

Genoppi (v1.0.0) Welcome Guide

Table of contents

Getting started	pg. 2
Input format	pg. 3
Basic plotting	pg. 5
Integrated plotting	pg. 7
Gene set annotations	pg. 13
Tissue enrichment	pg. 15
Multiple files comparison	pg. 18
Downloads	pg. 21
References	pg. 22

Getting started

Genoppi is an open-source software for performing quality control and analyzing quantitative human proteomic data. Genoppi streamlines the integration of user-inputted proteomic data with external datasets such as known protein-protein interactions in published literature, data from population genetic studies, gene set annotations, tissue-specific RNA or protein expression, or other user-defined inputs. This protocol provides documentation for using the interactive Genoppi web application, which is available at www.lagelab.org/genoppi. Descriptions for external datasets incorporated into Genoppi are provided in the “Data Documentation” page of the application. Source code for the application and the stand-alone Genoppi R package can be downloaded at github.com/lagelab/Genoppi for local installation.

Please direct questions and comments to Kasper Lage (lage.kasper@mgh.harvard.edu).

Input format

Genoppi can be used to analyze quantitative proteomic data that contain protein \log_2 fold change (FC) values between studied conditions, such as bait vs. control immunoprecipitations followed by mass spectrometry (IP-MS). Protein quantification results generated using labeling-based (e.g., iTRAQ, TMT, or SILAC) or label-free MS methods can be inputted into Genoppi following the input file format described below.

The input file must be a tab-delimited text file. At minimum, the file must contain three columns, with one column specifying protein identifiers and two columns listing protein \log_2 FC values for two or more experimental replicates. More specifically:

Column 1: protein identifiers as either HUGO Gene Nomenclature Committee¹ (HGNC) [www.genenames.org] approved symbols (with “gene” as column header), or UniProt² [www.uniprot.org] accession numbers (with “accession_number” as column header).

Columns 2, 3, (+ optional additional columns): \log_2 FC values for \geq two replicates, with “rep1”, “rep2”, and so on as column headers for the replicates.

OR

Columns 2, 3, 4: average \log_2 FC across replicates (“logFC”) with corresponding *P*-value (“pvalue”) and false discovery rate (“FDR”) calculated using a statistical test (e.g., a moderated t-test).

Missing values are not allowed; any rows with missing values would be disregarded with no error message.

Examples of accepted input format with correct column headers:

1. HGNC symbol and \log_2 FC for two replicates

gene	rep1	rep2
FOXP2	-0.496	-0.546
RB1	0.402	0.265
SHH	0.08	0.104

2. UniProt accession number and \log_2 FC for three replicates

accession_number	rep1	rep2	rep3
O15409	-0.496	-0.546	-0.447
P06400	0.402	0.265	0.410
Q15465	0.08	0.104	0.125

3. HGNC symbol and pre-calculated results of statistical test

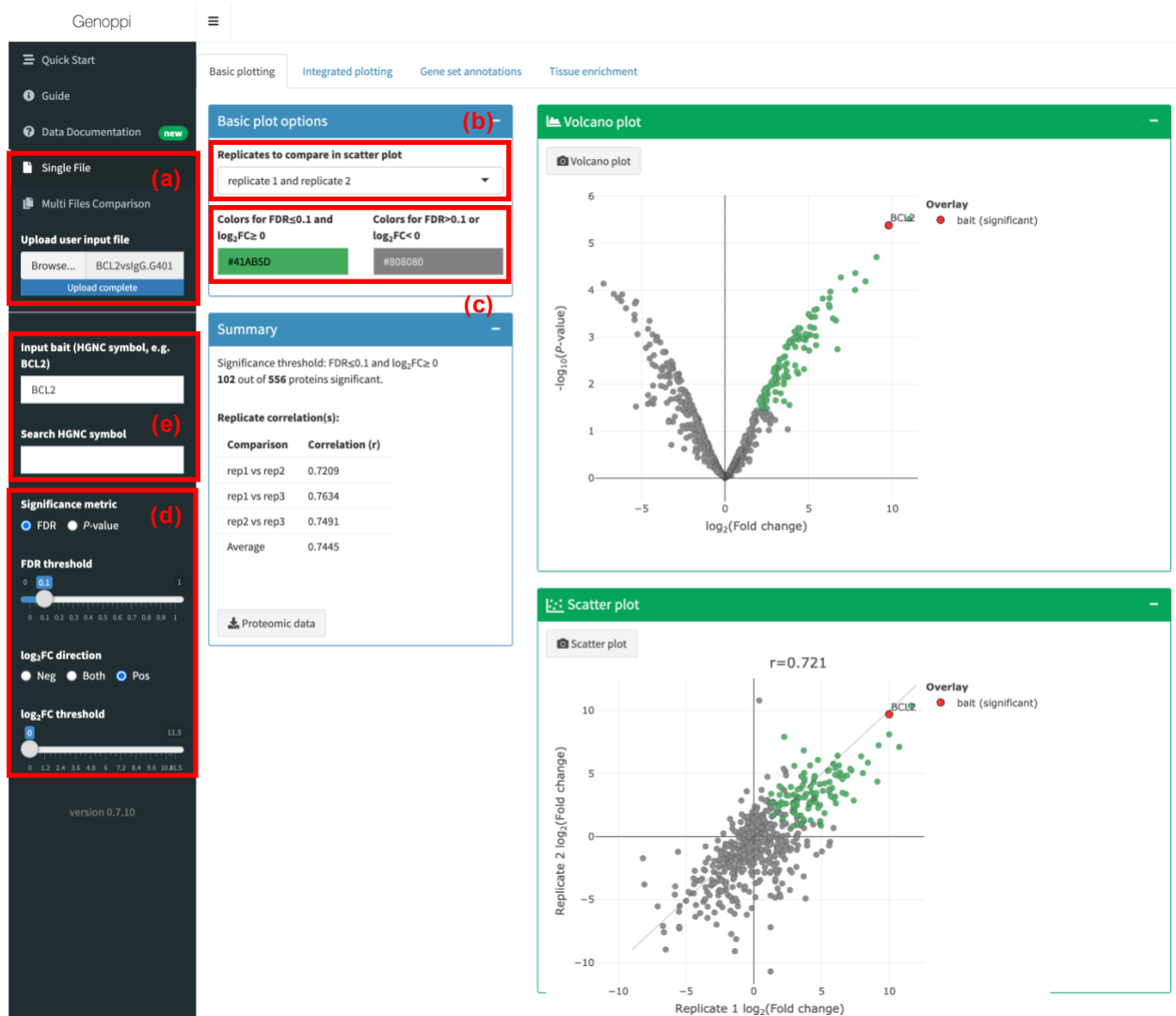
gene	logFC	pvalue	FDR
FOXP2	-0.521	1.64e-3	4.14e-3
RB1	0.334	0.0118	0.0189
SHH	0.092	0.211	0.242

