snpXplorer output description

Niccolo' Tesi

December 19, 2024

Dear user, thank you very much for using snpXplorer. We hope you found our tool useful for your analysis. If so, please do not forget to cite our paper in *Nucleic Acid Research* at this link.

Should you have question, comment or feedback, please email us at snpxplorer@gmail.com or n.tesi@amsterdamumc.nl.

In the next page you will find a detailed description of each of the **snpXplorer** outputs.

1 annotateMe_input_xxxxx.txt

This file reports the input as pasted into **snpXplorer** web-server and can be used for troubleshooting with the developers.

2 snp_annotation.txt

This file reports the complete annotation of the input variants. The column codes are as it follows:

- **locus**: chromosome:position of the variant
- **chr**: chromosome of the variant
- **pos**: genomic position of the variant. This position is typically aligned to the Human Reference GRCh37
- **ID**: variant identifier
- **ALT_FREQS**: minor allele frequency of the variant as calculated in all samples of the 1000Genome
- **cadd_ref**: reference allele in CADD database
- **cadd_alt**: alternative allele in CADD database
- **snp_conseq**: variant consequence as predicted by CADD annotation
- **snp_conseq_gene**: gene(s) predicted to be affected by the variant according to CADD annotation
- **phred**: pathogenicity score as predicted by CADD annotation, relative to the alternative allele
- **coding_snp**: whether the variant is a coding variant (*yes*) or not (*no/NA*)
- **chr_hg38**: chromosome of the variant according to the Human Reference GRCh38
- **pos_hg38**: genomic position of the variant. This position is typically aligned to the Human Reference GRCh38
- **code_gtex**: unique identifier used to match against GTEx database regarding eQTL (expression-quantitative-trait-loci) and sQTL (splicing-quantitative-trait-loci)
- **eqtl**: gene(s) whose expression is significantly associated with SNP genotype from GTEx database. In case multiple eQTLs are found (either because the same SNP has multiple eQTLs or because multiple tissues were selected), this field reports a comma-separated list of genes. Genes (from *eqtl* column) and tissues (from *eqtl_tissue* column) are linked by their index.
- **eqtl_tissue**: tissue(s) in which the eQTL was found. In case multiple eQTLs are found (either because the same SNP has multiple eQTLs or because multiple tissues were selected), this field reports a comma-separated list of tissues. Genes (from *eqtl* column) and tissues (from *eqtl_tissue* column) are linked by their index.
- **sqtl**: gene(s) whose splicing-expression is significantly associated with SNP genotype from GTEx database. In case multiple sQTLs are found (either because the same SNP has multiple sQTLs or because multiple tissues were selected), this field reports a comma-separated list of genes. Genes (from *sqtl* column) and tissues (from *sqtl_tissue* column) are linked by their index.
- **sqtl_tissue**: tissue(s) in which the sQTL was found. In case multiple sQTLs are found (either because the same SNP has multiple sQTLs or because multiple tissues were selected), this field reports a comma-separated list of tissues. Genes (from *sqtl* column) and tissues (from *sqtl_tissue* column) are linked by their index.
- **positional_mapping**: genes closest to the variant based on genomic position. This annotation is only reported when the variant is *not* coding *or* when the variant has *no* quantitative-trait-loci associations (either eQTLs or sQTLs).

- **source_finalGenes**: the source that was used for the annotation. The value is *coding* for variants annotated only based on CADD annotation; *sqtl+eqtl+cadd* for non-coding variants with quantitative-trait-loci. In this case the genes from CADD, eQTL and sQTL are combined; *position* for non-coding variants with *no* quantitative-trait-loci (eQTL and sQTL).
- **geneList**: final list of genes likely affected by the variant

3 snp_annotation_genelist.txt

This file reports the list of all genes likely associated with the variant. In case the analysis type chosen was *enrichment analysis*, this is the list of genes that was used for the gene-set enrichment analysis.

4 snp_gene_mapping.pdf

This file shows a summary of the annotation procedure. Three plots are reported:

- **pie-plot**: top-left plot, shows the source of annotation of all variants.
- **barplot**: top-right barplot, shows the number of genes associated with each variant.
- **barplot**: center barplot, shows the chromosomal distribution of the mapped genes.
- **circular-plot**: bottom plot, shows the final circular summary including chromosomal distribution of the input variants, their minor allele frequency and the source of the annotation used.

5 gwas_cat_genes_overlap.pdf

This file represents a figure showing the fraction of genes associated with the input SNPs for which a previous association was reported in the GWAS-catalog.

