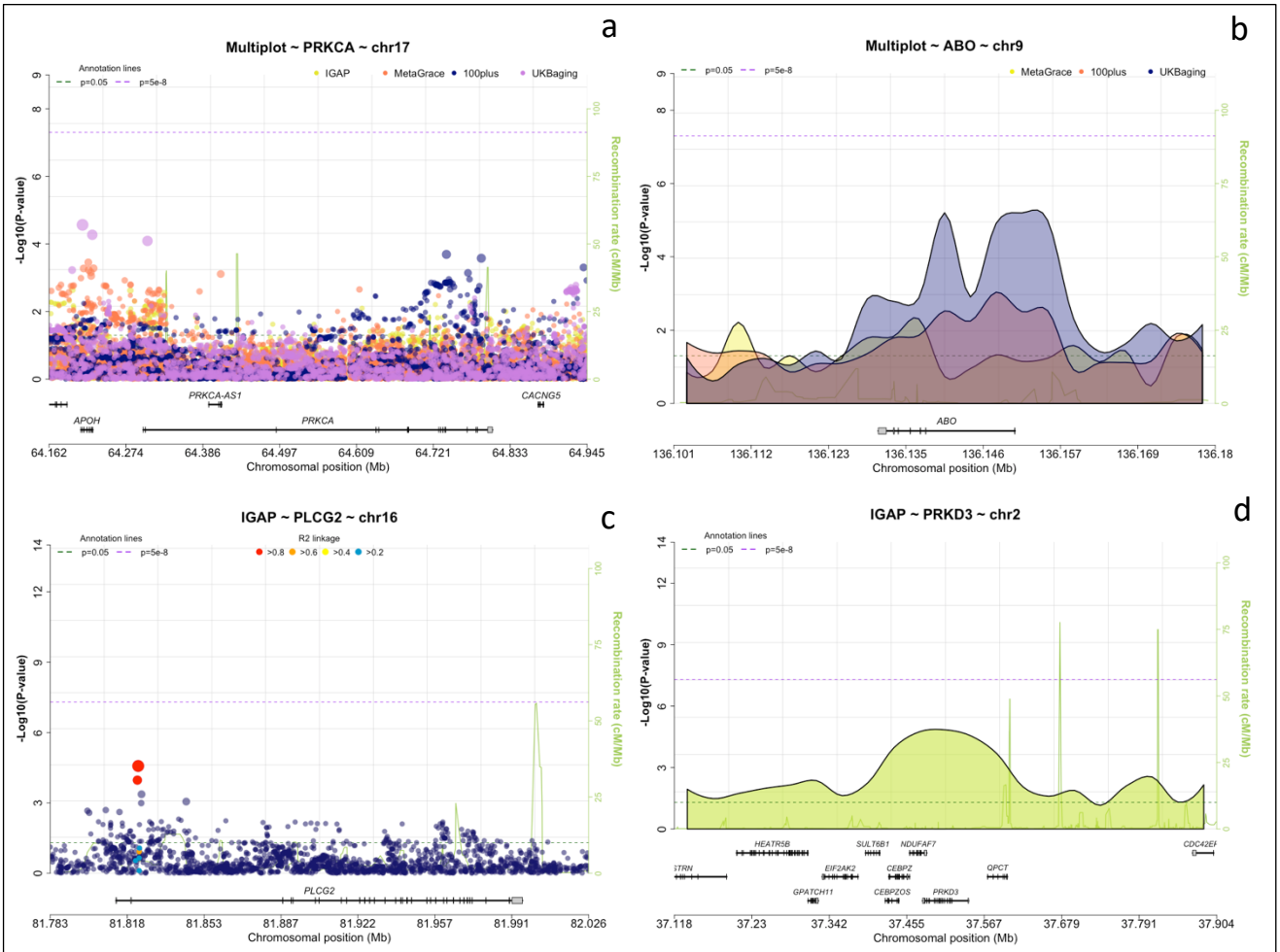

Figure 5: SNP association statistics. The figure shows examples of the SNP summary statistics representation. Four different studies are being plotted in the panel a using points as plot type. Different colors code for different studies. In panel b, three studies are plotted on top of each other using 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.

