XiaLab Analytics

Empowering researchers through trainings, tools and AI

https://www.xialab.ca • contact@xialab.ca

Omics Data Science Training Course

Winter 2024

About user registration

- To access the pro websites, you need to create an account using the same email you originally contacted us for subscription
- Depending on your subscription, same account is used for across all 6 pro tools
- Same for your OmicsForum account – the VIP label is based on the same email address

Our Tasks

- **Basic Protocol 5:** How to perform statistical and functional analysis of a cross-species RNA-seq count table generated by ortholog mapping with Seq2Fun.
- **Basic Protocol 7:** How to upload, process, and normalize a set of gene expression tables for meta-analysis.
- **Basic Protocol 8:** How to perform statistical and functional meta-analysis of gene expression data.
- **Basic Protocol 9:** How to analyze single or multiple gene expression signatures.
- **Basic Protocol 10:** How to perform dose-response and time-series analysis.

Recap: microarray core tasks

Input: probe intensity file

1. Processing
   - ○ Mapping probe IDs to gene/transcript IDs
   - ○ Quality checking
2. Data normalization
   - ○ Make data comparable across different arrays/samples
     - ❖ **Quantile normalization**
3. Differentially expression (DE) analysis
   - ○ Find genes that are significantly different between conditions
     - ❖ **Limma**
4. Clustering
   - ○ Find genes with similar expression patterns
5. Enrichment analysis

Recap: RNAseq core tasks

Input: FASTQ files

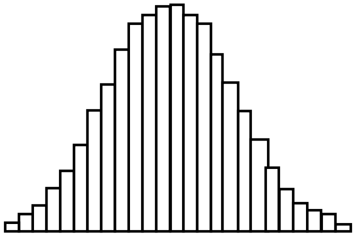
1. Reads are **mapped** to reference genome or transcriptome
2. Mapped reads are **counted** per gene or per transcript
3. Counts are **tested statistically** for significant difference
   - ➢ Limma, edgeR, DEseq2
4. **Enrichment analysis** are applied for functional insights

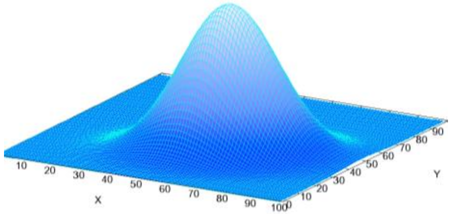
Comments on data “normality”

Formal testing for normality is **not** widely used in omics data

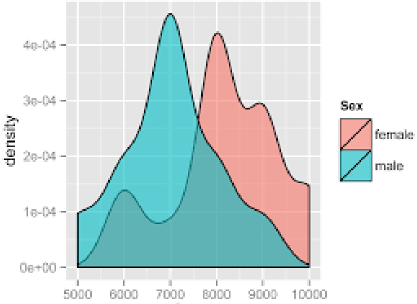
- ➢ Mainly defined for single variable, while omics data is inherently multivariate in nature
- ➢ Individual feature normal does not guarantee multi-variate normal
- ➢ Testing multivariate normal requires a lot of data points – impractical
- ➢ Central limit theorem

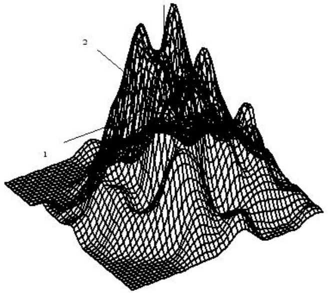
Theory



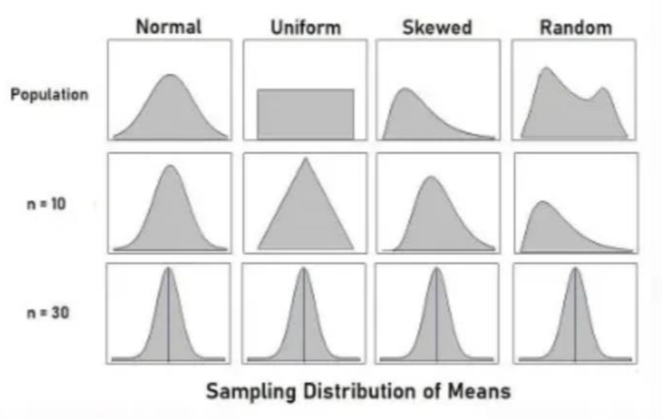


Reality





Central Limit Theorem

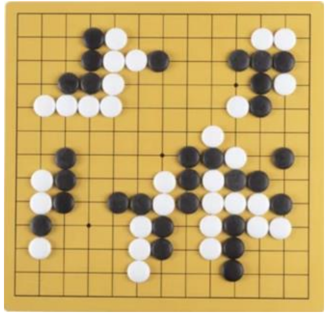


Given a sufficiently large sample size from a population with a finite level of variance, the **mean of all samples** from the same population will be approximately equal to the mean of the population.

- ❖ If the sample is **normal**, then the sampling distribution of sample means will also be normal, no matter what the sample size.
- ❖ When the sample population is approximately **symmetric**, the distribution becomes approximately normal for relatively small values of *n.*
- ❖ When the sample population is **skewed**, the sample size must be **at least 30** before the sampling distribution of the mean becomes approximately normal.

Comments on “best parameter”

- Omics data analysis is composed of multiple steps.
- We should aim for global optimal
  - The overall expectation, hypothesis, etc
- There are some criteria
  - QC plot
  - Some internal data consistency
    - House keeping genes, within group variance, sig #
  - Functional or biological end points
  - Literature context



Comments on "statistical significance"

Statistical significance should not be the goal. It is only a way to prioritize (by sifting through big data) the features of interest towards **biological significance**

- ➢ Sticking to strict statistics rules will cause data loss (nothing to work on later)
- ➢ Validations and context (other evidence) are required to confirm the biological significance
- ➢ Omics mainly for hypothesis generation, not decision making (validation studies)
- ➢ Should be rich, informative and inspiring!

Pipelining => Report => Explore => Iterate

Our Syllabus

Topic	Date	Lecture	Lab
Omics Data Science Foundations	Jan. 6	Omics data processing, statistics and visualization	--
	Jan. 13	From raw data to functional insights	--
Transcriptomics	Jan. 20	Gene expression data analysis (part I)	ExpressAnalyst & NetworkAnalyst
	Jan. 27	Gene expression data analysis (part II)	ExpressAnalyst & Seq2Fun
miRNAs & non-coding RNAs	Feb. 3	MicroRNAs, noncoding RNAs and biological networks	miRNet & NetworkAnalyst
Proteomics	Feb. 10	Proteomics data analysis and interpretation	ExpressAnalyst & NetworkAnalyst
Metabolomics	Feb. 17	Targeted metabolomics data analysis	MetaboAnalyst
	Feb. 24	LC-MS untargeted metabolomics data analysis	MetaboAnalyst
Microbiomics	Mar. 2	Marker gene data analysis	MicrobiomeAnalyst
	Mar. 9	Shotgun metagenomics data analysis	MicrobiomeAnalyst
Multi-omics	Mar. 16	Knowledge-driven multi-omics integration	OmicsNet
	Mar. 23	Data-driven multi-omics integration	OmicsAnalyst

Schedule for today

Time	Topics
9:00 – 9:10	Introduction
9:10 – 9:40	Network analysis of gene list data & Live Demo
9:40 – 10.00	Dose response analysis & Live Demo
10:05 – 10:40	RNAseq in non-model species & Live Demo
10:45 – 11:25	Meta-analysis of multiple expression data & Live Demo
Summary and discussion	

Biological Networks

Biological functions are executed through a series of molecular interaction networks

- **Protein interaction network:** proteins that are connected in physical interactions
  - Proteins and their interaction partners
- **Gene regulatory network**: a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and protein
  - miRNA, transcription factors, non-coding genes, etc
- **Metabolic network:** metabolic products and substrates that participate in biochemical reactions
  - Metabolites, enzymes, co-factors

Protein-protein interaction (PPI) network

Experimental

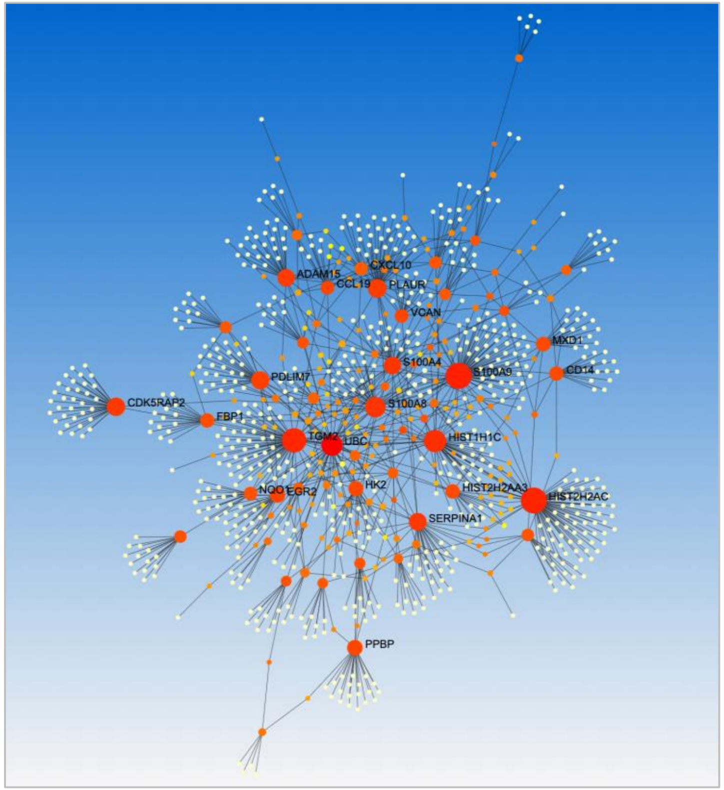
- InnataDB/IMEx Interactome
- Rolland Interactome (2014)
- HuRi (human reference interactome, 2020): Pairwise combinations of human protein-coding genes are tested systematically using high throughput yeast two-hybrid screens to detect protein-protein interactions.

Predicted

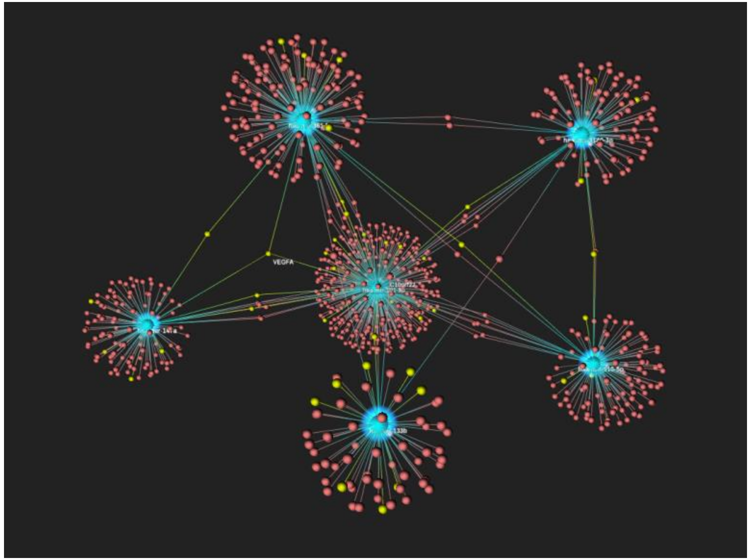
- Binding motifs, evolution, etc

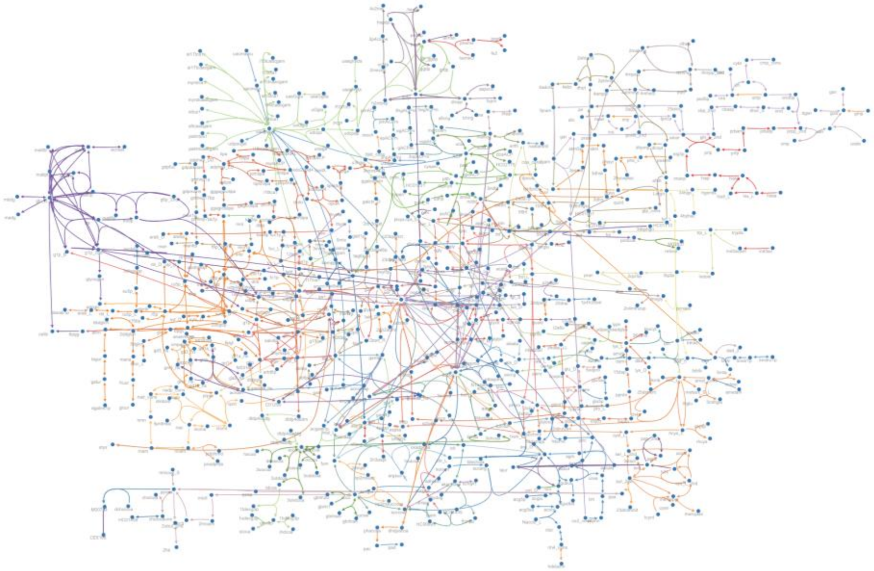
Both:

- STRING:
  - Experimentally derived interactions through literature curation;
  - Computationally predicted interactions through text mining and inference from other organisms
  - Supporting > 10,000 organisms



Different types of networks have different structures





miRNA-gene target

Metabolic networks

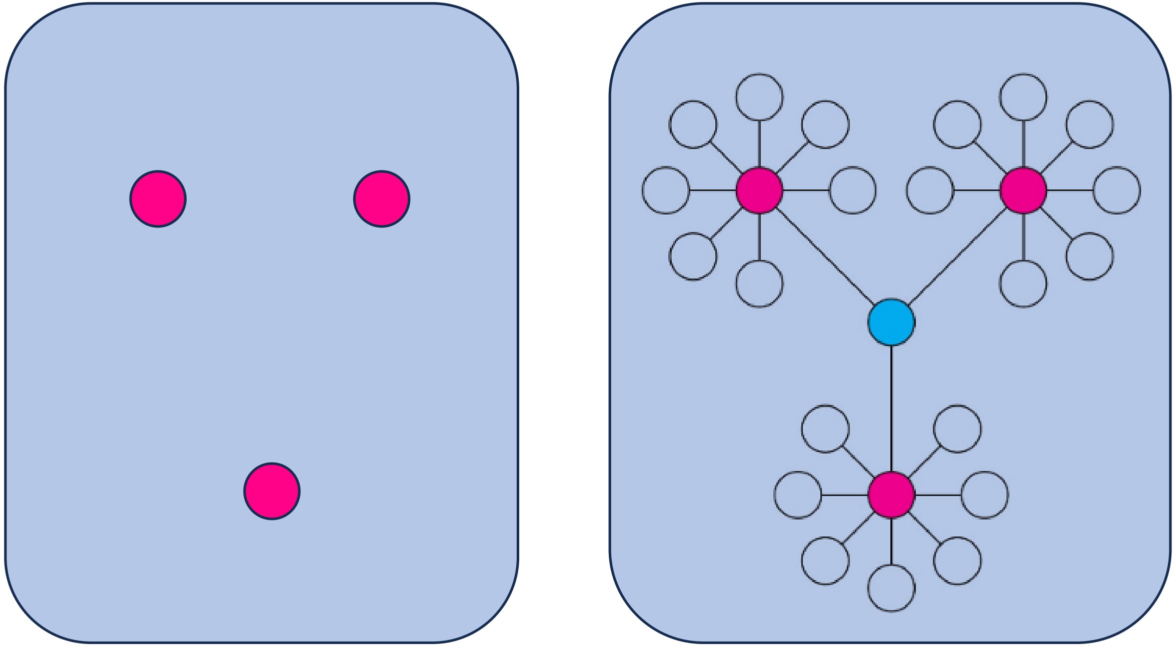
How to detect robust functional changes?

- Functions are group behavior – a balanced size to be specific yet informative
  - ~100s
- Too small will become less powerful to detect functions
  - ~10s
  - ➔ Signal amplification
- Too large will become too general not specific
  - ~1000s
  - ➔ Signal distillation

PPI Network for Signal Amplification

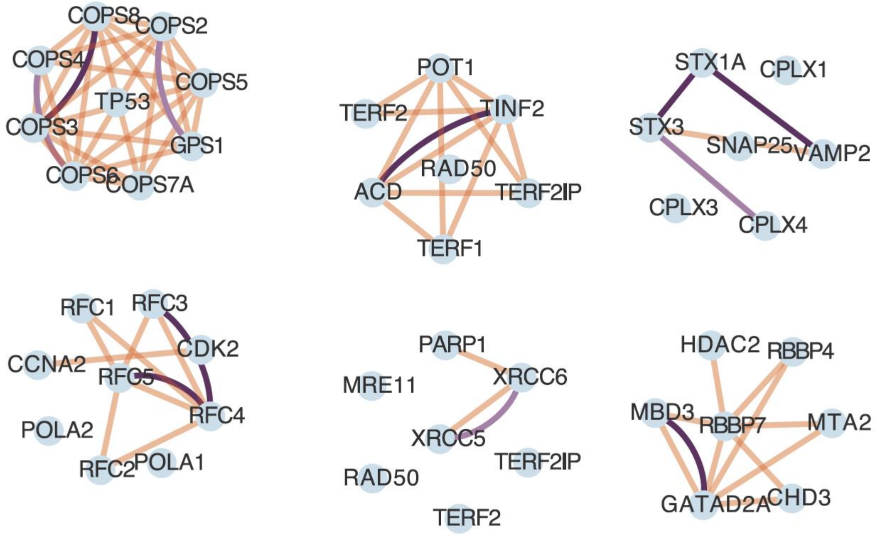
- ➢ Use seeds (i.e., a list of significant features) to query a database to obtain their close neighbors.
- ➢ Introducing those new members will amplify the signals

Assumption: location proximity ≈ functional similarity



PPI Network for Signal Distillation

- To refine the context by introducing known relationships among the large number features to facilitate identification of functional themes
    - Adding edges (known relationships) between the large number of significant genes
    - Detecting modules (i.e., densely connected subnetworks)



Different types of PPI network

- ➢ A **first-order network** (default) returns all seed genes and all nodes **directly** connected to them in the database.

