Welcome Guide

Table of contents

Getting started

Genoppi is an open-source software for performing quality control and analyzing quantitative proteomic data. Genoppi streamlines the integration of proteomic data with external datasets such as known protein-protein interactions in published literature, data from genetic studies, gene set annotations, or other user-defined inputs. This protocol provides documentation for using the interactive Genoppi web application, which is available at www.lagelab.org/genoppi. 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 enrichment 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 enrichment values for two or more experimental replicates. More specifically:

Column 1: protein identifiers as either HUGO Gene Nomenclature Committee[1] (HGNC) [www.genenames.org] approved symbols (with "gene" as column header), or UniProt[2] [www.uniprot.org] accession numbers (with "accession_number" as column header).

Columns 2, 3, (+ optional additional columns): $\log_2$ fold change (FC) values for ≥ 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

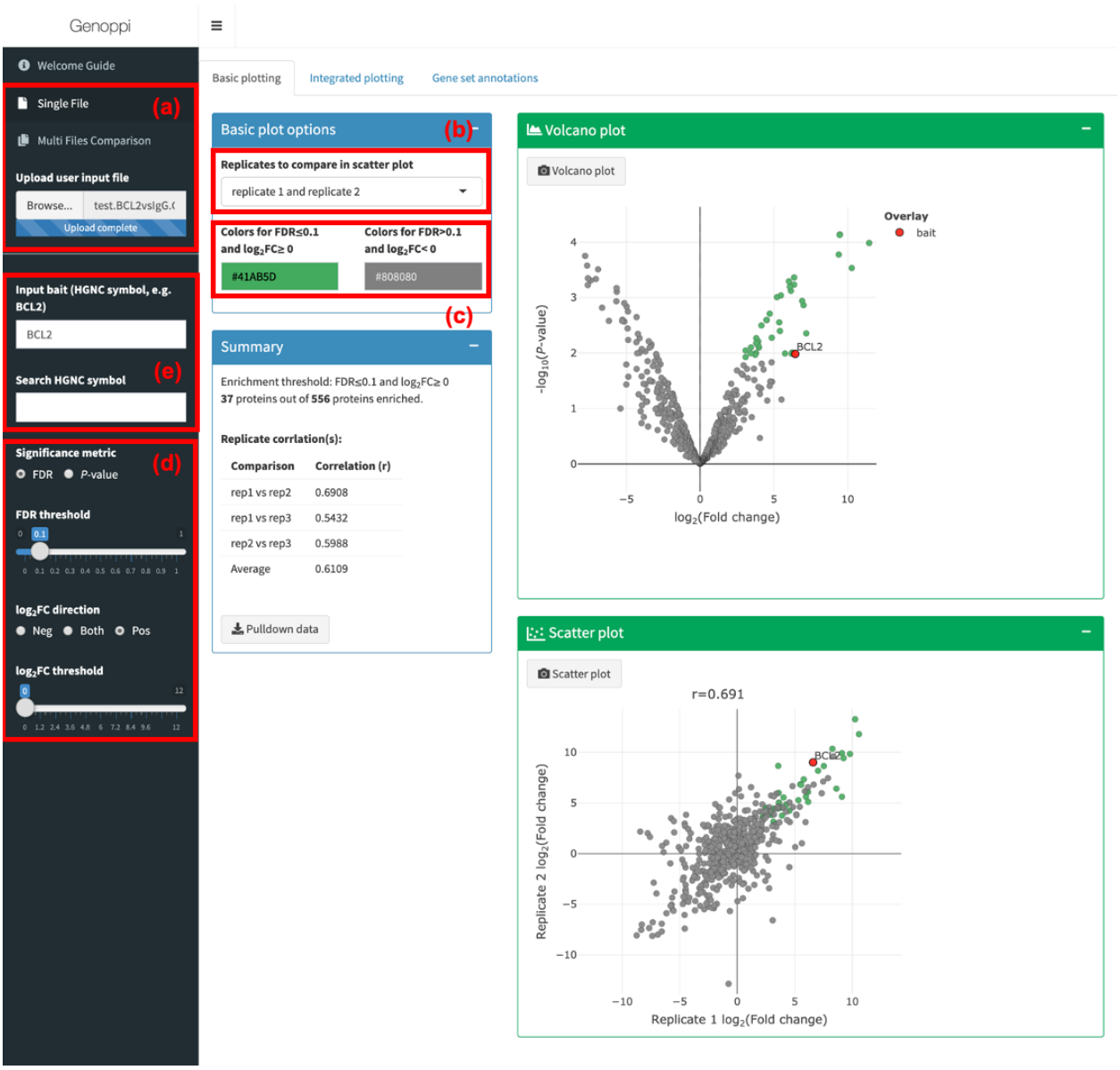
gene	logFC	pvalue	FDR
FOXP2	-0.521	1.64e-3	4.14e-3
RB1	0.334	0.0118	0.0189
SHH	0.092	0.211	0.242

