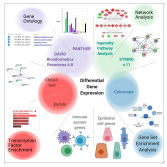


Version 2

Mar 28, 2021

Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type V.2

Evan Maestri[1], Vladimir Kuznetsov[1]

[1]SUNY Upstate Medical Center

1 Works for me dx.doi.org/10.17504/protocols.io.btqfnmtn

Evan Maestri

SUBMIT TO PLOS ONE

ABSTRACT

This approach was utilized for microarray-based gene expression profiling of duodenum mucosa in mice to conduct bioinformatics and network analysis. However, it is also applicable to any differential gene expression analysis, including RNA-seq datasets. Furthermore, the general method structure can be applied to other species, including human. For individuals with limited bioinformatics experience, many of the databases and software in this protocol allow simple inputs for gene list queries, allowing easily understandable analysis. This systems biology protocol can enhance transcriptome data analysis aiding in the generation of hypothesis-driven research and generating testable bioinformatics predictions.

EXTERNAL LINK

https://doi.org/10.1101/2021.03.10.433216

DOI

dx.doi.org/10.17504/protocols.io.btqfnmtn

EXTERNAL LINK

https://doi.org/10.1101/2021.03.10.433216

PROTOCOL CITATION

Evan Maestri, Vladimir Kuznetsov 2021. Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.btqfnmtn
Version created by Evan Maestri

KEYWORDS

RNA-seq, bioinformatics, mixture cell type, microarray, immune

CREATED

Mar 28, 2021

LAST MODIFIED

Mar 28, 2021

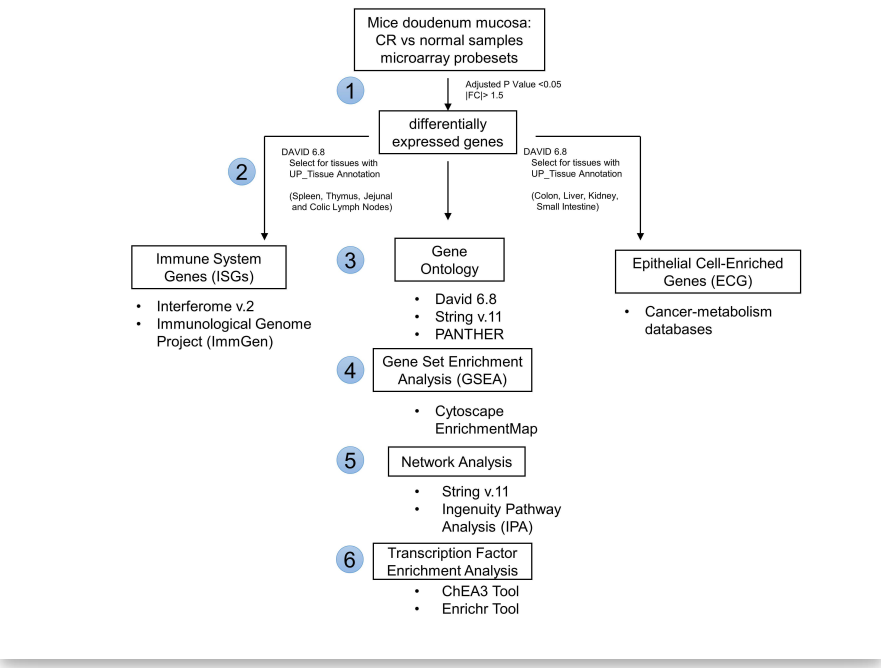
PROTOCOL INTEGER ID

48615

Differential Gene Expression Analysis of Microarray Data

1 The steps in this protocol assume a standard protocol for differential gene expression analysis has been previously followed for microarray or RNA-seq data. The goal is to highlight methodology for gaining biological insight into key networks and pathways exhibited in the expression data. For additional RNA-seq protocols please see this guide and the DESeq2 package through the Bioconductor website.

1.1 Here is an example protocol of the steps and tools utilized in this procedure for the linked calorie restriction (CR) publication.



Enrichment analysis of tissue-associated proteins

2 **Conduct enrichment analysis of tissue-associated proteins in DAVID Bioinformatics 6.8.**

Tissue-associated protein enrichment analysis can be completed using the Uniprot tissue (UP_tissue) annotation database in DAVID Bioinformatics 6.8. and its Functional Annotation Tool. Through selection of enriched tissues, gene subsets can be generated with tissue-specificity. For example, consider an immune and epithelial cell-type mixture. For immune system genes select jejunal and colic lymph nodes, spleen, activated spleen, and thymus. For epithelial genes select colon, liver, kidney, and small intestine. This will separate gene lists for further analysis with immune and epithelial cell-type annotations.