Structural variants panel

The second main visualization panel shows any structural variant present in the region of interest. These are extracted from three studies that represent the state-of-the-art regarding the estimation of major structural variations across the genome using third-generation sequencing technologies (*i.e.* long-read sequencing).(Chaisson *et al.*, 2019; Audano *et al.*, 2019; Linthorst *et al.*, 2020) Structural variants are represented as segments: the size of the segment codes for the maximum difference in allele sizes of the SVs as observed in the selected studies (Figure 6). Depending on the different studies, structural variations are annotated as insertions, deletions, inversions, duplications, copy number alterations, mini-, micro- and macro-satellites, and mobile element insertions (Alu elements, LINE1 elements, and SVA).

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 (Figure 7). Dendrograms regarding hierarchical clustering of the genes as well as the tissues is reported on the side of the heatmap.

Top associations in the region of interest

In order to guide the user through all the available inputs and options, help messages automatically appear upon hovering over items. The side panel allows not only the user to interact with the exploration section, but also reports valuable informations for the user. The side panel reports (*i*) the top 10 variants with highest significance (together with the trait they belong to, in case multiple studies were selected), and (*ii*) the top eQTLs associations (by default, eQTLs in blood are shown, and this can be optionally changed), and cross-references including GeneCards, GWAS-catalog, and LD-hub. Finally, download buttons allow to download the tables reporting the top SNP and eQTL associations, the SVs in the selected genomic window, and the LD table.

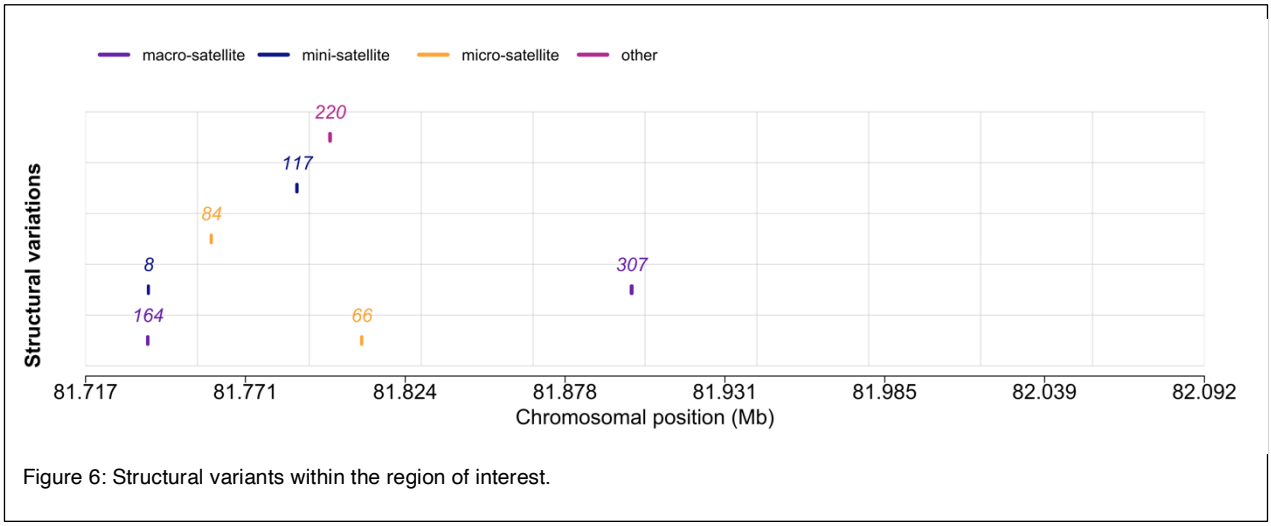

Figure 6: Structural variants within the region of interest.