➔ **signal amplification**

- ➢ A **zero-order network** will only introduce new edges between seed genes. In addition, it can reduce the number of seed genes because it retains only genes that are connected to each other within the underlaying database (i.e. orphan genes will be removed). This can help simplify your gene list to highlight the biological themes

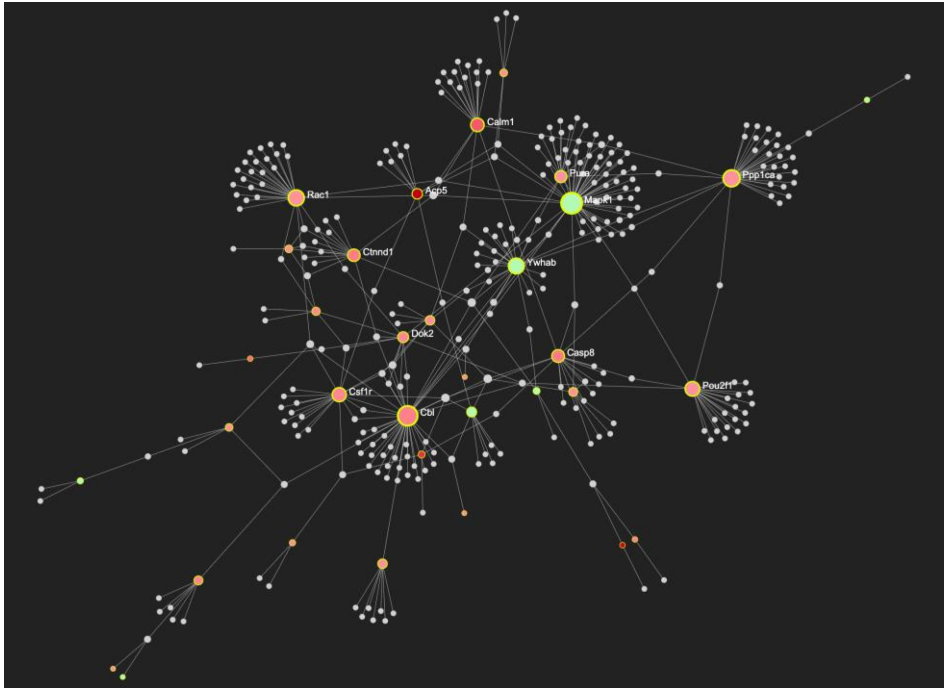
➔ **signal distillation**

- ➢ A **second-order network** increases the size because it returns all seed genes and all nodes that are within two connections in the database. Commonly known as “friends of their friends”.

➔ Usually leads to giant network, rarely used.

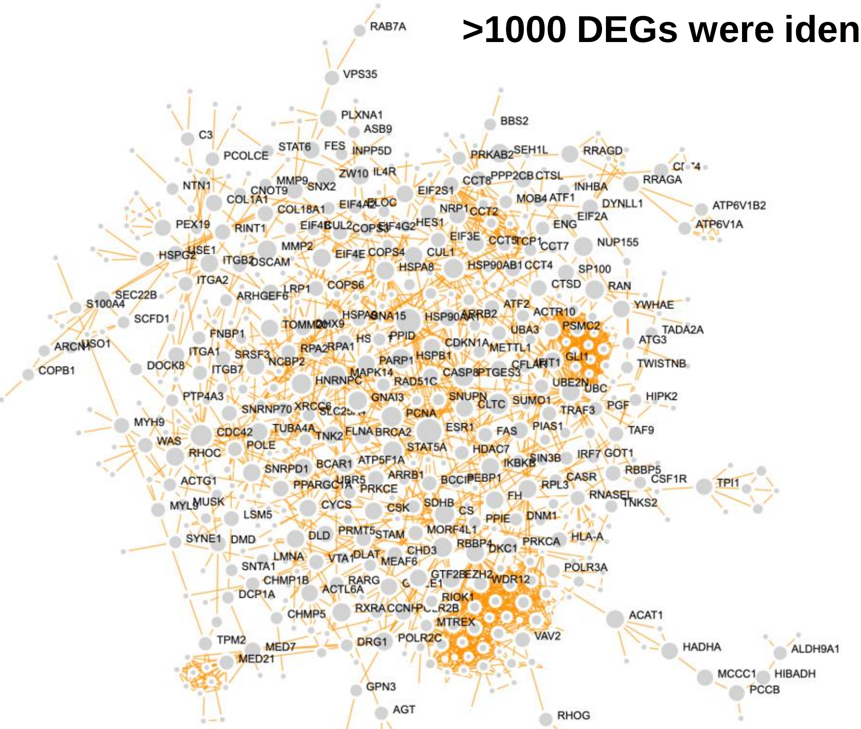
First-order network

- ~20 significant features (seeds) to query a comprehensive PPI database to obtain their interaction partners.
- Introducing those new members will improve the signals
- Colored nodes are “seeds”

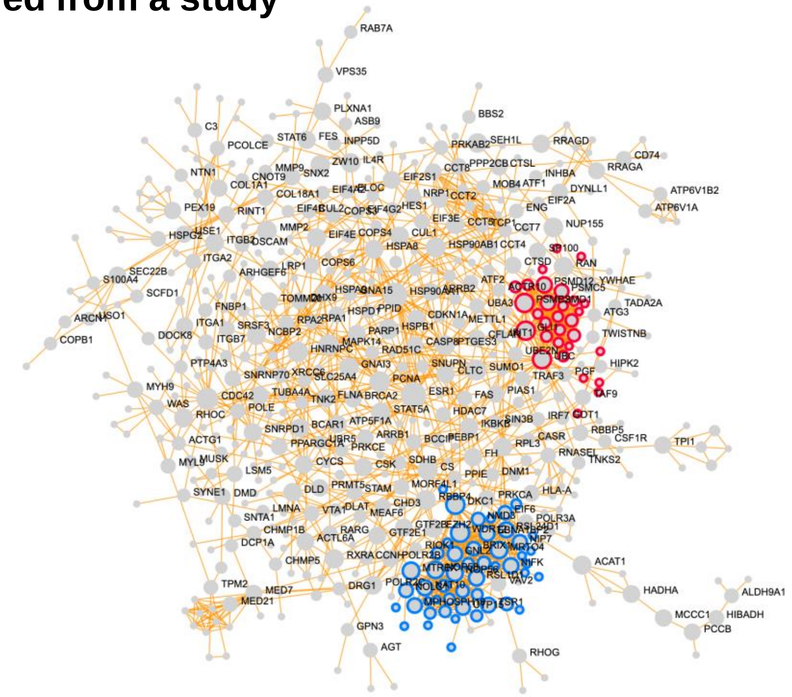


Zero-order network

>1000 DEGs were identified from a study



Identify relationships among DEGs

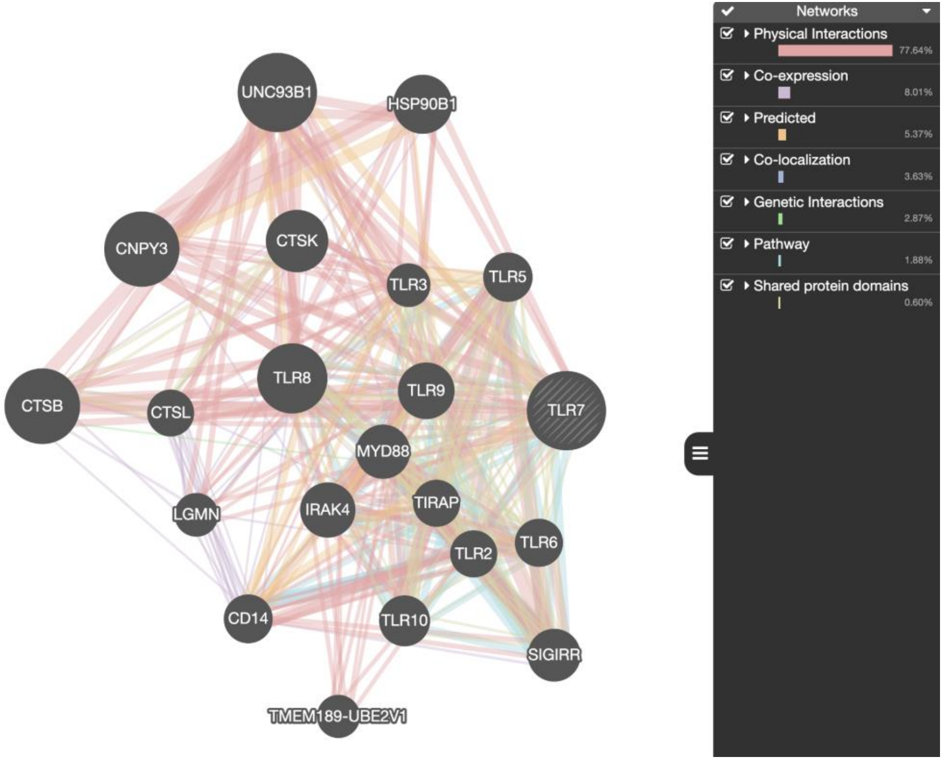


Detect functional modules

Improve the quality of the PPI network

High-quality connection => more accurate interpretation

- Choose different libraries
  - Experimental vs computational
  - Different approaches with different false positives
  - Multiple evidence
- Add cell-specific or tissue-specific filters
  - Single-cell sequencing repo
  - Many genes and proteins are tissue specific. They can only interact when both present



Networks

☑ Physical Interactions 77.64%

☑ Co-expression 8.01%

☑ Predicted 5.37%

☑ Co-localization 3.63%

☑ Genetic Interactions 2.87%

☑ Pathway 1.88%

☑ Shared protein domains 0.60%

https://genemania.org/search/homo-sapiens/TLR7/NFKB

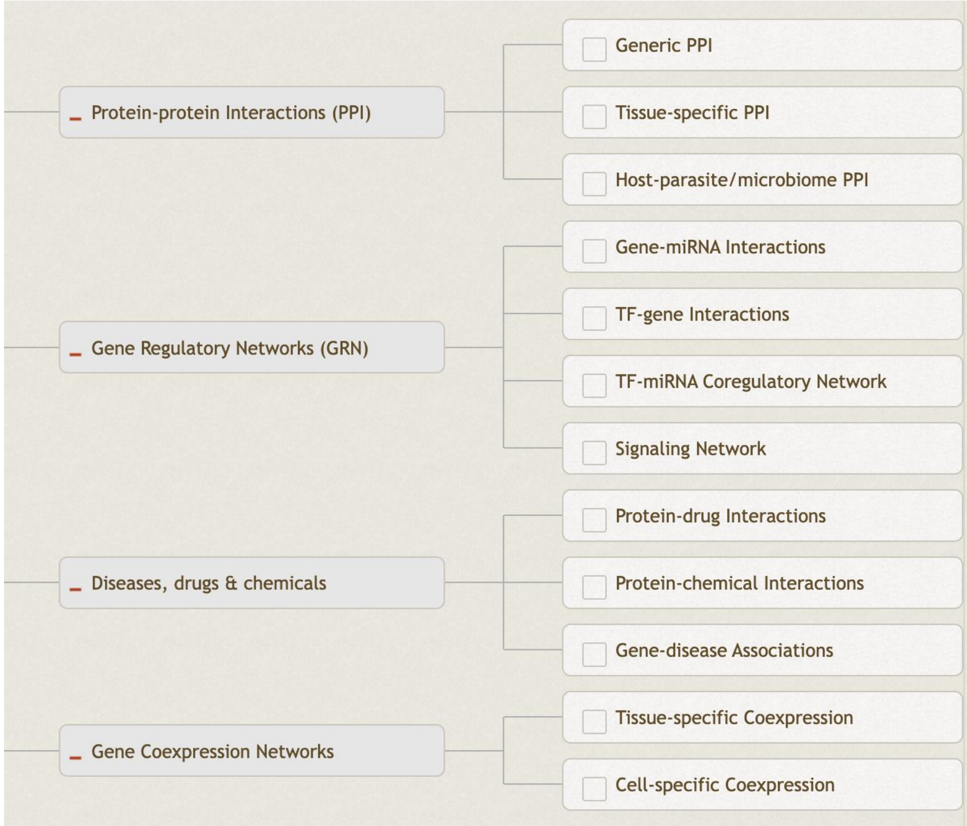
NetworkAnalyst

Comprehensive understanding a gene list within network context

- PPI network
- miRNA-gene
- TF-gene
- Signaling network
- Protein-drugs
- …..

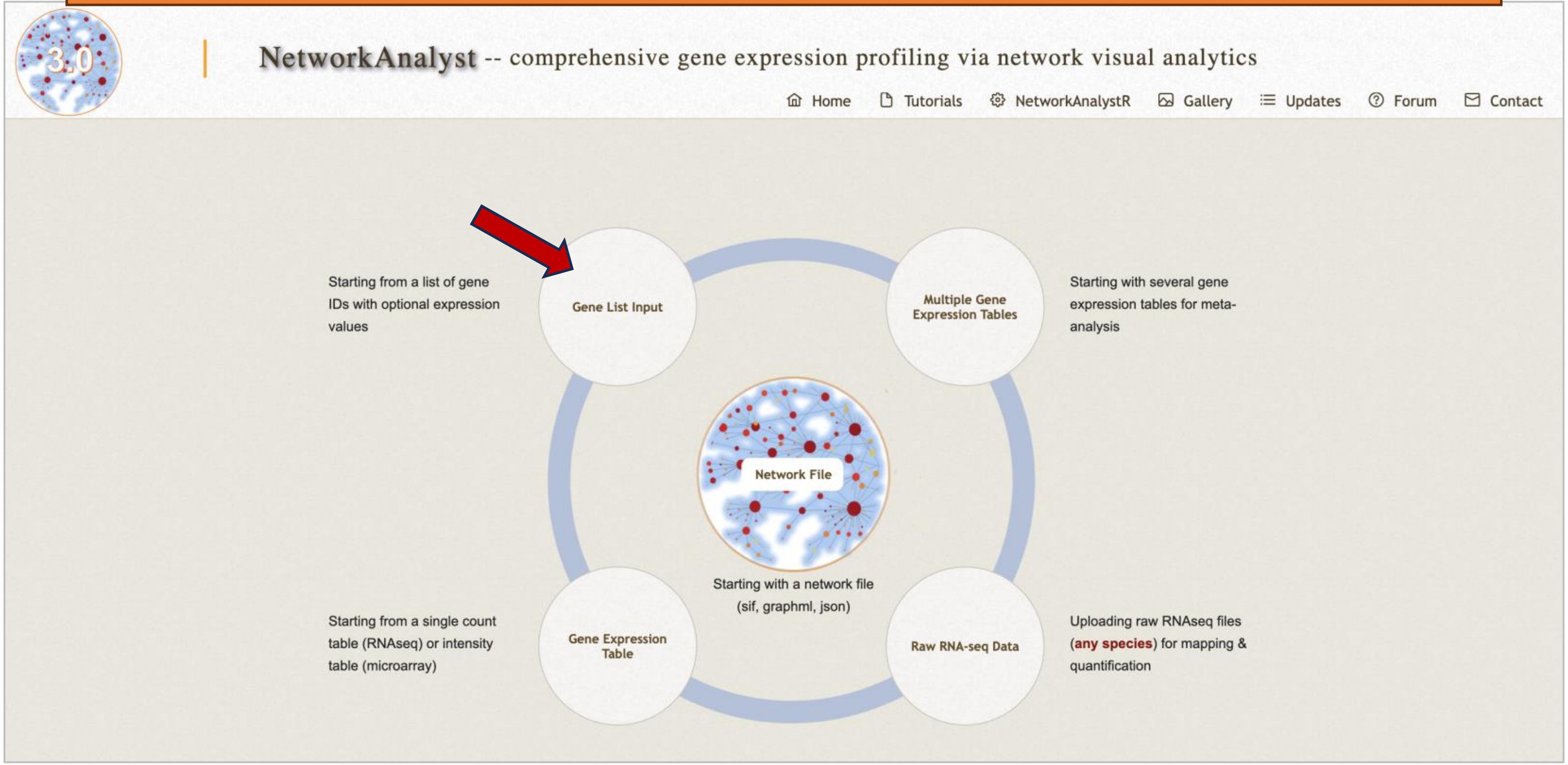
Build-in network visual analytics system

- 2D / 3D
- Different algorithms for layout and module detection
- Enrichment analysis



- − Protein-protein Interactions (PPI)
  - □ Generic PPI
  - □ Tissue-specific PPI
  - □ Host-parasite/microbiome PPI
- − Gene Regulatory Networks (GRN)
  - □ Gene-miRNA Interactions
  - □ TF-gene Interactions
  - □ TF-miRNA Coregulatory Network
  - □ Signaling Network
- − Diseases, drugs & chemicals
  - □ Protein-drug Interactions
  - □ Protein-chemical Interactions
  - □ Gene-disease Associations
- − Gene Coexpression Networks
  - □ Tissue-specific Coexpression
  - □ Cell-specific Coexpression