Upload your query gene list of interest. Select the accession identifier (e.g., Ensembl, Refseq, Genbank) and specify the appropriate genetic background. Within the annotation summary of results select Tissue_Expression. Display the chart for the Up_tissue annotations. Download the gene list per selected tissue of interest.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	UP_TISSUE	Liver	RT		125	28.0	1.5E-7	2.3E-5
☑	UP_TISSUE	Colon	RT		58	13.0	2.0E-6	1.5E-4
☐	UP_TISSUE	Activated spleen	RT		30	6.7	6.6E-5	2.7E-3
☐	UP_TISSUE	Spleen	RT		44	9.8	7.1E-5	2.7E-3
☐	UP_TISSUE	Mammary tumor metastatized to lung. Tumor arose spontaneously	RT		26	5.8	9.5E-4	2.9E-2
☐	UP_TISSUE	Cecum	RT		14	3.1	2.9E-3	7.4E-2
☐	UP_TISSUE	Mammary tumor. Metallothionien-TGF alpha model. 10 month old virgin mouse. Taken by biopsy.	RT		17	3.8	4.9E-3	1.0E-1
☑	UP_TISSUE	Small intestine	RT		22	4.9	5.3E-3	1.0E-1
☐	UP_TISSUE	Kidney	RT		76	17.0	9.7E-3	1.6E-1
☑	UP_TISSUE	Intestinal mucosa	RT		4	0.9	1.2E-2	1.7E-1
☐	UP_TISSUE	Bone marrow	RT		51	11.4	2.4E-2	3.3E-1
☐	UP_TISSUE	Mammary tumor. WAP-Tag model. 5 months old	RT		6	1.3	2.9E-2	3.6E-1
☐	UP_TISSUE	Mammary tumor. Brca1-/fl	RT		7	1.6	6.3E-2	7.3E-1
☐	UP_TISSUE	Jejunal and colic lymph node	RT		2	0.4	7.4E-2	8.0E-1
☐	UP_TISSUE	Oviduct	RT		7	1.6	9.8E-2	9.9E-1

Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn

DAVID Bioinformatics Resources 6.8

Huang da W, Sherman BT, Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.. Nature protocols.
https://doi.org/10.1038/nprot.2008.211

Gene Ontology Analysis

3 Conduct Gene Ontology (GO) Analysis.

GO analysis is useful to identify enriched or over-represented annotations within a gene set.

First run GO analysis on the entire list of genes in your dataset. Next, select a smaller gene subset list from your entire expression data. For example, separate your expression data by only upregulated genes and only downregulated genes. Run GO analysis of both subsets. This may yield additional insights into the treatment condition you are studying. For a mixture cell-type dataset, the downregulated gene sets may have annotations with certain tissue-specific pathways, and the upregulated gene sets with others.

3.1 Option 1: DAVID Bioinformatics.

From the annotation summary results of your input gene list entered in , select the category Gene_Ontology. By default, biological process, cellular component, and molecular function will be selected.

