XiaLab Analytics

Empowering researchers through trainings, tools and AI

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

Omics Data Science Training Course

Winter 2024

Schedule for today

Time	Topics
9:00 – 9:10	Introduction
9:10 – 10:10	Gene expression data analysis workflow
10:15 – 11:25	Live demo & hands-on practice
Summary and discussion	

Our Resources

Recordings & slides

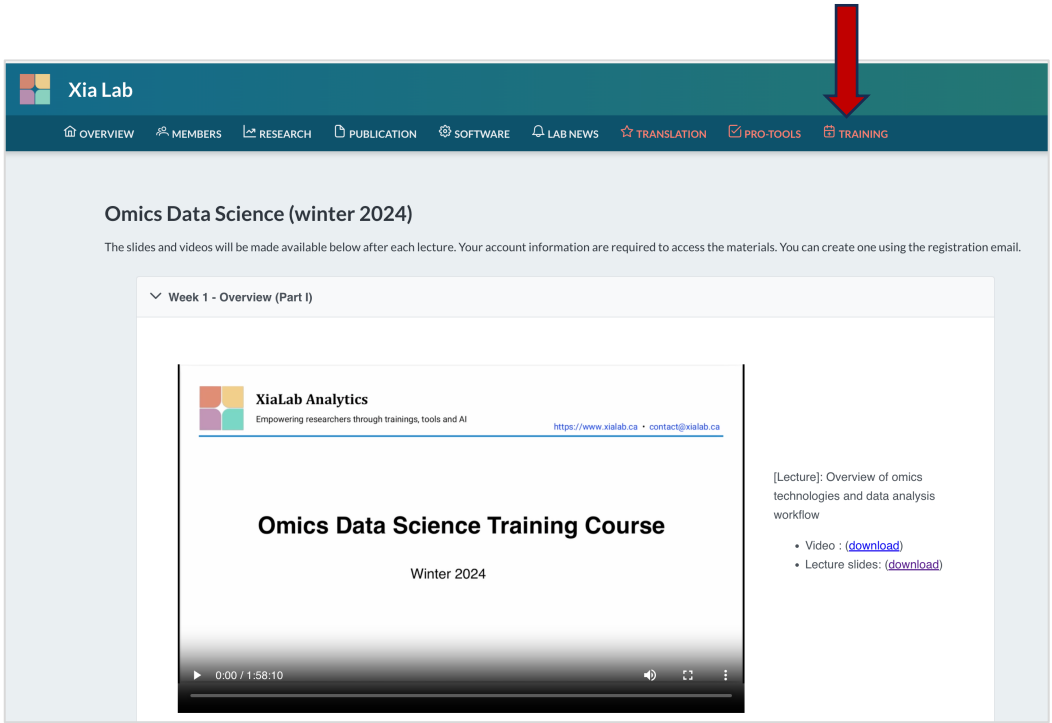
- https://www.xialab.ca under the "**Training**" tab within 3 days after lecture
- Faster with Firefox

Community tool:

- https://www.####.ca

Pro Tools:

- https://pro.####.ca
  - You will be assigned to one of the three nodes

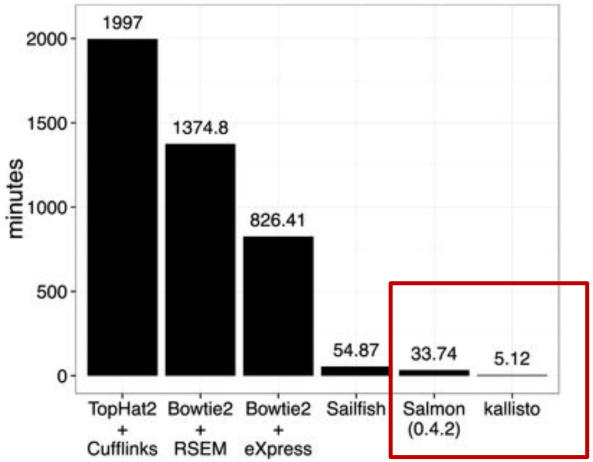


ExpressAnalyst Docker

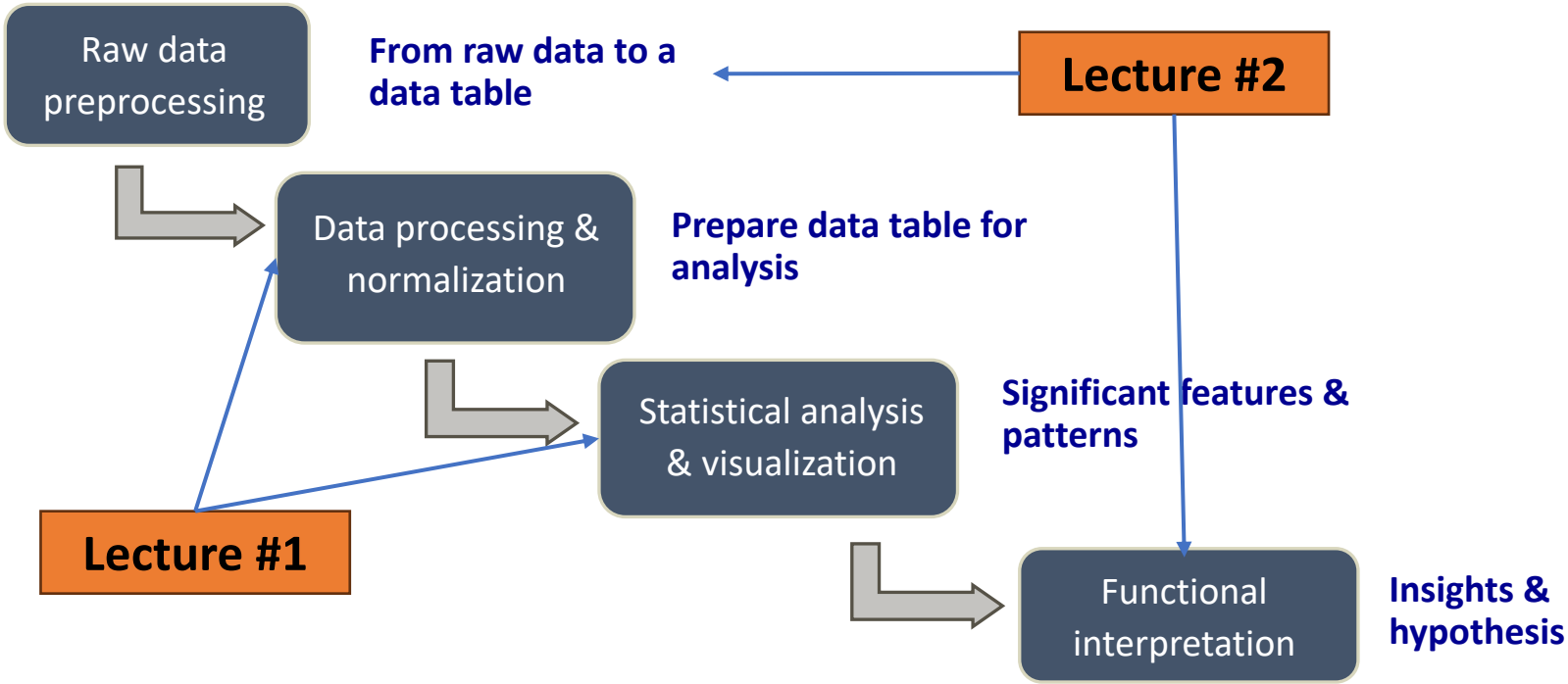
Some of you reported that Kallisto has some issues running inside Docker under Windows OS.

We are in the process of

- Adding support for the Salmon algorithm as an alternative
- Building a VM image to better control the environment
  - Recommend 32G RAM, 8 CPU cores



What we have covered so far

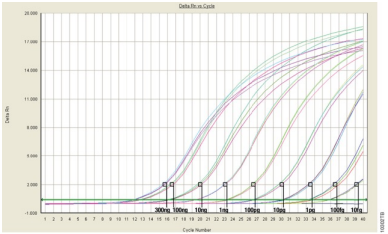


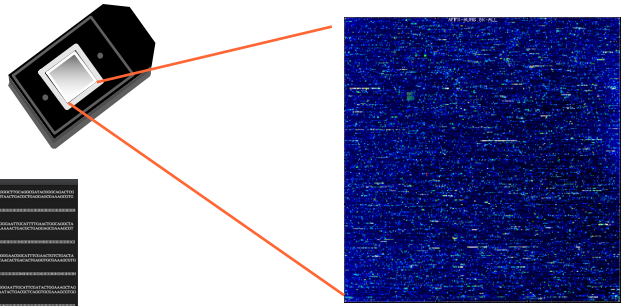
Our Syllabus

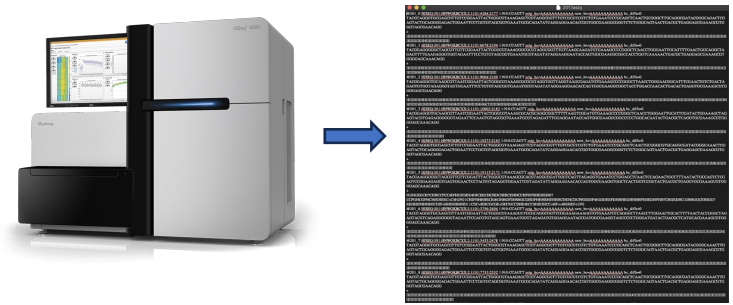
Topic	Date	Lecture	Lab
Omics Data Science Foundations	Jan. 6	Omics data processing, statistics and visualization	--
	Jan. 13	From raw data to functional insights	--
Transcriptomi	Jan. 20	Gene expression data analysis (part I)	ExpressAnalyst & NetworkAnalyst
	Jan. 27	Gene expression data analysis (part 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

Measuring Gene Expression

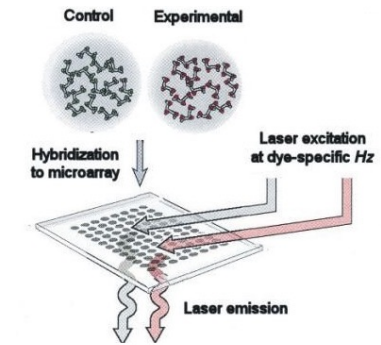
- Q-PCR
  - 10s~100s (costly for large-scale)
- Largely deprecated:
  - ESTs (expressed sequence tags)
  - SAGE (serial analysis of gene expression)
- Microarray
- RNAseq

