(w)HOL(e)ISTIC gene ontology and pathway analysis of data using open source web tools

Damarius S. Fleming
USDA-ARS National Animal Disease Center
Laura Miller (laura.miller@ars.usda.gov)
USDA Agricultural Research Service https://orcid.org/0000-0002-8946-9416

Research note

Keywords: Data analysis, next generation sequencing, omics approach, gene ontology, pathway analysis, annotation, visualization

Posted Date: January 8th, 2020

DOI: https://doi.org/10.21203/rs.2.20371/v1

Abstract

Objective Downstream analysis of next generation sequencing (NGS) experiments provides researchers a means of deciphering their results. These downstream analyses elucidate clusters of genes or networks of biological interest under the conditions being studied. One convention for examining gene interactions is to conduct downstream investigations based on gene ontology (GO), pathway, and network analyses of statistically significant genes of interest. Unfortunately, the software available for these types of analyses is expensive, not species specific, and subject to gaps in annotation. These difficulties can cause studies to omit this downstream step, limiting the utility of the data. In order to facilitate pathway and network analyses of candidate gene lists from NGS studies, a workflow was constructed based on the use of open-sourced freely available software and genomic databases termed the “(w)HOL(e)ISTIC GO enrichment” approach.

Results Overlap of multiple open-source software was used to annotate, analyze, and visualize biological networks. It is a 3-stage process in which stage 1 (Annotation) is the generation of alias identifiers. Stage 2 (Analysis) is a two-part process generating ontologies and networks with statistical inferences. Stage 2 relies on information from databases such as Reactome, KEGG, and InterPro. Stage 3 (Visualization) allows for figure creation.

Introduction

The data used for the example in this article is from a differential gene expression analysis of monocyte derived cells infected with either the highly pathogenic or low pathogenic Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) designated as HP-PRRSV and VR-2332 respectively. The candidate genes generated from this analysis were based on comparisons of the HP-PRRSV or VR-2332 infected cells that had been stimulated with one of five immune function cytokines. The samples were derived from blood monocytes from at least five outbred piglets that have similar genetic background. Total RNA was extracted from the pooled cells of four replicates, and used to construct sequencing libraries for the transcriptomic analysis. Sequence libraries were used to generate 100 bp paired-end reads using the Illumina® 2500 HiSEq. Procedures for read processing, mapping, and alignment have been previously described [1]. Normalization of gene counts was carried out using the rlog transformation function and calculations of differential gene expression for each comparison were based upon the average log expression and the regularized $\log_2$ fold change (rLogFC) using DESeq2 [2]. Genes were annotated using the Ensembl database [3] and g:Profiler [4]. The fold changes reported are the difference in expression values for each gene based on the comparison and a threshold of equal to or greater than +/- 2 rLogFC. Ensembl identifier (ID), rLogFC, and normalized count data for each treatment group and replicate were compared. These gene lists were used as the query lists for network prediction.

Main Text

In order to facilitate pathway and network analyses of candidate gene lists from high-throughput studies, an in-house method referred to as the (w)HOL(e)ISTIC GO and pathway analysis workflow was constructed based on the use of open source freely available software and genomic databases. The example data is from “Comparative analysis of signature genes in PRRSV-infected porcine monocyte-derived cells to different stimuli” published by Miller et al., 2017 [5]. The (w)HOL(e)ISTIC GO enrichment approach is useable with multiple species, and allows use of species specific annotations when available.

A list of differentially expressed genes with a rLogFC ≥ 2 or ≥ -2 was supplied from each comparison to the (w)HOL(e)ISTIC GO and pathway analysis work flow. The workflow can be used with data from any analysis method. The workflow takes as input a user-defined gene list. Application of the workflow to a gene list consists of several steps: annotation, analysis, and lastly visualization.