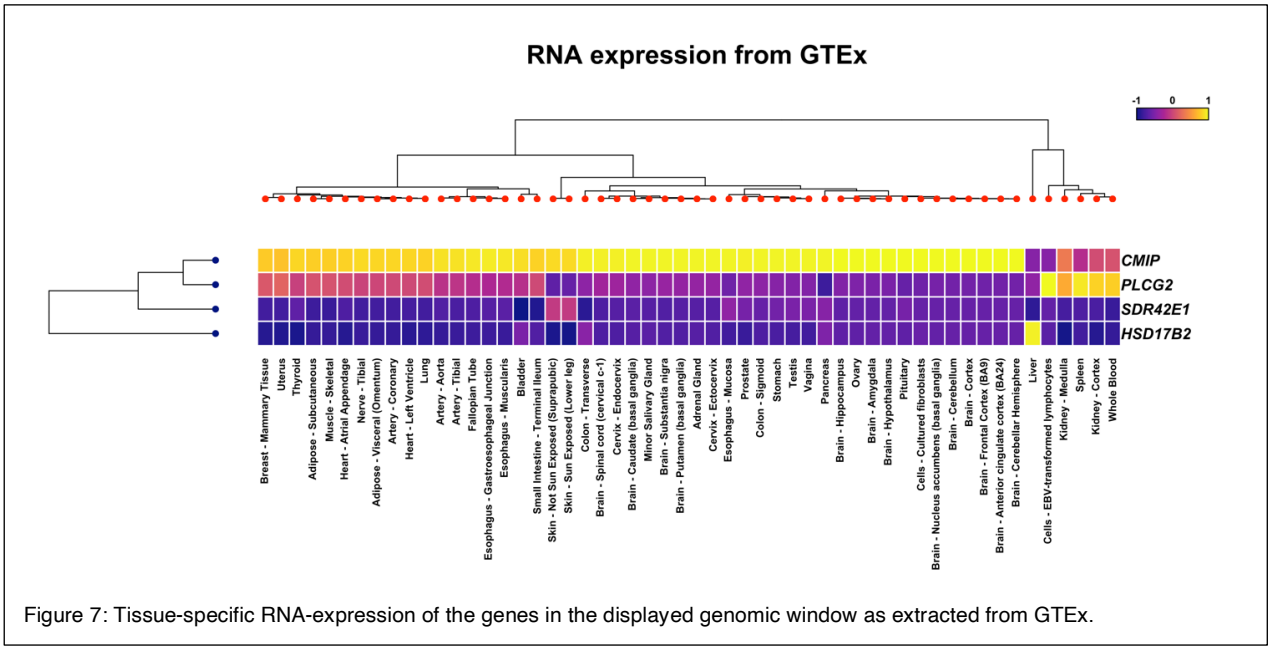

Figure 7: Tissue-specific RNA-expression of the genes in the displayed genomic window as extracted from GTEx.

4. Annotation section

The *functional annotation* section 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 the functional annotation 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) and the input-type should be specified the input data type. Additional settings for the annotation can be set: reference genome of interest (in case input type was not rsid), gene-sets for the enrichment analysis and tissues to consider for the eQTL analysis (from GTEx). After inputting data and all available options, 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 submit the annotation request and the analysis will start. The analysis will run in background and will send results by email when completed. Usually, an analysis including 100 variants takes less than 30 minutes to perform.

Variant-gene mapping

The functional annotation pipeline consists of a two-step procedure: firstly, genetic variants are linked to likely affected genes (*variant-gene mapping*); and, secondly, the likely affected genes are tested for pathway enrichment (*gene-pathway mapping*). In the *variant-gene mapping*, genetic variants are linked to the most likely affected gene(s) by (*i*) associating a variant to a gene when the variant is annotated to be coding by the Combined Annotation Dependent Depletion (CADD, v1.3),(Rentzsch *et al.*, 2019) (*ii*) annotating a variant to genes resulting from found expression-quantitative-trait-loci (eQTL) from the Genotype-expression consortium (GTEx, v8), or (*iii*) mapping a variant to genes that are within 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 (from RefSeq v98). Note that this procedure might map multiple genes to a single variant, depending on the effect and position of each variant.