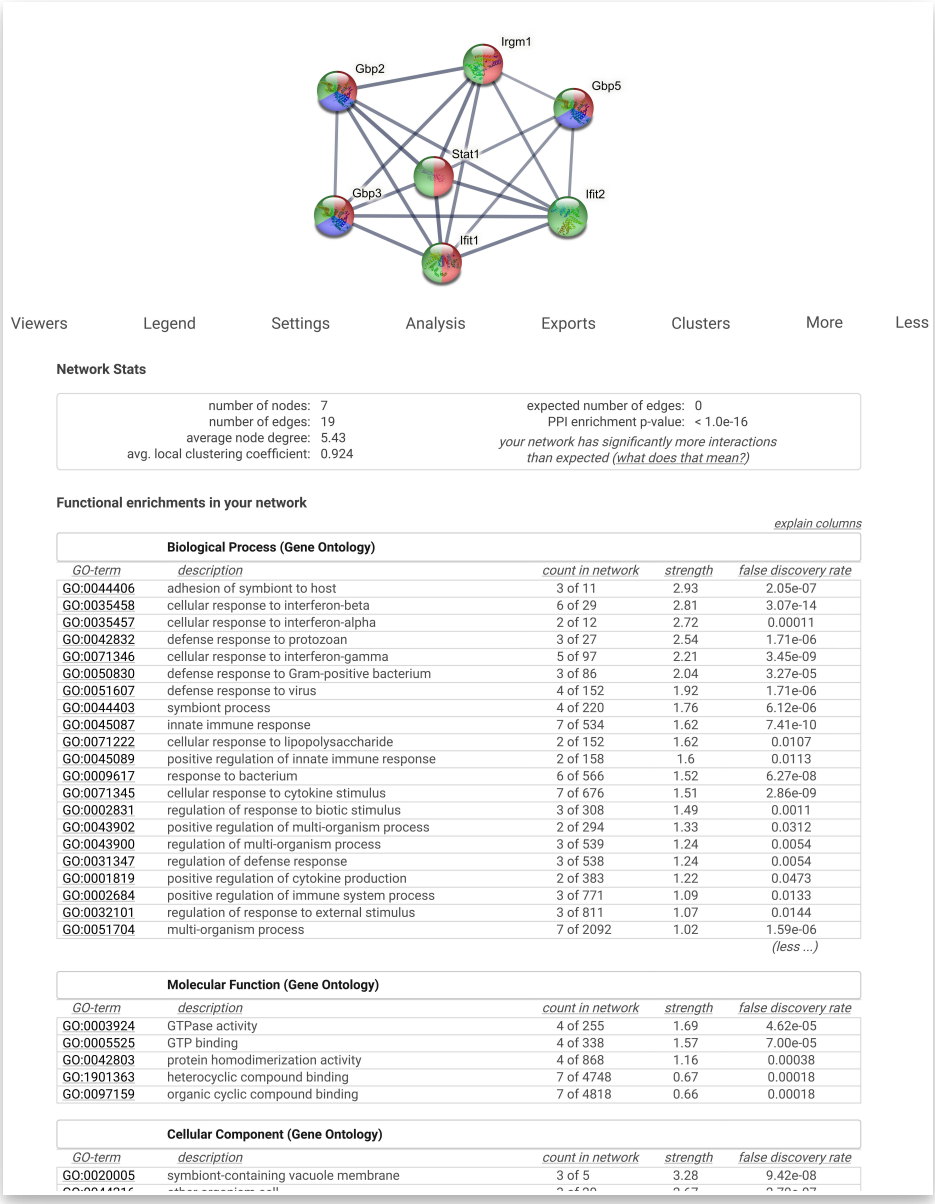
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☑	GOTERM_BP_DIRECT	defense response to virus	RT		21	4.7	4.0E-10	7.4E-7
☑	GOTERM_BP_DIRECT	immune system process	RT		30	6.7	3.2E-9	2.9E-6
☑	GOTERM_BP_DIRECT	response to virus	RT		14	3.1	2.2E-8	1.3E-5
☑	GOTERM_BP_DIRECT	innate immune response	RT		28	6.3	1.2E-7	5.5E-5
☐	GOTERM_BP_DIRECT	oxidation-reduction process	RT		37	8.3	3.7E-7	1.4E-4
☐	GOTERM_BP_DIRECT	response to nutrient	RT		11	2.5	3.0E-6	9.2E-4
☐	GOTERM_BP_DIRECT	metabolic process	RT		27	6.0	6.5E-6	1.7E-3
☐	GOTERM_BP_DIRECT	glutathione metabolic process	RT		9	2.0	7.8E-6	1.8E-3
☑	GOTERM_BP_DIRECT	immune response	RT		19	4.3	2.1E-5	4.0E-3
☑	GOTERM_BP_DIRECT	cellular response to interferon-beta	RT		8	1.8	2.2E-5	4.0E-3
☐	GOTERM_BP_DIRECT	sodium ion transport	RT		12	2.7	6.5E-5	1.1E-2
☐	GOTERM_BP_DIRECT	acyl-CoA metabolic process	RT		6	1.3	4.4E-4	6.8E-2
☐	GOTERM_BP_DIRECT	lipoprotein metabolic process	RT		6	1.3	5.2E-4	6.9E-2
☐	GOTERM_BP_DIRECT	cellular response to lipopolysaccharide	RT		14	3.1	5.3E-4	6.9E-2
☑	GOTERM_BP_DIRECT	inflammatory response	RT		18	4.0	1.1E-3	1.2E-1

3.2 Option 2: STRING v.11.

The STRING database provides a resource for enrichment analysis, implementing classification for Gene Ontology, KEGG, pathways, and domains. Submit a query of a set of proteins. Annotations that are enriched in the network of input proteins compared to the background will be determined. See Section 5.1 for more details on STRING networks.

Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn

Color your graph by selecting functional enrichments in Gene Ontology terms for your network.



3.3 Option 3: PANTHER.

The PANTHER (Protein ANalysis THrough Evolutionary Relationships) classification system v. 16.0 contains 15635 protein families. Utilizing this tool will classify proteins (and their genes) according to family and subfamily, molecular function (biochemical level), biological process (larger protein network), and pathway.

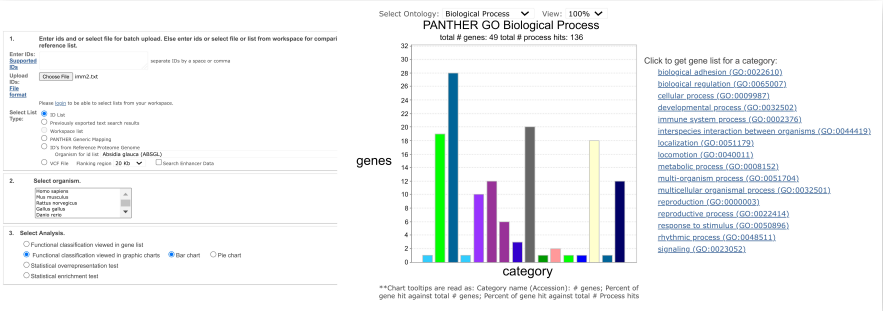
Enter IDs for gene list analysis. This can be completed for a variety of organisms including mouse and human.

Select analysis type. Options include functional classification, statistical overrepresentation testing, and statistical enrichment testing.

Select functional classification viewed in graphic charts for quick visualization of your dataset. Toggle between the Ontology classifications of molecular function, biological process, cellular component,

Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn

protein class, and pathway. Select an annotation category to identify the total number of hits and a gene list for each category.



PANTHER 16.0

Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, Thomas PD (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API.. Nucleic acids research. https://doi.org/10.1093/nar/gkaa1106

Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, Thomas PD (2019). Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0).. Nature protocols. https://doi.org/10.1038/s41596-019-0128-8

Gene Set Enrichment Analysis

4 Perform Gene Set Enrichment Analysis (GSEA).

Create an expression dataset txt file with genes or probes and expression value for each feature. The GSEA User Guide defines how to run a GSEA after installation. Additionally, it provides options to run GSEAPreranked, which is GSEA run on a list of genes supplied by the user already ranked (e.g., by fold change). This metric would rank strong upregulated genes at the top of the list and strong downregulated genes at the bottom.

Gene Set Enrichment Analysis 4.1.0

by UC San Diego, Broad Institute

GSEA requires a gene set file defining the gene set name and list of genes in the set in a tab-delimited text file in gmx or gmt format. Current releases of the gene set annotations can be downloaded from the BaderLab for human, mouse, and rat. The BaderLab is involved in many collaborative open-source bioinformatics projects for biological pathway

Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn

data visualization (Cytoscape, EnrichmentMap).

Gene Ontology Biological Processes (BP), All Pathways, Drug Targets, and Disease Phenotypes can be downloaded in gmt format. By selecting files with notation "no GO iea", the file will exclude GO annotation evidence codes 'IEA' (inferred from electronic annotation), 'ND' (No biological data available), 'RCA' (inferred from reviewed computational analysis). This will improve the evidence codes indicating that there is evidence from an experiment directly supporting the annotation. Use default settings and set "collapse dataset to gene symbols" to "false".