6 gwas_cat_snps_overlap.pdf

This file represents a figure showing the fraction of the input SNPs for which a previous association was reported in the GWAS-catalog.

7 gwas_cat_snps_overlap.txt

This tab-separated file reports a summary table of the GWAS-catalog analysis in the context of variants. The column codes can be found at this link.

8 gwas_cat_genes_overlap.txt

This tab-separated file reports a summary table of the GWAS-catalog analysis in the context of genes. The column codes can be found at this link. Note that a sampling framework is applied to this analysis, to allow for multiple genes to be associated with the same variant. Briefly, every iteration (500 iterations in total), one gene is sampled from the list of genes associated with each variant. The resulting gene-set is used to query the GWAS catalog for previous associations of the gene(s) with other traits. This results in a number of genes associated with each trait, for each iteration. The column *SUM* include then the sum of all matches across all iterations, while the *MEAN* column reports the average number of genes matching previous annotations in the GWAS catalog. Column *genes_max_overlap* reports all genes overlapping the reported trait, from the input genes.

9 SNP_and_SV_overlap.txt

This tab-separated file shows the structural variants (SV) that lie in the vicinity (10kb) of the input variants. Columns are coded as it follows:

- **chr_hg38**: chromosome according to Human Reference GRCh38.
- **start_pos_SV_hg38**: starting position of the SV according to GRCh38.
- **end_pos_SV_hg38**: end position of the SV according to GRCh38.
- **diff_allele_size_SV**: maximum difference in allele sizes observed for this SV.
- **SV_type**: the type of structural variant.
- **SV_source**: the source of the structural variant.
- **SV_SNP_relation**: whether the SNP is located upstream, downstream or inside the relative SV.
- **Distance_SNP_SV**: distance between the SNP position and the SV position (closest between the start and end position of the SV).
- **SNP_locus_hg38**: SNP position according to Human Reference GRCh38.
- **geneList**: the genes likely associated/affected by the relative SNP and potentially the SV.

In case you requested a **gene-set enrichment analysis**, the following additional files will be present in your results folder:

10 geneSet_enrichment_results_and_clusters.txt

This file shows the results of the gene-set enrichment analysis after sampling and averaging p-values (please click here for further details how this is done). Results from our term-based clustering analysis as included as well. The columns in this files code are as it follows:

- **term_name**: the name of the term tested for enrichment.
- **term_id**: the unique identifier of the term tested for enrichment. This code is specific for each term and depends on the gene-set chosen for enrichment analysis.
- **avgP**: average p-value across the iterations (N=500 iterations).
- **log10P**: $-\log_{10}(p_{avg})$.
- **source_gset**: main gene-set of the relative term.
- **cluster_in_dendro**: cluster that is assigned to the term according to our term-based clustering analysis of enriched functional terms.

11 Enrichment_results.pdf

This file shows the enrichment results for all selected gene-set of interest *except* for Gene-Ontology, which is reported as a separate file.

12 revigo_inp.txt

This file represents the input for REVIGO analysis, and reports the list of GO terms and respective p-values (same as the geneSet_enrichment_results_and_clusters.txt file). This file can be used to replicate REVIGO analysis, or in case this failed (*e.g.* if REVIGO was under maintenance).

13 revigo_out.txt

This file represents the output table of the REVIGO analysis. As for reference, REVIGO performs a semantic-similarity-based reduction of GO terms to easier the interpretation of gene-set enrichment analysis results.

14 clustering_GO_terms.pdf

This file represents the visual output of REVIGO analysis. The plot is drawn from the revigo_out.txt file using a home-made script that is available at this Github page.

15 alternative_Lin_distance.txt

This file report the semantic-similarity-based distance matrix. The rows and columns of the matrix represent the different enriched terms. This distance matrix is done using *Lin* as semantic similarity measure.

16 pheatmap_lin_distance.png

As the title suggests, this file shows a heatmap of the semantic-similarity-based distance matrix. Note that this file and alternative_Lin_distance.txt are not produced by REVIGO, but are based on our term-based clustering of enriched terms.

17 dendrogram_GOterms.png

This figure shows the dendrogram as obtained through hierarchical clustering of the semantic-similarity-based distance matrix (pheatmap_lin_distance.png). The number of clusters is calculated using a dynamic cut tree algorithm. Please click here for further details how this is done.

18 cluster_x_wordcloud.png

Depending on the number of cluster that our clustering approach finds, you will get one wordcloud plot for each cluster. The wordcloud is meant to help in the interpretation of the clusters in the dendrogram. Most recurring words in GO terms descriptions are removed, thus the wordcloud shows only the most frequent terms within each cluster.