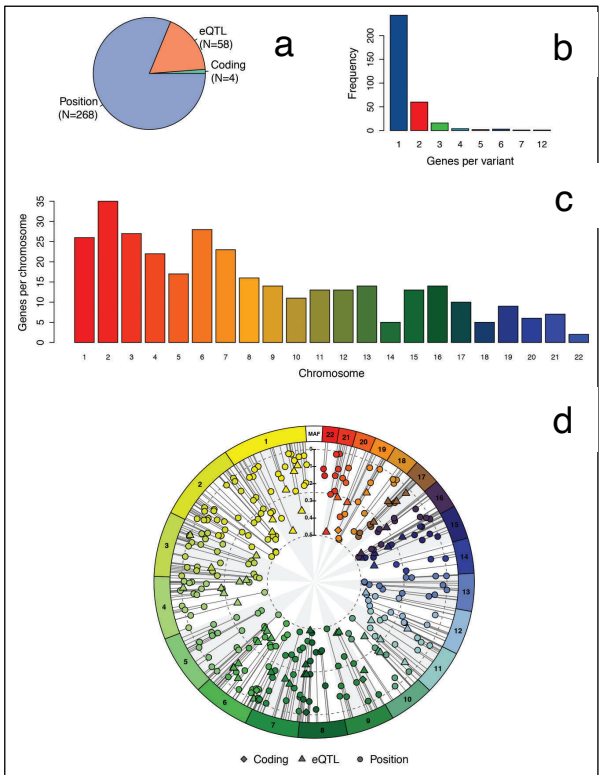

Figure 8: Variant-gene mapping from **AnnotateMe** tool.

We represent *variant-gene mapping* annotation with several plots showing the source of annotation of all variants (Figure 8a), the number of genes associated with each variant (Figure 8b), the distribution of the mapped genes across chromosomes (Figure 8c) and a circular summary visualization showing the source of

variant mapping, variant frequency and chromosomal distribution (Figure 8d). All plots will be sent to the user by email.

Previous associations of the variants and gene

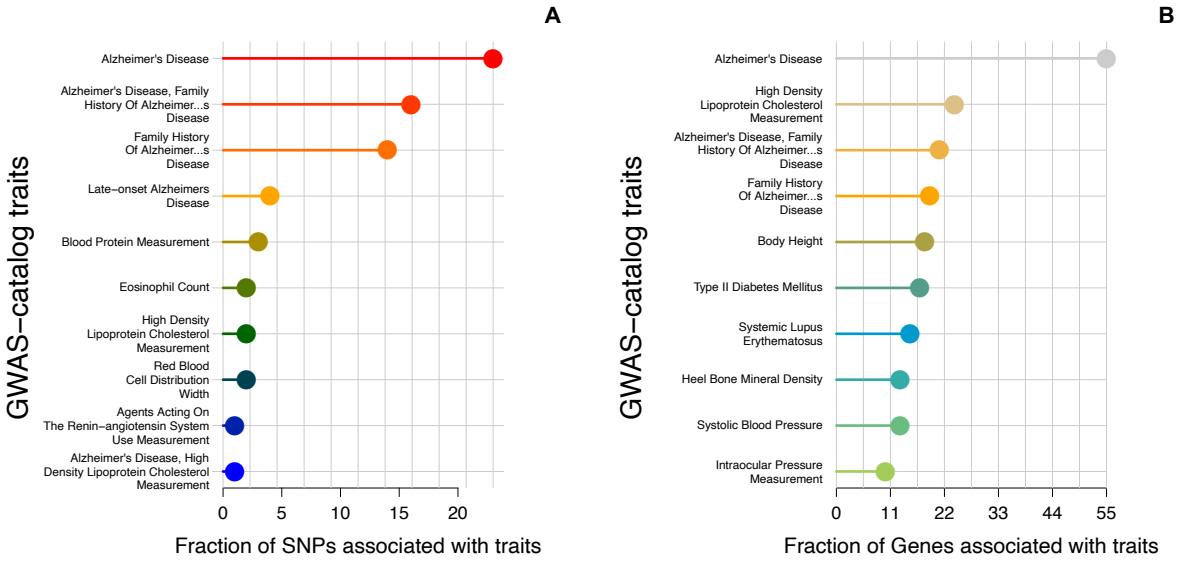

Figure 9: Overlap with GWAS catalog. The plot shows how we represent overlapping associations in the context of GWAS catalog. The overlap is done both at the level of variants and genes.