The first step is the annotation of the query list from the gene expression or other experiment (Fig. 1). This step calls for a researcher to take their output list of genes from their analysis and generate a list of gene name aliases for each gene of interest. This step is carried out in the first step to assuage issues with the lack of robust gene and protein annotation in many livestock and non-model species. This step is carried out by starting with the query list in your preferred nomenclature [i.e. Ensembl, Human Genome Organisation Gene Nomenclature Committee (HGNC)]. Next copy and paste the query list into any or all of the following web-based software: Ensembl Biomart [3], g:Profiler (g:convert tool) [4], HGNC [HGNC Comparison of Orthology Predictions tool (HCOP)] [6], or other. These software programs will search various biological databases to find any additional gene/protein names associated with the query list of results. The order in which the software is used doesn’t matter because all terms will be aggregated at the end prior to moving on to step 2. The main goal of this step is to provide the researcher with multiple gene names to better ensure that the results of interest can examined regardless of up-to-date or out-of-date gene curation. The gene aliases also help to examine syntenic regions and orthologues in case a researcher needs to rely on sequence homology as part of their analyses.

The second step will use the query list that has been populated with the gene aliases to carry out a two-part analysis step. Part one of this step uses the expanded query list to perform a gene ontology (GO) analysis using several web based programs. The use of multiple GO analysis software programs is done to allow a researcher the ability to compare consensus in any terms, pathways/networks, or statistical models shared between the programs. Examination of consensuses pathways and functions is done to afford the researcher repeatability, which in turn leads to greater confidence in an experiments results. The GO analysis portion of step 2 is carried out using the web-based programs: GOtermFinder [7], PantherDB [8], DAVID6.8 [9–11], and the g:Profiler (g:GoSt Tool) [4].

Part two of this step employs the programs STITCH [12]and STRING [13] to predict any possible gene-gene or gene-chemical interactions, in order to help uncover possible gene networks or interactions related to the results. This portion of step 2 is also based upon the expanded query list from step one and as output provides a visual representation of nodes (genes) and edges (the interaction) that exist as

networks within the data. The software does not differentiate between the expression levels of the genes in the list, but does draw in information from various databases to predict the effect (positive, negative, unknown) a gene is expected to exert on another in the network. In the predicted network outputs, a red line connecting nodes represents inhibition, green lines represent activation; dark blue lines represent binding; purple lines represent catalysis; yellow lines represent transcriptional regulation, light blue represents phenotype, and black lines are representative of reaction.

The last step of the method is the visualization step, which can be carried over from the pathway/network analysis portion of step 2. This is because the software STITCH [12] and STRING [13] produces a visual output that can be manipulated and downloaded as an image file.

Limitations

Annotation

- Some ontology programs will eliminate duplicate genes and aliases, others won't, be sure to check input type accepted, compare starting list with output, to avoid duplicate counts.
- Multiple databases and IDs might be needed. Ensembl Biomart will only allow 3 external sources for name annotation, g:Profiler will only do one at a time.
- Check versions and updates of software. The most recent may not always be the best as it may be a beta version.
- Order of which software is used doesn't matter as you will aggregate all terms and aliases at the end prior to step 2.

Analysis

- 2 stage step (ontology and pathway)
- Be mindful of usefulness of statistics vs. network or pathway.
- In the PantherDB GO analysis only one dataset will be output at a time. Check the results that the input list numbers match.
- Be aware of most current updates (reference genome and annotation).
- All programs give a text and/or html option for output.

Visualization

- STRING/STITCH images can be manipulated to give the best or additional views.
- STRING uses a spring model to generate the network images. Nodes are modelled as masses and edges as springs; the final position of the nodes in the image is computed by minimizing the 'energy' of the system.
- Be aware that if you want to use prediction software to make statistical determinations- use only your query. The best option may be to query with and without shell predictors and only use predictors

as a marker of potential downstream/upstream effects based upon changes in the query genes.