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 of enrichment on the y-axis. If $\log_2$ FC values from ≥ two replicates are provided in the input file, a moderated *t*-test from the limma[3] R package is applied to calculate the average $\log_2$ FC, nominal *P*-value, and FDR (see **Online Methods**); 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, enriched 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 enriched proteins based on different FDR, *P*-value, and $\log_2$ FC cutoffs (**Screenshot 1d**). The "Summary" box shows the number of enriched 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 database[4], NHGRI-EBI GWAS catalog[5], gnomAD database[6], or user-uploaded SNP or gene lists.

InWeb_InBioMap

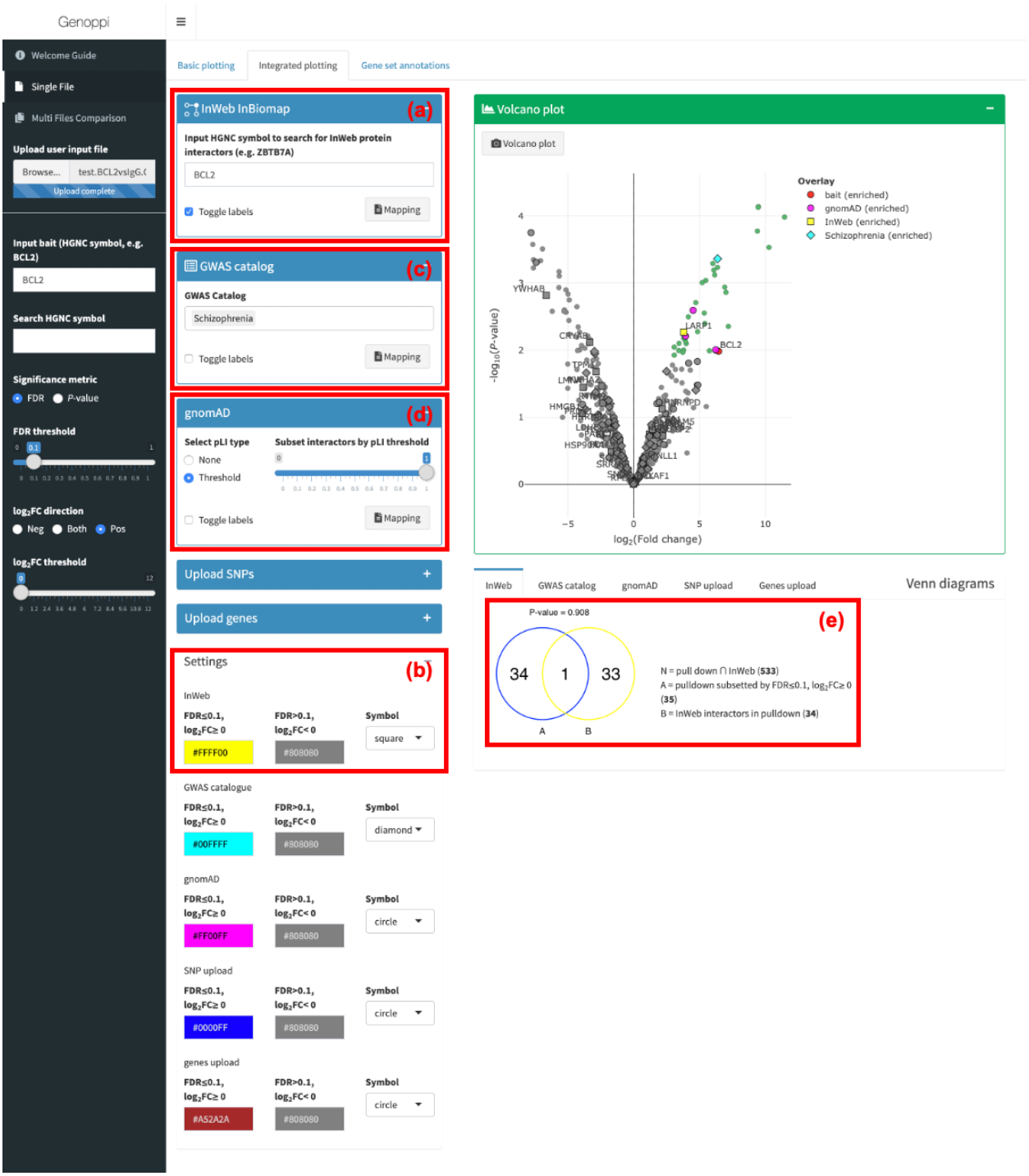
Genoppi can overlay the proteomic data with data from InWeb_InBioMap (InWeb) [www.intomics.com/inbio/map.html#downloads], which contains human protein-protein interactions compiled from > 40,000 published articles. This integration enables the user to easily distinguish new interactions from those already reported in the literature. The user can search for InWeb interactors of a specific protein to visualize their overlap with enriched proteins in the proteomic data in an overlaid volcano plot (**Screenshot 2a**). In this plot, InWeb 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 Project[7] [www.internationalgenome.org], using pre-calculated pairwise linkage disequilibrium (LD) measures between SNPs to identify all genes in LD regions (see **Online Methods**). 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 also 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**).