For Mac users exporting data from Excel format, please convert it to text file by selecting “File” > “Save As...” > “File Format” > “Tab delimited Text (.txt)”. This would avoid generating a file that terminates each line with a carriage return character, which is incompatible with subsequent analysis in Genoppi.

Basic plotting

Screenshot 1 illustrates the basic user interface of the Genoppi application. After the user uploads a “Single File” input in the left panel (**Screenshot 1a**), the “Basic plotting” module will generate an interactive volcano plot, depicting the average \log_2 FC of proteins on the x-axis and the $-\log_{10}$ P -value on the y-axis. If \log_2 FC values from \geq two replicates are provided in the input file, a moderated t -test from the limma³ R package is applied to calculate the average \log_2 FC, nominal P -value, and FDR; otherwise, Genoppi uses the user-supplied statistics to generate the plot. In addition, a scatter plot showing replicate \log_2 FC correlation is generated if the input file includes separate replicates; when there are > 2 replicates, the user can select from a drop-down menu to show the scatter plot corresponding to each pair of replicates (**Screenshot 1b**).

In the default coloring scheme, significant proteins with \log_2 FC ≥ 0 and FDR ≤ 0.1 are in green, and other detected proteins are in grey. The user can change the colors (**Screenshot 1c**) or modify the significance threshold for defining significant proteins based on different FDR, P -value, and \log_2 FC cutoffs (**Screenshot 1d**). The adjustable cutoffs allow Genoppi to account for various types of proteomic experiments. For instance, when identifying the interactome of a bait protein compared to control, the user should look for significant proteins with positive \log_2 FC; when identifying proteins with differential abundance in two experimental conditions, significant proteins with either positive or negative \log_2 FC are both of interest. The “Summary” box shows the number of significant proteins (and total number of detected proteins) based on the specified threshold, as well as the correlation between replicates when appropriate. Hovering over each protein’s data point in either the volcano or scatter plot would show its corresponding HGNC symbol. The user can also query specific HGNC symbols to label bait and other proteins in the plots (**Screenshot 1e**).



Screenshot 1. Basic plotting interface showing volcano and replicate correlation scatter plots generated from input proteomic data.

Integrated plotting

In the “Integrated plotting” module, Genoppi enables integration of the proteomic data with data from the InWeb_InBioMap^{4, 5}, iRefIndex⁶, BioPlex^{7, 8}, NHGRI-EBI GWAS catalog⁹, gnomAD¹⁰, GTEx^{11, 12}, Human Protein Atlas¹³ (HPA), or user-uploaded SNP or gene lists.

InWeb_InBioMap, iRefIndex, or BioPlex

Genoppi can overlay the proteomic data with published human protein-protein interactions in PPI databases, including InWeb_InBioMap (InWeb) [www.intomics.com/inbio/map.html#downloads], iRefIndex [irefindex.vib.be/wiki/index.php/iRefIndex], and BioPlex [bioplex.hms.harvard.edu]. This integration enables the user to easily distinguish new interactions from those already reported in the literature. The user can search for known interactors of a specific protein in the selected database to visualize their overlap with significant proteins in the proteomic data in an overlaid volcano plot (**Screenshot 2a**). In addition, the user can choose to subset the published interactions based on the confidence metric provided by each database. For instance, InWeb interactors can be subsetted to “gold-standard” interactors curated from pathway databases, or “high-confidence” interactors with high confidence scores in the database, thereby excluding noisier data in the literature. In the volcano plot, published interactors (as well as other data types described below) detected in the proteomic data are labeled using an adjustable color and shape scheme (**Screenshot 2b**).