List Of Abbreviations

NGS: next generation sequencing

GO: gene ontology

KEGG: Kyoto Encyclopedia of Genes and Genomes

PRRSV: porcine reproductive and respiratory syndrome virus

HP-PRRSV: highly pathogenic porcine reproductive and respiratory syndrome virus

VR-2332: prototypical North American porcine reproductive and respiratory syndrome virus isolate

RNA: ribonucleic acid

rLogFC: regularized log2 fold change

ID: identifier

HGNC: Human Genome Organisation Gene Nomenclature Committee

HCOP: Human Genome Organisation Gene Nomenclature Committee Comparison of Orthology Predictions

STITCH: search tool for interactions of chemicals

STRING: search tool for the retrieval of interacting genes/proteins

Declarations

Ethics approval and consent to participate

The animal use protocol was reviewed and approved by the Institutional Animal Care and Use Committee (IACUC) of the National Animal Disease Center-USDA-Agricultural Research Service. Experiments involving animals and viruses were approved by the Kansas State University Institutional Animal Care and Use and Biosafety Committees. Consent to participate was not applicable to this study.

Availability of data and material

Within the article Miller LC, Fleming DS, Li X, Bayles DO, Blecha F, et al. (2017) Comparative analysis of signature genes in PRRSV-infected porcine monocyte-derived cells to different stimuli. PLOS ONE 12(7):

e0181256. https://doi.org/10.1371/journal.pone.0181256

Direct URL to data: https://doi.org/10.1371/journal.pone.0181256.s001

Funding

This work was mainly supported by the USDA NIFA AFRI 2013-67015-21236, and in part by USDA NIFA AFRI 2015-67015-23216.

Acknowledgments

We would like to thank Barbara Lutjemeier, Qinfang Liu, and Sarah Anderson for their excellent technical support.

Consent for publication

Not applicable.

Competing interests

The author(s) declare(s) that they have no competing interests.

Author's contributions

DSF and LCM designed the study; LCM collected the data; DSF conducted data analysis and interpretation. DSF and LCM interpreted results, wrote, revised the initial and final manuscript. All authors read and approved the final manuscript.

References

1. Miller LC, Fleming D, Arbogast A, Bayles DO, Guo B, Lager KM, et al. Analysis of the swine tracheobronchial lymph node transcriptomic response to infection with a Chinese highly pathogenic strain of porcine reproductive and respiratory syndrome virus. BMC veterinary research. 2012;8:208.
2. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

3. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44(D1):D710-6.
4. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016;44(W1):W83-9.
5. Miller LC, Fleming DS, Li X, Bayles DO, Blecha F, Sang Y. Comparative analysis of signature genes in PRRSV-infected porcine monocyte-derived cells to different stimuli. PLoS One. 2017;12(7):e0181256.
6. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. genenames.org: the HGNC resources in 2011. Nucleic acids research. 2011;39(Database issue):D514-9.
7. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 2004;20(18):3710-5.
8. Mi H, Thomas P. PANTHER Pathway: An Ontology-Based Pathway Database Coupled with Data Analysis Tools. In: Nikolsky Y, Bryant J, editors. Protein Networks and Pathway Analysis. Totowa, NJ: Humana Press; 2009. p. 123-40.
9. Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome biology. 2007;8(9):R183.
10. Huang da W, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic acids research. 2007;35(Web Server issue):W169-75.
11. Sherman BT, Huang da W, Tan Q, Guo Y, Bour S, Liu D, et al. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics. 2007;8:426.
12. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, von Mering C, Jensen LJ, et al. STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res. 2014;42(Database issue):D401-7.
13. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016;44(D1):D380-4.
14. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. (1362-4962 (Electronic)).