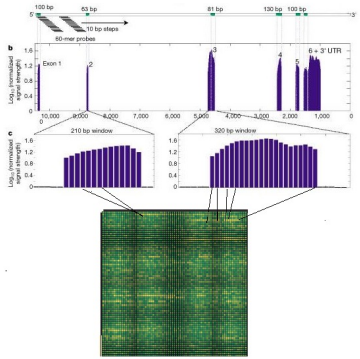


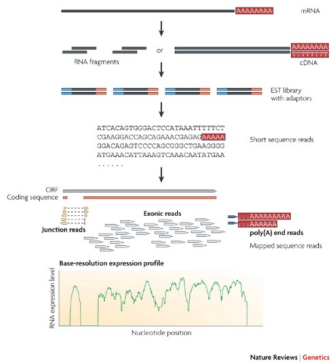




Evolution of transcriptomics technologies







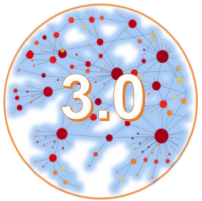
1995 P. Brown, et. al. Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

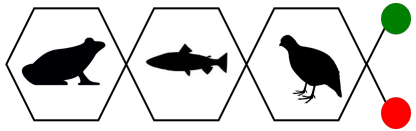
2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)

XiaLab Tools for gene expression analysis

ExpressAnalyst A unified platform for expression analysis



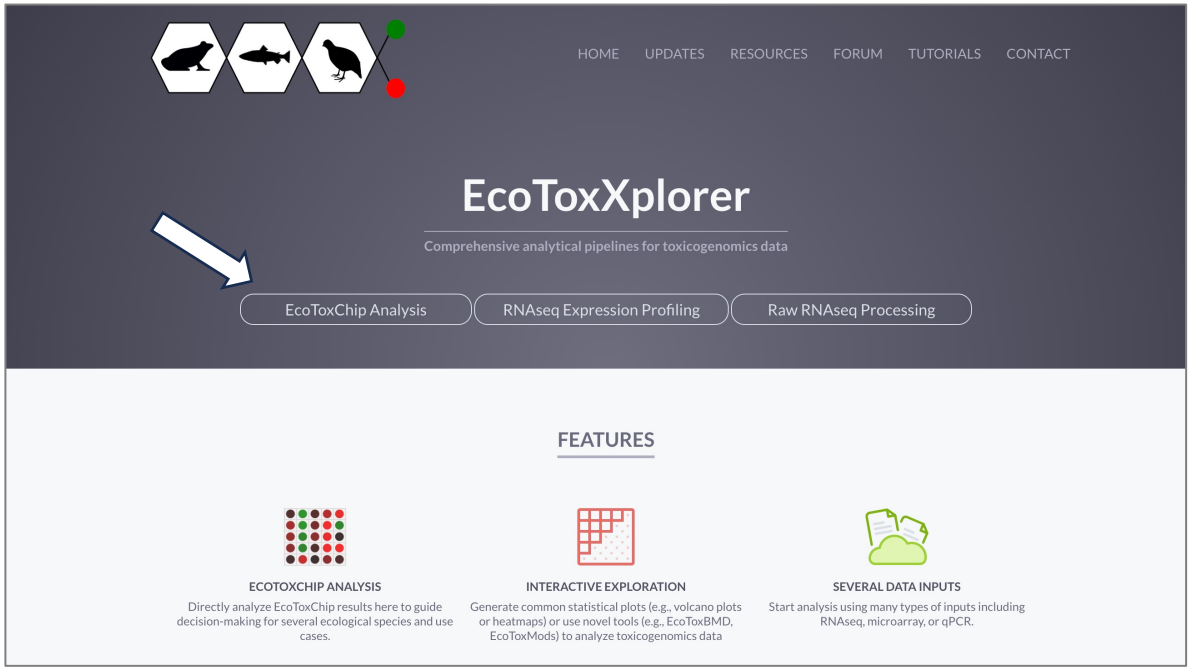
NetworkAnalyst Understanding gene lists for model species



EcoToxXplorer Quantitative PCR (qPCR) 96/384-well plates

A web-based tool for qPCR data analysis

- 384-well plate
- One file per plate
- Two column format
  - well ID
  - raw Ct values



https://www.ecotoxxplorer.ca

Schedule for today

Time	Topics
9:00 – 9:10	Introduction
9:10 – 10:10	Gene expression data analysis workflow
10:15 – 11:25	Live demo & hands-on practice
Summary and discussion	

Microarray Data Analysis

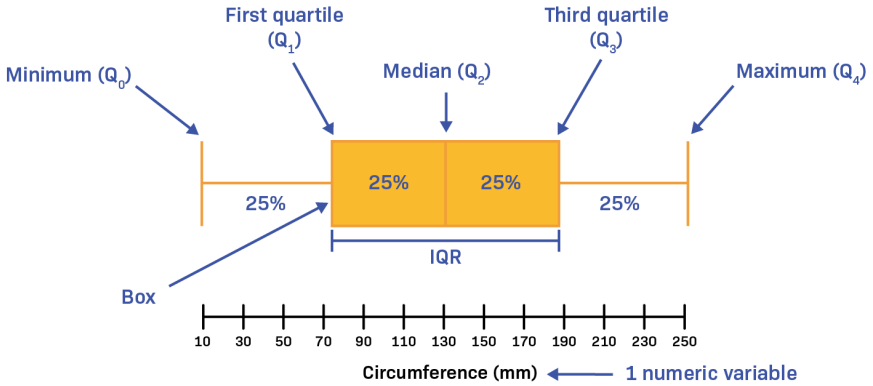
Microarray core tasks

Input: probe intensity file

1. Processing
   - Mapping probe IDs to gene/transcript IDs (sum / average)
   - Quality checking to make sure data are suitable for analysis
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. Interpreting the results
   - Find which pathways or biological processes are changed

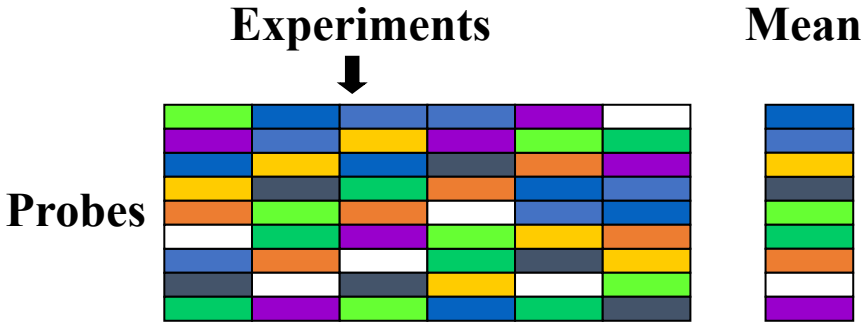
Quantile Normalization (I)

- Considered the best normalization method for **microarray** data
- Quantile: a distribution based on the rank order of values in that distribution.
  - You can find any quantile by sorting the sample. The middle value of the sorted sample (middle quantile, 50th percentile) is known as the median. The limits are the minimum and maximum values. Any other locations between these points can be described in terms of centiles/percentiles.



Quantile Normalization (II)

- Assume most of the probes/genes don't change between samples
- Calculate mean for each quantile and reassign each probe by the quantile mean
  - No experiments retain value, but all experiments have exact same distribution
  - Only their orders are different



Quantile Normalization (III)

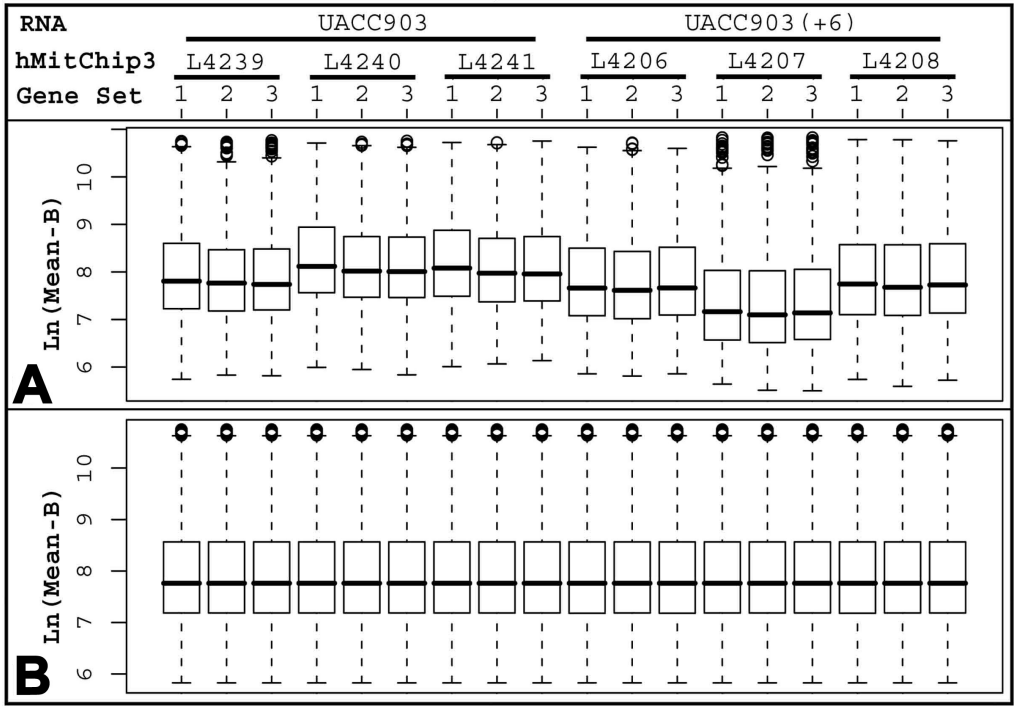
Raw data			
2	4	4	5
5	14	4	7
4	8	6	9
3	8	5	8
3	9	3	5

Order values within each sample (or column)			
2	4	3	5
3	8	4	5
3	8	4	7
4	9	5	8
5	14	6	9

Average across rows and substitute value with average			
3.5	3.5	3.5	3.5
5.0	5.0	5.0	5.0
5.5	5.5	5.5	5.5
6.5	6.5	6.5	6.5
8.5	8.5	8.5	8.5

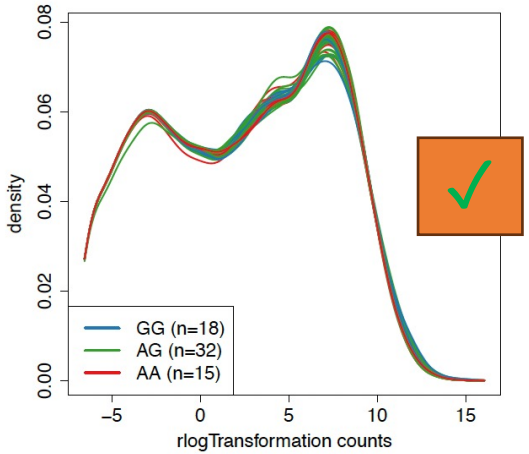
Re-order averaged values in original order			
3.5	3.5	5.0	5.0
8.5	8.5	5.5	5.5
6.5	5.0	8.5	8.5
5.0	5.5	6.5	6.5
5.5	6.5	3.5	3.5