Live Demo



NetworkAnalyst -- comprehensive gene expression profiling via network visual analytics

Starting from a list of gene IDs with optional expression values

Starting with several gene expression tables for meta-analysis

Starting with a network file (sif, graphml, json)

Starting from a single count table (RNAseq) or intensity table (microarray)

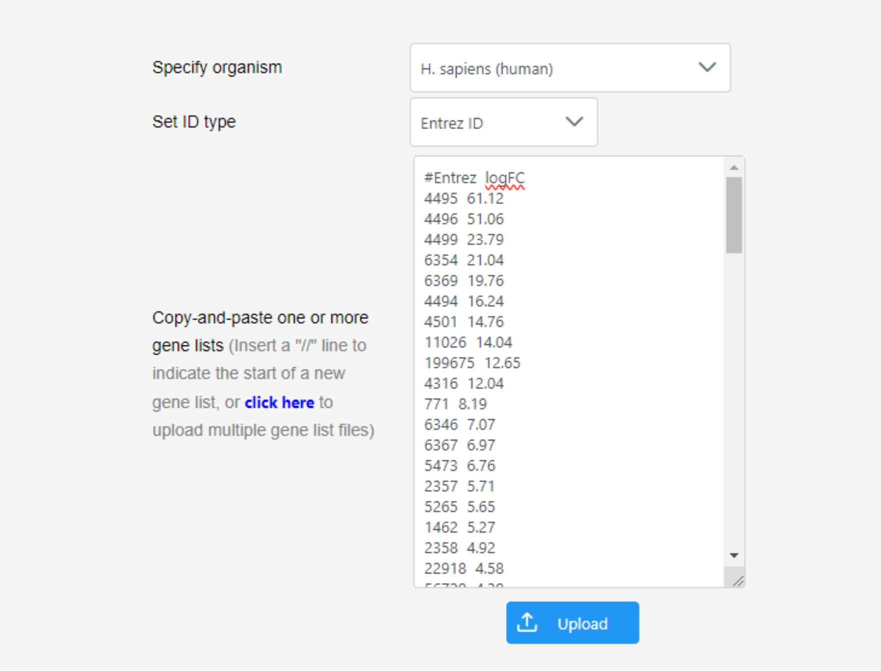
Uploading raw RNAseq files (**any species**) for mapping & quantification

XiaLab.ca

Empowering researchers through trainings, tools, and AI

Gene signature – Data format

- List of genes with optionally a numerical value (i.e. logFC) as second column



Specify organism: H. sapiens (human)

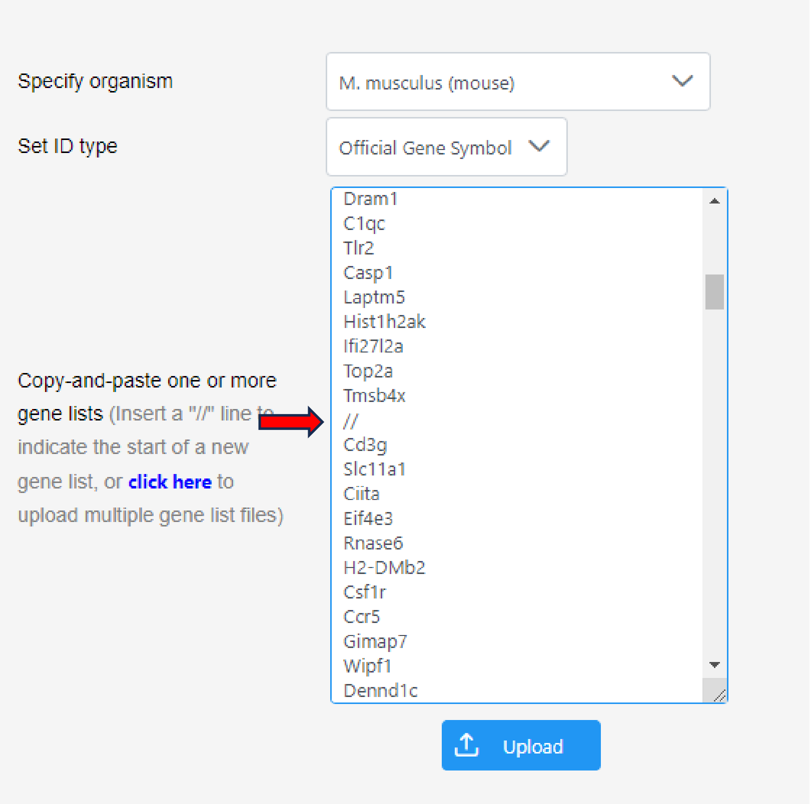
Set ID type: Entrez ID

Copy-and-paste one or more gene lists (Insert a "//" line to indicate the start of a new gene list, or **click here** to upload multiple gene list files)

```
#Entrez  logFC
4495  61.12
4496  51.06
4499  23.79
6354  21.04
6369  19.76
4494  16.24
4501  14.76
11026  14.04
199675  12.65
4316  12.04
771  8.19
6346  7.07
6367  6.97
5473  6.76
2357  5.71
5265  5.65
1462  5.27
2358  4.92
22918  4.58
```

Gene signature – Data format

- For multiple lists, insert "//" in a line to indicate the start of new list.



Specify organism: M. musculus (mouse)

Set ID type: Official Gene Symbol

Copy-and-paste one or more gene lists (Insert a "//" line to indicate the start of a new gene list, or **click here** to upload multiple gene list files)

```
Dram1
C1qc
Tlr2
Casp1
Laptm5
Hist1h2ak
Ifi27l2a
Top2a
Tmsb4x
//
Cd3g
Slc11a1
Ciita
Eif4e3
Rnase6
H2-DMb2
Csf1r
Ccr5
Gimap7
Wipf1
Dennd1c
```

Schedule for today

Time	Topics
9:00 – 9:10	Introduction
9:10 – 9:40	Network analysis of gene list data & Live Demo
9:40 – 10.00	Dose response analysis & Live Demo
10:05 – 10:40	RNAseq in non-model species & Live Demo
10:45 – 11:25	Meta-analysis of multiple expression data & Live Demo
	Summary and discussion

What/why dose response analysis?

- Widely used in toxicology & pharmacology
  - To establish the quantitative relationship between exposure (dose) – effects (response)
  - Risk assessment (the effect at a specified dose)
    - ✓ “The right dose differentiates a poison from a remedy”

- Effect can be measured by
  - Organism-level – apical outcome such as reproductive failure, or developmental dysfunction
  - Molecular levels such as gene expression

- To provide a quick assessment of the mechanistic processes by which test substances modify biological systems, and more importantly, the doses at which these changes occur – known as benchmark doses (BMD)

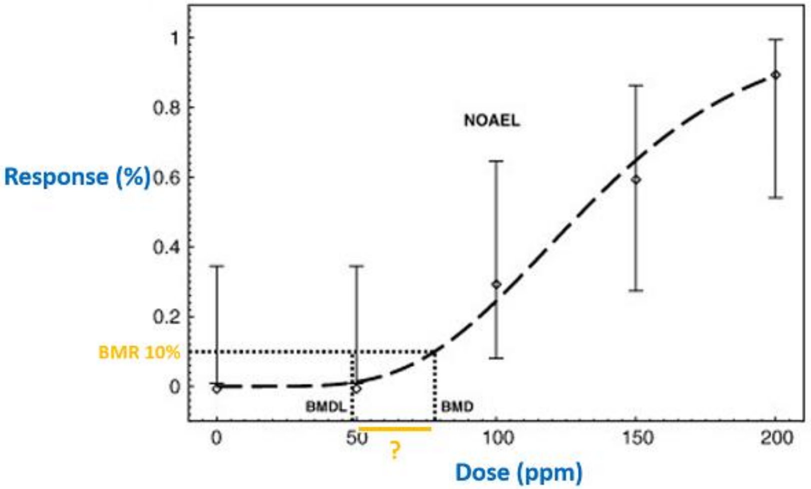
Dose response analysis is critical for evaluating “alternatives to animal testing”

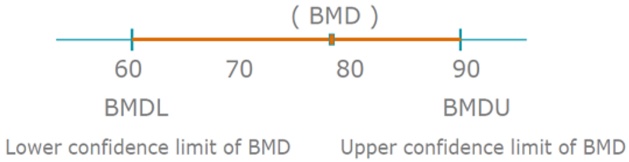
Dose response study design

1. Dose-response experimental designs typically include a control group (dose = 0) and at least three different dose groups, typically with the same number of replicates in each group.
2. To perform transcriptomics dose-response analysis, the data are processed and normalized according to standard protocols, and then differential analysis is used to identify genes that have a relationship with dose.
3. All genes that pass the DEA filters are used for dose-response curve fitting, in which a suite of linear and non-linear curves is fitted to the expression of each gene, and the best fit model for each gene is kept.
4. The curve is analyzed to determine the precise concentration at which the fitted curve departs from the expression values in the control group (called the gene benchmark dose, or BMD).

Benchmark dose (BMD) and Benchmark Response (BMR)

- A benchmark dose (BMD) is a dose or concentration that produces a predetermined change in the response rate of an adverse effect.
- This predetermined change in response is called the benchmark response (BMR). The default BMR is 5% (continuous) or 10% (dichotomous) change in the response rate of an adverse effect relative to the response of control group.
- Usually we use benchmark dose (lower confidence limit) (BMDL) to calculate human health guidance





Dose response curve fitting

Dose-response curves:

- X axis plots concentration (or dose) of a drug or chemical.
- Y axis plots response, such as gene expression levels
- Can assume many different shapes (models)

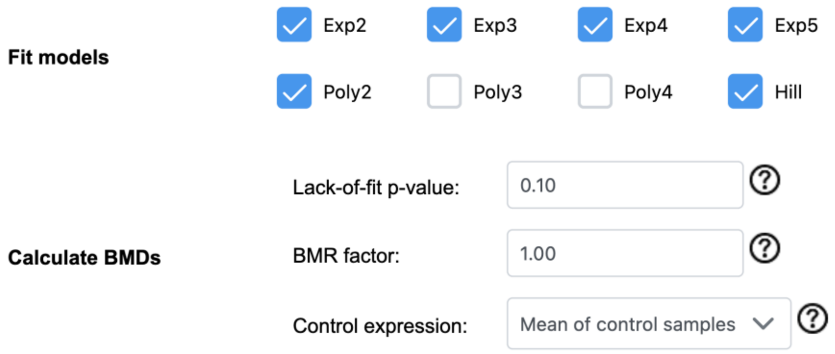
The response is often gene / feature specific (different genes respond differently to the same dose)

➔ Automated fitting procedure for dose-response curves

  - For each gene / feature
  - For each model

Model Fitting Process

- Current support 10 widely used models
- Use “Lack-of-fit p-value” to filter out those do not fit the given models (large p-values are good candidates, as it tests the opposite)



Fit models

☑ Exp2 ☑ Exp3 ☑ Exp4 ☑ Exp5 ☑ Linear

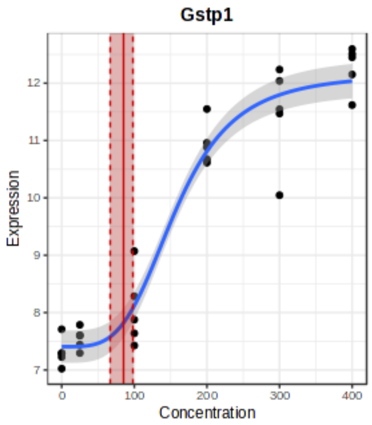
☑ Poly2 □ Poly3 □ Poly4 ☑ Hill ☑ Power

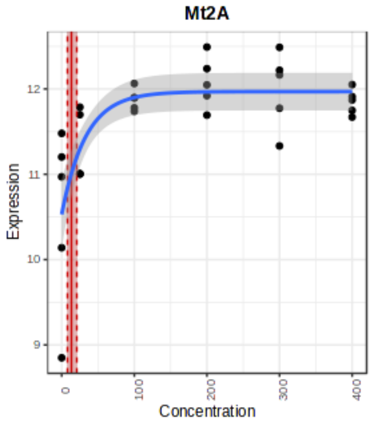
Calculate BMDs

Lack-of-fit p-value: 0.10 ?

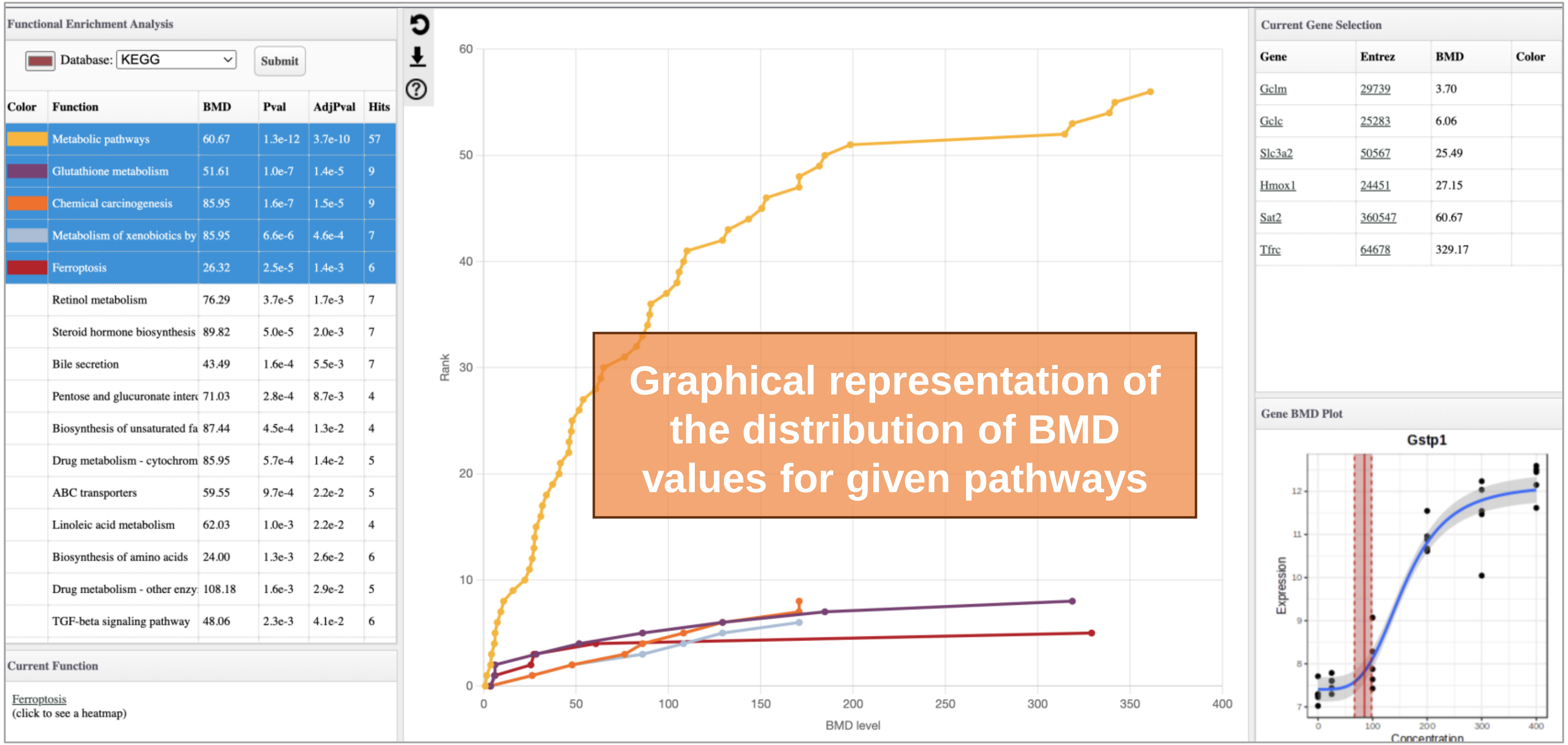
BMR factor: 1.00 ?

Control expression: Mean of control samples ?





