Genoppi: A web application for interactive integration of experimental proteomics results with genetic datasets.

April Kim[1,2], Edyta Malolepsza[1,2], Justin Lim[1,3], Kasper Lage[1,2]

Affiliations:
1. Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
2. Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts, USA.
3. Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Abstract
Summary: Integrating protein-protein interaction experiments and genetic datasets can lead to new insight into the cellular processes implicated in diseases, but this integration is technically challenging. Here, we describe Genoppi, a web application that integrates quantitative interaction proteomics data and results from genome-wide association studies or exome sequencing projects, to highlight biological relationships that might otherwise be difficult to discern. Genoppi also facilitates data sharing in cross-disciplinary collaborations. Written in Python, Bash script and R using Shiny framework, Genoppi is a user-friendly framework for integrative genetic and proteomic analyses that can be easily deployed across Mac OS and Linux distributions.
Availability: Genoppi is open source and available at https://github.com/lagelab/Genoppi
Contact: aprilkim@broadinstitute.org and lage.kasper@mgh.harvard.edu

Introduction
With recent advances in stem cell technologies and quantitative proteomics methods, it is now possible to experimentally interrogate the physical interactions of proteins in a tissue- or cell-type-specific manner at scale. In parallel, the ongoing genomic revolution has enabled the identification of common variant loci significantly associated with diseases through genome-wide association studies (GWAS). Exome sequencing technologies have also led to the identification of specific genes or protein-coding mutations that are linked to particular diseases. Since it has been shown that genes implicated in a common complex traits often interact at the level of proteins (Rossin *et al.*, 2011; Lundby *et al.*, 2014; Lage, 2014) it is desirable to be able to integrate the results of quantitative proteomics experiments and user-defined genetic datasets in an interactive workflow that can be shared with collaborators.

Towards that aim, we developed an open-source application Genoppi. It provides (i) quality control (QC) of mass spectrometry-based quantitative interaction proteomics datasets and (ii) on-the-fly integrative analyses of the proteomic data and user-defined GWA or exome sequencing studies. By allowing users to archive the code – and the results of analyses and visualizations – seasoned R users and non-programmers alike can generate, store and share R session results.

Features
In Genoppi, users can i) upload experimental quantitative interaction proteomics data from immunoprecipitations (IP) followed by tandem mass spectrometry (MS/MS) or ii) the MS/MS results from a full proteome analysis. Upon loading a file, the software

plots interactive graphical representations of the data as volcano and scatter plots (**Fig. 1**). As part of the QC workflow, known protein interaction partners from InWeb (Rossin *et al.*, 2011; Lage, 2014; Li *et al.*, 2016) can be identified. InWeb contains known protein-protein interactions from >40,000 articles and overlaying this information with the experimental data serves as a QC and enables users to easily distinguish new interactions from those already reported in the literature. Analysis thresholds and visualization features can be defined and adjusted by users during the analyses. The changes are reflected immediately in the interactive plots.

Genoppi is designed to provide integration of proteomics results with genetic data. Users can (i) overlay proteomics and GWAS data by simply inputting SNPs from genetic studies relevant to the proteomics experiment, (ii) map UniProt identifiers (The Uniprot Consortium, 2015) to HUGO Gene Nomenclature Committee (HGNC) symbols (Gray *et al.*, 2015), (iii) execute protein family (PFE) analyses, (iv) identify proteins encoded by genes that are intolerant to loss-of-function mutations as determined by the Exome Aggregation Consortium (ExAC) data (Lek *et al.*, 2016), and (v) upload gene names of interest to the users; corresponding proteins are immediately highlighted in the applications plots (e.g., a set of interesting genes from the Genotype Tissue Expression [GTEx] project).

When different datasets are integrated with the proteomics data, overlaps are statistically tested and the results are made available to the users. As such, the overlap of interaction partners of their protein of interest and genes genetically linked to a particular disease is determined.

