bioRxiv preprint doi: https://doi.org/10.1101/115576; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Gene expression

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] and Kasper Lage[1,2,*]

[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA, [2]Department of Surgery, Massachusetts General Hospital, Boston, MA 02114, USA and [3]Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*To whom correspondence should be addressed.

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 present 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. Written in R, Python and Bash script, Genoppi is a user-friendly framework 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

1 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 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 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 interaction proteomics datasets and (ii) on-the-fly integrative analyses of the proteomic data and user-defined GWAS or exome sequencing studies. By allowing users to store the results of analyses and visualizations, Genoppi facilitates data sharing in cross-disciplinary collaborations.

2 Features

In Genoppi, users can upload experimental quantitative interaction proteomics data from immunoprecipitations (IP) or full proteome analysis followed by tandem mass spectrometry (MS/MS). 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 human 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 enables users to easily distinguish new interactions from those already reported in the literature.

Further, Genoppi allows users to (i) overlay GWAS data by inputting single nucleotide polymorphisms (SNP), (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 intolerant to loss-of-function mutations as determined by the Exome Aggregation Consortium (ExAC) data (Lek et al., 2016), and (v) highlight specific proteins corresponding to users' uploaded genes of interest (e.g., a set of genes from the Genotype Tissue Expression [GTEx] project).

When different datasets are integrated with proteomics results, overlaps are statistically tested and visualized. For example, the overlap between interaction partners of bait identified in MS/MS experiment and genes linked to a particular disease is determined.

A significant barrier to using the results from GWAS is that SNP-to-gene mapping is nontrivial. For this reason, the application automatically maps SNPs to genes using haplotype information from 1KGP (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.

bioRxiv preprint doi: https://doi.org/10.1101/115576; this version posted March 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

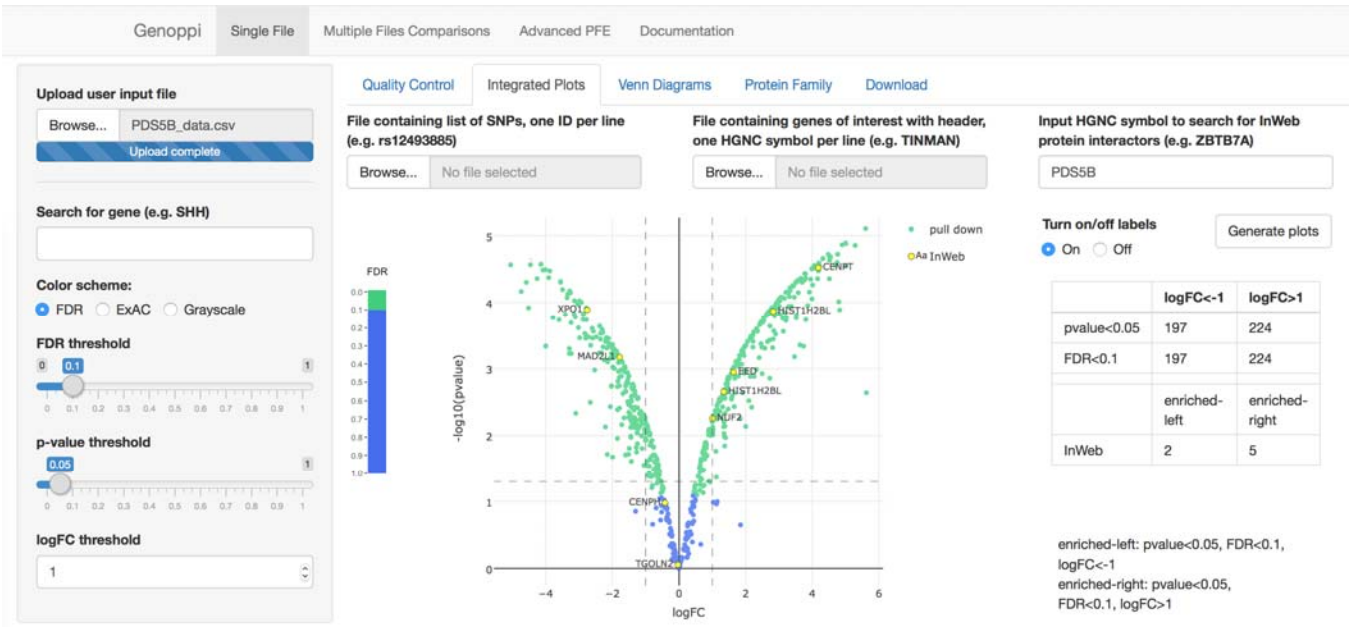

Fig.1 | Genoppi interface. Figure shows a volcano plot of mock proteomics results. Log_2 fold change (logFC) values of identified proteins are plotted on the x-axis and the negative log_{10} transformed P values of that enrichment on the y-axis. Proteins with a false discovery rate (FDR) < 0.1 are green, and those with FDR >= 0.1 are blue. Known interaction partners of the bait determined by InWeb are highlighted in yellow. Different color schemes are available: grayscale and the intolerance to loss-of-function variation (pLI) of a given gene.

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 cells with various mutations). The latter format allows users to identify human 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.

Users can download generated data in text format and interactive plots as HTML reports. Genoppi can be archived and easily modified to suit users' custom needs.

3 Implementation

Genoppi is written using R standard packages (R Core Team, 2016) using Shiny framework (Chang et al., 2016) and Bash script. Automated assignment of human UniProt accession numbers to HGNC symbols is written in Python. Median normalization of user dataset and 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 phase 3 of 1KGP. R graphing library plotly (Sievert et al., 2016) was used to create interactive graphs.

Acknowledgements

EM thanks Zuzana Tothova, Josephine Kahn, Siddhartha Jaiswal, and Srinivas Viswanathan for discussion on data analysis, and the Proteomics Platform at the Broad Institute for explaining output structures.

Author Contribution

AK, EM, JL developed algorithms with supervision from KL. AK developed Genoppi application. AK, EM, KL wrote paper. KL initiated and led project.

Funding

AK, EM and KL are supported by a grant from the Stanley Center at the Broad Institute, a Broadnext10 grant from the Broad Institute, 1R01MH109903 from NIMH and a grant from the Lundbeck Foundation.

Conflict of Interest: none declared.

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.