GWAS catalog

Genoppi can perform SNP-to-gene mapping for SNPs found in the 1000 Genomes Project¹⁴ [www.internationalgenome.org], using pre-calculated pairwise linkage disequilibrium (LD) measures between SNPs to identify all genes in LD regions. Therefore, the user can query diseases and traits found in the NHGRI-EBI GWAS catalog [www.ebi.ac.uk/gwas] to identify genes mapped from published trait-associated SNPs in the catalog (**Screenshot 2c**). Proteins encoded by the mapped genes would be labeled in the interactive volcano plot, and hovering over each of these proteins would show the SNP(s) that map to it.

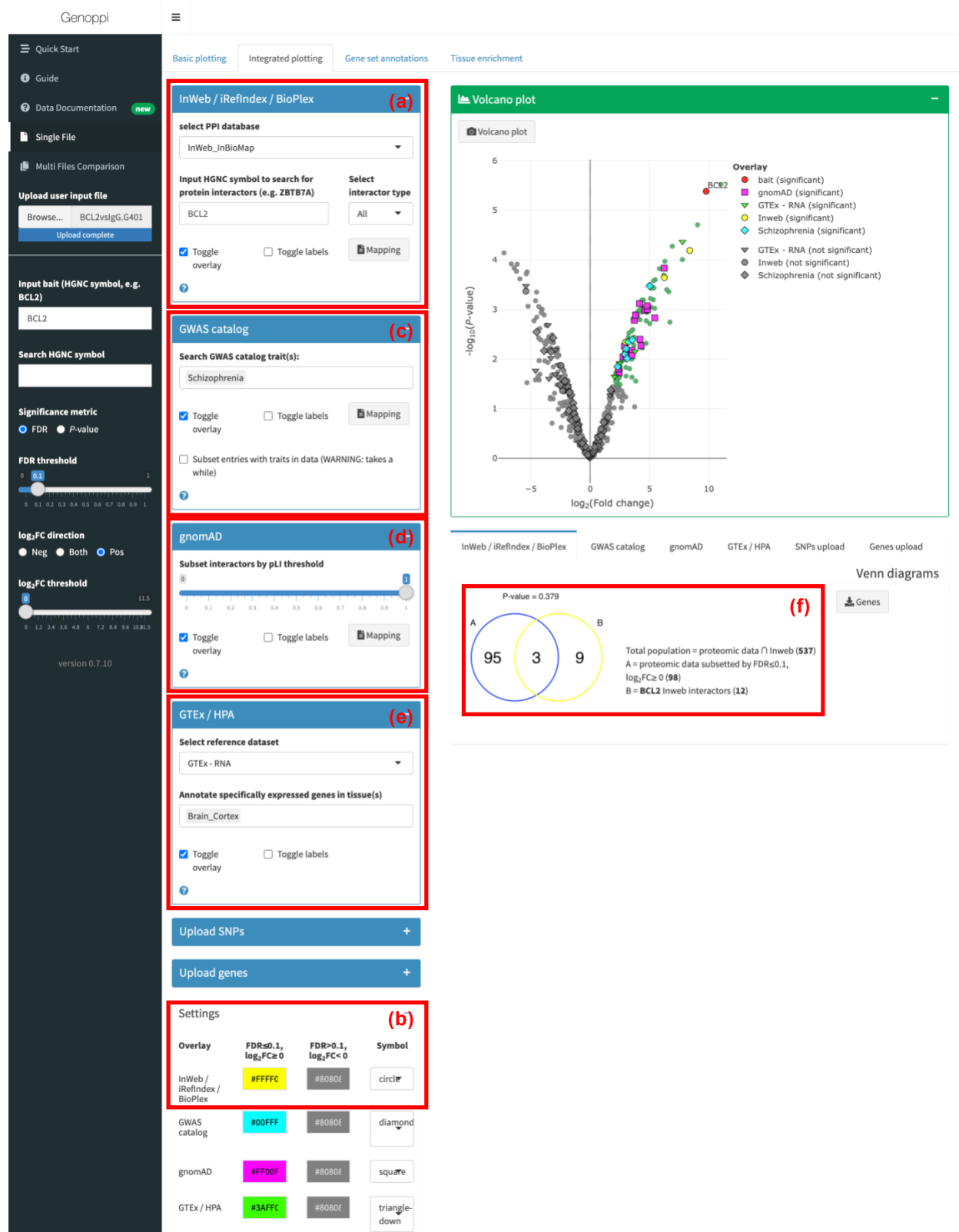
gnomAD

Genoppi can identify proteins encoded by genes that are likely intolerant of loss-of-function (LoF) mutations using constraint data from gnomAD [gnomad.broadinstitute.org]. The user can label proteins with pLI scores (i.e., probability of intolerance to LoF mutations) greater than an

adjustable threshold to visualize the most intolerant proteins in the overlaid volcano plot (**Screenshot 2d**).

GTEX or HPA

Genoppi can identify proteins encoded by tissue-specific genes derived from one of three GTEX [gtexportal.org] or HPA [www.proteinatlas.org] datasets: (1) “GTEX - RNA”: tissue-specific genes defined by Finucane *et al.*¹¹ using GTEX RNA-seq data, (2) “GTEX - protein”: tissue-enriched genes defined by Jiang *et al.*¹² using GTEX protein expression data, and (3) “HPA - RNA”: tissue-elevated genes defined by Uhlén *et al.*¹³ using HPA RNA-seq data. After picking one of the datasets in a drop-down menu, the user can select tissue(s) found in the dataset to visualize the corresponding tissue-specific proteins in the overlaid volcano plot (**Screenshot 2e**).