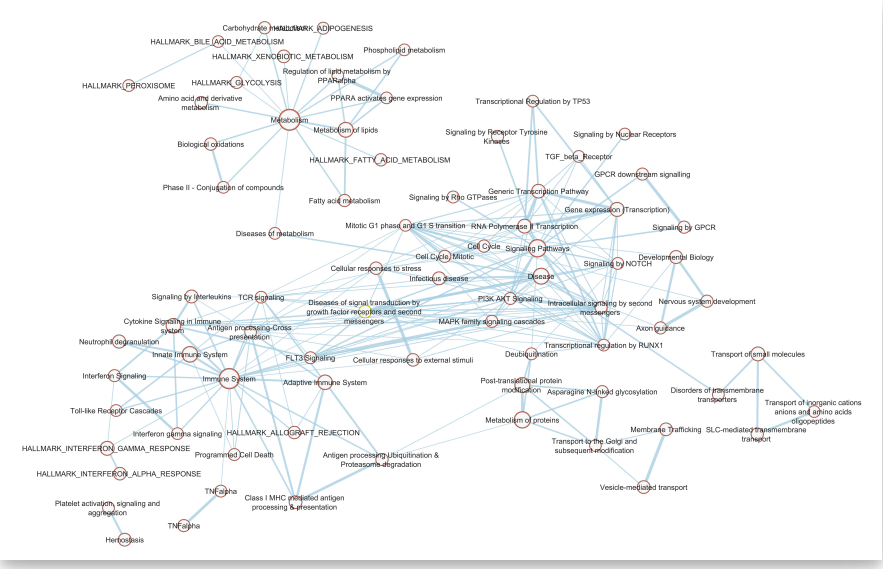
A ranked list prior to GSEA analysis can be calculated by the p-value and direction of the fold change for the differentially expressed genes. Significant gene sets can be visualized in Cytoscape v3.8.2 EnrichmentMap.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.. Proceedings of the National Academy of Sciences of the United States of America.

4.1 Visualize gene sets using Cytoscape's Enrichment Map Tool.

Cytoscape 3.8.2

The EnrichmentMap Cytoscape App which allows visualization of the results of gene-set enrichment as a network can be downloaded from the Cytoscape app store. The Cytoscape User Manual and EnrichmentMap User Guide provide step-by-step instructions for installation and network creation. Enrichments must be generated outside of EnrichmentMap. Gene-sets (including pathways and Gene Ontology terms) with overlapping genes between categories cluster together for better visualization. Select significant gene sets with specified p-value and FDR cutoffs for display in EnrichmentMap.



Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn

Merico D, Isserlin R, Stueker O, Emili A, Bader GD (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation.. PloS one.
https://doi.org/10.1371/journal.pone.0013984

Network Analysis

5 Networks can be generated separately for each individual gene subset (e.g., immune system, then epithelial). Key hubs can be identified per gene subset exerting major influence over the network. Next, examination of how the network subsets interact can yield different crucial regulators modulating the interconnections between cell-type networks.

5.1 Option 1: STRING v11.

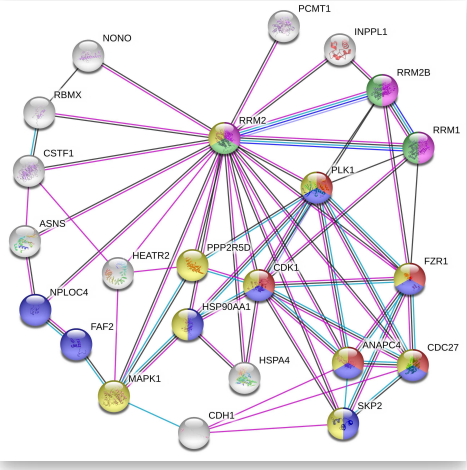
STRING integrates publicly available information on protein–protein interactions, allowing visualization of gene-set enrichment and interaction networks. Search in STRING by single protein name, multiple proteins or by inputting an amino acid sequence. Specify the organism of interest.

For the subset of input proteins, a network will be generated based on the predicted associations. Nodes represent proteins. Edges between proteins represent predicted functional associations.

Within the settings, change the active interaction sources by selecting which evidence types will contribute to the prediction score. Sources include: fusion, neighborhood, cooccurrence, experimental, textmining, database, and coexpression evidence. Select experimental with high (0.7) or highest confidence (0.9) for best reliability of the predicted interaction and network.

Network statistics include: number of nodes, number of edges, average node degree, average local clustering coefficient, expected number of edges, and PPI enrichment p-value.

Disconnected nodes can be hidden from the network. Download the network as a static png or scalable vector graphic.



STRING 11
by Swiss Institute of Bioinformatics, Novo Nordisk Foundation Center Protein Research, European Molecular Biology Laboratory

Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.. Nucleic acids research. https://doi.org/10.1093/nar/gky1131

5.2 Option 2: Ingenuity Pathway Analysis (IPA).

IPA is a commercial software package. QIAGEN Knowledge Base aids in the biological contexts of expression analysis experiments. Generate enriched canonical pathways, drug targets in networks, and disease-associations of your expression data. IPA will identify significant pathways (indicate activation/inhibition prediction) and discover regulatory networks.