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

Upload SNPs or genes

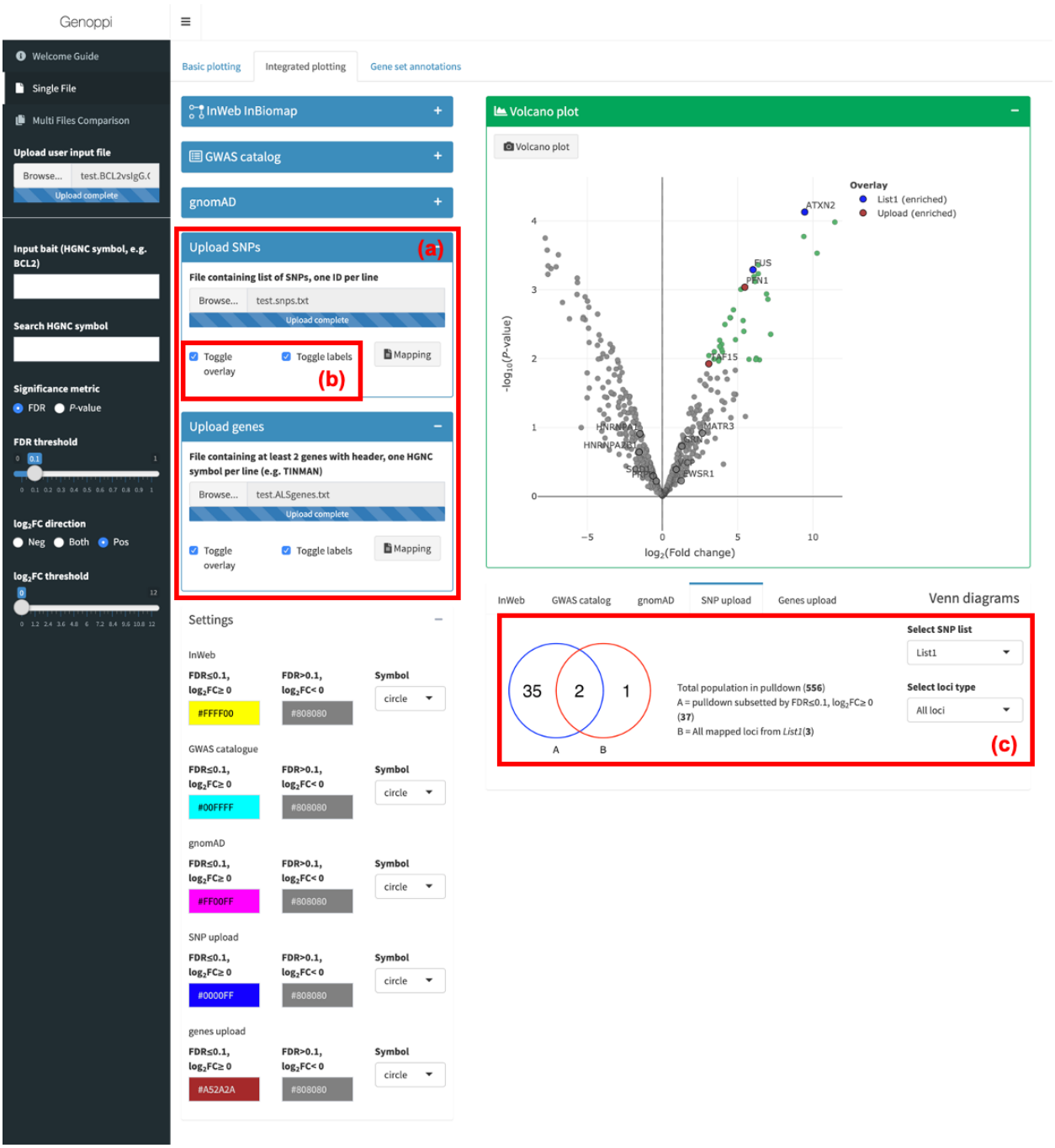
Besides incorporating the public datasets described above, the user may also upload one or more custom SNP or gene lists (e.g. disease-causing genes curated from literature review or genes implicated by gene-based burden testing) to assess their overlaps with the proteomic data (**Screenshot 3a**). As 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, and 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 text labels for the proteins using the "Toggle labels" option for each data type (**Screenshot 3b**).