Screenshot 2. Integrated plotting interface showing integration of proteomic data with InWeb, GWAS catalog, gnomAD, and GTEx or HPA data.

Upload SNPs or genes

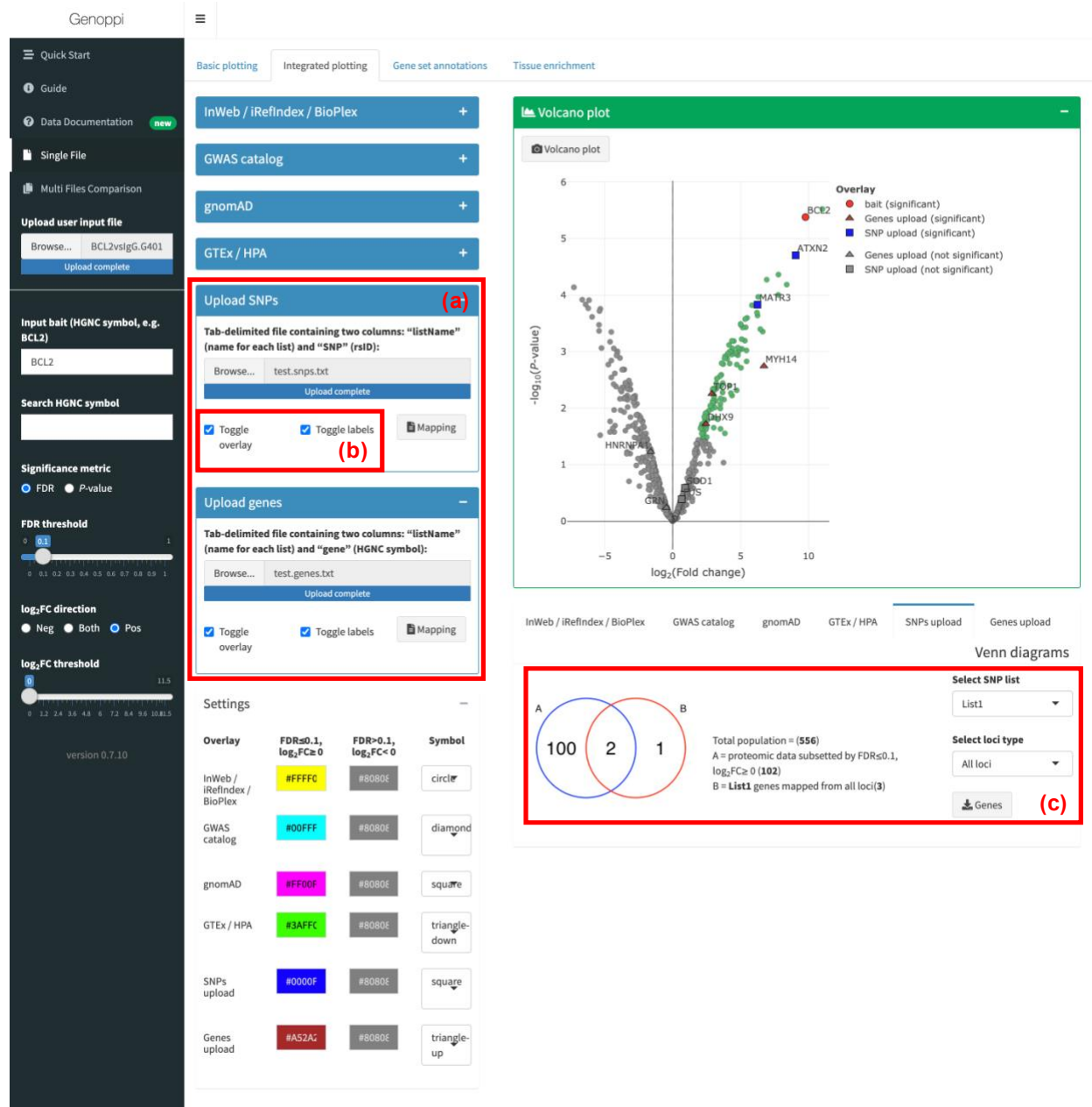
Besides incorporating the public datasets described above, the user may also upload one or more custom SNP or gene lists (e.g., disease-causing genes curated from literature review or genes implicated by gene-based burden testing) to assess their overlaps with the proteomic data (**Screenshot 3a**). As described in the “GWAS catalog” section, uploaded SNPs would be mapped to genes in LD using Genoppi’s built-in SNP-to-gene mapping functionality. The SNP list(s) must be uploaded as a tab-delimited plain text file containing two columns: “listName” (name for each list) and “SNP” (rsID). For example:

listName	SNP
List1	rs848132
List1	rs244285
List1	rs10757278
List2	rs3892097
List2	rs539515

Similarly, the gene list(s) must be uploaded as a tab-delimited text file consisting of two columns: “listName” (name for each list) and “gene” (HGNC symbol). For example:

listName	gene
ListA	SHH
ListA	UBC
ListB	FOXP2
ListB	RB1
ListB	KRAS

In the “Integrated plotting” module, the user may input any combination of InWeb interaction partners, GWAS catalog mapped genes, gnomAD constrained genes, GTEx or HPA tissue-specific genes, and custom SNP and gene lists. The resulting volcano plot would highlight all identified proteins from these inputs. Overlaying multiple datasets could result in a densely labeled plot, in which case the user can choose to remove the overlay or the protein text labels for each data type using the “Toggle overlay” or “Toggle labels” option, respectively (**Screenshot 3b**).