A significant barrier to using the results from GWAS in other fields is that SNP-to-gene mapping is nontrivial. For this reason, the application automatically maps SNPs to genes using haplotype information from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) and highlights those genes in the proteomic data. If multiple protein-coding genes are present in genetic locus defined by a SNP of interest, all genes are mapped as candidate genes for the associated SNP. Together, this allows users to identify interaction partners of bait that are encoded in loci genetically associated with phenotype relevant to the bait.

PFE analyses can be applied to either single bait versus control experiment or comparison between different experimental conditions (e.g., with and without drug treatment on various mutated cells). The latter format allows users to identify protein families defined by proteins enriched in the conditions of choice that can serve as an additional QC or to guide hypothesis generation or follow up experiments. PFE is performed using protein families dataset of human protein-coding genes curated by HGNC.

Users can download the generated data in text format. Furthermore, publication-quality figures and interactive plots can be downloaded as a report in the HTML format. By allowing users to archive the code, the application can be easily extended and modified to suit custom needs by the user.

Implementation

Genoppi is written in code based on R (R Core Team, 2016) standard packages using Shiny framework (Chang *et al.*, 2016). Automated assignment of UniProt protein accession numbers to HGNC symbols with detailed annotation is written in Python.

Analyses for GWAS results, PFE and InWeb overlap are written in Bash script to reduce programming runtime. Median normalization of user dataset and application of single sample moderated t-test is performed using Bioconductor limma package (Ritchie *et al.*, 2015). Pairwise measure of LD was calculated using VCFtools (Danecek *et al.*, 2011) on VCF files containing genotype data from the final phase (phase 3) of 1000 Genomes Project. R graphing library plotly (Sievert *et al.*, 2016) was used to create interactive graphs.

Acknowledgments
EM would like to thank Zuzana Tothova, Josephine Kahn, Siddhartha Jaiswal, and Srinivas Viswanathan for discussion on data analysis and visualization, and Monica Schenone, Benjamin Tanenbaum, Christina Hartigan, Karsten Krug, and DR Mani for explaining the structure of proteomics outputs.

Author Contributions
Developed algorithms: AK, EM, JL with supervision from KL
Developed Genoppi web application: AK
Wrote paper: AK, EM, KL
Initiated and led project: KL