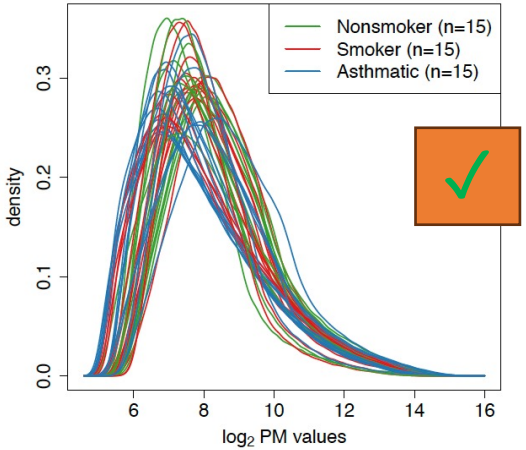
Identical Distribution after Quantile Normalization

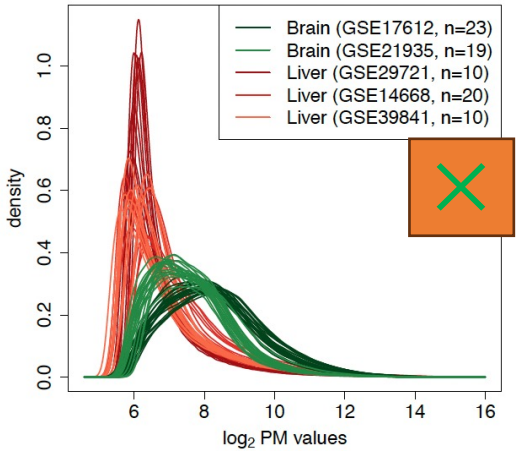


Thoughts on normalization

- Improve signals & reduce noises
- Compare things that are “comparable”







doi: http://dx.doi.org/10.1101/012203

Differential Expression Analysis (DEA)

Simple Approaches

- Fold change
- T-test /ANOVA

nature biotechnology

Rat toxicogenomic study reveals analytical consistency across microarray platforms

Lei Guo[1], Edward K Lobenhofer[2], Charles Wang[3], Richard Shippy[4], Stephen C Harris[1], Lu Zhang[5], Nan Mei[1], Tao Chen[1], Damir Herman[6], Federico M Goodsaid[7], Patrick Hurban[2], Kenneth L Phillips[2], Jun Xu[3], Xutao Deng[3], Yongming Andrew Sun[8], Weida Tong[1], Yvonne P Dragan[1] & Leming Shi[1]

To validate and extend the findings of the MicroArray Quality Control (MAQC) project, a biologically relevant toxicogenomics data set was generated using 36 RNA samples from rats treated with three chemicals (aristolochic acid, riddelliine and comfrey) and each sample was hybridized to four microarray platforms. The MAQC project assessed concordance in intersite and cross-platform comparisons and the impact of gene selection methods on the reproducibility of profiling data in terms of differentially expressed genes using distinct reference RNA samples. The real-world toxicogenomic data set reported here showed high concordance in intersite and cross-platform comparisons. Further, gene lists generated by fold-change ranking were more reproducible than those obtained by *t*-test *P* value or Significance Analysis of Microarrays. Finally, gene lists generated by fold-change ranking with a nonstringent *P*-value cutoff showed increased consistency in Gene Ontology terms and pathways, and hence the biological impact of chemical exposure could be reliably deduced from all platforms analyzed.

Microarray data analysis: from disarray to consolidation and consensus

David B. Allison[‡§], Xiangqin Cui*[§], Grier P. Page* and Mahyar Sabripour**

Abstract | In just a few years, microarrays have gone from obscurity to being almost ubiquitous in biological research. At the same time, the statistical methodology for microarray analysis has progressed from simple visual assessments of results to a weekly deluge of papers that describe purportedly novel algorithms for analysing changes in gene expression. Although the many procedures that are available might be bewildering to biologists who wish to apply them, statistical geneticists are recognizing commonalities among the different methods. Many are special cases of more general models, and points of consensus are emerging about the general approaches that warrant use and elaboration.

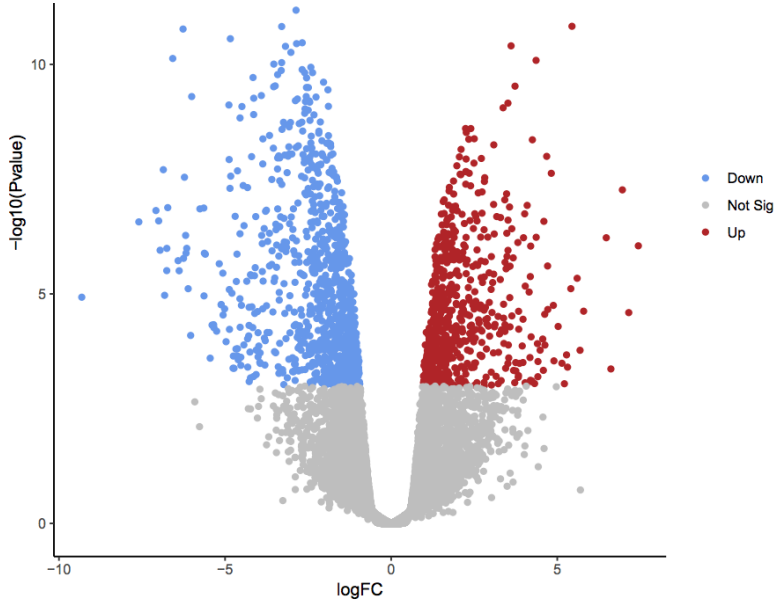
Fold change
A metric for comparing a gene's mRNA-expression level between two distinct experimental conditions. Its arithmetic definition differs between investigators.

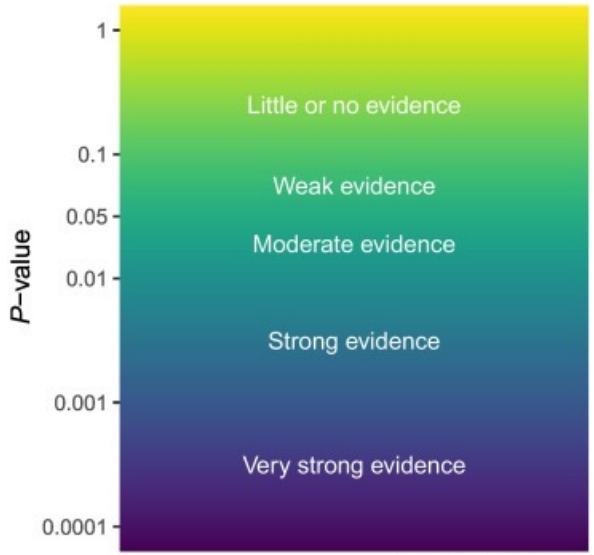
Gene-expression microarrays have become almost as widely used as measurement tools in biological research as western blots (BOX 1). A wide range of methods for microarray data analysis have evolved, ranging from simple fold-change (FC) approaches to testing for differential expression, to many complex and computationally demanding techniques[1]. The result might seem like most design strategies. The relative merits of specific designs are discussed elsewhere[1,4–7].

Consensus point 1: Biological replication is essential. In microarray analysis, two types of replication can be carried out: technical replication, when mRNA from a single biological case is used on multiple microarrays,

Consensus point 1: Using fold change alone as a differential expression test is not valid. FC was the first method used to evaluate whether genes are differentially expressed, and is a reasonable measure of effect size. However, it is widely considered to be an inadequate test statistic[38,39] because it does not incorporate variance and offers no associated level of 'confidence'[38,40]. Using FC alone with a fixed cut off, regardless of sample size or variance, results in type 1 error rates that are either unknown or depend on sample size, and even tests for which power can decrease with increasing sample size.

Volcano plot: best of both worlds?

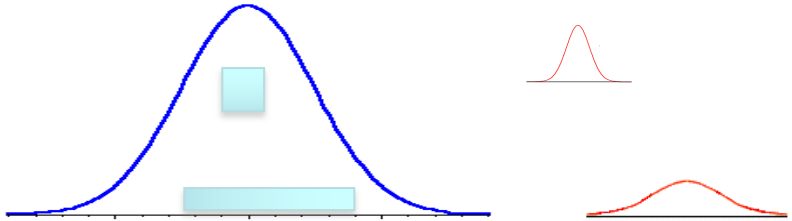


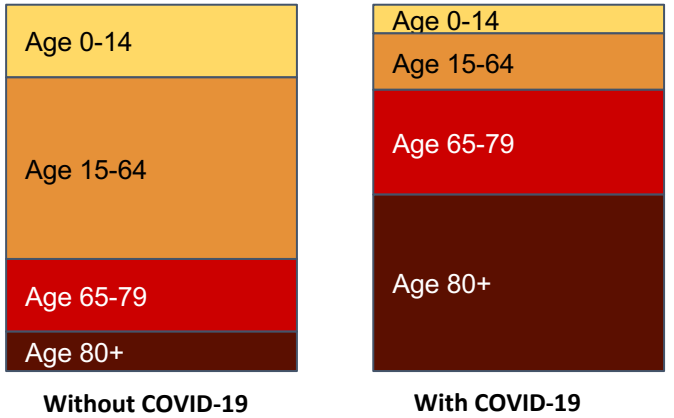


https://doi.org/10.1016/j.tree.2021.10.009

Issues with t-tests / ANOVA

1. When the sample size is small (e.g. 2-3 replicates in each condition), t-tests will not work well
   - Variance estimation becomes unstable for very few data points!
2. Observational studies contain many uncontrolled variables (covariates):
   - Biological: sex, age, disease status
   - Environmental: location, lifestyle, temperature





Empirical Bayes

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

t = Student's t-test
m = mean
μ = theoretical value
s = standard deviation
n = variable set size

"Borrow information" to improve variance estimation for individual gene

Predicted variance based on **all genes** (i.e. weighted average, trimmed means)

Group difference

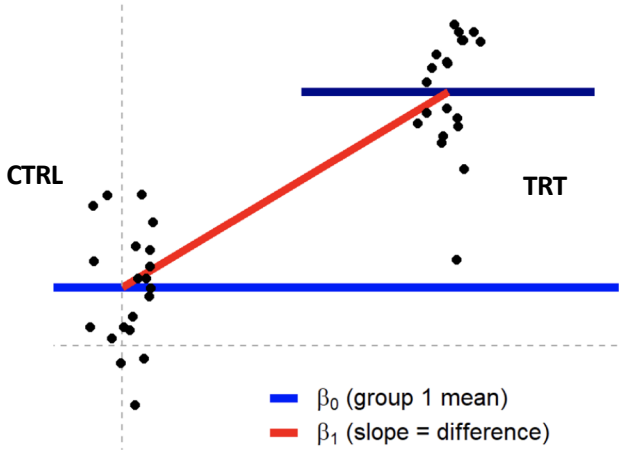
$$\frac{\hat{\delta}}{\sqrt{B\tilde{\sigma}^2 + (1-B)\hat{\sigma}^2}}$$