Screenshot 3. Integrated plotting interface showing integration of proteomic data with user-uploaded SNP and gene lists.

Venn diagrams

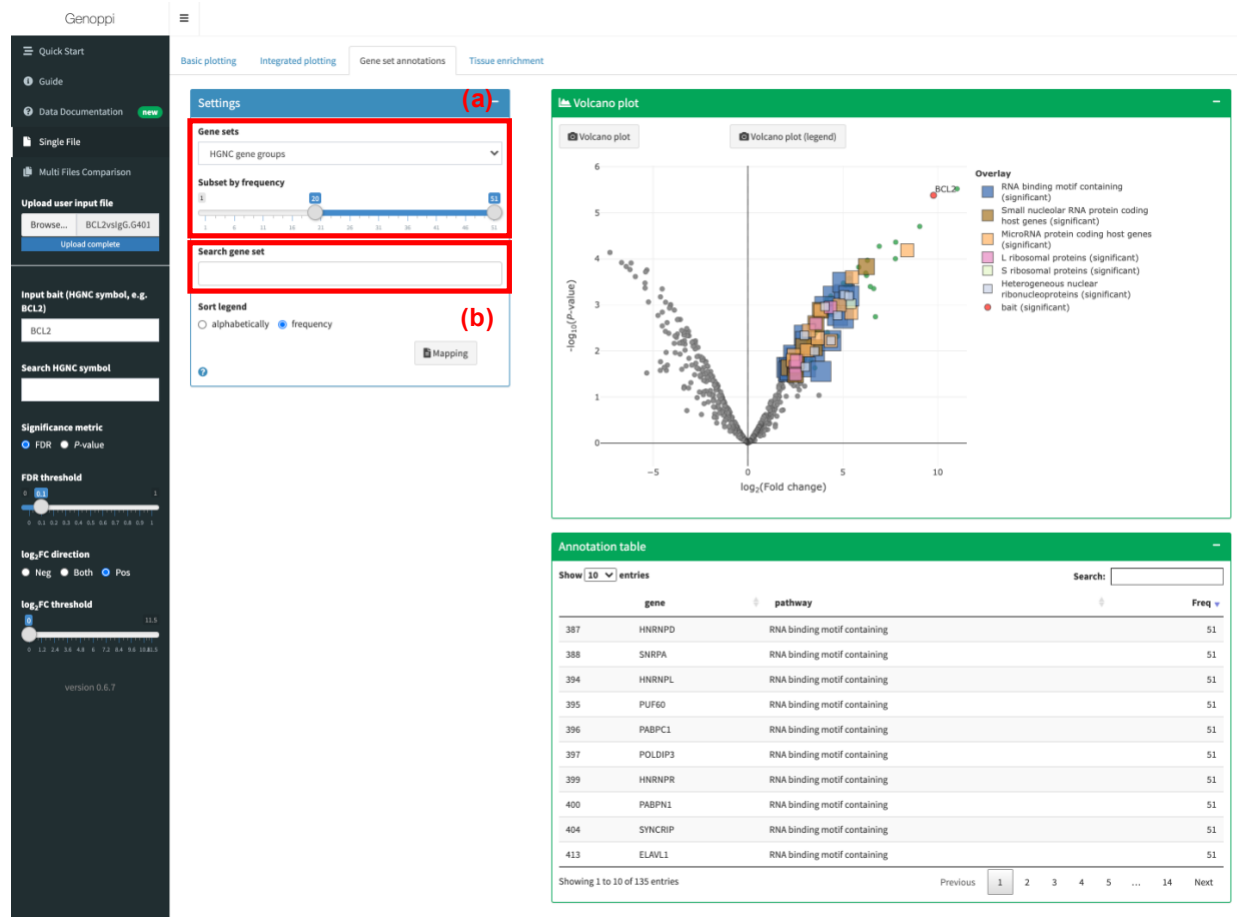
In the “Integrated plotting” module, Genoppi also summarizes the overlaps between the significant proteins in the proteomic data and the various data types described above using Venn diagrams. When showing overlap with PPI database interactors, gnomAD constrained genes, GTEx or HPA tissue-specific genes, or user-uploaded gene lists, Genoppi also assesses the overlap enrichment by calculating a hypergeometric *P*-value, which is displayed above the Venn diagram (**Screenshot 2f**). For genes mapped from GWAS catalog or user-uploaded SNP lists, this calculation is not performed as the statistic is not robust when each SNP could be mapped to multiple genes in LD. In addition, for user-uploaded SNP lists (which should contain only independent SNPs), Genoppi generates three Venn diagrams to show the overlap of significant proteins with all mapped genes, genes in single-gene loci, or genes in multi-gene loci, respectively; these diagrams can be selected using a drop-down menu (**Screenshot 3c**).