Accumulation Curve



Functional Enrichment Analysis

🟥 Database: KEGG Submit

Color	Function	BMD	Pval	AdjPval	Hits
🟨	Metabolic pathways	60.67	1.3e-12	3.7e-10	57
🟪	Glutathione metabolism	51.61	1.0e-7	1.4e-5	9
🟧	Chemical carcinogenesis	85.95	1.6e-7	1.5e-5	9
🟦	Metabolism of xenobiotics by	85.95	6.6e-6	4.6e-4	7
🟥	Ferroptosis	26.32	2.5e-5	1.4e-3	6
	Retinol metabolism	76.29	3.7e-5	1.7e-3	7
	Steroid hormone biosynthesis	89.82	5.0e-5	2.0e-3	7
	Bile secretion	43.49	1.6e-4	5.5e-3	7
	Pentose and glucuronate interc	71.03	2.8e-4	8.7e-3	4
	Biosynthesis of unsaturated fa	87.44	4.5e-4	1.3e-2	4
	Drug metabolism - cytochrom	85.95	5.7e-4	1.4e-2	5
	ABC transporters	59.55	9.7e-4	2.2e-2	5
	Linoleic acid metabolism	62.03	1.0e-3	2.2e-2	4
	Biosynthesis of amino acids	24.00	1.3e-3	2.6e-2	6
	Drug metabolism - other enzy	108.18	1.6e-3	2.9e-2	5
	TGF-beta signaling pathway	48.06	2.3e-3	4.1e-2	6

Current Function

Ferroptosis
(click to see a heatmap)

Current Gene Selection

Gene	Entrez	BMD	Color
Gclm	29739	3.70	
Gclc	25283	6.06	
Slc3a2	50567	25.49	
Hmox1	24451	27.15	
Sat2	360547	60.67	
Tfrc	64678	329.17	

Gene BMD Plot

Graphical representation of the distribution of BMD values for given pathways

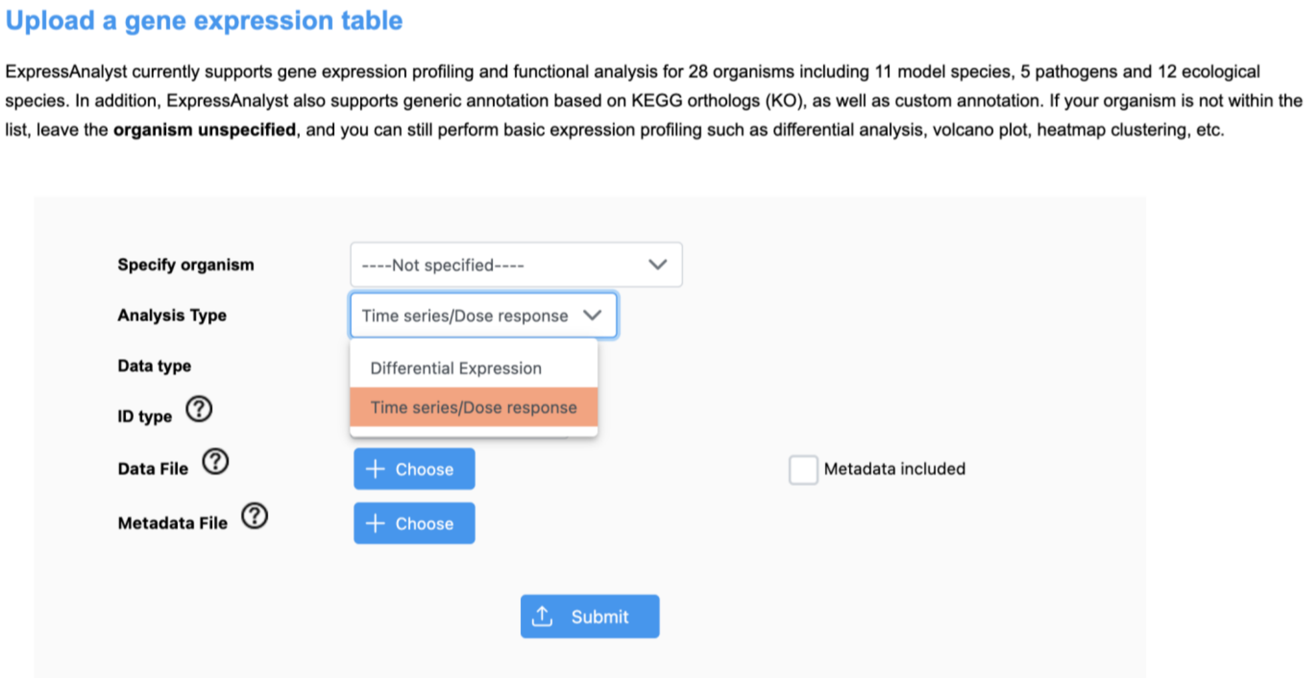
Analyzing data with ordered design (time, dose)

Both time series and dose responses require a series of experiments and require specific analysis (i.e., order) in addition to differential expression analysis, such as

- ❖ Time/dose profiling
- ❖ Cause-effect analysis
- ❖ Mode of action

➔ Visualization and modeling in addition to DEA are useful for both dose response and time series data

Live Demo (ExpressAnalyst)



Upload a gene expression table

ExpressAnalyst currently supports gene expression profiling and functional analysis for 28 organisms including 11 model species, 5 pathogens and 12 ecological species. In addition, ExpressAnalyst also supports generic annotation based on KEGG orthologs (KO), as well as custom annotation. If your organism is not within the list, leave the **organism unspecified**, and you can still perform basic expression profiling such as differential analysis, volcano plot, heatmap clustering, etc.

Specify organism ----Not specified----

Analysis Type Time series/Dose response

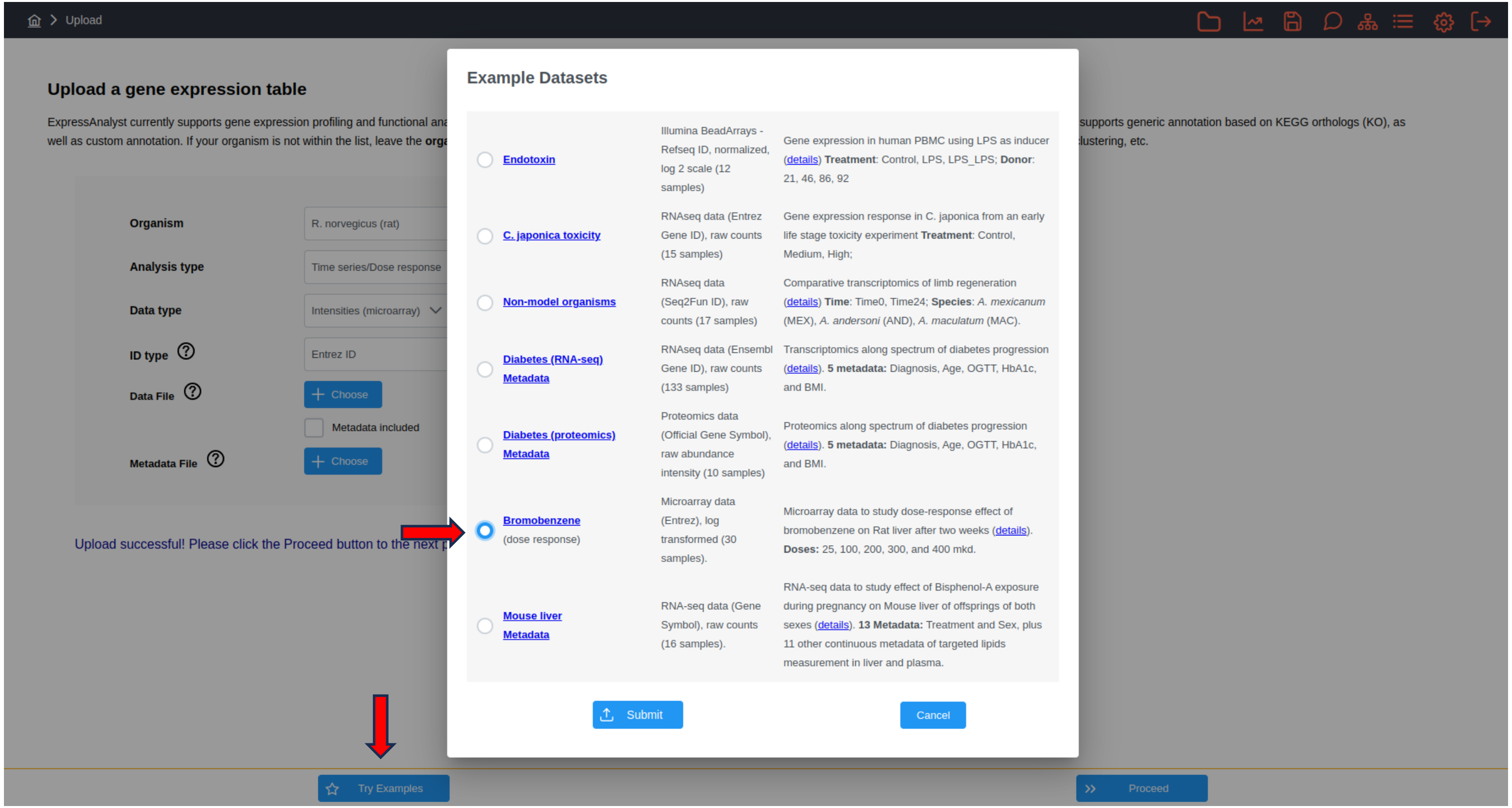
Data type

ID type

Data File

Metadata File

Dose response analysis – example data



Upload a gene expression table

ExpressAnalyst currently supports gene expression profiling and functional ana … supports generic annotation based on KEGG orthologs (KO), as well as custom annotation. If your organism is not within the list, leave the **orga** … lustering, etc.

Organism	R. norvegicus (rat)
Analysis type	Time series/Dose response
Data type	Intensities (microarray)
ID type	Entrez ID
Data File	+ Choose
	□ Metadata included
Metadata File	+ Choose

Upload successful! Please click the Proceed button to the next p

Example Datasets

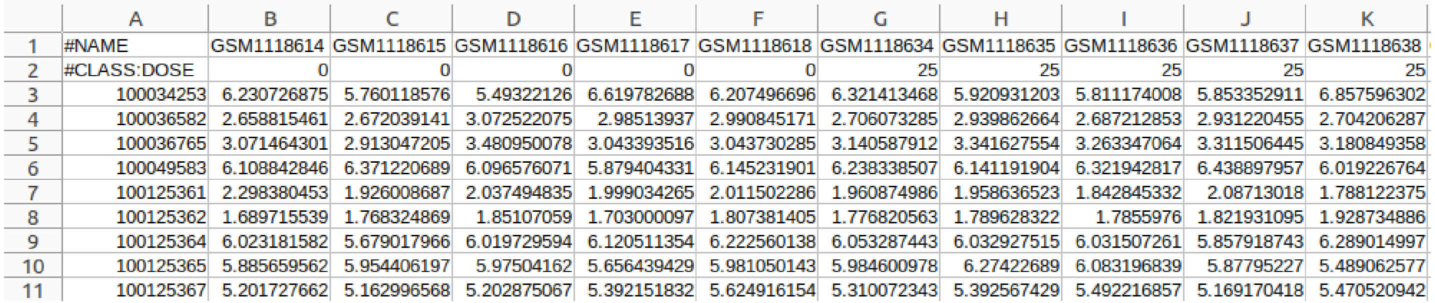
○	Endotoxin	Illumina BeadArrays - Refseq ID, normalized, log 2 scale (12 samples)	Gene expression in human PBMC using LPS as inducer (details) **Treatment**: Control, LPS, LPS_LPS; **Donor**: 21, 46, 86, 92
○	C. japonica toxicity	RNAseq data (Entrez Gene ID), raw counts (15 samples)	Gene expression response in C. japonica from an early life stage toxicity experiment **Treatment**: Control, Medium, High;
○	Non-model organisms	RNAseq data (Seq2Fun ID), raw counts (17 samples)	Comparative transcriptomics of limb regeneration (details) **Time**: Time0, Time24; **Species**: *A. mexicanum* (MEX), *A. andersoni* (AND), *A. maculatum* (MAC).
○	Diabetes (RNA-seq) Metadata	RNAseq data (Ensembl Gene ID), raw counts (133 samples)	Transcriptomics along spectrum of diabetes progression (details). **5 metadata:** Diagnosis, Age, OGTT, HbA1c, and BMI.
○	Diabetes (proteomics) Metadata	Proteomics data (Official Gene Symbol), raw abundance intensity (10 samples)	Proteomics along spectrum of diabetes progression (details). **5 metadata:** Diagnosis, Age, OGTT, HbA1c, and BMI.
●	Bromobenzene (dose response)	Microarray data (Entrez), log transformed (30 samples).	Microarray data to study dose-response effect of bromobenzene on Rat liver after two weeks (details). **Doses:** 25, 100, 200, 300, and 400 mkd.
○	Mouse liver Metadata	RNA-seq data (Gene Symbol), raw counts (16 samples).	RNA-seq data to study effect of Bisphenol-A exposure during pregnancy on Mouse liver of offsprings of both sexes (details). **13 Metadata:** Treatment and Sex, plus 11 other continuous metadata of targeted lipids measurement in liver and plasma.

Submit

Cancel

Dose response analysis – example data

Meta-data within Data table



	A	B	C	D	E	F	G	H	I	J	K
1	#NAME	GSM1118614	GSM1118615	GSM1118616	GSM1118617	GSM1118618	GSM1118634	GSM1118635	GSM1118636	GSM1118637	GSM1118638
2	#CLASS:DOSE	0	0	0	0	0	25	25	25	25	25
3	100034253	6.230726875	5.760118576	5.49322126	6.619782688	6.207496696	6.321413468	5.920931203	5.811174008	5.853352911	6.857596302
4	100036582	2.658815461	2.672039141	3.072522075	2.98513937	2.990845171	2.706073285	2.939862664	2.687212853	2.931220455	2.704206287
5	100036765	3.071464301	2.913047205	3.480950078	3.043393516	3.043730285	3.140587912	3.341627554	3.263347064	3.311506445	3.180849358
6	100049583	6.108842846	6.371220689	6.096576071	5.879404331	6.145231901	6.238338507	6.141191904	6.321942817	6.438897957	6.019226764
7	100125361	2.298380453	1.926008687	2.037494835	1.999034265	2.011502286	1.960874986	1.958636523	1.842845332	2.08713018	1.788122375
8	100125362	1.689715539	1.768324869	1.85107059	1.703000097	1.807381405	1.776820563	1.789628322	1.7855976	1.821931095	1.928734886
9	100125364	6.023181582	5.679017966	6.019729594	6.120511354	6.222560138	6.053287443	6.032927515	6.031507261	5.857918743	6.289014997
10	100125365	5.885659562	5.954406197	5.97504162	5.656439429	5.981050143	5.984600978	6.27422689	6.083196839	5.87795227	5.489062577
11	100125367	5.201727662	5.162996568	5.202875067	5.392151832	5.624916154	5.310072343	5.392567429	5.492216857	5.169170418	5.470520942

- Microarray data from Rat liver
- Study dose-response effect of Bromobenzene at concentration 0 (control), 25, 100, 200, 300, 400 mkd (mg/kg/day)
- Make sure at least **3 replicates** per dose/time point

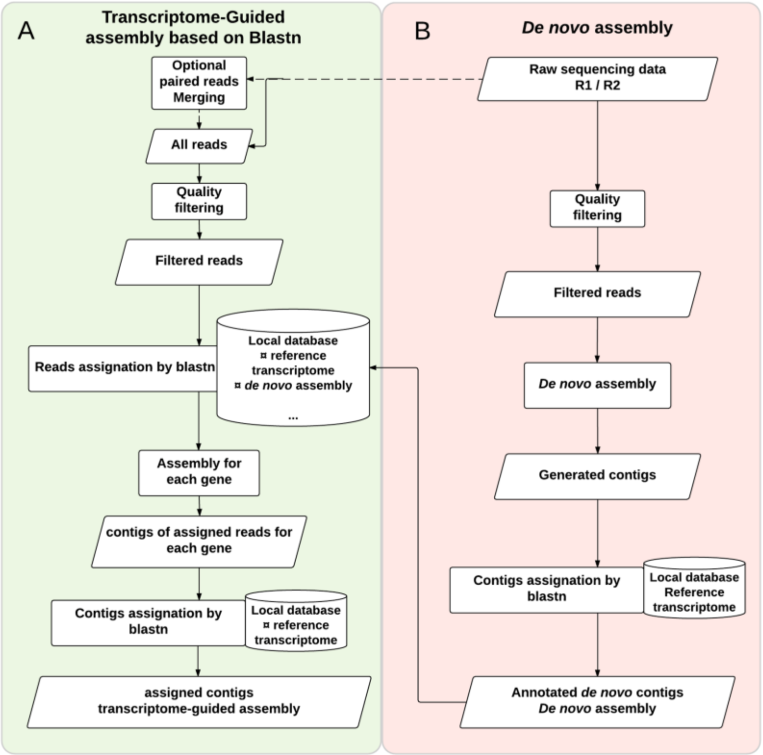
Schedule for today

Time	Topics
9:00 – 9:10	Introduction
9:10 – 9:40	Network analysis of gene list data & Live Demo
9:40 – 10.00	Dose response analysis & Live Demo
10:05 – 10:40	RNAseq in non-model species & Live Demo
10:45 – 11:25	Meta-analysis of multiple expression data & Live Demo
Summary and discussion	

RNAseq for non-model species

How to do RNAseq analysis without reference genome?

1. Perform transcriptome assembly and annotation
2. Mapping RNAseq reads to the annotated transcriptome
3. Perform regular gene expression analysis


A
Transcriptome-Guided assembly based on Blastn
Optional paired reads Merging
All reads
Quality filtering
Filtered reads
Reads assignation by blastn
Local database ¤ reference transcriptome ¤ *de novo* assembly …
Assembly for each gene
contigs of assigned reads for each gene
Contigs assignation by blastn
Local database ¤ reference transcriptome
assigned contigs transcriptome-guided assembly
B
De novo assembly
Raw sequencing data R1 / R2
Quality filtering
Filtered reads
De novo assembly
Generated contigs
Contigs assignation by blastn
Local database Reference transcriptome
Annotated *de novo* contigs *De novo* assembly

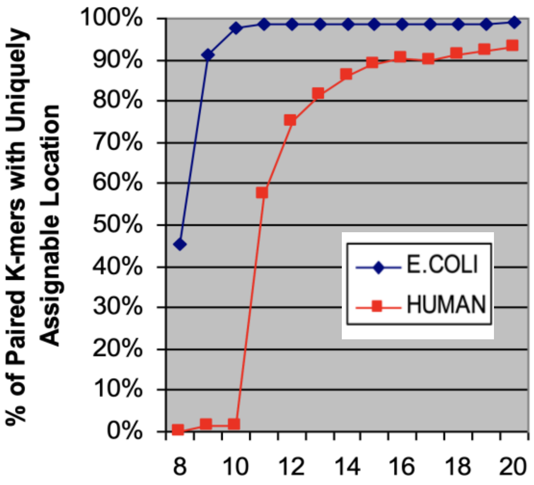
https://doi.org/10.1371/journal.pone.0185020

How do we annotate genome / transcriptome?