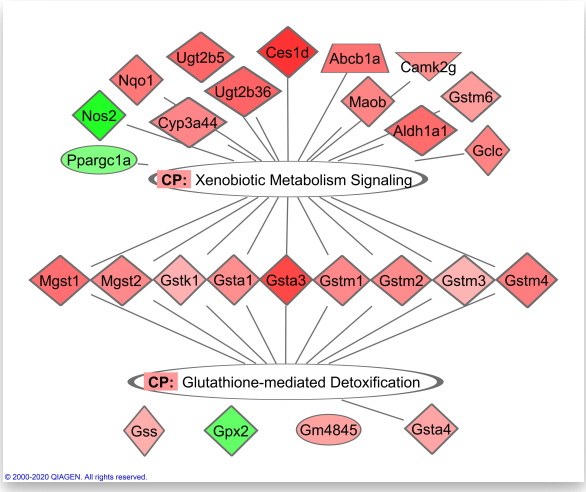
The User Manual and **IPA-Data-Analysis-training-slides.pdf** provide protocols for proper data entry and best practices for network generation.

Run a new core expression analysis in IPA. This will assess signaling pathways, regulators, molecular interaction networks, and disease and biological functions modulated based on expression changes in the input dataset.

Upload data (gene expression, protein expression, metabolomics). Select core analysis settings. Specify the reference set (e.g. Mouse Gene 1.0 ST Array). Select network relationship type (direct and indirect or only direct). Set data cutoff filters by fold-change and p-value. Fischer's exact test in IPA will determine enriched canonical pathways for each gene subset entered.

Select individual canonical pathways of interest to be overlapped and displayed as a network. For example, in an immune-system dataset B-cell, T-cell, NK-cell pathway annotations can be overlapped to identify major regulators. In cancer contexts, canonical pathways regulating epithelial neoplasms, abdominal carcinomas, digestive organ tumors, apoptosis, and proliferative pathways may yield predictive drug targets and novel regulators.

Overlaps between gene subsets in enriched canonical pathways can be selected and displayed as a network. Overlaying an analysis can color the genes by their up or down regulation (color intensity by fold change). See the user manual for overlaying datasets, analyses, and lists.



Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn

Ingenuity Pathway Analysis
by QIAGEN

Krämer A, Green J, Pollard J Jr, Tugendreich S (2014). Causal analysis approaches in Ingenuity Pathway Analysis.. Bioinformatics (Oxford, England).
https://doi.org/10.1093/bioinformatics/btt703

5.3 Network statistics and additional visualization.

Though the resources in 5.1 and 5.2 provide useful statistics, Cytoscape also provides resources for visualization of networks. STRING networks can be directly imported into Cytoscape. This requires installation of the stringApp. STRING functionality within Cytoscape includes protein query, PubMed query, disease query, protein/compound query by entering a list of proteins/compound names. Similarly, Gene Ontology, KEGG Pathways, and protein domain enrichments can be identified at specific significance levels. Networks from IPA can also be imported into Cytoscape for further analysis.

Use the network analyzer tool to calculate Cytoscape network statistics using directed analysis. This identifies the number of proteins, edges (protein interactions), average number of neighbors, clustering coefficient, and network density. Select major network regulators with greatest edge connectivity (highest number of input/output protein connections). Remove unconnected nodes for better visualization of networks.

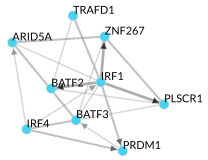
Transcription Factors Controlling Networks

6 ChEA3 Tool.

ChEA3 is a web-based transcription factor (TF) enrichment analysis tool. This tool identifies the transcription factors associated with your gene set. When comparing a control to a perturbation state, ChEA3 can help determine the transcription factors involved in the changes. The ChEA3 TF-target sets incorporate ChIP-seq experiments (ENCODE, ReMap) and processed RNA-seq (GTEx, ARCHS4).

Simply input a list of human or mouse gene symbols to conduct transcription factor enrichment analysis.