If the user uploads multiple SNP or gene lists, Venn diagrams and overlap statistics for each list would be generated separately for each list, and the individual list results can be accessed through clicking on the list name in a drop-down menu (**Screenshot 3c**). Note that when a bait protein has been indicated in the “Input bait” search box in the left panel, Genoppi would exclude the bait when calculating the numbers and statistics in the *Venn diagrams* section.

Gene set annotations

In the “Gene set annotations” module, Genoppi enables annotation of the proteomic data with gene sets from various databases, including HGNC gene groups, Gene Ontology^{15, 16} (GO) [geneontology.org] terms (molecular function, cellular component, and biological process), and MSigDB^{17, 18} [www.gsea-msigdb.org/gsea/msigdb/index.jsp] gene sets (H and C1-C7 collections), allowing the user to explore the diversity of protein functions in the proteomic results.

The user can annotate significant proteins in their volcano plot by selecting a collection of gene sets from a drop-down menu (**Screenshot 4a**). Proteins belonging to different gene sets are annotated using square markers of distinct colors; the marker size is scaled with the frequency of each gene set (i.e., number of proteins assigned to each set), providing quick visualization of overrepresentation trends in the data. The volcano plot can display up to 100 most recurrent gene sets at once; the user can further filter these top gene sets using the frequency slider (**Screenshot 4a**). Hovering over each marker in the resulting volcano plot would show the protein’s gene set annotations. Alternatively, the table below the volcano plot lists all the gene set annotations without the 100 gene sets limitation. Finally, the user can also query specific gene sets using a search box (**Screenshot 4b**), and the proteins belonging to the queried sets would be labeled with diamond markers in the volcano plot.



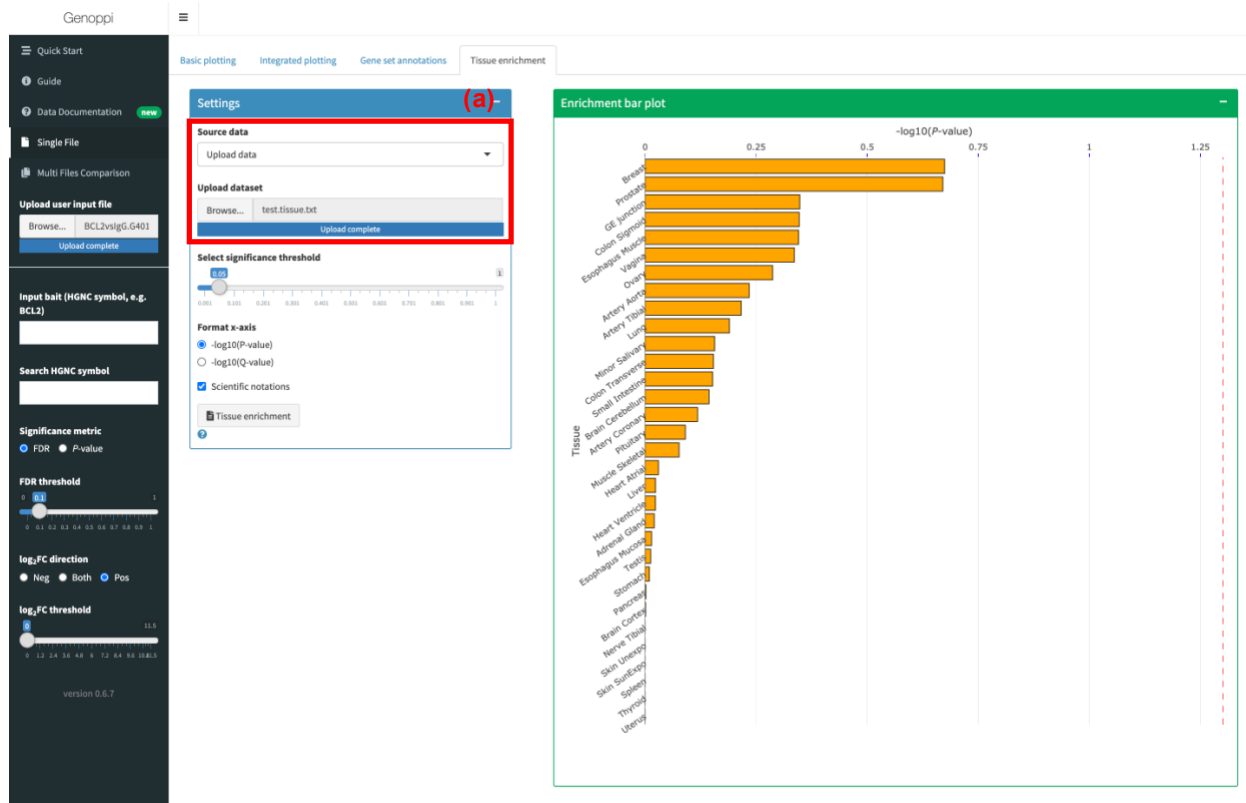
Screenshot 4. Gene set annotations interface showing significant proteins in proteomic data annotated with most recurrent HGNC gene groups.