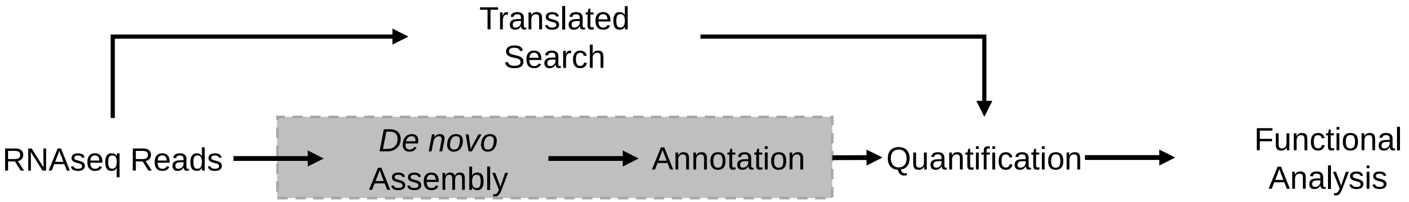
- **Blast2GO** uses BLAST searches to find similar sequences to one or several input sequences.
- It extracts the GO terms associated to each of the obtained hits and returns an evaluated GO annotation for the query sequence(s).
- NGS short reads should be first assembled into longer reads (i.e. contigs).
- Taking very long time to run ….

http://docs.blast2go.com/user-manual/

Need new-generation algorithm for short-reads



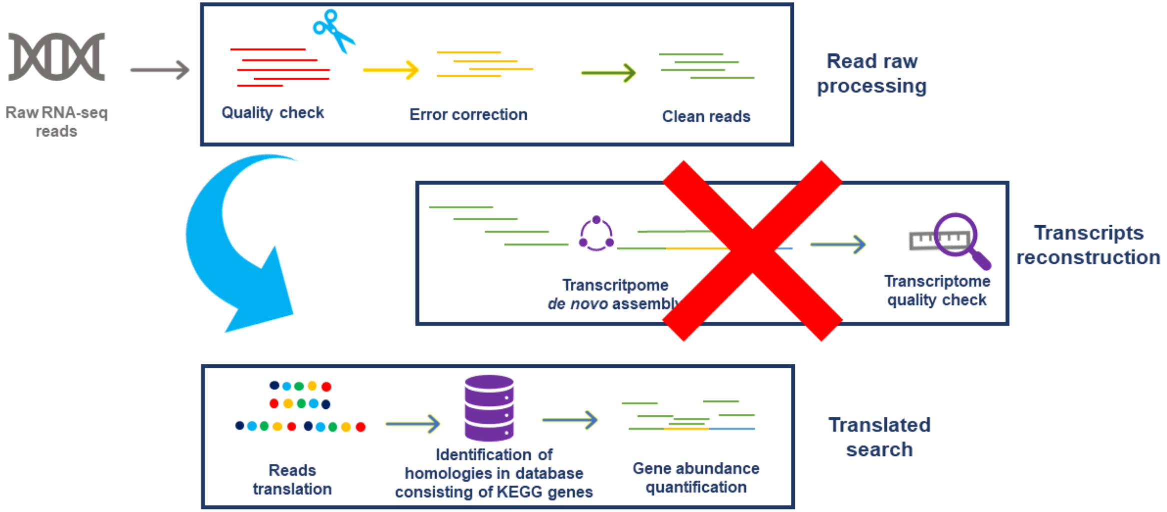
Can we directly detect orthologs and quantify them from NGS short-reads data without performing assembly?



Translated Search

RNAseq Reads → *De novo* Assembly → Annotation → Quantification → Functional Analysis

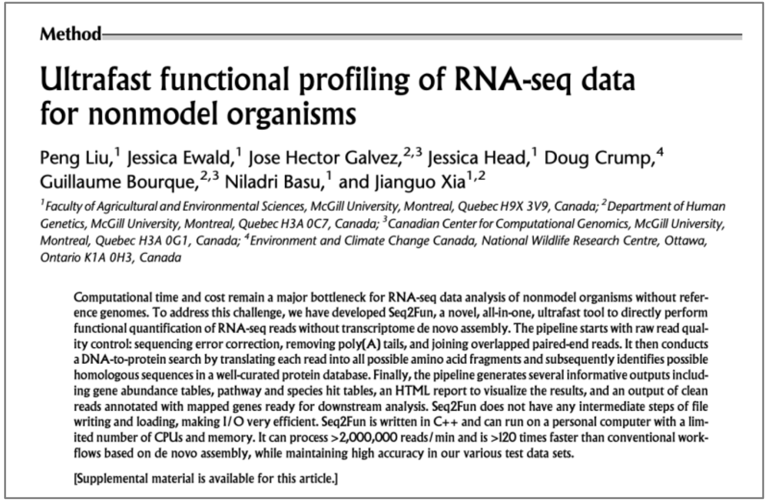
Reference-free RNAseq analysis


Raw RNA-seq reads
Quality check
Error correction
Clean reads
Read raw processing
Transcriptome de novo assembly
Transcriptome quality check
Transcripts reconstruction
Reads translation
Identification of homologies in database consisting of KEGG genes
Gene abundance quantification
Translated search

No assembly. Performing mapping & annotation simultaneously

Seq2Fun: direct annotation & quantification without transcriptome assembly

- **Ultra-fast:** Seq2Fun is > 120 times faster (~ 2 million reads / minute) than the conventional RNA-seq workflow.
- **Extremely low memory cost:** Seq2Fun consumes as little as ~2.27 GB memory
- **Reference-free:** Seq2Fun does not require the genome or transcriptome reference of the organism



Method

Ultrafast functional profiling of RNA-seq data for nonmodel organisms

Peng Liu,[1] Jessica Ewald,[1] Jose Hector Galvez,[2,3] Jessica Head,[1] Doug Crump,[4] Guillaume Bourque,[2,3] Niladri Basu,[1] and Jianguo Xia[1,2]

[1] Faculty of Agricultural and Environmental Sciences, McGill University, Montreal, Quebec H9X 3V9, Canada; [2] Department of Human Genetics, McGill University, Montreal, Quebec H3A 0C7, Canada; [3] Canadian Center for Computational Genomics, McGill University, Montreal, Quebec H3A 0G1, Canada; [4] Environment and Climate Change Canada, National Wildlife Research Centre, Ottawa, Ontario K1A 0H3, Canada

Computational time and cost remain a major bottleneck for RNA-seq data analysis of nonmodel organisms without reference genomes. To address this challenge, we have developed Seq2Fun, a novel, all-in-one, ultrafast tool to directly perform functional quantification of RNA-seq reads without transcriptome de novo assembly. The pipeline starts with raw read quality control: sequencing error correction, removing poly(A) tails, and joining overlapped paired-end reads. It then conducts a DNA-to-protein search by translating each read into all possible amino acid fragments and subsequently identifies possible homologous sequences in a well-curated protein database. Finally, the pipeline generates several informative outputs including gene abundance tables, pathway and species hit tables, an HTML report to visualize the results, and an output of clean reads annotated with mapped genes ready for downstream analysis. Seq2Fun does not have any intermediate steps of file writing and loading, making I/O very efficient. Seq2Fun is written in C++ and can run on a personal computer with a limited number of CPUs and memory. It can process >2,000,000 reads/min and is >120 times faster than conventional workflows based on de novo assembly, while maintaining high accuracy in our various test data sets.

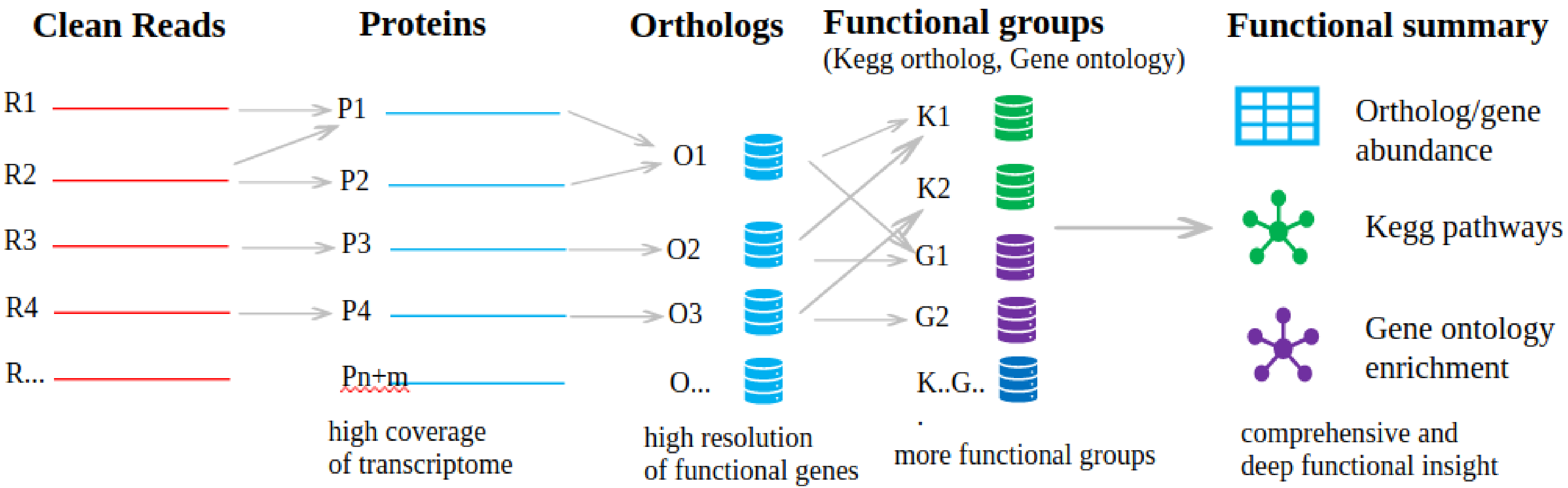
[Supplemental material is available for this article.]

Genome Research (2021)

The main idea behind

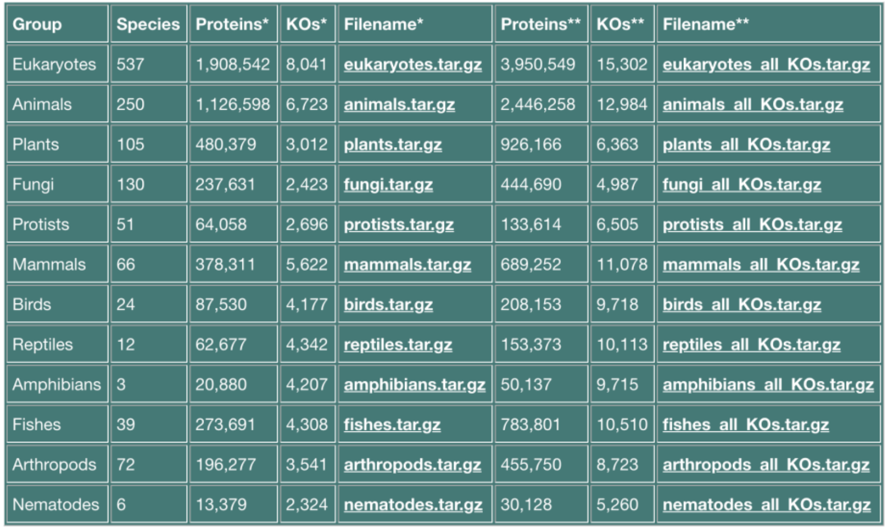
- Translated search (protein level not DNA level for ortholog detection)
  - 100 bp => 33 a.a. max
  - Matching score based on BLOSUM62 to consider evolutionary distance
- Two modes
  - Maximum exact match (MEM) mode: allows exact matches between query and reference sequences for organisms that have very closely related species in the database.
  - Greedy mode: allows mismatches between query and reference sequences to help overcome evolutionary divergence for organisms that do not have a closely related reference genome in the database

Directly from reads to functions

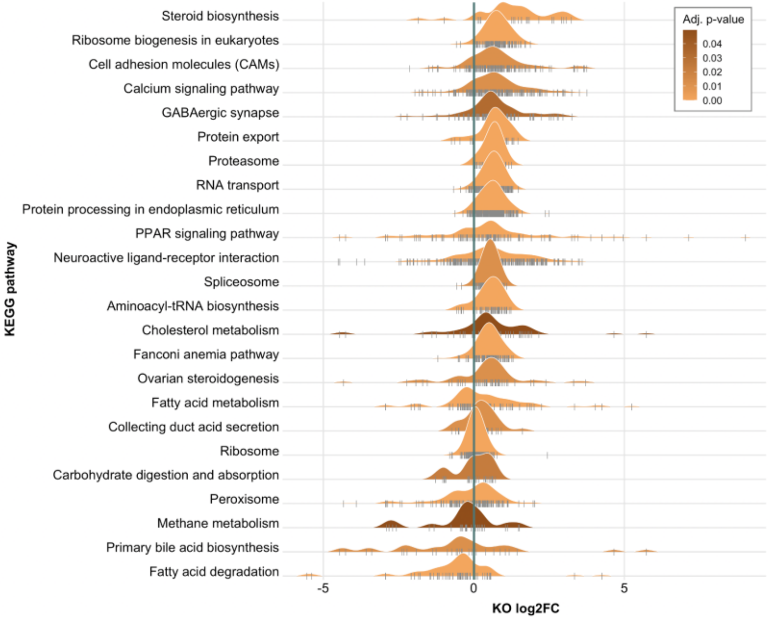


Seq2Fun v2.0

RNAseq analysis across domains of life



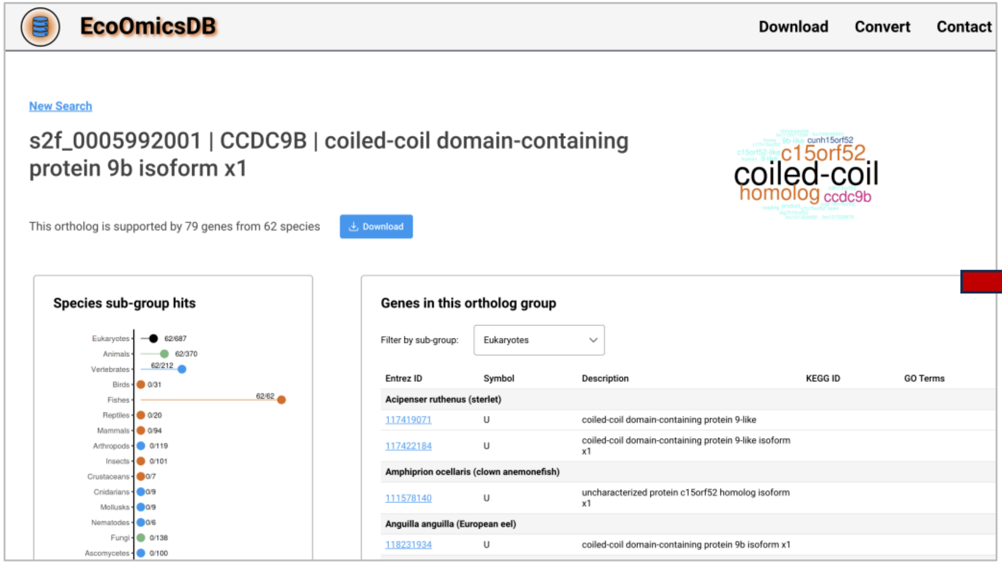
Group	Species	Proteins*	KOs*	Filename*	Proteins**	KOs**	Filename**
Eukaryotes	537	1,908,542	8,041	eukaryotes.tar.gz	3,950,549	15,302	eukaryotes_all_KOs.tar.gz
Animals	250	1,126,598	6,723	animals.tar.gz	2,446,258	12,984	animals_all_KOs.tar.gz
Plants	105	480,379	3,012	plants.tar.gz	926,166	6,363	plants_all_KOs.tar.gz
Fungi	130	237,631	2,423	fungi.tar.gz	444,690	4,987	fungi_all_KOs.tar.gz
Protists	51	64,058	2,696	protists.tar.gz	133,614	6,505	protists_all_KOs.tar.gz
Mammals	66	378,311	5,622	mammals.tar.gz	689,252	11,078	mammals_all_KOs.tar.gz
Birds	24	87,530	4,177	birds.tar.gz	208,153	9,718	birds_all_KOs.tar.gz
Reptiles	12	62,677	4,342	reptiles.tar.gz	153,373	10,113	reptiles_all_KOs.tar.gz
Amphibians	3	20,880	4,207	amphibians.tar.gz	50,137	9,715	amphibians_all_KOs.tar.gz
Fishes	39	273,691	4,308	fishes.tar.gz	783,801	10,510	fishes_all_KOs.tar.gz
Arthropods	72	196,277	3,541	arthropods.tar.gz	455,750	8,723	arthropods_all_KOs.tar.gz
Nematodes	6	13,379	2,324	nematodes.tar.gz	30,128	5,260	nematodes_all_KOs.tar.gz