Rank	TF	Mean Rank	Overlapping Genes	Library
1	BATF2	4.333	7	ARCHS4 Coexpression,8;Enrichr Queries,2;GTEx Coexpression,3
2	BATF3	9.667	5	ARCHS4 Coexpression,15;Enrichr Queries,10;GTEx Coexpression,4
3	TRAFD1	20.67	5	ARCHS4 Coexpression,4;Enrichr Queries,12;GTEx Coexpression,46
4	PLSCR1	23.0	3	ARCHS4 Coexpression,10;GTEx Coexpression,36
5	ARID5A	38.0	1	ARCHS4 Coexpression,43;GTEx Coexpression,33
6	IRF1	39.33	5	Literature ChIP-seq,15;ARCHS4 Coexpression,6;ENCODE ChIP-seq,46;Enrichr Queries,8;ReMap ChIP-seq,160;GTEx Coexpression,1
7	PRDM1	42.6	6	ARCHS4 Coexpression,20;ENCODE ChIP-seq,7;Enrichr Queries,137;ReMap ChIP-seq,1;GTEx Coexpression,48
8	ZNF267	43.0	5	ARCHS4 Coexpression,3;Enrichr Queries,18;GTEx Coexpression,108
9	ARID3C	46.5	1	ARCHS4 Coexpression,64;GTEx Coexpression,29
10	IRF4	59.2	4	ARCHS4 Coexpression,63;ENCODE ChIP-seq,20;Enrichr Queries,43;ReMap ChIP-seq,71;GTEx Coexpression,99

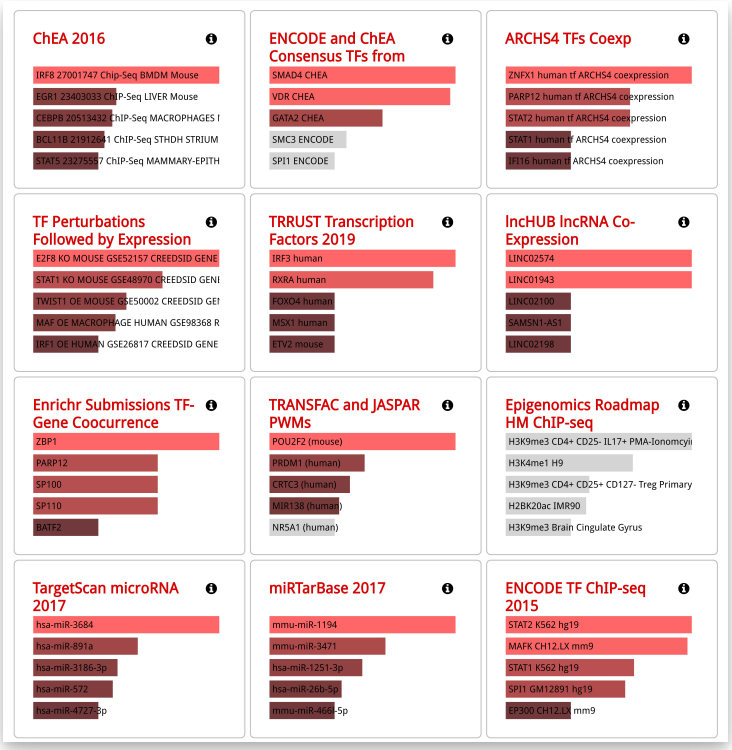


Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn

Keenan AB, Torre D, Lachmann A, Leong AK, Wojciechowicz ML, Utti V, Jagodnik KM, Kropiwnicki E, Wang Z, Ma'ayan A (2019). ChEA3: transcription factor enrichment analysis by orthogonal omics integration.. Nucleic acids research.
https://doi.org/10.1093/nar/gkz446

6.1 Enrichr Tool.

Additionally, a list of Entrez gene symbols or in BED format can be submitted to the tool Enrichr. Enrichments in co-occurrence of your gene set with other genes or signatures resulting from single TF perturbations in genome-wide gene expression experiments is determined.



Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.. Nucleic acids research.
https://doi.org/10.1093/nar/gkw377

Tissue-specific resources for decomposition

7 Immune-system resources

Additional tissue-specific resources can aid in deeper analysis and decomposition of a cellular mixture. Consider a dataset with immune system genes from the GO, network, and pathway analysis in Steps 1-6.