B is the shrinkage factor:
- B= 0 => standard t tests
- B=1 => fold change

Variance observed on the specific genes

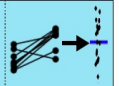
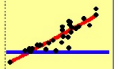
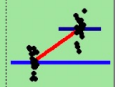
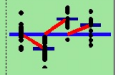
T-test vs. linear regression

- You can do t-test with linear regression:
- $y = B_0 + B_1 * x$
  - y: level of metabolite A
  - x: variable of interest



Essentially, a hypothesis test on the slope is asking whether or not the slope is 0

Common statistical tests are linear models

Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
y is independent of x P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise $\mathbf{y_2}$-$\mathbf{y_1}$ differences. - (Same, but it predicts the *signed rank* of $\mathbf{y_2}$-$\mathbf{y_1}$.)	
y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=…[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +…+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +…+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	

https://lindeloev.github.io/tests-as-linear/

Limma: linear models for microarray data

The *de facto* method for differential expression data analysis

- Support simple and complex design
- Empirical Bayes for small sample size
- Efficient & high performance

[HTML] **limma** powers differential expression analyses for RNA-sequencing and microarray studies

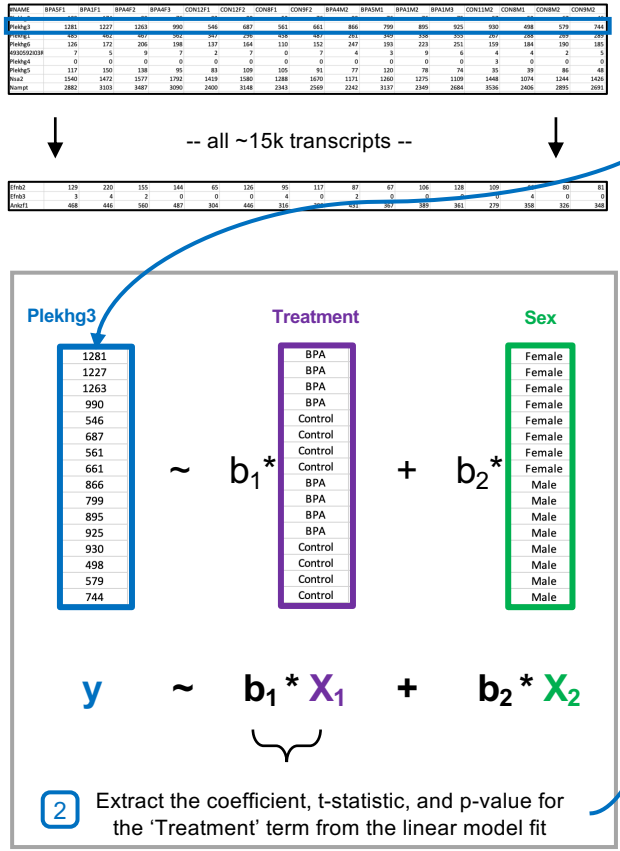
ME Ritchie, B Phipson, DI Wu, Y Hu… - Nucleic acids …, 2015 - academic.oup.com

… Over the past decade, **limma** has been a popular choice for … Recently, the capabilities of **limma** have been significantly … This article reviews the philosophy and design of the **limma** …

☆ Save Cite Cited by 26407 Related articles All 22 versions

Limma for DEA

- Treatment (Control/BPA)
- Sex (male/female)
- Features associated with ***treatment*** while considering ***sex***
  - Treatment = primary
  - Sex = covariate
- Can include many covariates
- Primary/covariate can be continuous or discrete



Symbols	BPA-Control	t	P.Value	adj.P.Val
Maz	0.18999	2.7563	0.015091	0.052129
Slc11a2	0.27426	2.7561	0.015096	0.052133
Cyhr1	0.1355	2.7557	0.01511	0.052166
Tbc1d2b	0.26692	2.7554	0.015117	0.052177
Sat1	0.61953	2.7552	0.015124	0.052188
Cers2	0.16232	2.7549	0.015131	0.052188
Cox11	0.31729	2.7549	0.015133	0.052188
Ccdc117	0.49021	2.7548	0.015136	0.052188
Plekhg3	0.31496	2.754	0.01516	0.052246
Lmo2	-0.44288	-2.754	0.015161	0.052246
Pgam2	1.7962	2.7534	0.015178	0.052275
Ublcp1	-0.16546	-2.7534	0.015178	0.052275
Polr2g	0.20745	2.7533	0.015181	0.052275
Nrxn1	-0.22709	-2.7523	0.015212	0.052371
Uap1l1	0.49729	2.7521	0.015217	0.052371
Mgst1	0.24975	2.7518	0.015226	0.052389
Zfp93	0.24983	2.7517	0.015229	0.052389
Bola2	-0.23162	-2.7513	0.015242	0.052419
Eif2d	0.17861	2.7507	0.01526	0.052463
Nadk	0.18645	2.7506	0.015263	0.052463
Csnk1g1	-0.24522	-2.75	0.015281	0.052513
Marveld2	0.32542	2.7496	0.015292	0.052537
Grb7	0.21623	2.7487	0.015321	0.052621
Cdk20	0.4224	2.748	0.015342	0.052679
Akt1s1	0.15424	2.7468	0.015378	0.052791
Ppm1g	-0.22411	-2.7463	0.015394	0.052804
Suds3	0.16652	2.7462	0.015395	0.052804
Cep97	0.35979	2.7449	0.015436	0.052847
Slc35a1	0.20103	2.7448	0.015438	0.052847
Aifm3	0.90739	2.7433	0.015484	0.05299

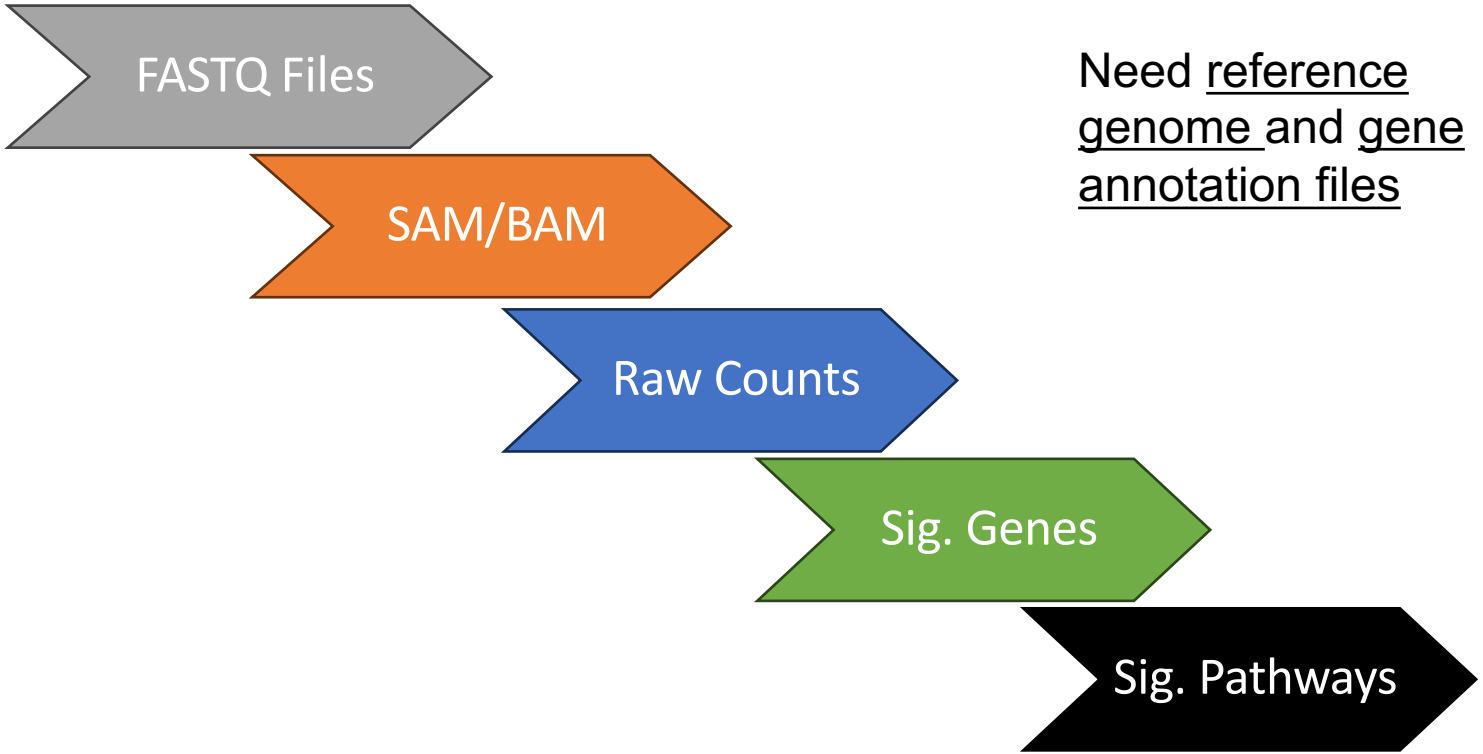
RNAseq Data Analysis

RNAseq core tasks

Input: FASTQ files

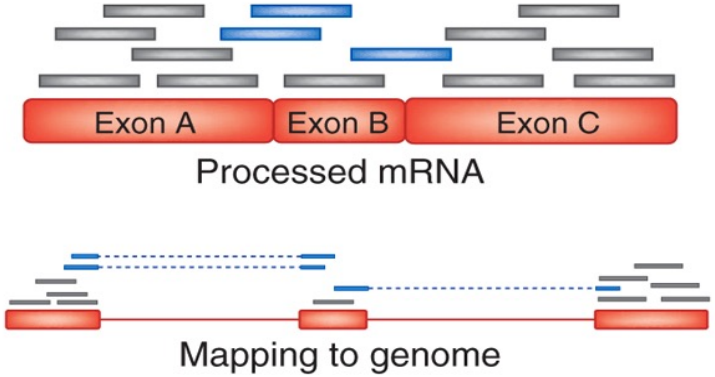
1. Reads are **mapped** to reference genome or transcriptome
2. Mapped reads **counted** per gene or per transcripts
3. Counts are **tested statistically** for significant difference
4. **Enrichment analysis** are applied for functional insights