XiaLab.ca

Empowering researchers through trainings, tools, and AI

EcoOmicsDB for viewing evidence



New Search

s2f_0005992001 | CCDC9B | coiled-coil domain-containing protein 9b isoform x1

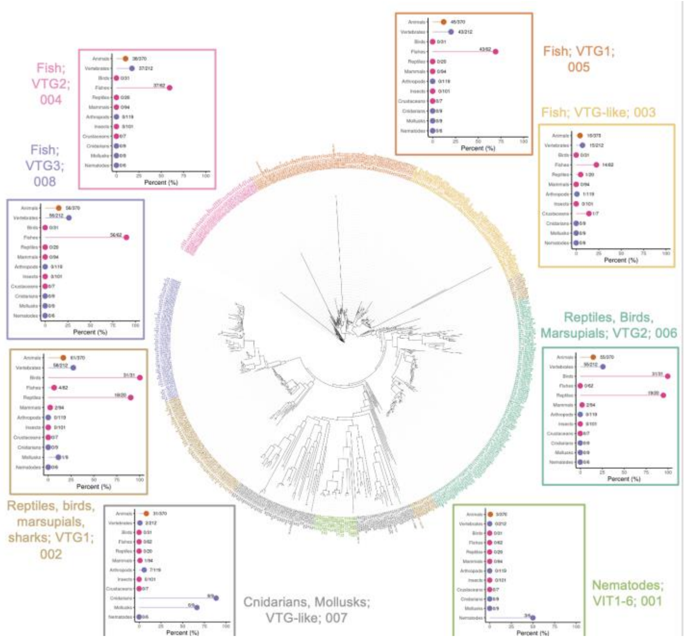
This ortholog is supported by 79 genes from 62 species

Species sub-group hits

Genes in this ortholog group

Filter by sub-group: Eukaryotes

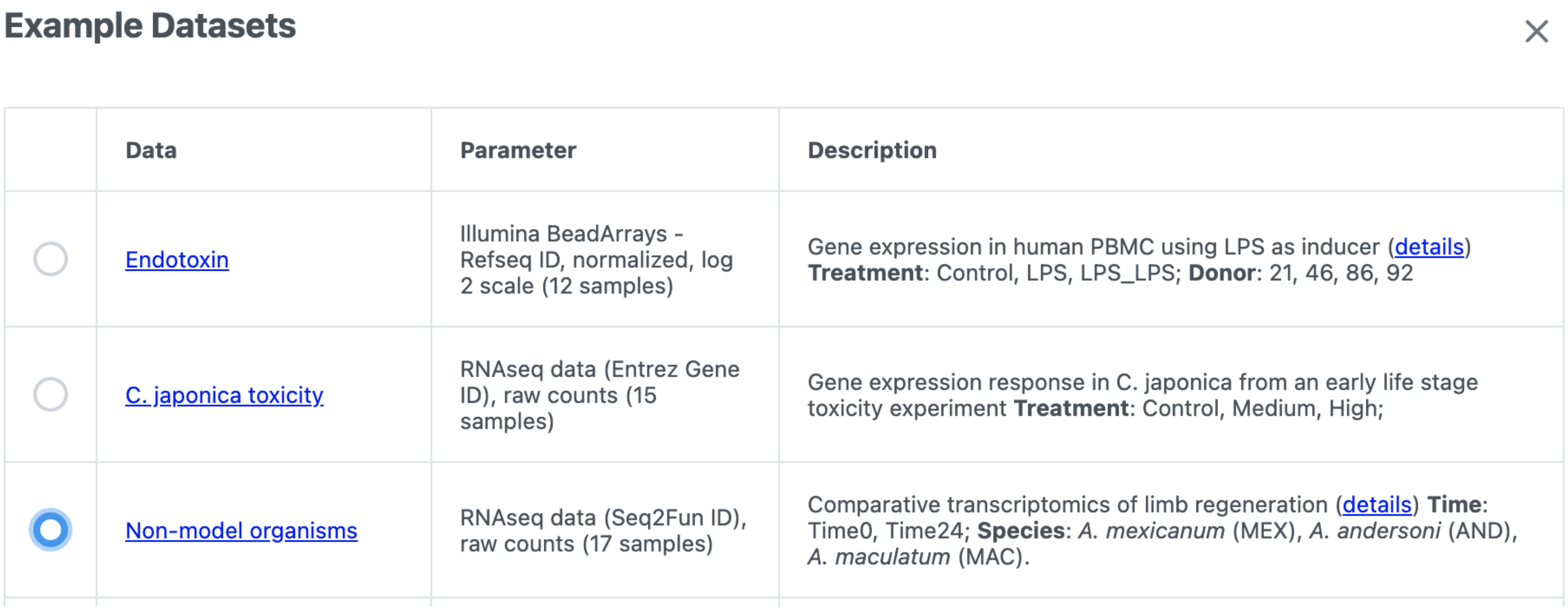
Entrez ID	Symbol	Description	KEGG ID	GO Terms
Acipenser ruthenus (sterlet)				
117419071	U	coiled-coil domain-containing protein 9-like		
117422184	U	coiled-coil domain-containing protein 9-like isoform x1		
Amphiprion ocellaris (clown anemonefish)				
111578140	U	uncharacterized protein c15orf52 homolog isoform x1		
Anguilla anguilla (European eel)				
118231934	U	coiled-coil domain-containing protein 9b isoform x1		



Summary of RNAseq for non-model species

1. Perform RNASeq reads mapping using Seq2Fun (ExpressAnalyst Docker)
2. Upload the data table together with metadata table to ExpressAnalyst
3. Perform different statistical analysis, visualization and functional enrichment analysis (KEGG and GO)
4. Inspecting the specific-specific evidence in EcoOmicsDB for key annotated genes

Live Demo (ExpressAnalyst)



Example Datasets

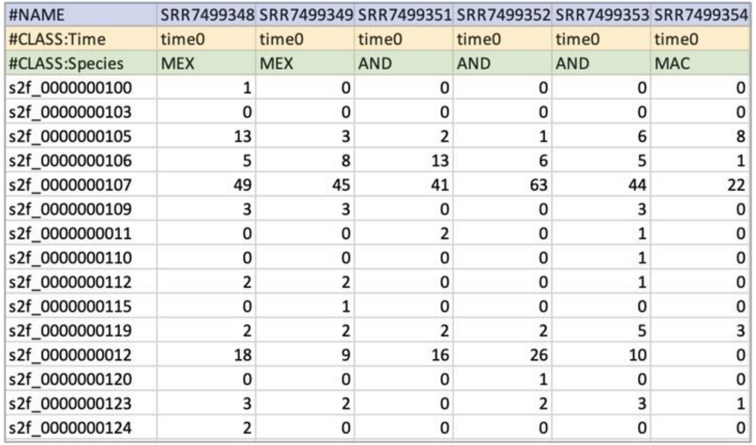
	Data	Parameter	Description
○	Endotoxin	Illumina BeadArrays - Refseq ID, normalized, log 2 scale (12 samples)	Gene expression in human PBMC using LPS as inducer (details) **Treatment**: Control, LPS, LPS_LPS; **Donor**: 21, 46, 86, 92
○	C. japonica toxicity	RNAseq data (Entrez Gene ID), raw counts (15 samples)	Gene expression response in C. japonica from an early life stage toxicity experiment **Treatment**: Control, Medium, High;
●	Non-model organisms	RNAseq data (Seq2Fun ID), raw counts (17 samples)	Comparative transcriptomics of limb regeneration (details) **Time**: Time0, Time24; **Species**: *A. mexicanum* (MEX), *A. andersoni* (AND), *A. maculatum* (MAC).

Dataset overview

- Study gene signature associated with tissue regeneration across three different salamander species
- Amputate a limb and get tissue sample at time point 0 and 24 hours
- 3 replicates for each species/time point for a total of 18 samples.

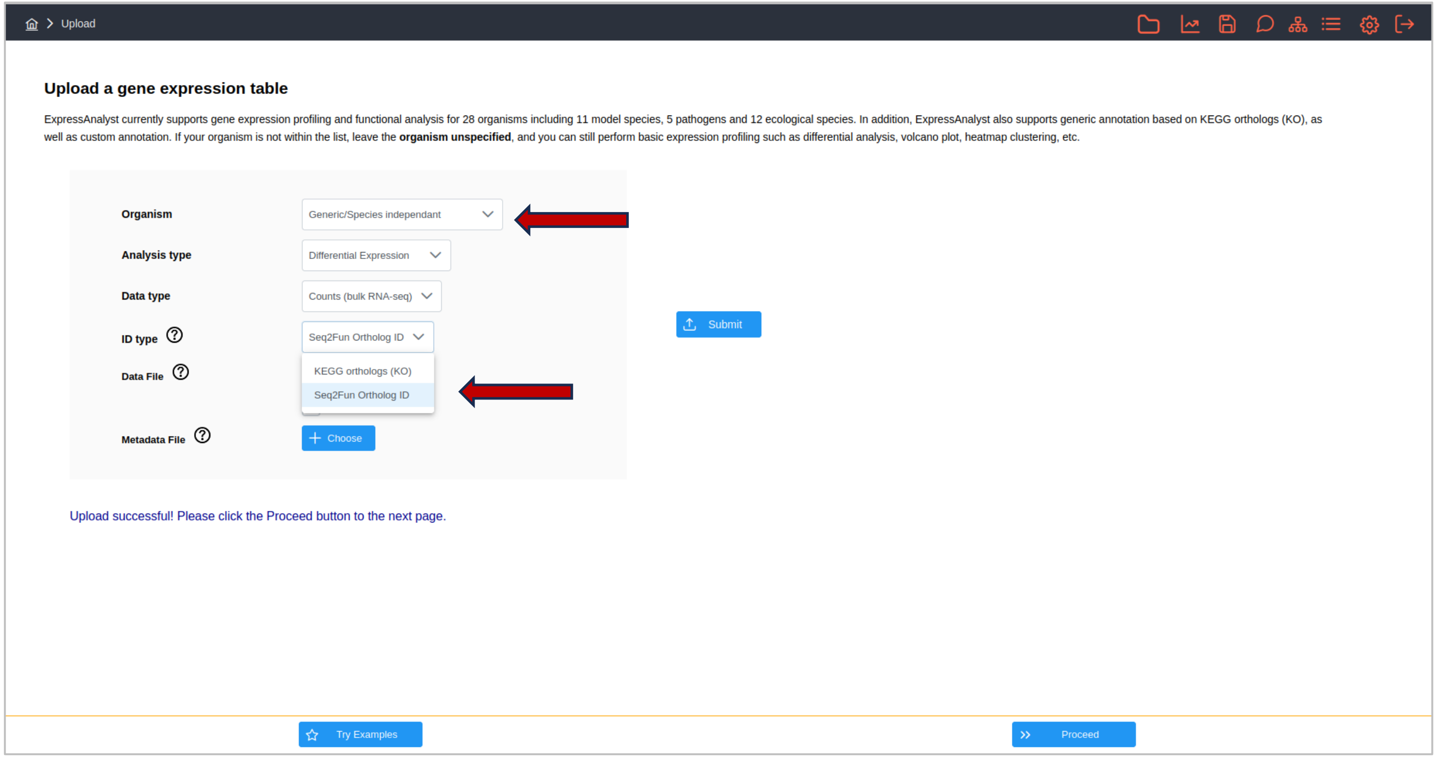
Sample IDs labeled with **#NAME**

Each metadata labeled with **#CLASS:** followed by the variable name



#NAME	SRR7499348	SRR7499349	SRR7499351	SRR7499352	SRR7499353	SRR7499354
#CLASS:Time	time0	time0	time0	time0	time0	time0
#CLASS:Species	MEX	MEX	AND	AND	AND	MAC
s2f_0000000100	1	0	0	0	0	0
s2f_0000000103	0	0	0	0	0	0
s2f_0000000105	13	3	2	1	6	8
s2f_0000000106	5	8	13	6	5	1
s2f_0000000107	49	45	41	63	44	22
s2f_0000000109	3	3	0	0	3	0
s2f_0000000011	0	0	2	0	1	0
s2f_0000000110	0	0	0	0	1	0
s2f_0000000112	2	2	0	0	1	0
s2f_0000000115	0	1	0	0	0	0
s2f_0000000119	2	2	2	2	5	3
s2f_0000000012	18	9	16	26	10	0
s2f_0000000120	0	0	0	1	0	0
s2f_0000000123	3	2	0	2	3	1
s2f_0000000124	2	0	0	0	0	0

Data upload



Upload a gene expression table

ExpressAnalyst currently supports gene expression profiling and functional analysis for 28 organisms including 11 model species, 5 pathogens and 12 ecological species. In addition, ExpressAnalyst also supports generic annotation based on KEGG orthologs (KO), as well as custom annotation. If your organism is not within the list, leave the **organism unspecified**, and you can still perform basic expression profiling such as differential analysis, volcano plot, heatmap clustering, etc.

Organism	Generic/Species independant
Analysis type	Differential Expression
Data type	Counts (bulk RNA-seq)
ID type	Seq2Fun Ortholog ID
Data File	
Metadata File	

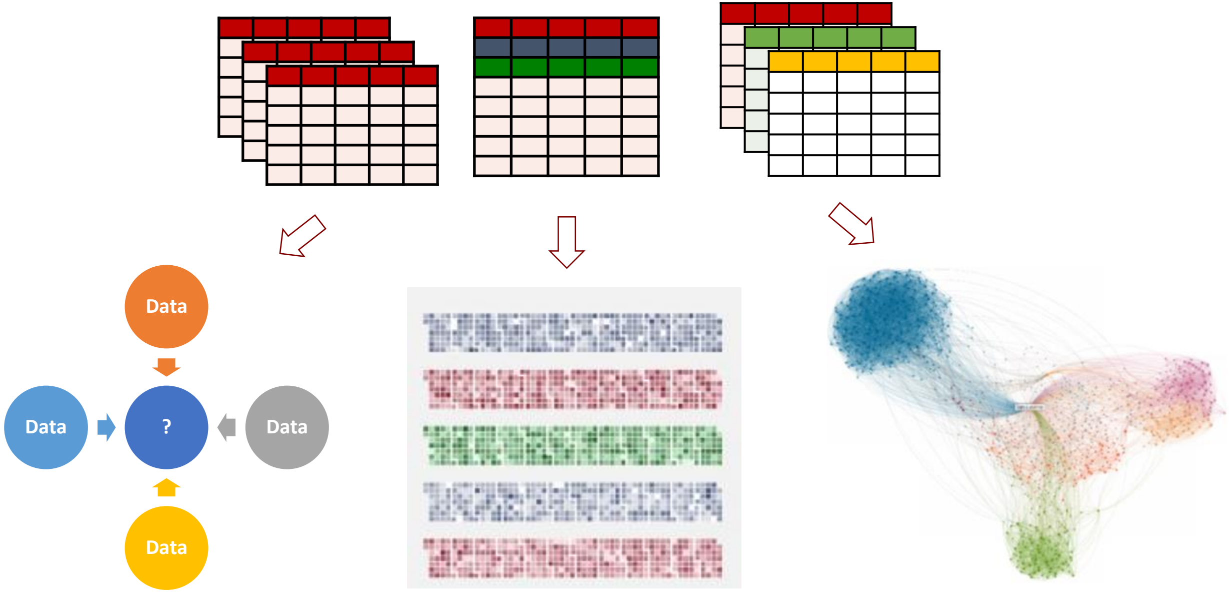
Upload successful! Please click the Proceed button to the next page.