Figures

Start	Ensemble Biomart					gProfiler	
Experimental list	Associated Gene Name	HGNC symbol	EntrezGene ID	WikiGene Name	Refseq_mRNA	EntrezGene (name)	Description
ENSSSCG00000002471	ISG12(A)		100153902	ISG12(A)	NM_001198921.1	ISG12(A)	interferon, alpha-inducible protein 27 [Source:HGNC Symbol;Acc:HGNC:5397]
ENSSSCG00000008648	RSAD2		396752	RSAD2	NM_213817.1	RSAD2	radical S-adenosyl methionine domain containing 2 [Source:HGNC Symbol;Acc:HGNC:30908]
ENSSSCG00000008977	CXCL10		494019	CXCL10	NM_001008691.1	CXCL10	C-X-C motif chemokine 10 precursor [Source:RefSeq peptide;Acc:NP_001008691]
ENSSSCG00000009004	SFRP2		100516027	SFRP2	NM_001244395.1	SFRP2	secreted frizzled-related protein 2 [Source:HGNC Symbol;Acc:HGNC:10777]
ENSSSCG00000009361	POSTN	POSTN	100152401	POSTN	NM_001206347.1	POSTN	periostin [Source:HGNC Symbol;Acc:HGNC:16953]
ENSSSCG00000010451	IFIT2	IFIT2	100155467	IFIT2		IFIT2	interferon induced protein with tetratricopeptide repeats 2 [Source:HGNC Symbol;Acc:HGNC:5409]
ENSSSCG00000010452	IFIT3		100154248	IFIT3	NM_001204395.1	IFIT3	interferon induced protein with tetratricopeptide repeats 3 [Source:HGNC Symbol;Acc:HGNC:5411]
ENSSSCG00000010453	IFIT1		100153038	IFIT1	NM_001244363.1	IFIT1	interferon induced protein with tetratricopeptide repeats 1 [Source:HGNC Symbol;Acc:HGNC:5407]
ENSSSCG00000012077	MX1	MX1	397128	MX1	NM_214061.1	MX1	MX dynamin like GTPase 1 [Source:HGNC Symbol;Acc:HGNC:7532]
ENSSSCG00000015326	COL1A2		100626716	COL1A2	NM_001243655.1	COL1A2	N/A

		HUGO: HCOP Tool						
Primary species	Ortholog species	Primary Symbol	Ortholog symbol	Ortholog species DBID	Primary Ensembl ID	Ortholog Ensembl ID	Primary NCBI Gene ID	Ortholog NCBI Gene ID
Pig	Human	ISG12(A)	IFI27	HGNC:5397	ENSSSCG00000002471	ENSG00000165949	100153902	3429
Pig	Human	ISG12(A)	IFI27L1	HGNC:19754	ENSSSCG00000002471	ENSG00000165948	100153902	122509
Pig	Human	IFIT3	IFIT3	HGNC:5411	ENSSSCG00000010452	ENSG00000119917	100154248	3437
Pig	Human	IFIT3	IFIT2	HGNC:5409	ENSSSCG00000010452	ENSG00000119922	100154248	3433
Pig	Human	LOC100518083	HERC5	HGNC:24368	ENSSSCG00000030548	ENSG00000138646	100518083	51191
Pig	Human	LOC100518083	HERC6	HGNC:26072	ENSSSCG00000030548	ENSG00000138642	100518083	55008
Pig	Human	MX1	MX1	HGNC:7532	ENSSSCG00000012077	ENSG00000157601	397128	4599
Pig	Human	MX1	MX2	HGNC:7533	ENSSSCG00000012077	ENSG00000183486	397128	4600
Pig	Human	COL1A2	COL1A2	HGNC:2198	ENSSSCG00000015326	ENSG00000164692	100626716	1278
Pig	Human	SFRP2	SFRP2	HGNC:10777	ENSSSCG00000009004	ENSG00000145423	100516027	6423
Pig	Human	CXCL10	CXCL10	HGNC:10637	ENSSSCG00000008977	ENSG00000169245	494019	3627
Pig	Human	IFIT2	IFIT2	HGNC:5409	ENSSSCG00000010451	ENSG00000119922	100155467	3433
Pig	Human	IFIT1	IFIT1	HGNC:5407	ENSSSCG00000010453	ENSG00000185745	100153038	3434
Pig	Human	IFIT1	IFIT1B	HGNC:23442	ENSSSCG00000010453	ENSG00000204010	100153038	439996
Pig	Human	RSAD2	RSAD2	HGNC:30908	ENSSSCG00000008648	ENSG00000134321	396752	91543
Pig	Human	POSTN	POSTN	HGNC:16953	ENSSSCG00000009361	ENSG00000133110	100152401	10631

Figure 1

Step 1: Annotation of query list A. Start with your input list in your preferred nomenclature [14], e.g. fold change, variations. B. Next copy and paste list into following databases: A. Ensembl Biomart B. g:Profiler (g:Convert tool) C. HGNC (HCOP tool) Set-up several columns with headings for the external references.