Data Flow



Need reference genome and gene annotation files

Read Mapping

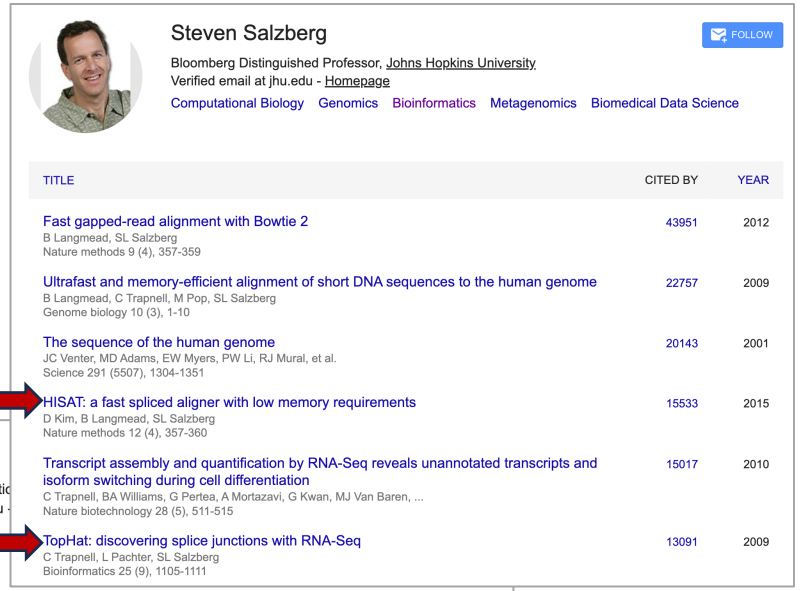


Aligns reads to the whole genome AND to exon-junctions

Gapped Read Aligner

Gapper Read Aligners & Mapper

- 1st generation
  - TopHat
- 2nd generation
  - HISAT/STAR
- 3rd generation
  - Kallisto/Salmon



Steven Salzberg

FOLLOW

Bloomberg Distinguished Professor, Johns Hopkins University
Verified email at jhu.edu - Homepage

Computational Biology Genomics Bioinformatics Metagenomics Biomedical Data Science

TITLE	CITED BY	YEAR
Fast gapped-read alignment with Bowtie 2 B Langmead, SL Salzberg Nature methods 9 (4), 357-359	43951	2012
Ultrafast and memory-efficient alignment of short DNA sequences to the human genome B Langmead, C Trapnell, M Pop, SL Salzberg Genome biology 10 (3), 1-10	22757	2009
The sequence of the human genome JC Venter, MD Adams, EW Myers, PW Li, RJ Mural, et al. Science 291 (5507), 1304-1351	20143	2001
HISAT: a fast spliced aligner with low memory requirements D Kim, B Langmead, SL Salzberg Nature methods 12 (4), 357-360	15533	2015
Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation C Trapnell, BA Williams, G Pertea, A Mortazavi, G Kwan, MJ Van Baren, ... Nature biotechnology 28 (5), 511-515	15017	2010
TopHat: discovering splice junctions with RNA-Seq C Trapnell, L Pachter, SL Salzberg Bioinformatics 25 (9), 1105-1111	13091	2009

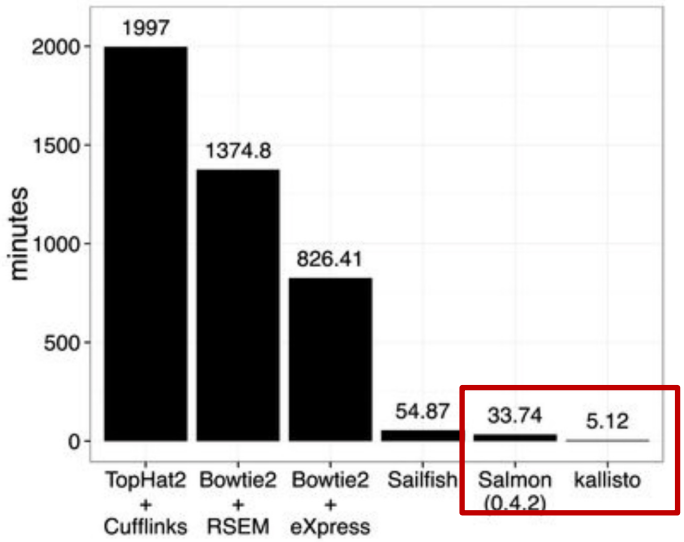
Lior Pachter

Bren Professor of Computati
Verified email at caltech.edu

TITLE	CITED BY	YEAR
Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation C Trapnell, BA Williams, G Pertea, A Mortazavi, G Kwan, MJ Van Baren, ... Nature biotechnology 28 (5), 511-515	15019	2010
TopHat: discovering splice junctions with RNA-Seq C Trapnell, L Pachter, SL Salzberg Bioinformatics 25 (9), 1105-1111	13091	2009
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks C Trapnell, A Roberts, L Goff, G Pertea, D Kim, DR Kelley, H Pimentel, ... Nature protocols 7 (3), 562-578	12536	2012
Initial sequencing and comparative analysis of the mouse genome European Bioinformatics Institute: Birney Ewan 3 Goldman Nick 3 Kasprzyk ... Nature 420 (6915), 520-562	8443	2002
Near-optimal probabilistic RNA-seq quantification NL Bray, H Pimentel, P Melsted, L Pachter Nature biotechnology 34 (5), 525-527	7356	2016

Salmon & Kallisto (mapper not aligner)

- Salmon is based on the concept of quasi-mapping. It uses a suffix array to discover shared substrings of any length between a read and the complete set of transcripts. Mismatches are handled with chains of maximally exact matches (MEM).
- Kallisto does not perform *alignment* or use a reference genome. It performs *pseudoalignment* to determine the *compatibility* of reads with targets (transcript sequences in this case). Based on k-mer and Targeted de Bruijn Graph (T-DBG)



Kallisto and Salmon assign reads to genes (mapping) without perform exact alignments.

RNAseq Common Files

Sequence files

- FASTA
- FASTQ

Alignment files

- SAM
- BAM

Feature annotations

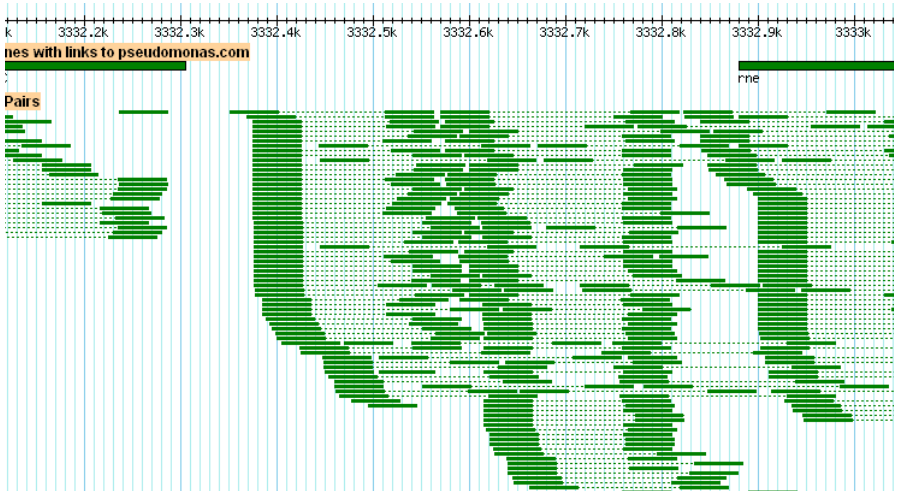
- GFF (general feature format)
- GTF (gene transfer format)

SAM: Sequence Alignment Map

SAM – a tab-delimited text file that contains a compact and index-able representation of nucleotide sequence alignments

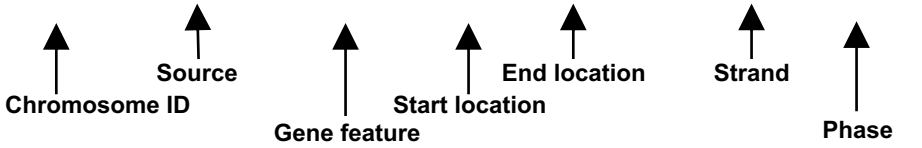
BAM – binary version of SAM

- Save space
- Faster I/O



GFF (general feature format)

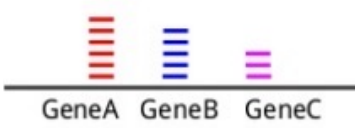
```
Chr1  amel_OGSv3.1    gene    204921  223005  .       +       .       ID=GB42165
Chr1  amel_OGSv3.1    mRNA    204921  223005  .       +       .       ID=GB42165-
RA;Parent=GB42165
Chr1  amel_OGSv3.1    3'UTR   222859  223005  .       +       .       Parent=GB42165-RA
Chr1  amel_OGSv3.1    exon    204921  205070  .       +       .       Parent=GB42165-RA
Chr1  amel_OGSv3.1    exon    222772  223005  .       +       .       Parent=GB42165-RA
```



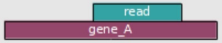
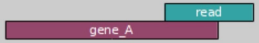
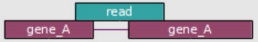
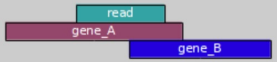
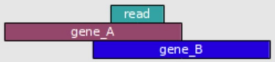
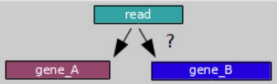
1. **seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seqname must be one used with a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **feature** - feature type name, e.g. Gene, Variation, Similarity
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **frame** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

Task #2: generate gene counts

From “aligned” reads to gene counts



Things can get “complicated”

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

https://htseq.readthedocs.io/en/latest/htseqcount.html

Calculating gene expression from read counts

- The expression level of the gene
  - This is what we want to estimate
- The total number of the reads generated
  - If one library is sequenced at 20M reads and the other library is sequenced at 40M reads, then most genes will ~double their counts
- The length of the transcript
  - Larger genes will generate more fragments
- GC-content of the gene
  - GC-rich and GC-poor fragments tend to be under-represented in RNA-Seq