Schedule for today

Time	Topics
9:00 – 9:10	Introduction
9:10 – 9:40	Network analysis of gene list data & Live Demo
9:40 – 10.00	Dose response analysis & Live Demo
10:05 – 10:40	RNAseq in non-model species & Live Demo
10:45 – 11:25	Meta-analysis of multiple expression data & Live Demo
Summary and discussion	

Meta-analysis: Integrate multiple evidence for decision making

Omics Data
+
Metadata



Statistical integration

Visual integration

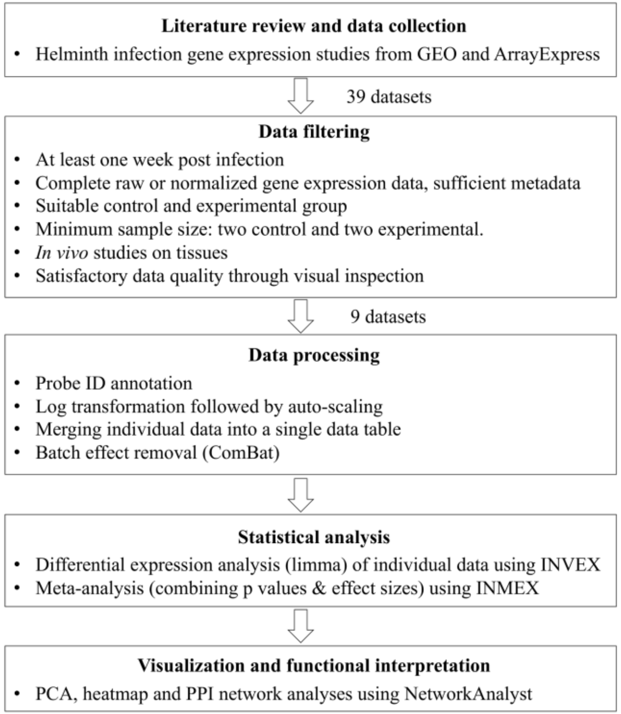
Network integration

Statistical Meta-analysis

Motivation: to improve statistical power by increasing sample size, reduce potential bias

1. Identify studies of interest
   - Different gene expression data generated by different groups to study the same phenotype
2. Determine eligibility of studies
   - Inclusion: which ones to keep
   - Exclusion: which ones to throw out
3. Obtain data from the studies
4. Analyze data in the studies statistically
5. Functionally interpret the significant features identified from meta-analysis

Examples: transcriptomics (I)



Literature review and data collection

- Helminth infection gene expression studies from GEO and ArrayExpress

39 datasets

Data filtering

- At least one week post infection
- Complete raw or normalized gene expression data, sufficient metadata
- Suitable control and experimental group
- Minimum sample size: two control and two experimental.
- *In vivo* studies on tissues
- Satisfactory data quality through visual inspection

9 datasets

Data processing

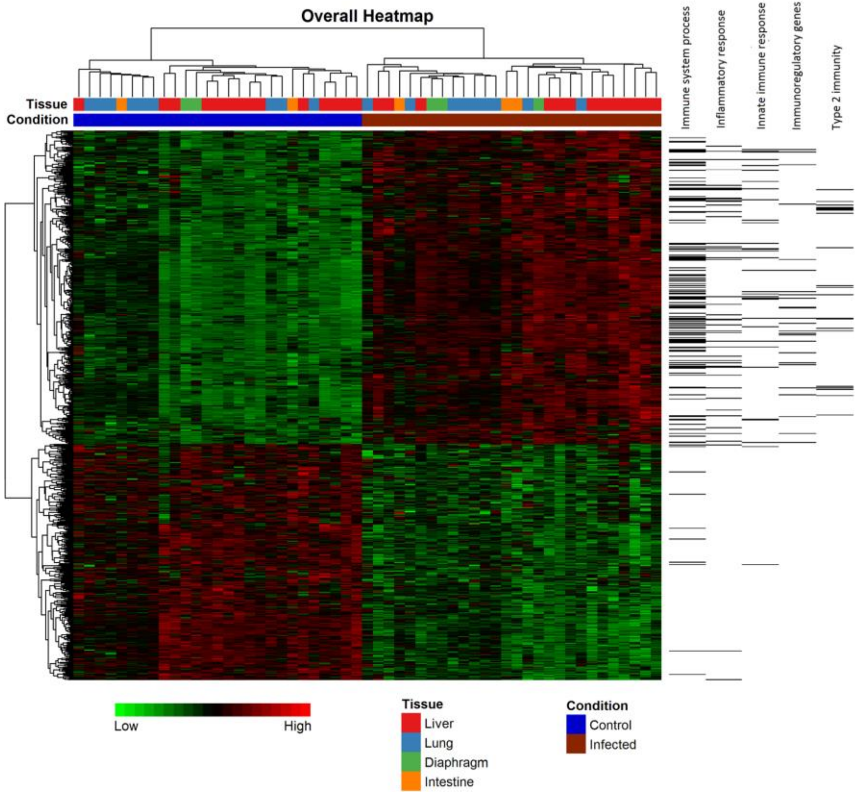
- Probe ID annotation
- Log transformation followed by auto-scaling
- Merging individual data into a single data table
- Batch effect removal (ComBat)

Statistical analysis

- Differential expression analysis (limma) of individual data using INVEX
- Meta-analysis (combining p values & effect sizes) using INMEX

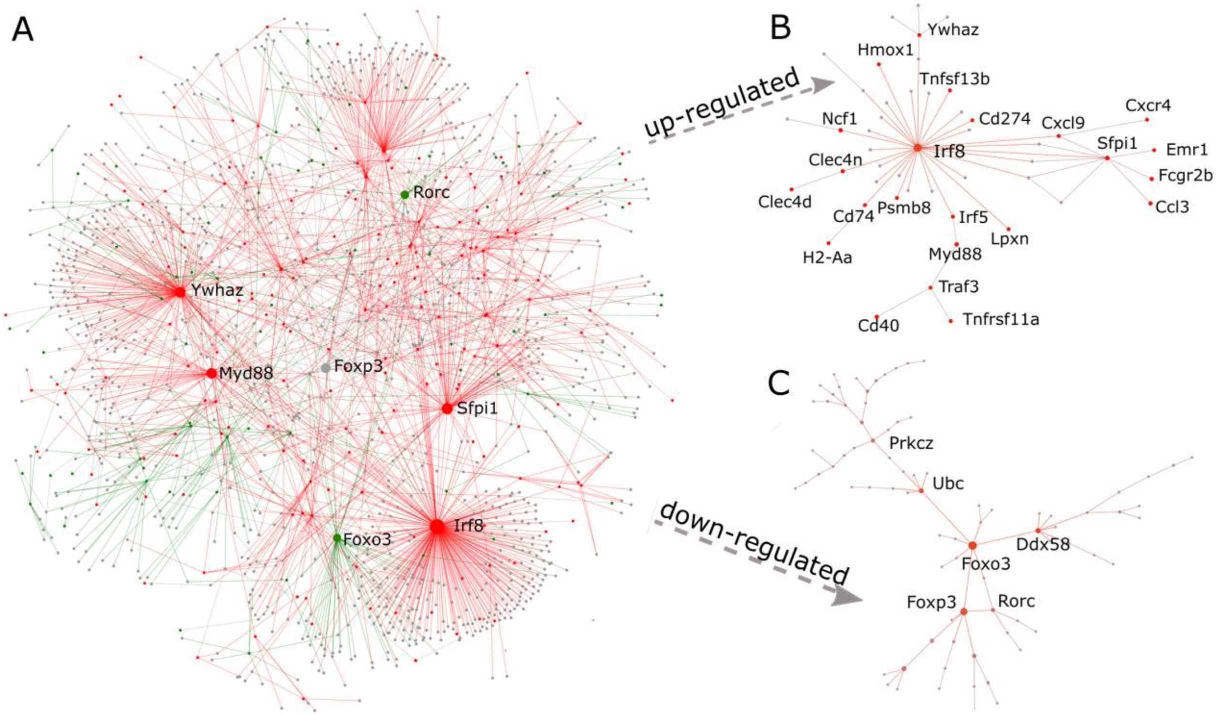
Visualization and functional interpretation

- PCA, heatmap and PPI network analyses using NetworkAnalyst

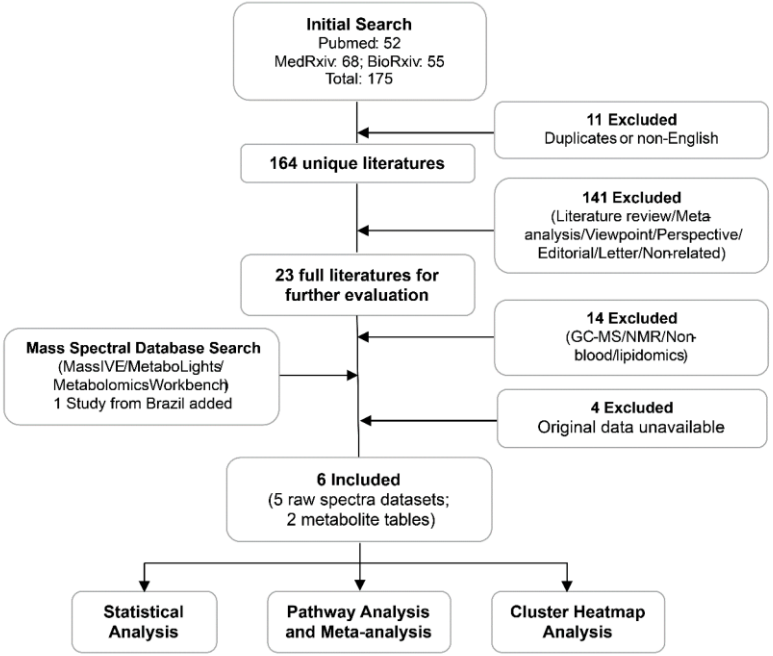


https://doi.org/10.1371/journal.pntd.0004624

Examples: transcriptomics (II)



Examples: metabolomics (I)



Initial Search
Pubmed: 52
MedRxiv: 68; BioRxiv: 55
Total: 175

11 Excluded
Duplicates or non-English

164 unique literatures

141 Excluded
(Literature review/Meta-analysis/Viewpoint/Perspective/Editorial/Letter/Non-related)

23 full literatures for further evaluation

14 Excluded
(GC-MS/NMR/Non-blood/lipidomics)

Mass Spectral Database Search
(MassIVE/MetaboLights/MetabolomicsWorkbench)
1 Study from Brazil added

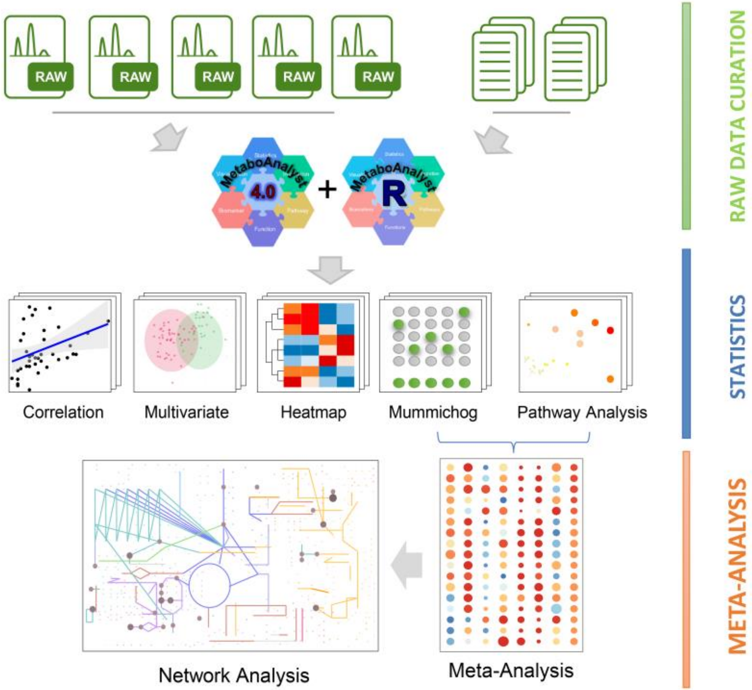
4 Excluded
Original data unavailable

6 Included
(5 raw spectra datasets;
2 metabolite tables)

Statistical Analysis

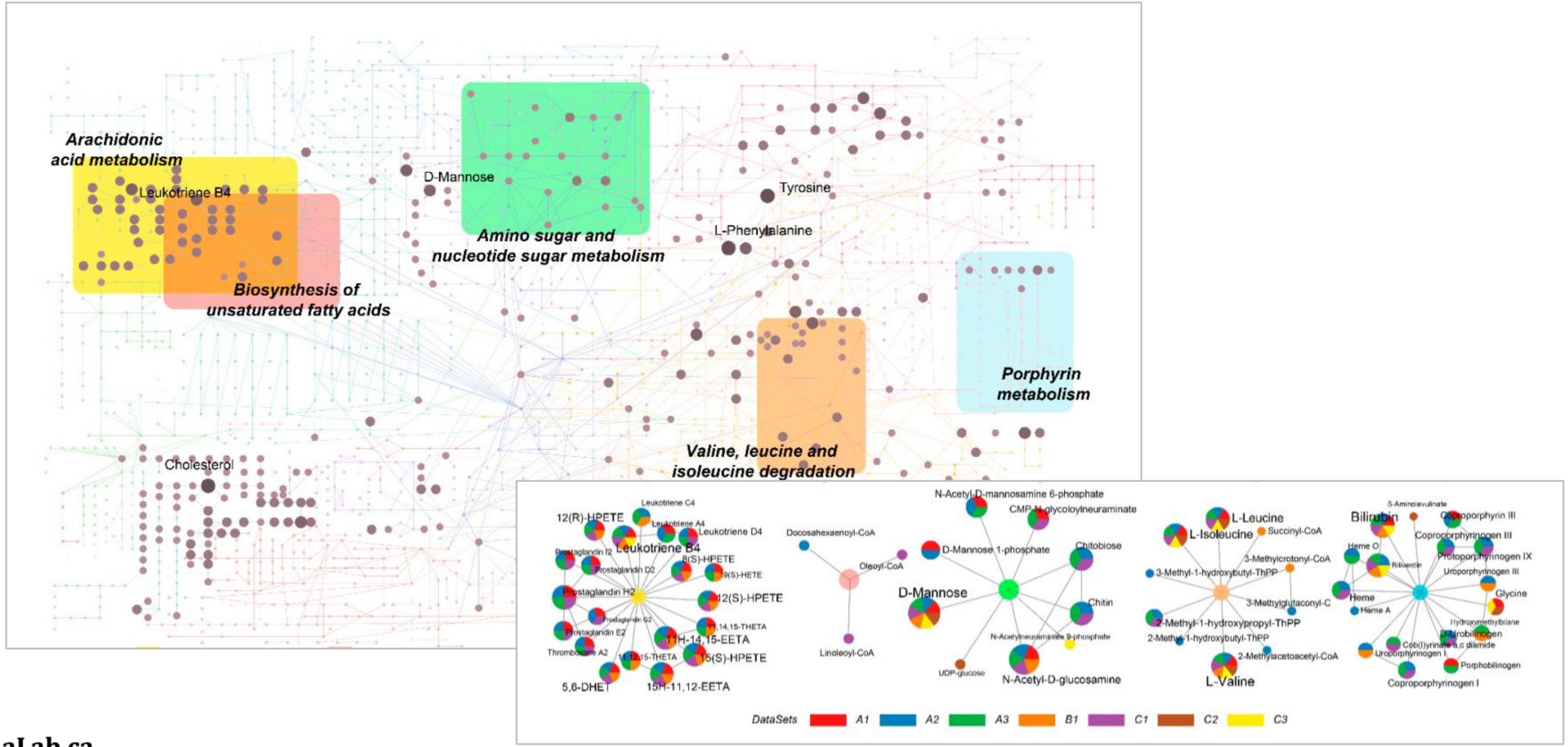
Pathway Analysis and Meta-analysis

Cluster Heatmap Analysis



Metabolites 2021, *11*(1), 44; https://doi.org/10.3390/metabo11010044

Examples: metabolomics (II)



Common methods for statistical integration

1. Combining p-values
2. Combining effect sizes
3. Vote counting,
   - Simply counts the number of datasets that a gene is differentially expressed and keeps those genes that pass a specified count threshold;
4. Direct merging
   - Concatenate different data into a big matrix

The first two are recommended, as they use summary statistics that are less susceptible to study-specific effects.

Summary statistics: p values & effect size

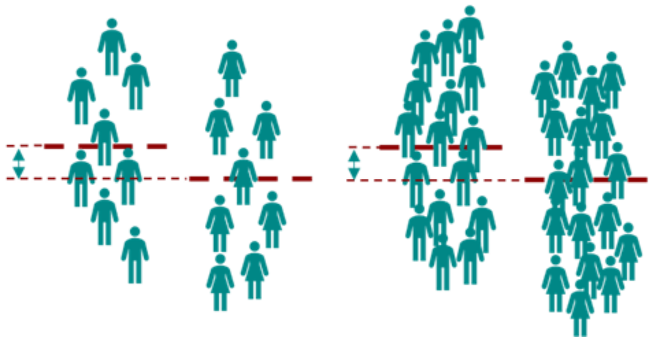
P values: says nothing about the size of the effect or difference and depends much on the sample size

✓ Any small difference can become significant when sample size are large enough

Effect size: standardized mean difference (SMD) between two populations

✓ SMD values of 0.2 to 0.5 are considered small, 0.5 to 0.8 are considered medium, and greater than 0.8 are considered large

Same difference different p-value



Mean value of sample 1

Mean value of sample 2

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}} \rightarrow d = t\sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}$$

Estimated standard deviation of the population from the sample

t value from the independent t-test

Size of the first sample

Size of the second sample

https://datatab.net/tutorial/effect-size-independent-t-test

Combining p values

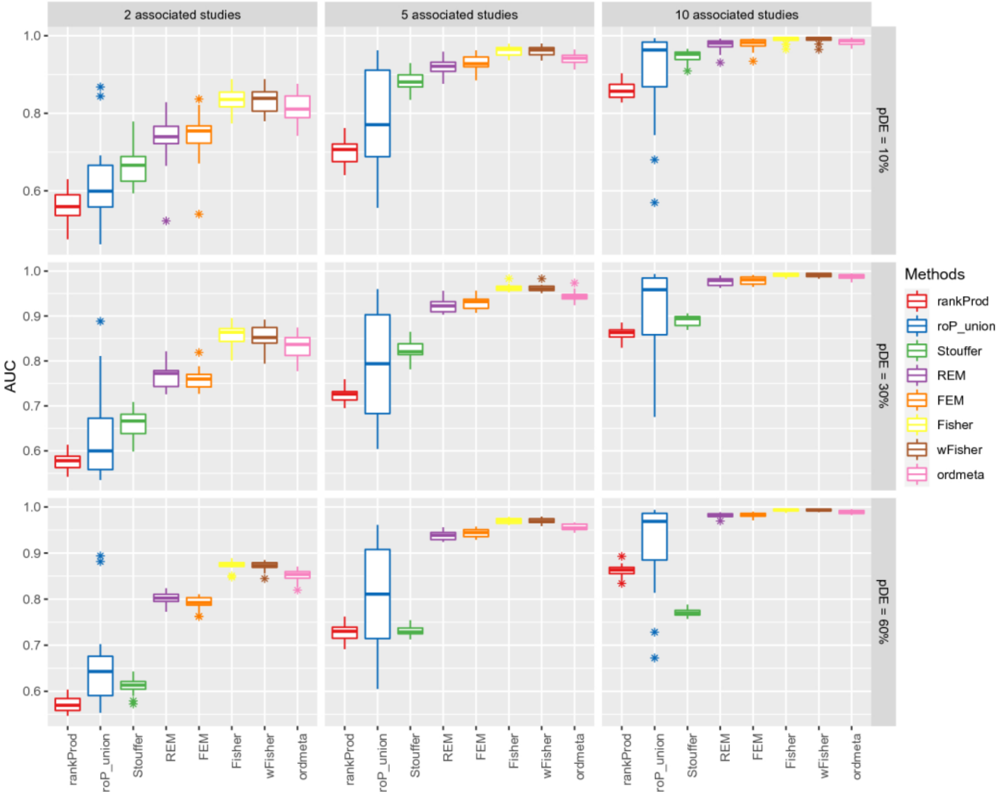
- Fisher's method is weight-free (-2*∑Log(p))
- Stouffer's method
  - incorporates weight (i.e. based on sample sizes) into the calculation;

➢ However, larger sample size does not warrant larger weights as the quality of each study can be variable.