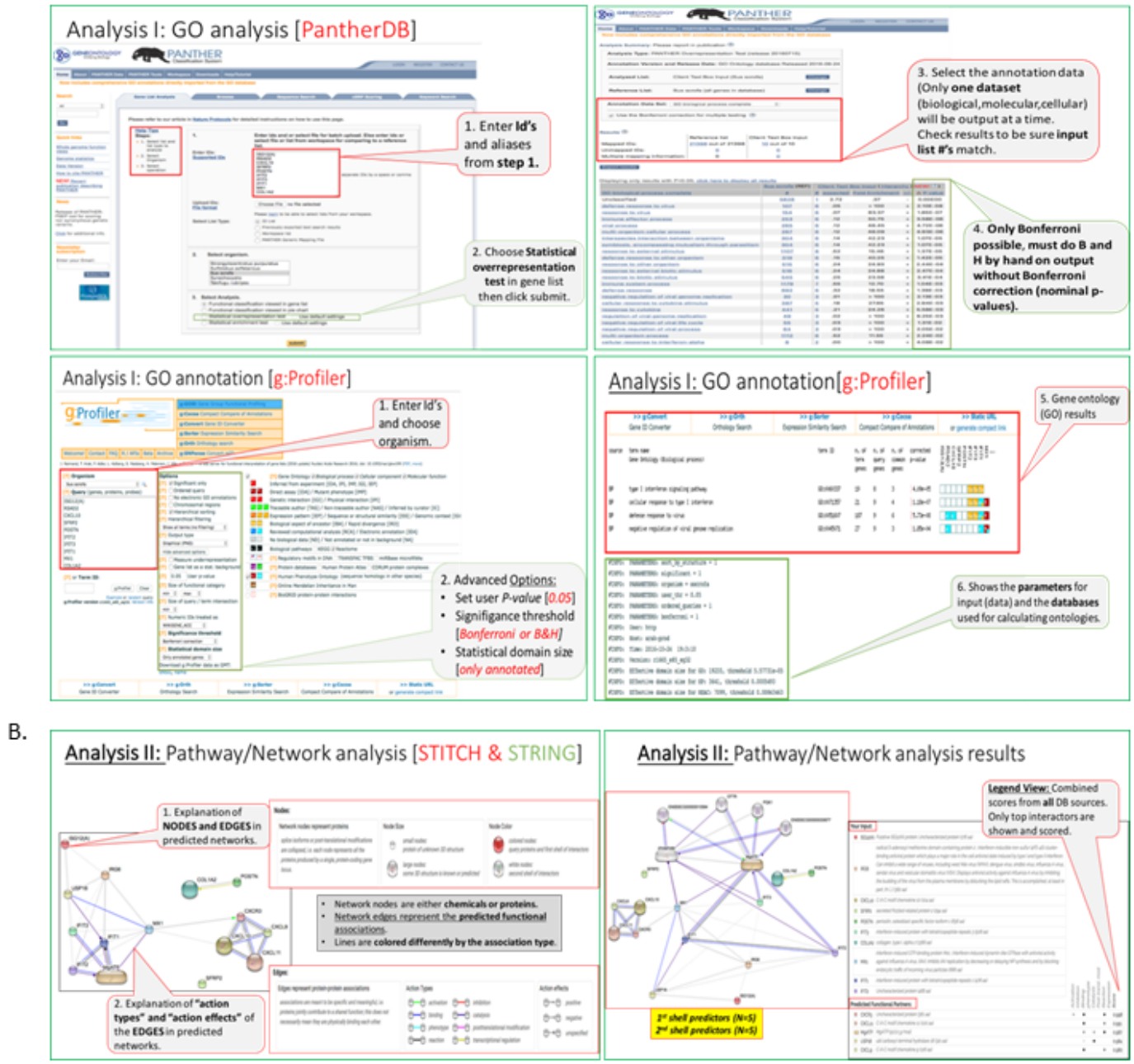

Figure 2

Step 2: Analysis of gene ontologies, pathways and networks A. Analysis I. Software used to generate Gene Ontology (GO): GotermFinder (http://go.princeton.edu), PantherDB, g:Profiler (gGost Tool), DAVID 6.8. B. Analysis II. Software used to generate pathways and networks: STITCH and STRING.