Quantifying expression from read mapping

Normalize for gene size and sequencing depth

- RPKM: Reads per kilobase of transcript per million mapped reads (single-end).
- FPKM: Fragments per kilobase of transcript per million mapped reads (paired-end).
- TPM: Transcripts per kilobase of transcript per million mapped reads

$$RPKM_i \text{ or } FPKM_i = \frac{q_i}{\frac{l_i}{10^3} * \frac{\sum_j q_j}{10^6}} = \frac{q_i}{l_i * \sum_j q_j} * 10^9$$

$$TPM_i = \frac{q_i / l_i}{\sum_j (q_j / l_j)} * 10^6$$

q_i denotes reads mapped to transcript, l_i is the transcript length

Comments on different quantification

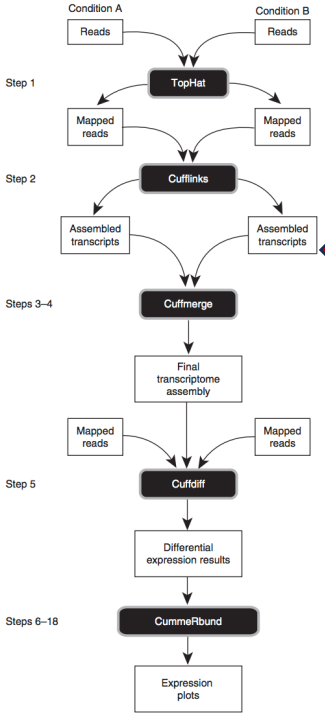
RPKM/FPKM/TPM

- Good for the comparison of RNA transcript expression within a single sample

Raw Counts

- Input for more robust statistical methods for differential expression to compare the same genes across samples
  - The effect of the gene length is the same!
  - We are only concerned with relative difference
- Accommodates more sophisticated experimental designs

Two widely used protocols



PROTOCOL

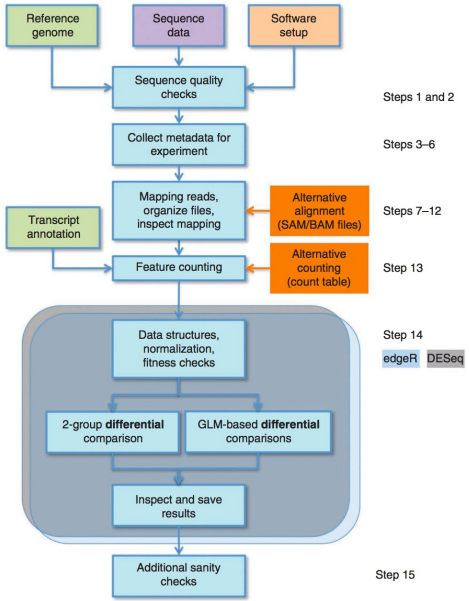
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]

PROTOCOL

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]



A gene count table

NAME	X454039	X454051	X454057	X454066	X454081	X454090	X454102	X454108
ENSMUSG00000023868	219	49	42	50	6	17	22	21
ENSMUSG00000029538	424	376	191	436	529	1172	914	1014
ENSMUSG00000064141	32	4	4	22	76	165	163	170
ENSMUSG00000020986	198	167	100	203	158	475	406	369
ENSMUSG00000024811	475	255	194	442	384	1365	1148	1219
ENSMUSG00000038781	2	2	2	2	2	5	8	3
ENSMUSG00000022159	96	31	29	59	102	222	154	171
ENSMUSG00000002103	888	505	298	638	541	1570	1351	1517
ENSMUSG00000024608	1504	868	613	1521	1083	3077	2111	4917
ENSMUSG00000022037	57	32	4	12	0	0	1	0
ENSMUSG00000027108	410	323	184	392	396	1145	752	965
ENSMUSG00000033540	359	138	90	279	483	1382	1462	1524
ENSMUSG00000021910	4407	2380	1497	3471	7348	19703	15845	18788
ENSMUSG00000039156	2518	1181	706	1821	383	948	821	966
ENSMUSG00000021959	345	228	134	277	700	1720	1274	1616
ENSMUSG00000020364	0	0	4	0	16	36	16	8
ENSMUSG00000020415	586	958	439	915	212	464	516	707
ENSMUSG00000015340	4749	2689	1880	4618	1582	5098	4652	4618
ENSMUSG00000025875	574	156	101	231	514	1389	1259	1438
ENSMUSG00000017734	122	46	27	64	231	450	441	509
ENSMUSG00000022787	110	66	32	66	178	336	284	248
ENSMUSG00000031834	25	17	15	29	259	700	564	649
ENSMUSG00000002428	85	61	44	109	246	763	642	705
ENSMUSG00000026064	123	79	30	93	31	111	76	96
ENSMUSG00000041420	44	13	5	3	116	214	254	161
ENSMUSG00000027297	0	6	0	5	0	6	0	0

Task #3: Differential Expression Analysis

RNAseq Differential Expression (I)

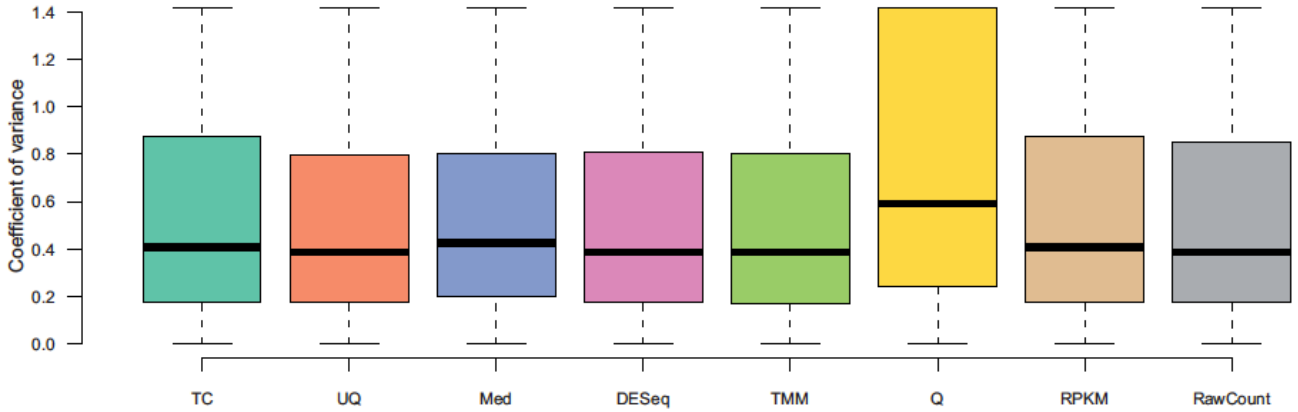
- Three widely used methods
  - Limma
  - EdgeR
  - DESeq2
- Main differences are in normalization approaches
  - Limma: Adapted to support RNAseq after “voom” transformation
  - DESeq and EdgeR are very similar and both assume that no genes are differentially expressed. DESeq2 uses a "geometric" normalisation strategy, whereas EdgeR is a weighted mean of log ratios-based method.

Normalization methods for gene count data

- Quantile (Q)
- RPKM
- Total Count (TC)
- Upper Quantile (UQ)
- Median (Med)
- Trimmed Mean of M-values (TMM)
  - EdgeR default
- Relative log expression (RLE)
  - DESeq default

Comparison of different normalizations (I)

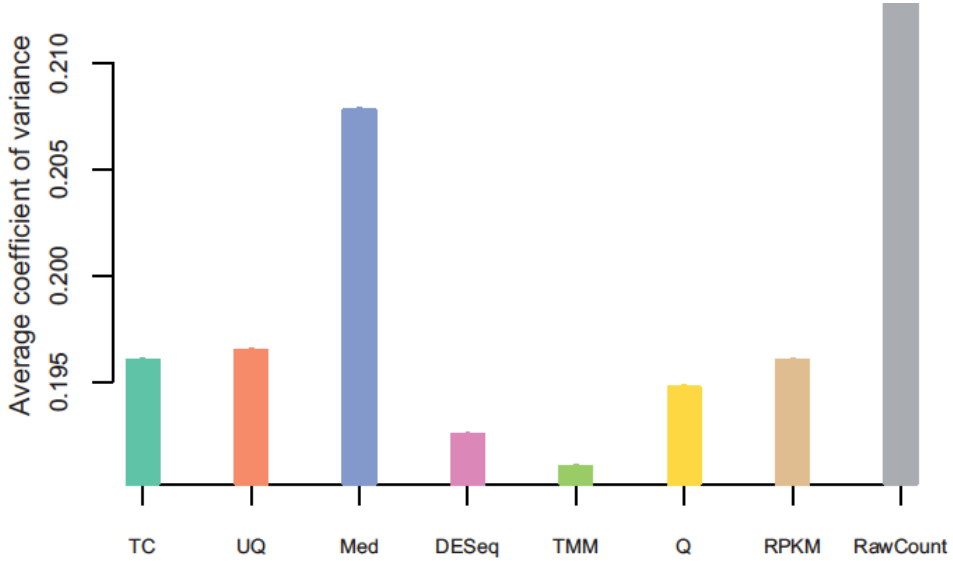
Within group variance (the smaller the better)



http://www.ncbi.nlm.nih.gov/pubmed/22988256

Comparison of different normalizations (II)

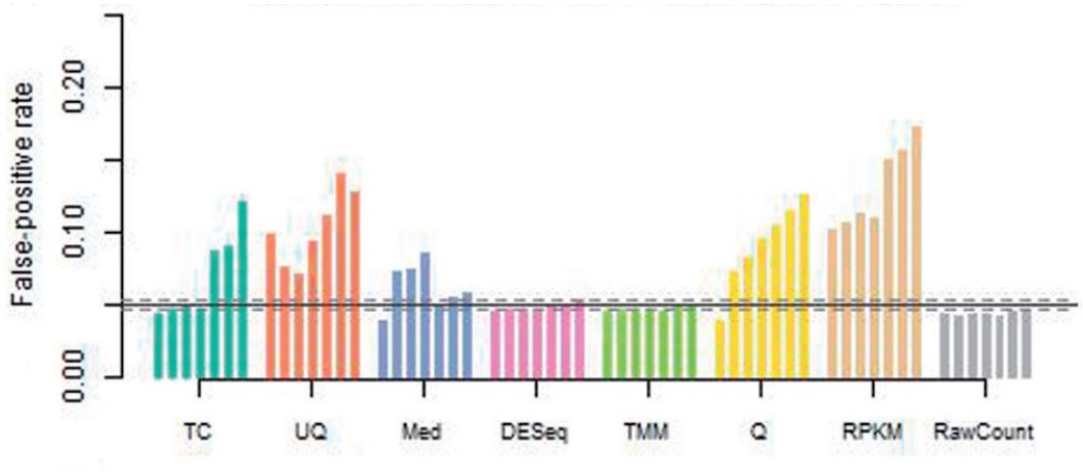
Variance of house keeping genes (the smaller the better)