Then, we first report whether the input SNPs as well as their likely associated genes were previously associated with any trait in the GWAS-Catalog (traits are coded by their Experimental Factor Ontology EFO). For this analysis, we downloaded all significant SNP-trait associations of all studies available in the GWAS-Catalog (v1.0.2, available at https://www.ebi.ac.uk/gwas/docs/file-downloads), which includes associations with $p<9x10^{-6}$. Given a set of input SNPs associated with a set of genes, this analysis results in a set of traits (provided that the SNPs and/or the genes were previously associated with a trait). Hereto, we plot the number of SNPs in the list of uploaded SNPs that associate with the trait (expressed as a fraction). To correct for multiple genes being associated with a single variant, we estimate these fractions by sampling (500 iterations) one gene from the pool of genes associated with each variant, and averaging the resulting fractions across the sampling. We introduce this sampling-based framework in order to correct for multiple genes being associated with each variant, as neighboring genes are often functionally related. Therefore, 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 9a) and at the level of the genes (Figure 9b). Summary tables of the GWAS-Catalog analysis, including also EFO URI links for cross-referencing are provided as additional output.

Structural variants report

Next, we report on the structural variations (SV) that line in the vicinity ($10kb$ upstream and downstream) of the input variants, and present information such as SV start and end position, SV type, maximum in allele size, and genes likely associated with the relative SNPs.