Tissue enrichment

The “Tissue enrichment” module is an extension of the “GTEx or HPA” functionality described in the “Integrated plotting” module. Here, instead of integrating the proteomic data with a single tissue-specific gene list, Genoppi calculates the overlap enrichment of significant proteins in the proteomic data and tissue-specific genes across all tissues found in the “GTEx - RNA”, “GTEx - protein” or “HPA - RNA” dataset, which can be selected from a drop-down menu (**Screenshot 5a**). The hypergeometric enrichment results are displayed in a bar plot; the user can choose to display the $-\log_{10}$ *P*-value or $-\log_{10}$ *Q*-value (i.e., Benjamini-Hochberg FDR) on the x-axis, as well as selecting the significance threshold to highlight in the plot (**Screenshot 5b**). Hovering over each tissue’s bar in the plot would show the *P*-value, *Q*-value, and the list of significant proteins that are specifically expressed in the tissue.

In this module, the user also has the option of uploading sets of tissue- or cell-type-specific genes derived from their own expression data to perform the enrichment analysis (e.g., single-cell RNA-seq dataset that allows cell type deconvolution; **Screenshot 6a**). The data must be uploaded as a tab-delimited plain text file containing 3 columns: “tissue” (tissue or cell type name), “gene” (HGNC symbol), and “significant” (“TRUE” or “FALSE” indicating whether a gene is specifically expressed in the tissue or cell type). For example:

tissue	gene	significant
Celltype1	SHH	TRUE
Celltype1	UBC	TRUE
Celltype1	FOXP2	FALSE
Celltype2	RB1	TRUE
Celltype2	KRAS	FALSE



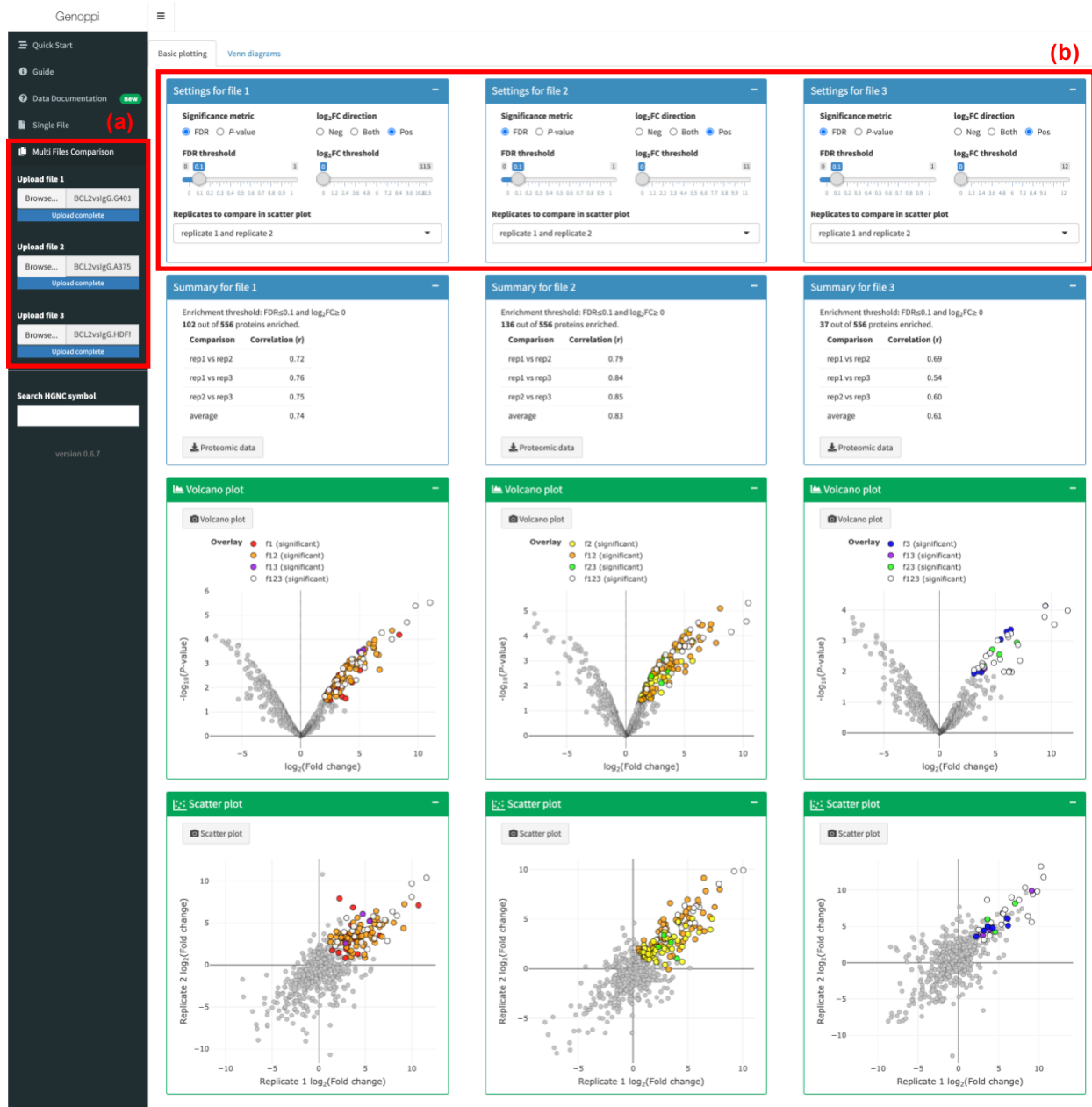
Screenshot 6. Tissue enrichment interface showing overlap enrichment of significant proteins in proteomic data and user-uploaded tissue- or cell-type-specific genes.