http://www.ncbi.nlm.nih.gov/pubmed/22988256

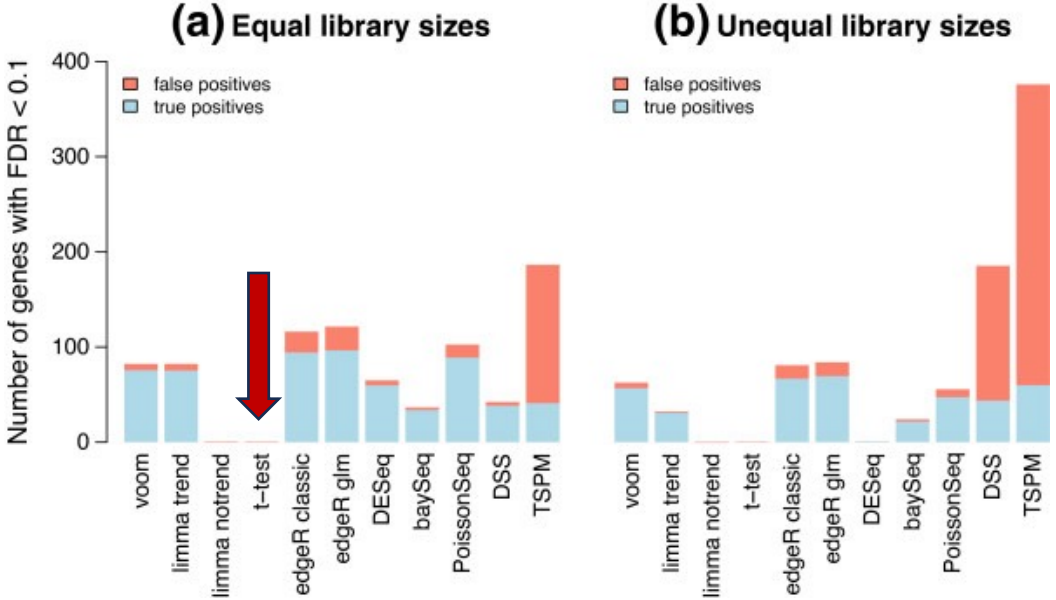
Comparison of different normalizations (III)

False discovery rate (the smaller the better)



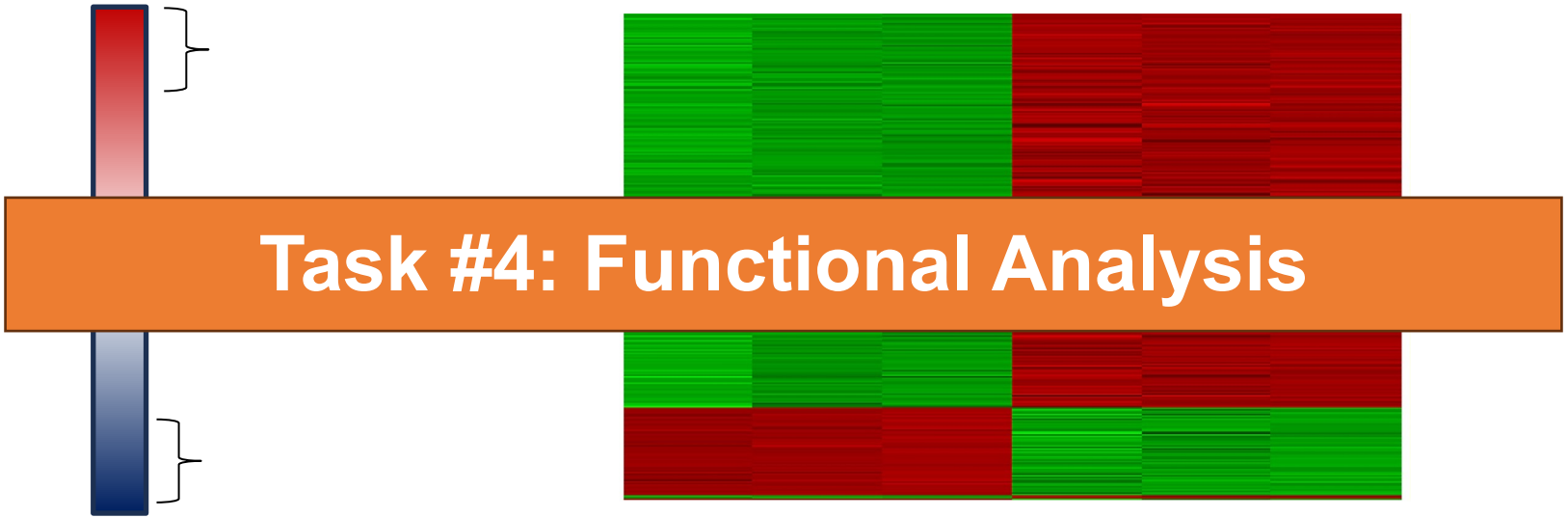
http://www.ncbi.nlm.nih.gov/pubmed/22988256

EdgeR, Limma (voom) and DESeq2



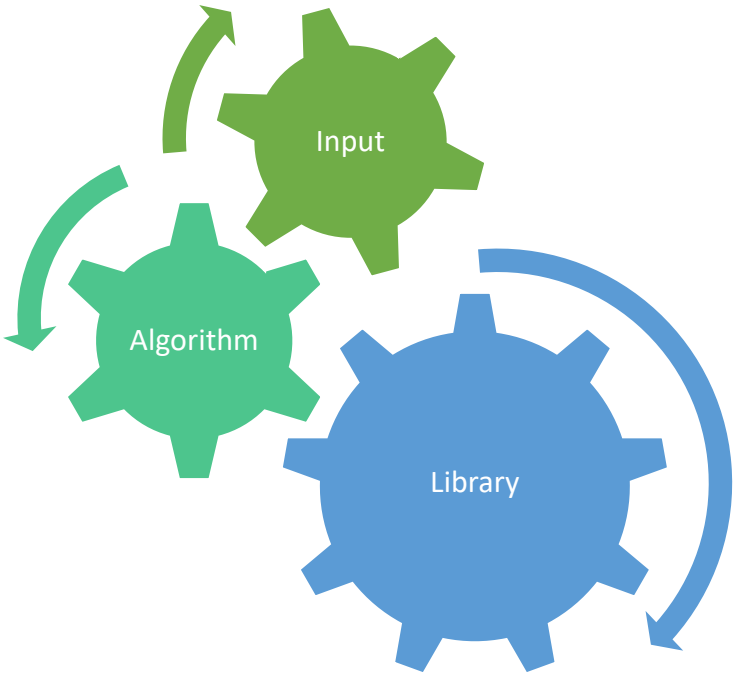
- EdgeR is more liberal (more DEGs)
- DESeq2 is computationally intensive and should only be used for small sample size (less than 50)
- Limma-voom is generally preferred for large sample size

http://www.ncbi.nlm.nih.gov/pubmed/24485249



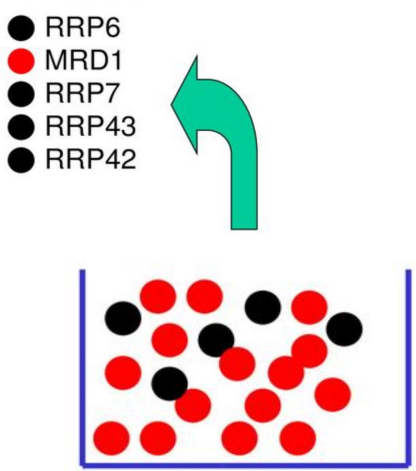
Task #4: Functional Analysis

Three key components



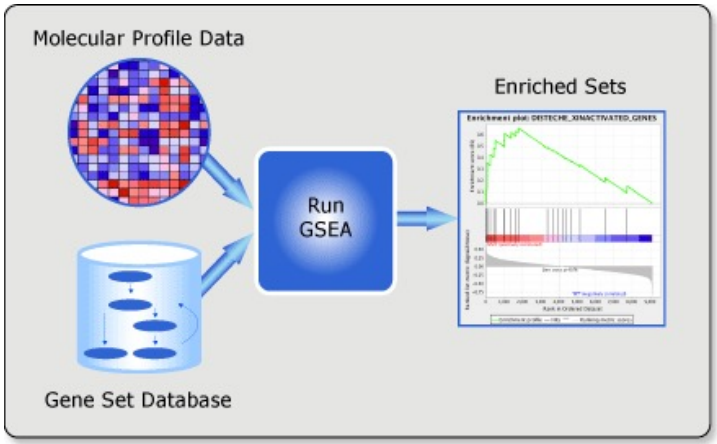
Issues with ORA approach

1. The input list is subject to some arbitrary cutoff
   - Different cutoff values can lead to different results
2. The “background universe” (all measurable features by the platform) may not be well defined
   - All those genes in the user uploaded data?
   - All those defined in the functional library?
3. Selection bias: all balls in the container should have equal chance of being selected) issue
   - Every gene has equal chance being sequenced by NGS?

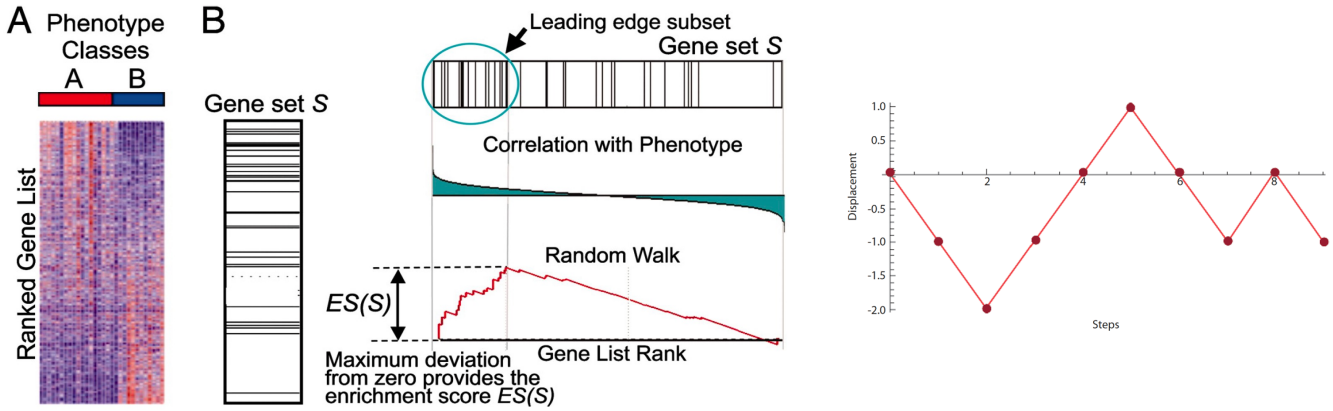


Gene Set Enrichment Analysis

- Cut-off free method
- Using the whole ranked gene list (not just those that are significant)
  - Are any gene sets (pathways) ranked surprisingly high or low?
  - Using permutation to evaluate whether this is surprise or not
- Claimed to be able to detect "subtle but consistent" changes



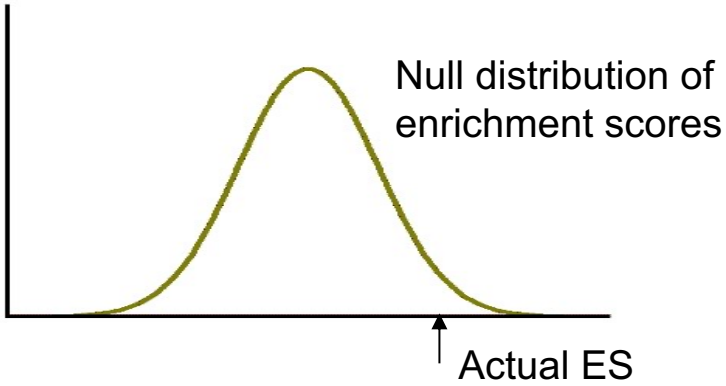
GSEA (Part I) – Enrichment Score (summary stat)



- GSEA walks down the ranked list of genes, increasing a running-sum statistic when a gene belongs to the set and decreasing it when the gene does not.
- The enrichment score (ES) is the maximum deviation from zero encountered during that walk. The ES reflects the degree to which the genes in a gene set are overrepresented at the top or bottom of the entire ranked list of genes.