Gene-set enrichment analysis

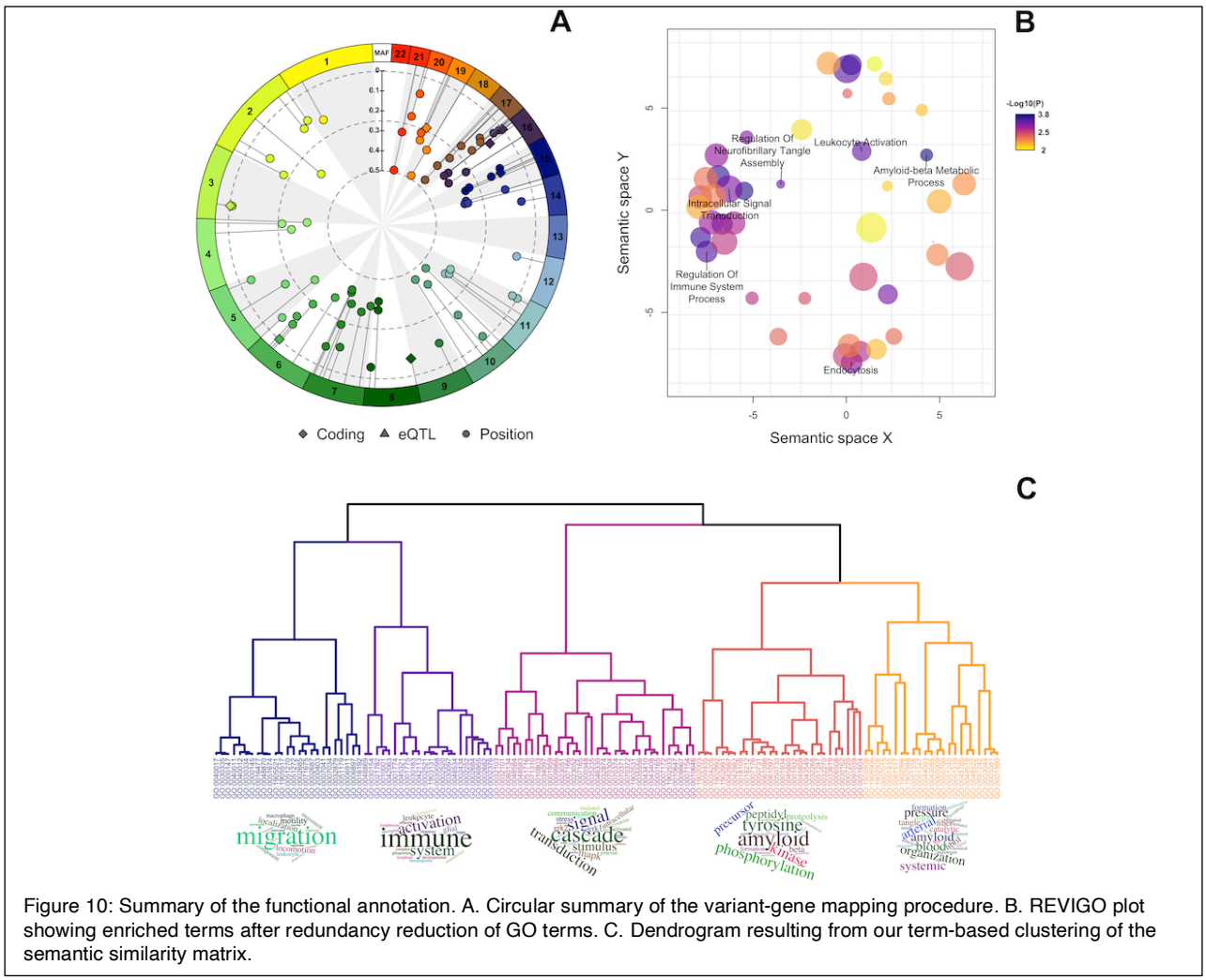

Figure 10: Summary of the functional annotation. A. Circular summary of the variant-gene mapping procedure. B. REVIGO plot showing enriched terms after redundancy reduction of GO terms. C. Dendrogram resulting from our term-based clustering of the semantic similarity matrix.

Finally, we perform a gene-set enrichment analysis to find molecular pathways enriched within the set of genes associated with the input variants. Also, here, we use the mentioned sampling technique to avoid a potential enrichment bias due to multiple genes being mapped to the same variant. However, this time the sampling is used to calculate *p*-values for each term, and these are averaged across the number of iterations. The gene-set enrichment analysis is performed using the *Gost* function from the R package *gprofiler2*. The user can specify several gene-set sources, such as Gene Ontology (release 2020-12-08), KEGG (release 2020-12-14), Reactome (release 2020-12-15), Wiki-pathways (release 2020-12-10).(Raudvere *et al.*, 2019) For each of the selected gene-set sources, the significant enriched terms are plotted (up to FDR<10%). In case the Gene Ontology is chosen as gene-set source, we additionally reduce the visual complexity of the enriched biological processes using (i) the REVIGO tool and (ii) a term-based clustering approach.(Supek *et al.*, 2011) We do so because the interpretation of gene-set enrichment analyses is typically difficult due to the large number of terms. Clustering enriched terms then helps to get an overview, and thus eases the interpretation of the results. Briefly, REVIGO masks redundant terms based on a semantic similarity measure,

and displays enrichment results in an embedded space via eigenvalue decomposition of the pairwise distance matrix. In addition to REVIGO, we developed a term-based clustering approach to remove redundancy between enriched terms. To do so, we first calculate a semantic similarity matrix between all enriched terms, and then apply hierarchical clustering on the obtained distance matrix. We estimate the optimal number of clusters using a dynamic cut tree algorithm and plot the most recurring words of the terms underlying each cluster using wordclouds. We use Lin as semantic distance measure for both REVIGO and our term-based clustering approach.(McInnes and Pedersen, 2013) Figures representing REVIGO results, the semantic similarity heatmap (showing relationships between enriched terms), the hierarchical clustering dendrogram, and the wordclouds of each clusters, are generated. Finally, all tables describing REVIGO analysis and our term-based clustering approach (including all enriched terms and their clustering scheme) are produced and sent as additional output to the user for further manipulation. Note that the initial significant GO terms are not removed and also included in the reporting.

5. Tutorial

The Help section of ***snpXplorer*** contains tutorial videos and sample files that can be directly use to get familiar with ***snpXplorer*** exploration and functional annotation sections. The Help section contains also the description of the datasets available in the exploration section (with relative references), the structural variation datasets and the populations included from 1000Genome project in order to calculate LD.