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

Venn diagrams

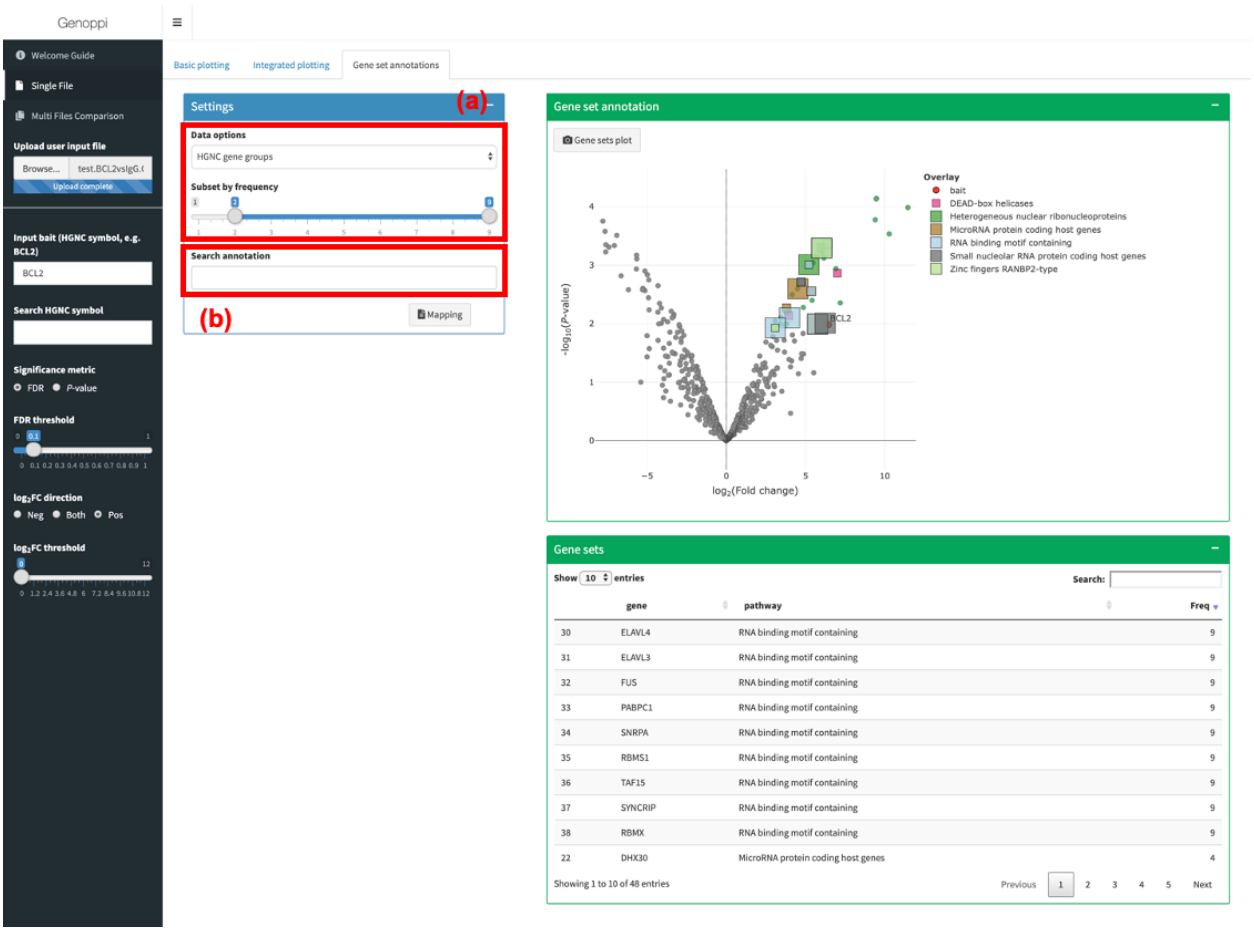
In the “Integrated plotting” module, Genoppi also summarizes the overlaps between the enriched proteins in the proteomic data and the various data types described above using Venn diagrams. When showing overlap with InWeb_InBioMap, gnomAD constrained genes, or user-uploaded gene lists, Genoppi also assesses the overlap enrichment by calculating a hypergeometric *P*-value, which is displayed above the Venn diagram (**Screenshot 2e**; see **Online Methods**). 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 enriched 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[8,9] (GO) [geneontology.org] terms (molecular function, cellular component, and biological process), and MSigDB[10,11] [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 enriched 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 enriched proteins in proteomic data annotated with most recurrent HGNC gene groups.

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. In the “Basic plotting”, “Integrated plotting”, and “Gene set annotations” modules, Genoppi would generate side-by-side volcano, scatter, and/or Venn diagram plots to compare the multiple datasets and offer customization features similar to that for the “Single File” input.

Protein comparison

In the “Protein comparison” module, the user can visualize the overlaps between enriched proteins identified in different datasets. The significance threshold for defining enriched proteins can be individually adjusted for each dataset before generating the comparison results. In the resulting volcano and scatter plots, each enriched protein is color-coded based on the combination of datasets that share this enriched protein. The number of proteins in each combination group is also summarized in a Venn diagram and the identities (i.e. HGNC symbols) of these proteins are displayed in a table.

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., et al. 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, M.E., et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
4. Li, T., et al. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat Methods* **14**, 61-64 (2017).
5. 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).
6. Karczewski, K.J., et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Pre-print in *bioRxiv* https://doi.org/10.1101/531210 (2020).
7. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
8. Ashburner, M., et al. Gene ontology: tool for the unification of biology. *Nat Genet* **25**, 25-29 (2000).
9. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338 (2019).
10. 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).
11. Liberzon, A., et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).