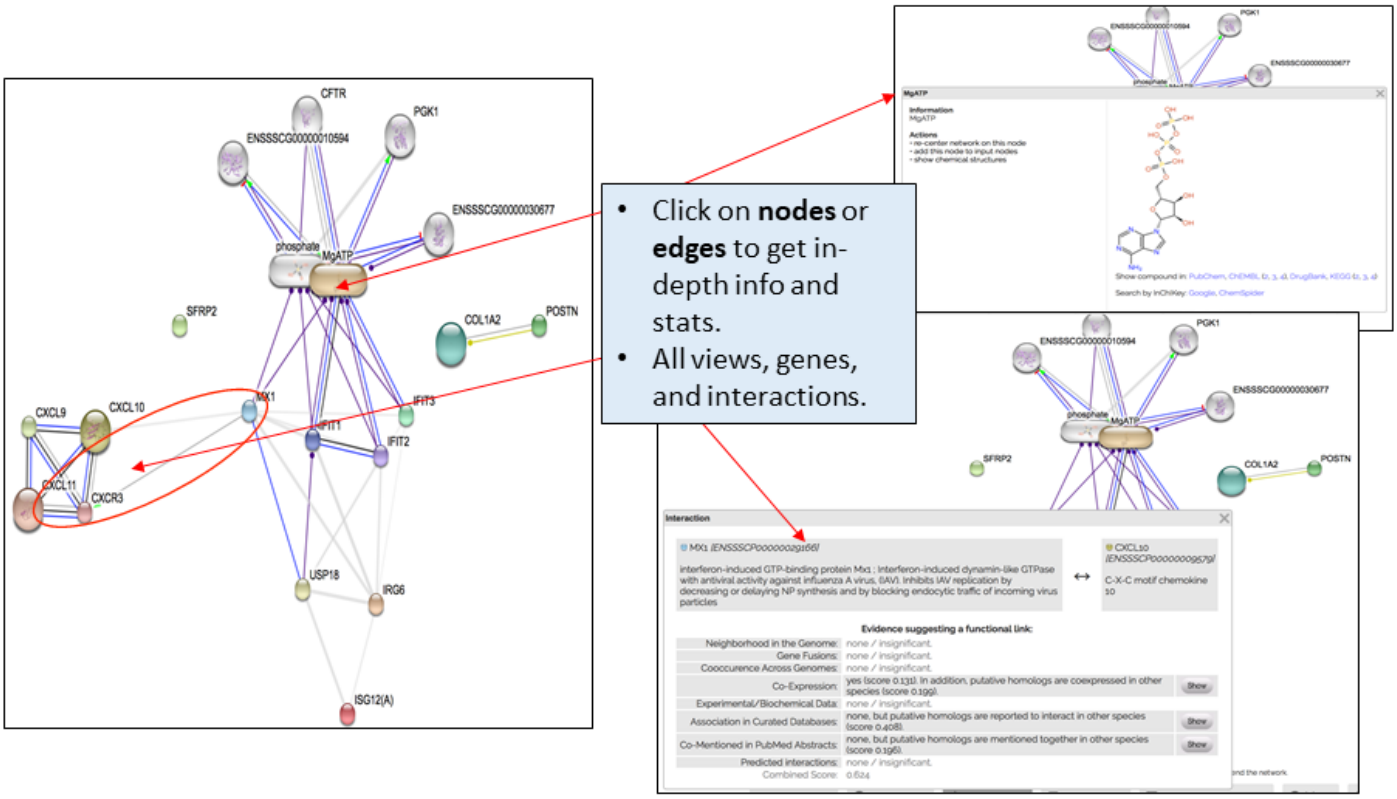

Figure 3

Step 3. Visualization.