6. Stand-alone version

snpXplorer is freely available online as a web-server at https://snpxplorer.eu.ngrok.io. 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). 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 or snpxplorer@gmail.com) 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. ***snpXplorer*** uses the following R packages: shiny, data.table, stringr, ggplot2, liftOver, colourpicker, rvest, plotrix, parallel, SNPlocs.Hsapiens.dfSNP144.GRCh37, lme4, ggsci, RColorBrewer, gprofiler2, GOSemSim, GO.db, org.Hs.eg.db, pheatmap, circlize, devtools, treemap, basicPlotteR, gwascat, GenomicRanges, rtracklayer, Homo.sapiens, BiocGenerics, and the following python libraries: re, werkzeug, robobrowser, pygosemsim, numpy, csv, networkx and sys.

7. 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 at https://www.biorxiv.org/content/10.1101/2020.11.11.377879v1. In addition, ***AnnotateMe*** tool was used before in publications, and we report here the citations.(Tesi *et al.*, 2020, 2021)

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.

8. References

1000 Genomes Project Consortium *et al.* (2015) A global reference for human genetic variation. *Nature*, ***526***, 68–74.
Alzheimer Disease Genetics Consortium (ADGC), *et al.* (2019) Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.*, ***51***, 414–430.
Audano,P.A. *et al.* (2019) Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, ***176***, 663-675.e19.
CARDIoGRAMplusC4D Consortium *et al.* (2013) Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.*, ***45***, 25–33.
Chaisson,M.J.P. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, ***10***, 1784.
FinnGen *et al.* (2021) An expanded analysis framework for multivariate GWAS connects inflammatory biomarkers to functional variants and disease. *Eur. J. Hum. Genet.*, ***29***, 309–324.
Forgetta,V. *et al.* (2020) Rare Genetic Variants of Large Effect Influence Risk of Type 1 Diabetes. *Diabetes*, ***69***, 784–795.
GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, ***45***, 580–585.
Han,Y. *et al.* (2020) Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat. Commun.*, ***11***, 1776.
Hinrichs,A.S. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, ***34***, D590–D598.
Hong,E.P. and Park,J.W. (2012) Sample size and statistical power calculation in genetic association studies. *Genomics Inform.*, ***10***, 117–122.
Jansen,I.E. *et al.* (2019) Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.*, ***51***, 404–413.
Kent,W.J. *et al.* (2002) The Human Genome Browser at UCSC. *Genome Res.*, ***12***, 996–1006.
Linthorst,J. *et al.* (2020) Extreme enrichment of VNTR-associated polymorphicity in human subtelomeres: genes with most VNTRs are predominantly expressed in the brain. *Transl. Psychiatry*, ***10***, 369.
Manousaki,D. *et al.* (2020) Genome-wide Association Study for Vitamin D Levels Reveals 69 Independent Loci. *Am. J. Hum. Genet.*, ***106***, 327–337.
Matoba,N. *et al.* (2020) Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. *Transl. Psychiatry*, ***10***, 265.
McInnes,B.T. and Pedersen,T. (2013) Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *J. Biomed. Inform.*, ***46***, 1116–1124.
MDD Working Group of the Psychiatric Genomics Consortium *et al.* (2020) Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat. Genet.*, ***52***, 437–447.
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.
Raudvere,U. *et al.* (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, ***47***, W191–W198.
Regeneron Genetics Center *et al.* (2020) MEPE loss-of-function variant associates with decreased bone mineral density and increased fracture risk. *Nat. Commun.*, ***11***, 4093.
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.
Tesi,N. *et al.* (2020) Polygenic Risk Score of Longevity Predicts Longer Survival Across an Age Continuum. *J. Gerontol. Ser. A*, glaa289.
Tesi,N. *et al.* (2021) The effect of Alzheimer's disease-associated genetic variants on longevity Genetic and Genomic Medicine.
The GenOMICC Investigators *et al.* (2020) Genetic mechanisms of critical illness in Covid-19. *Nature*.
The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, ***449***, 851–861.
the Million Veteran Program *et al.* (2018) Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.*, ***50***, 1412–1425.

Timmers,P.R. *et al.* (2019) Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife*, ***8***.
Vojinovic,D. *et al.* (2018) Genome-wide association study of 23,500 individuals identifies 7 loci associated with brain ventricular volume. *Nat. Commun.*, ***9***, 3945.
Wang,Y.-F. *et al.* (2021) Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups. *Nat. Commun.*, ***12***, 772.
Yengo,L. *et al.* (2018) Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.*, ***27***, 3641–3649.