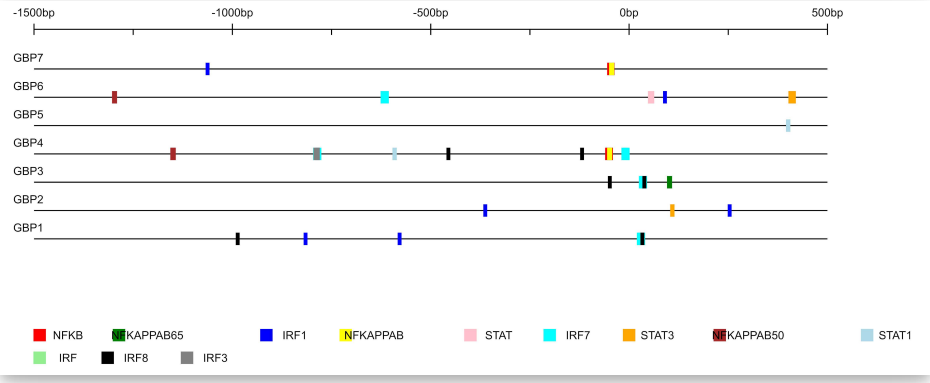
Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn

7.1 Interferome v.2

This database curates data sets from types I, II and III interferon (IFN)-treated cells and standardizes them through quantitative and statistical analyses.

Search by Gene Symbol, GenBank Accession, or Ensembl ID lists. Treatment concentration, treatment time, in vivo/in vitro studies, organ, cell-type, cell line, and fold change can be specified. Curated experiments include both human and mouse.

Experiment data for your queried genes of interest can be downloaded. Selecting TF Analysis will generate a figure of the promoter region within the -1500 to 500 base pair region with colored blocks depicting the predicted TF-binding elements.

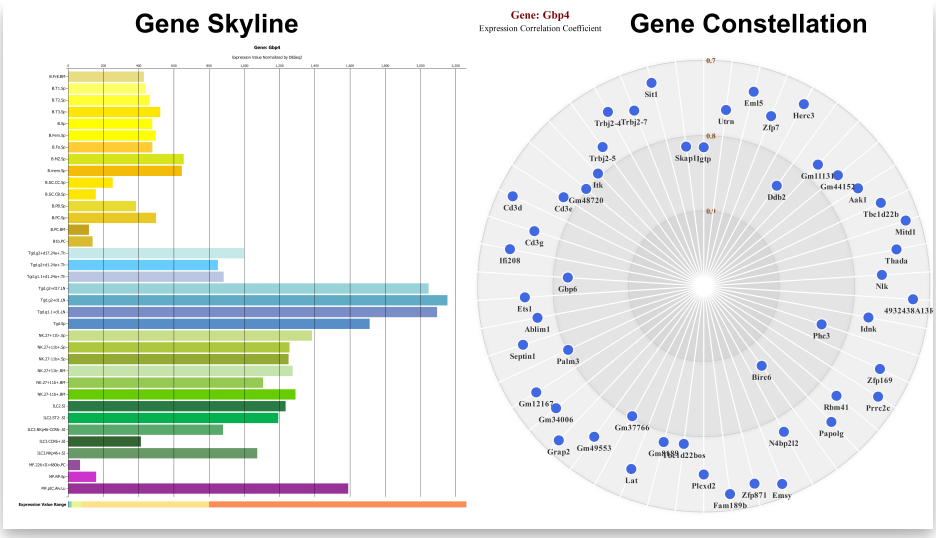


Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H, Chapman R, Hertzog PJ (2013). Interferome v2.0: an updated database of annotated interferon-regulated genes.. Nucleic acids research.
https://doi.org/10.1093/nar/gks1215

7.2 Immunological Genome Project (ImmGen)

The ImmGen DataBrowsers provide useful graphical visualizations of mice immune cell gene-expression. The Single Cell Portal displays single cell RNA-seq profiles of immunological organs and cell-types. Gene Skyline allows visualization of a selected gene across different immune cell type populations (e.g., B cells, T cells, macrophages). Gene Constellation creates a graphic with genes whose expression across the ImmGen landscape are most closely correlated with a queried target. From the input gene, toggle between the STRING analysis and Enrichr tabs.

Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn



Immunological Genome Project (2020). ImmGen at 15.. Nature immunology.
https://doi.org/10.1038/s41590-020-0687-4

Citation: Evan Maestri, Vladimir Kuznetsov (03/28/2021). Protocol for Bioinformatics and Network Analysis of Microarray Data from Mixture Cell Type. https://dx.doi.org/10.17504/protocols.io.btqfnmtn