Batch Annotation output description

snpXplorer_input_XXXX.txt: Raw input as pasted in snpXplorer website including SNPs to annotate.

invalid_queries.txt: Any invalid SNP part of the raw input that could not be parsed by snpXplorer.

variant_annotation_combined.tsv: the final annotation of the target SNPs. Includes the following columns:

Column	**Description**
Query Type	Query type identified by snpXplorer
RsID	RS identifier of the variant
Chromosome	Chromosome
Position (hg38)	Position with respect to GRCh38
Position (hg19)	Position with respect to GRCh37
Reference	Reference Allele (from TOPMed)
Alternative	Alternative Allele (from TOPMed)
Alternative Allele Frequency	Alternative Allele Frequency (from TOPMed)
Locus ID	Locus identifier
CADD Reference	Reference Allele (CADD)
CADD Alternative	Alternative Allele (CADD)
CADD Score	CADD Phred Score (CADD)
CADD Gene	Likely affected gene (CADD)
CADD Annotation Types	Type of annotation (CADD)
CADD Consequences	SNP Consequences (CADD)
CADD CpG Max	Maximum CpG context score (CADD)

CADD GC Max	Maximum GC content score (CADD)
CADD H3K27ac Max	Overlap with H3K27ac histone marks (CADD)
CADD H3K4me3 Max	Overlap with H3K4me3 marks (CADD)
CADD DNAase Max	DNAse I hypersensitivity (CADD)
CADD SpliceAI Acc Loss Max	Predicted loss of acceptor splice site (CADD)
CADD SpliceAI Don Loss Max	Predicted loss of donor splice site (CADD)
CADD SIFT Max	SIFT Score (CADD)
CADD PolyPhen Max	PolyPhen Score (CADD)
CADD GERPRS Max	Evolutionary context (CADD)
CADD PhyloP Max	Nucleotide-level conservation (CADD)
eQTL Reference	Reference Allele for eQTL (GTEx)
eQTL Alternative	Alternative Allele for eQTL (GTEx)
eQTL ensemble	Ensemble gene ID (GTEx)
eQTL Gene	eQTL Gene name (GTEx)
eQTL Tissue	eQTL Tissue (GTEx)
eQTL TSS Distance	Distance of SNP-TSS (GTEx)
eQTL P-value	eQTL P-value (GTEx)
eQTL Slope	eQTL effect size of Alternative allele (GTEx)
eQTL MAF	Minor allele frequency (GTEx)
sQTL Reference	Reference Allele for sQTL (GTEx)
sQTL Alternative	Alternative Allele for sQTL (GTEx)
sQTL ensemble	Ensemble gene ID (GTEx)
sQTL Gene	sQTL Gene name (GTEx)
sQTL Tissue	sQTL Tissue (GTEx)
sQTL TSS Distance	Distance of SNP-TSS (GTEx)
sQTL P-value	sQTL P-value (GTEx)
sQTL Slope	sQTL effect size of Alternative allele (GTEx)
sQTL MAF	Minor allele frequency (GTEx)
Most Likely Genes	Most likely gene(s) affecting variant (snpXplorer)
Source	Source(s) used to identify genes (snpXplorer)

target_annotations/cadd_annotation_target.tsv: Complete CADD annotation for each variant. Columns are the same as in *variant_annotation_combined.tsv*, however, multiple

annotations per variant may exist. In the *variant_annotation_combined.tsv*, the annotation with the largest CADD Score is prioritized and reported.

target_annotations/eqtl_annotation_target.tsv: Complete eQTL annotation for each variant. Columns are the same as in *variant_annotation_combined.tsv*.

target_annotations/sqtl_annotation_target.tsv: Complete sQTL annotation for each variant. Columns are the same as in *variant_annotation_combined.tsv.*

target_annotations/gwas_annotation_target.tsv: GWAS hits identified for each variant. Note that not all GWAS traits from OpenGWAS are queried (>10,000 traits). To redundancy between similar terms, trait names were embedded using a combination of MiniLD and SapBERT language models to capture semantic similarity. Pairwise cosine similarity between embeddings was then used to cluster related traits. Columns are as it follows:

Column	**Description**
GWAS Trait	Trait name
Effect Allele	Effect Allele
Non-effect Allele	Other Allele
EAF	Effect Allele Frequency
GWAS Beta	Effect size of Effect allele
GWAS SE	Standard error of effect size
GWAS P	P-value of association
Sample Size	Sample size of GWAS
GWAS ID	GWAS identifier (OpenGWAS)
RsID	Variant Identifier

target_annotations/sv_annotation_target.tsv: Structural variants in the vicinity of each variant. The set of structural variants was derived from *Tesi et al, 2024*. Columns are as it follows:

Column	**Description**
Repeat Class	Class of the repeat
Chromosome	Chromosome
Start Position (hg38)	Start Position in GRCh38
End Position (hg38)	End Position in GRCh38

Length	Length of the structural variant
Repeat Name	Repeat name
Repeat Family	Repeat family
Color	Color as shown in snpXplorer Exploration section

plots_variant_annotation/gwas_annotation_top_traits.png: Summary image of the GWAS traits (OpenGWAS) associated with each variant.

plots_variant_annotation/variant_annotation_summary.pdf: Summary image of each variant, frequency, type and chromosomal distribution.