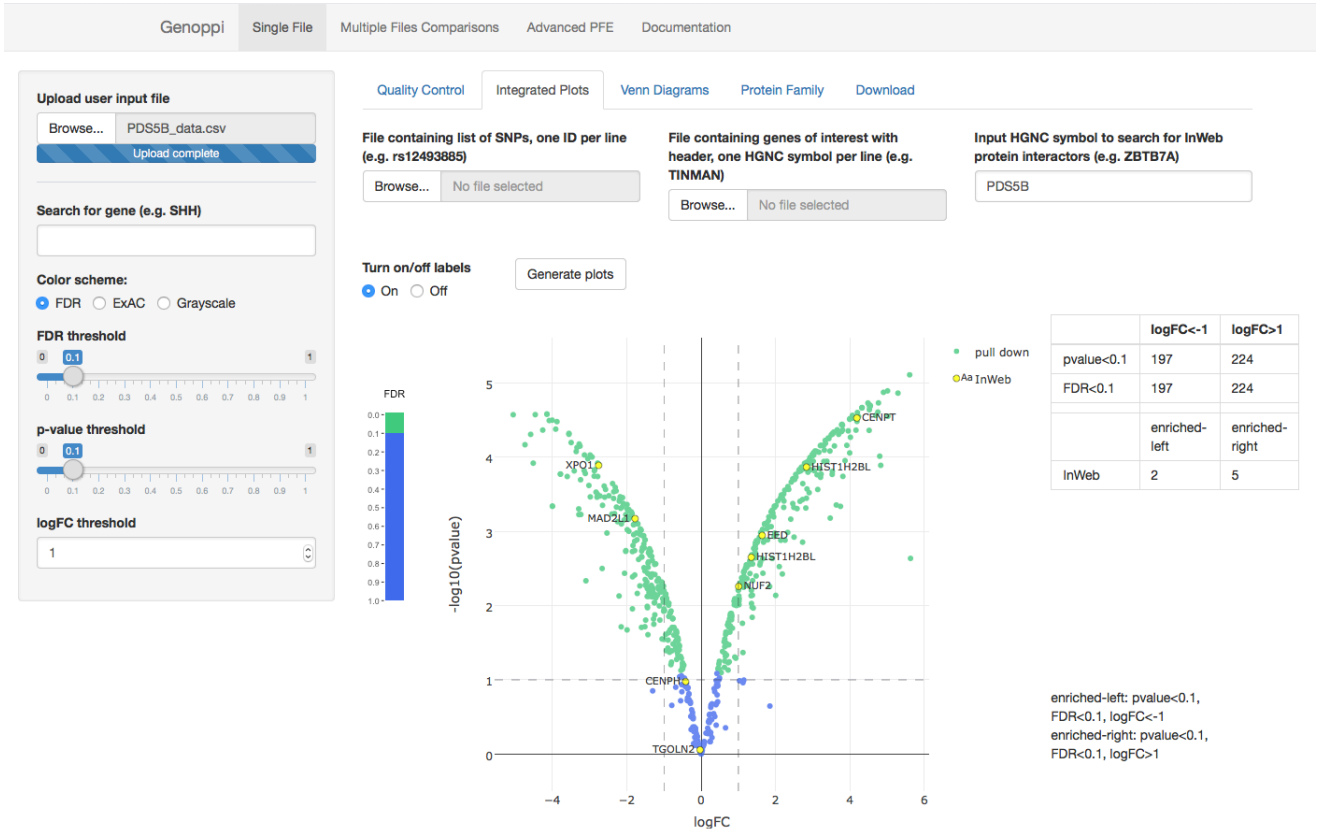

Fig. 1. | Example of Genoppi interface. In this example we are illustrating a plot of a quantitative interaction proteomics experiment. Results are shown as a volcano plot where $\log_2$ fold change (logFC) values of the bait versus control are plotted on the x-axis and the negative $\log_{10}$ transformed P values of that enrichment are plotted on the y-axis.

Interaction partners that are enriched in the experiment with a false discovery rate (FDR) < 0.1 are green, and interaction partners with an FDR >= 0.1 are blue. Green proteins with a positive logFC are significantly enriched in bait condition and green proteins with a negative logFC are enriched in the control condition. Known interaction partners of the bait protein (as determined by overlaying InWeb data) are highlighted in yellow enabling users to QC the data and to distinguish known and new interaction partners. When the input file includes two replicates, a scatter plot of the correlation between these datasets is shown as well. Different color schemes, analogous to the highlighted yellow proteins, are available to the users to represent, for example, the FDR values or the intolerance to loss-of-function variation (pLI) of a given gene in a dataset.

References

Chang,W. *et al.* (2016) shiny: Web Application Framework for R.

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Gray,K.A. *et al.* (2015) Genenames.org : the HGNC resources in 2015. *Nucleic Acids Res.*, **43**, 1079–1085.

Lage,K. (2014) Protein-protein interactions and genetic diseases: The interactome. *BBA - Mol. Basis Dis.*, **1842**, 1971–1980.

Lek,M. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 30338.

Li,T. *et al.* (2016) A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods*, **14**, 64535.

Lundby,A. *et al.* (2014) Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics. *Nat. Methods*, **11**, 868–874.

R Core Team (2016) R: A Language and Environment for Statistical Computing.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**.

Rossin,E.J. *et al.* (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.*, **7**.

Sievert,C. *et al.* (2016) plotly: Create Interactive Web Graphics via 'plotly.js'.

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

The Uniprot Consortium (2015) UniProt : a hub for protein information. *Nucleic Acids*

Res., **43**, 204–212.