➢ Users should choose to apply Stouffer’s method only when all studies are of similar qualities (i.e. same platform with similar levels of missing values).

➢ Fisher’s method works generally well

Benchmarking on p-value combination



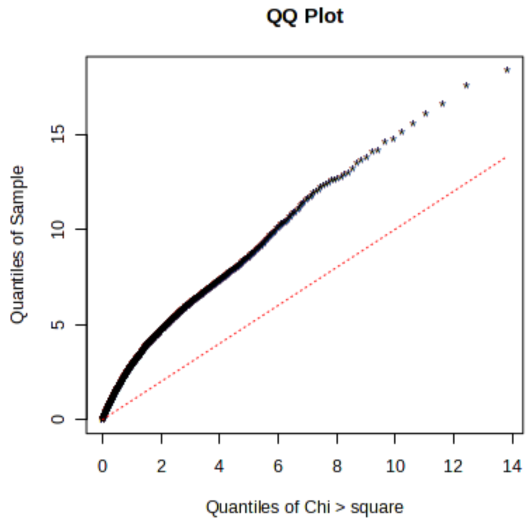
https://www.nature.com/articles/s41598-021-86465-y/figures/3

Effect Size and Cochran’s Q tests

- Fixed effect model (FEM)
  - The estimated effect size in each study is assumed to come from an underlying true effect size plus measurement error.
- Random effect model (REM),
  - Each study further contains a random effect that can incorporate unknown cross-study heterogeneities in the model (i.e. due to different platforms).

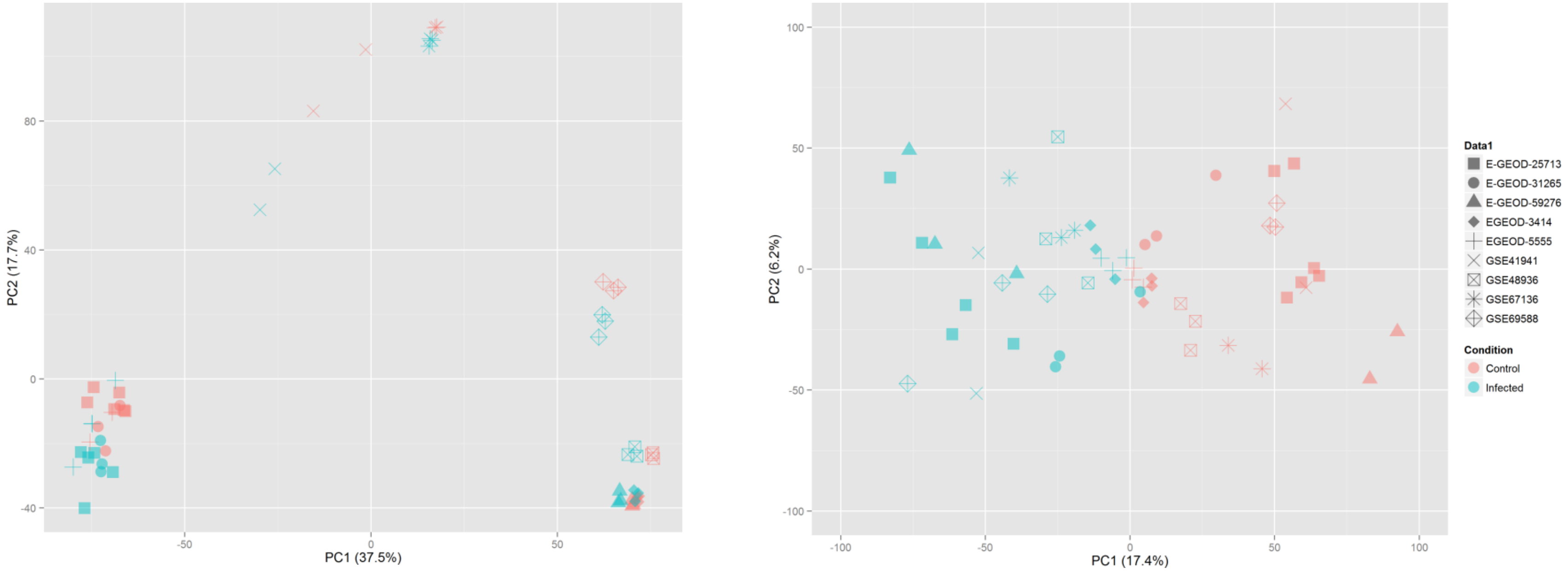
The method usually gives more conservative results (less DE features but more confident).

FEM/REM can be selected based on statistical heterogeneity estimated using **Cochran’s Q tests**.

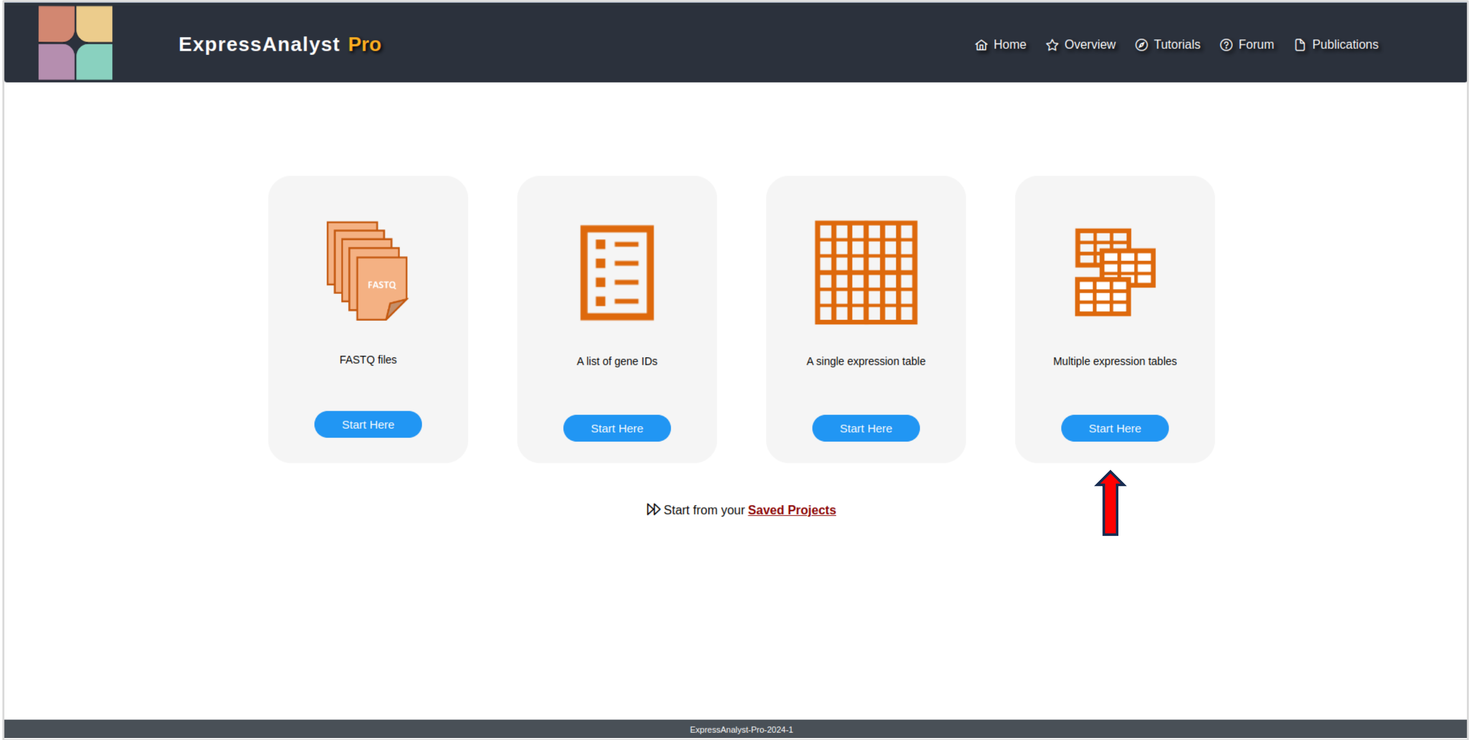


When the estimated Q values have approximately a chi-squared distribution, it suggests FEM assumption is appropriate. If it deviates significantly from a chi-squared distribution, REM should usually be used

Adjusting batch (“study-specific”) effect



Live Demo (ExpressAnalyst)



FASTQ files

Start Here

A list of gene IDs

Start Here

A single expression table

Start Here

Multiple expression tables

Start Here

Start from your **Saved Projects**

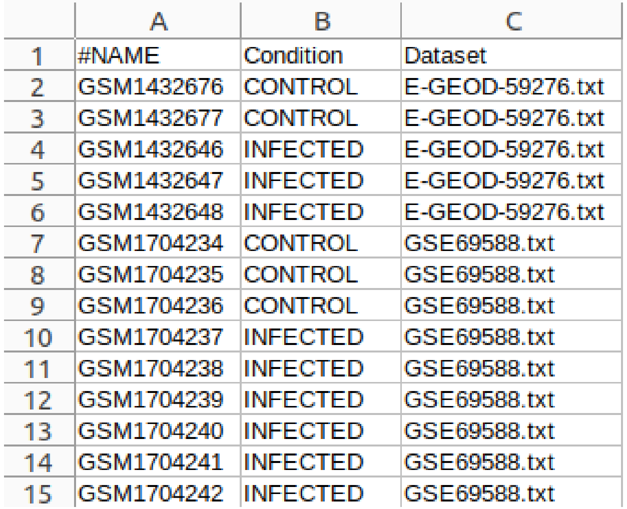
ExpressAnalyst-Pro-2024-1

Meta-analysis – example data

- Microarray data from mouse liver from three different datasets

- Study the effect of helminth (parasitic worm) infection

Dataset column



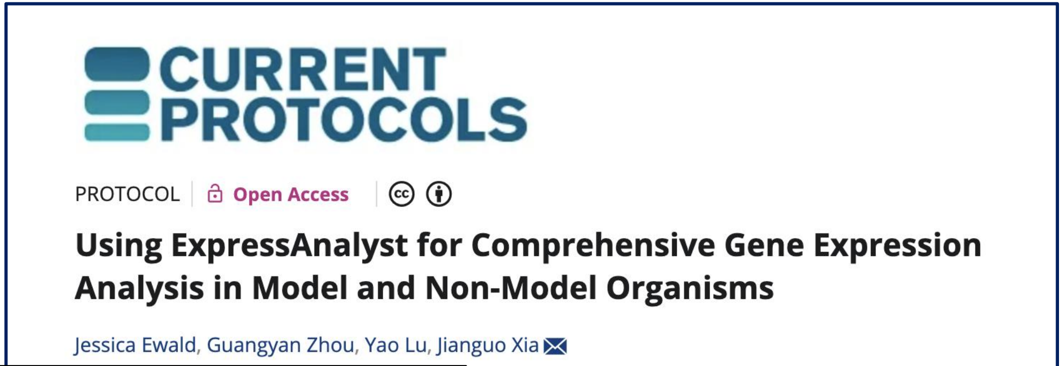
	A	B	C
1	#NAME	Condition	Dataset
2	GSM1432676	CONTROL	E-GEOD-59276.txt
3	GSM1432677	CONTROL	E-GEOD-59276.txt
4	GSM1432646	INFECTED	E-GEOD-59276.txt
5	GSM1432647	INFECTED	E-GEOD-59276.txt
6	GSM1432648	INFECTED	E-GEOD-59276.txt
7	GSM1704234	CONTROL	GSE69588.txt
8	GSM1704235	CONTROL	GSE69588.txt
9	GSM1704236	CONTROL	GSE69588.txt
10	GSM1704237	INFECTED	GSE69588.txt
11	GSM1704238	INFECTED	GSE69588.txt
12	GSM1704239	INFECTED	GSE69588.txt
13	GSM1704240	INFECTED	GSE69588.txt
14	GSM1704241	INFECTED	GSE69588.txt
15	GSM1704242	INFECTED	GSE69588.txt

Meta-data table

Next Lecture

Topic	Date	Lecture	Lab
Omics Data Science Foundations	Jan. 6	Omics data processing, statistics and visualization	--
	Jan. 13	From raw data to functional insights	--
Transcriptomics	Jan. 20	Gene expression data analysis (part I)	ExpressAnalyst & NetworkAnalyst
	Jan. 27	Gene expression data analysis (part II)	ExpressAnalyst & Seq2Fun
miRNAs & non-coding RNAs	Feb. 3	MicroRNAs, noncoding RNAs and biological networks	miRNet & NetworkAnalyst
Proteomics	Feb. 10	Proteomics data analysis and interpretation	ExpressAnalyst & NetworkAnalyst
Metabolomics	Feb. 17	Targeted metabolomics data analysis	MetaboAnalyst
	Feb. 24	LC-MS untargeted metabolomics data analysis	MetaboAnalyst
Microbiomics	Mar. 2	Marker gene data analysis	MicrobiomeAnalyst
	Mar. 9	Shotgun metagenomics data analysis	MicrobiomeAnalyst
Multi-omics	Mar. 16	Knowledge-driven multi-omics integration	OmicsNet
	Mar. 23	Data-driven multi-omics integration	OmicsAnalyst

Our Tutorials



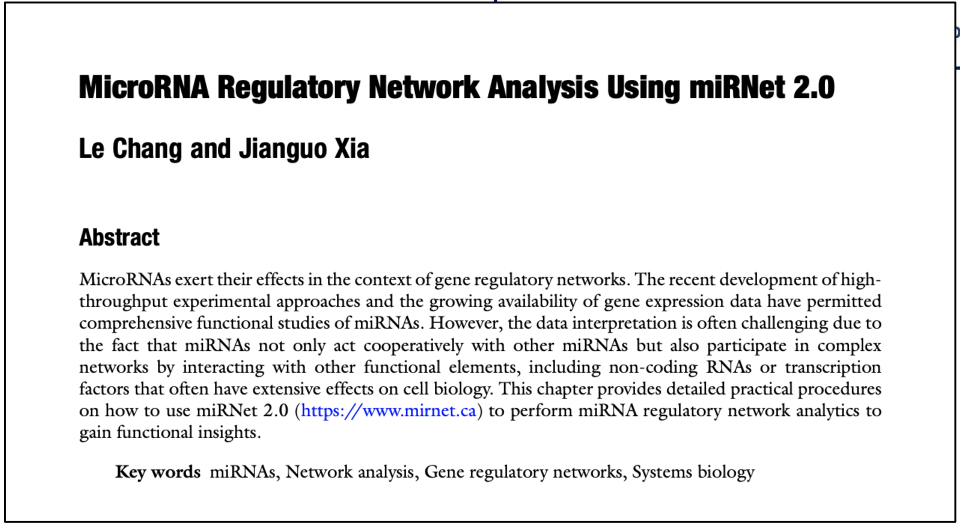
CURRENT PROTOCOLS

PROTOCOL | Open Access

Using ExpressAnalyst for Comprehensive Gene Expression Analysis in Model and Non-Model Organisms

Jessica Ewald, Guangyan Zhou, Yao Lu, Jianguo Xia

ps://doi.org/10.1002/cpz1.922



MicroRNA Regulatory Network Analysis Using miRNet 2.0

Le Chang and Jianguo Xia

Abstract

MicroRNAs exert their effects in the context of gene regulatory networks. The recent development of high-throughput experimental approaches and the growing availability of gene expression data have permitted comprehensive functional studies of miRNAs. However, the data interpretation is often challenging due to the fact that miRNAs not only act cooperatively with other miRNAs but also participate in complex networks by interacting with other functional elements, including non-coding RNAs or transcription factors that often have extensive effects on cell biology. This chapter provides detailed practical procedures on how to use miRNet 2.0 (https://www.mirnet.ca) to perform miRNA regulatory network analytics to gain functional insights.

Key words miRNAs, Network analysis, Gene regulatory networks, Systems biology

https://link.springer.com/protocol/10.1007/978-1-0716-2815-7_14

We would like to hear your comment & feedback

contact@xialab.ca

See (most of) you next week!