ld_partner_annotations/ld_annotation_target.tsv: Linkage disequilibrium (LD) between input variants and all their LD partners. For LD, we used TOPMed maps. Columns are as it follows:

Column	**Description**
Query Position	Position of query variant (input variant) in GRCh38
LD Partner Position (hg38)	Position of the partner variant in GRCh38
LD Partner RsID	RsID of the partner variant
LD R2	LD R2 value between variants (TOPMed)
LD Distance (bp)	Distance between query and partner
Chromosome	Chromosome
Position (hg38)	NA
Partner unique ID	Unique ID of the partner variant

ld_partner_annotations/cadd_annotation_LD_partners.tsv: CADD annotation of all LD partners (i.e. variants in LD with the input variants). Columns are the same as in *cadd_annotation_target.tsv.*

ld_partner_annotations/eqtl_annotation_LD_partners.tsv: eQTL annotation of all LD partners (i.e. variants in LD with the input variants). Columns are the same as in *eqtl_annotation_target.tsv.*

ld_partner_annotations/sqtl_annotation_LD_partners.tsv: sQTL annotation of all LD partners (i.e. variants in LD with the input variants). Columns are the same as in *sqtl_annotation_target.tsv.*

gene_set_enrichment/gene_set_enrichment_results.tsv: Only available when gene-set enrichment analysis was requested. This file contains enrichment results. Note that our SNP-gene annotation procedure can link a single SNP with multiple genes, depending on annotation uncertainty. To take this uncertainty into account, gene-set enrichment analysis is performed in a sampling-based framework (100 iterations). Each iteration, a single gene per SNP is taken. The resulting gene list is used for gene-set enrichment analysis. Enrichment results are corrected for multiple tests (false discovery rate, FDR). At the end of all iterations, enrichment p-value of all terms is averaged across iterations. The final Enrichment P-value is therefore an average enrichment across all iterations, and does not need to be further corrected for multiple tests. Columns are as it follows:

Column	**Description**
Enrichment Term ID	Term ID (gProfiler2)
Enrichment Term Name	Term Name (gProfiler2)
Enrichment Source	Enrichment source (gProfiler2)
Enrichment P-value	P-value of enrichment (gProfiler2)
Enrichment Term Size	Size of the Term (gProfiler2)
Enrichment Query Size	Query Size (gProfiler2)
Enrichment Precision	Precision (gProfiler2)
Enrichment Recall	Recall (gProfiler2)
Enrichment Term Description	Term Description (gProfiler2)
Enrichment Intersections	Intersection between Term and Query (gProfiler2)

gene_set_enrichment/semantic_similarity_matrix.tsv: To improve interpretation of gene-set enrichment analysis, we use semantic similarity between Terms. Note that this analysis is restricted to significant terms (P-value<0.05) from GeneOntology, as it follows a tree-like structure that improves similarity measurement. We use *Lin* as semantic similarity distance. This matrix shows pairwise similarity between all significant GO terms after gene-set enrichment analysis.

gene_set_enrichment/pheatmap_lin_distance.pdf: Image showing the similarity between GO terms.

gene_set_enrichment/clustering_max_{5,8,10,15}_clusters.tsv: The semantic similarity goes through a dynamic cut tree algorithm to identify clusters. We use different parameters to allow a maximum number of clusters ranging 5-15, where 5 gives less clusters (max 5) and 15 gives more fine-tuned clustering (max 15 clusters). Depending on the use, the user might prefer one clustering over another one. Note that in case two clustering parameters (e.g., max 8 clusters and max 10 clusters) produced the same output, only one is reported. Columns are as it follows:

Column	**Description**
Term	GO term ID of the term
Cluster	Cluster number
Cluster in dendrogram	Cluster as it appears on the dendrogram image
Native	GO term ID of the term
Name	Description of the term

gene_set_enrichment/clustering_max_{5,8,10,15}_dendrogram_clusters.png: Images of the dynamic cut tree algorithm showing the dendrogram of the significant terms, colored by their cluster assignment. One image for each clustering parameter.

gene_set_enrichment/wordclouds_{5,8,10,15}_clusters: To facilitate interpretation of the enrichment clusters, we use wordclouds of the term underlying each cluster. These folders contain an image per cluster, specifying the order in which they occur in the dendrogram (left to right). There is one folder for each clustering parameter (max 5, 8, 10, 15 clusters).

pathway_prs_weights/pathway_prs_weights_{number of cluster}_clusters.tsv: These files, one for each successful clustering run (similar and aligned to the clustering above and wordclouds), report weights to be used for pathway-specific polygenic risk scores (PRS). Specifically, for each input variant, depending on what genes it was associated with (SNP-gene annotation), and depending on which significant terms the gene was involved into (gene-set enrichment analysis), it is reported the effect of that variant on each of the pathways (in proportion, summing to 1 across clusters of pathways). Pathway-PRS

constructed in this way were published multiple times (*Tesi et al., 2021*, *Lorenzini et al., 2023*, *Tranfa et al., 2025*). The file shows one variant per line, and each column represent a cluster, with the relative proportion of effect of each variant on that cluster.
Using this file (with some minimal edits), users can calculate pathway-specific PRS in their cohort. We provide a tool for that, *jordan*, which can be run both as a command-line and with a user interface.