Multiple files comparison

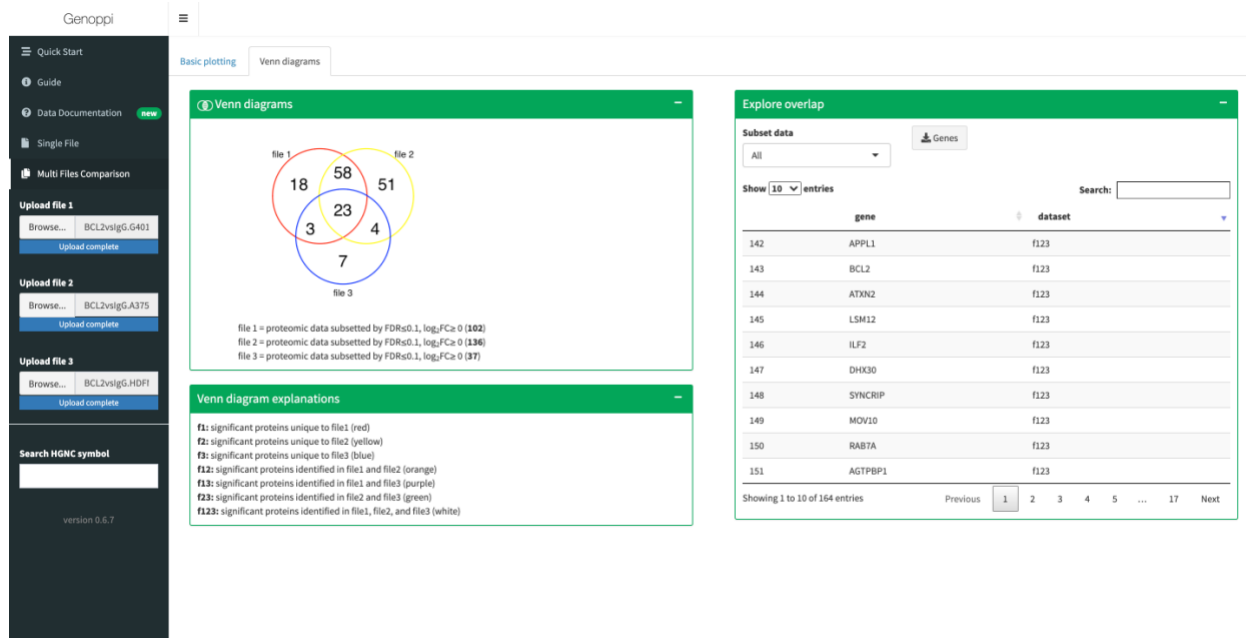
Besides performing analyses for a single proteomic dataset as described in the previous sections, Genoppi also allows comparison of multiple proteomic datasets at once. Using the “Multi Files Comparison” input option, the user can upload two to three proteomic datasets to perform comparative analyses (**Screenshot 7a**). In the “Basic plotting” module, Genoppi would generate side-by-side volcano and scatter plots to compare the multiple datasets. The threshold for defining significant proteins can be individually adjusted for each dataset (**Screenshot 7b**). In the resulting plots, each significant protein is color-coded based on the combination of dataset files that share this significant protein. The possible combination groups are:

- f1**: significant proteins unique to file1 (red)
- f2**: significant proteins unique to file2 (yellow)
- f3**: significant proteins unique to file3 (blue)
- f12**: significant proteins identified in file1 and file2 (orange)
- f13**: significant proteins identified in file1 and file3 (purple)
- f23**: significant proteins identified in file2 and file3 (green)
- f123**: significant proteins identified in file1, file2, and file3 (white)

Furthermore, the “Venn diagrams” module summarizes the number of proteins in each combination group in a Venn diagram and displays the identities (i.e., HGNC symbols) of these proteins in a table (**Screenshot 8**).



Screenshot 7. Basic plotting interface showing side-by-side volcano and replicate correlation scatter plots for multiple proteomic datasets.



Screenshot 8. Venn diagrams interface showing overlaps of significant proteins identified in multiple proteomic datasets.

Downloads

Individual plots and data files generated by Genoppi can be downloaded in their respective modules by clicking on the interactive download buttons. In general, plots are saved as PNG image files, while text files are saved in comma-separated CSV format. Download buttons are only active once the relevant data have been generated.

References

1. Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res* **45**, D619-D625 (2017).
2. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699 (2018).
3. Ritchie ME, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
4. Lage K, *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* **25**, 309-316 (2007).
5. Li T, *et al.* A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat Methods* **14**, 61-64 (2017).
6. Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**, 405 (2008).
7. Huttlin EL, *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425-440 (2015).
8. Huttlin EL, *et al.* Dual Proteome-scale Networks Reveal Cell-specific Remodeling of the Human Interactome. Preprint at <https://www.biorxiv.org/content/101101/20200119905109v1> (2020).
9. Buniello A, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
10. Karczewski KJ, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
11. Finucane HK, *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).
12. Jiang L, *et al.* A Quantitative Proteome Map of the Human Body. *Cell* **183**, 269-283 e219 (2020).

13. Uhlen M, *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
14. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
15. Ashburner M, *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
16. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338 (2019).
17. Subramanian A, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
18. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).