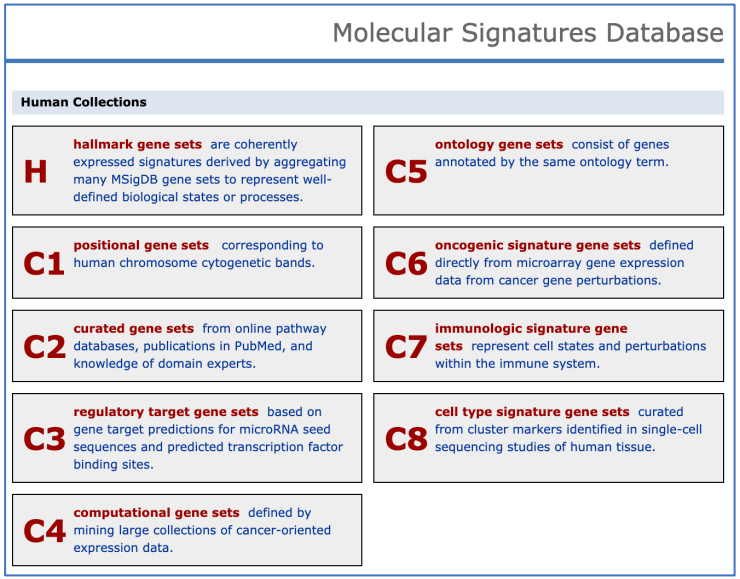
GSEA - Part II: Permutation Test

1. Randomise data (groups), rank genes again and repeat test 1000 times
2. Null distribution of 1000 ES for gene set
3. FDR q-value computed – corrected for gene set size and testing multiple gene sets



Molecular Signature Database (MSigDB)

- A companion database with GSEA software tool
- MSigDB offers gene sets based on various groupings
  - Pathways
  - GO terms
  - Chromosomal position
  - …



http://www.broadinstitute.org/gsea/msigdb

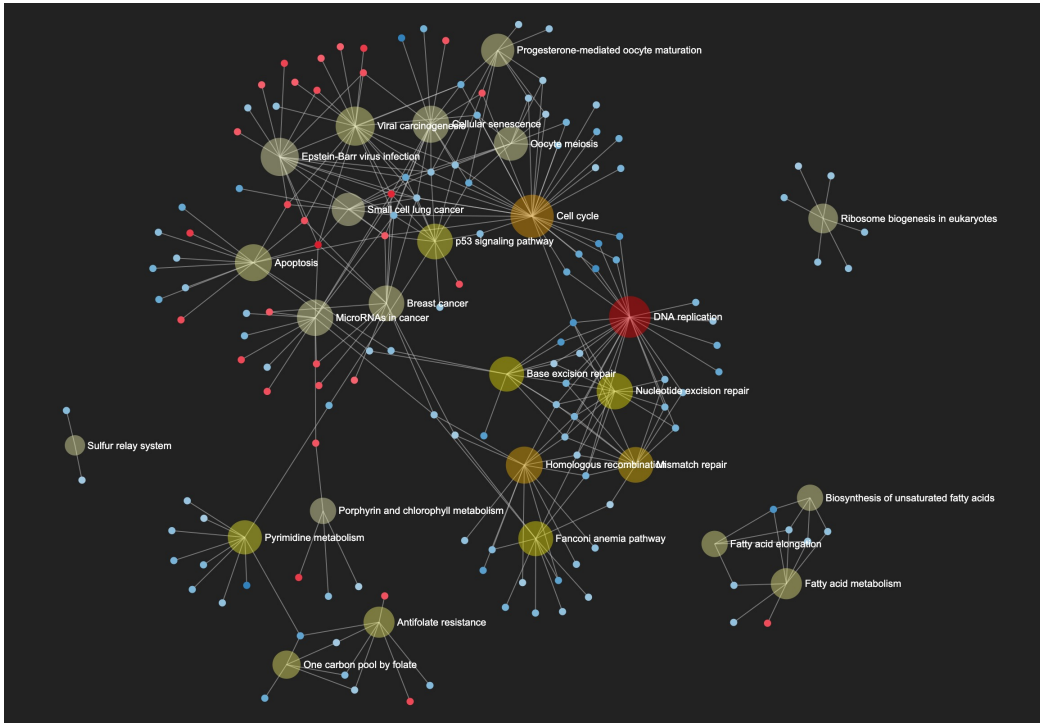
Issues with pathways or gene sets

- Many functions (pathways, gene sets) share some key genes
- Pathway boundaries are “fluid”
- Similar pathways are often of different sizes in different pathway databases
  - KEGG
  - Reactome
  - BioCyc
  - SMPDB
  - …..

➔ **Using Networks**

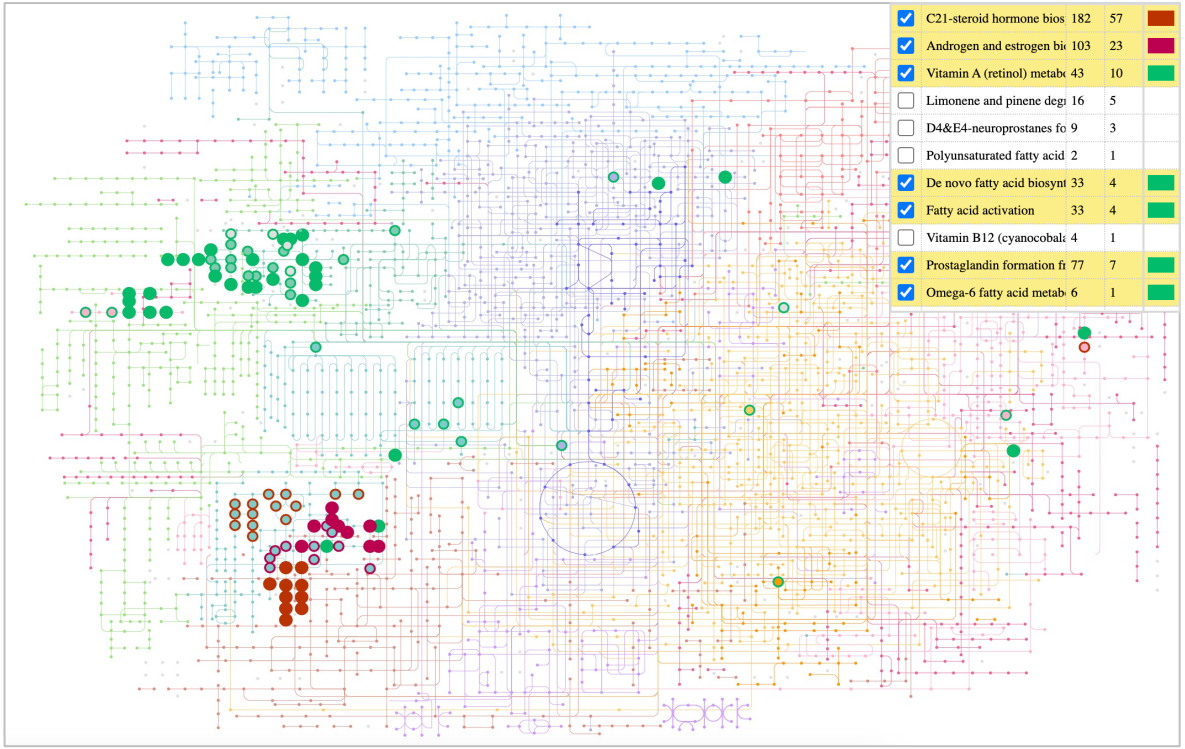
Enrichment network to visualize shared drivers

- Pathways overlap on the network
- Showing the key driver genes
- Similar pathways will be close to each other (pulled by same genes)



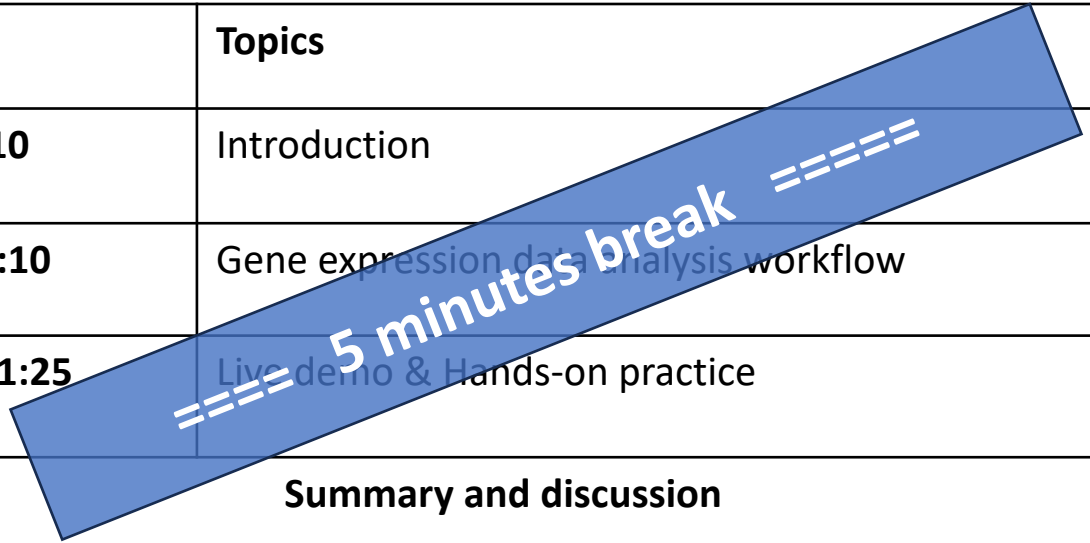
Global network to visualize converged functions

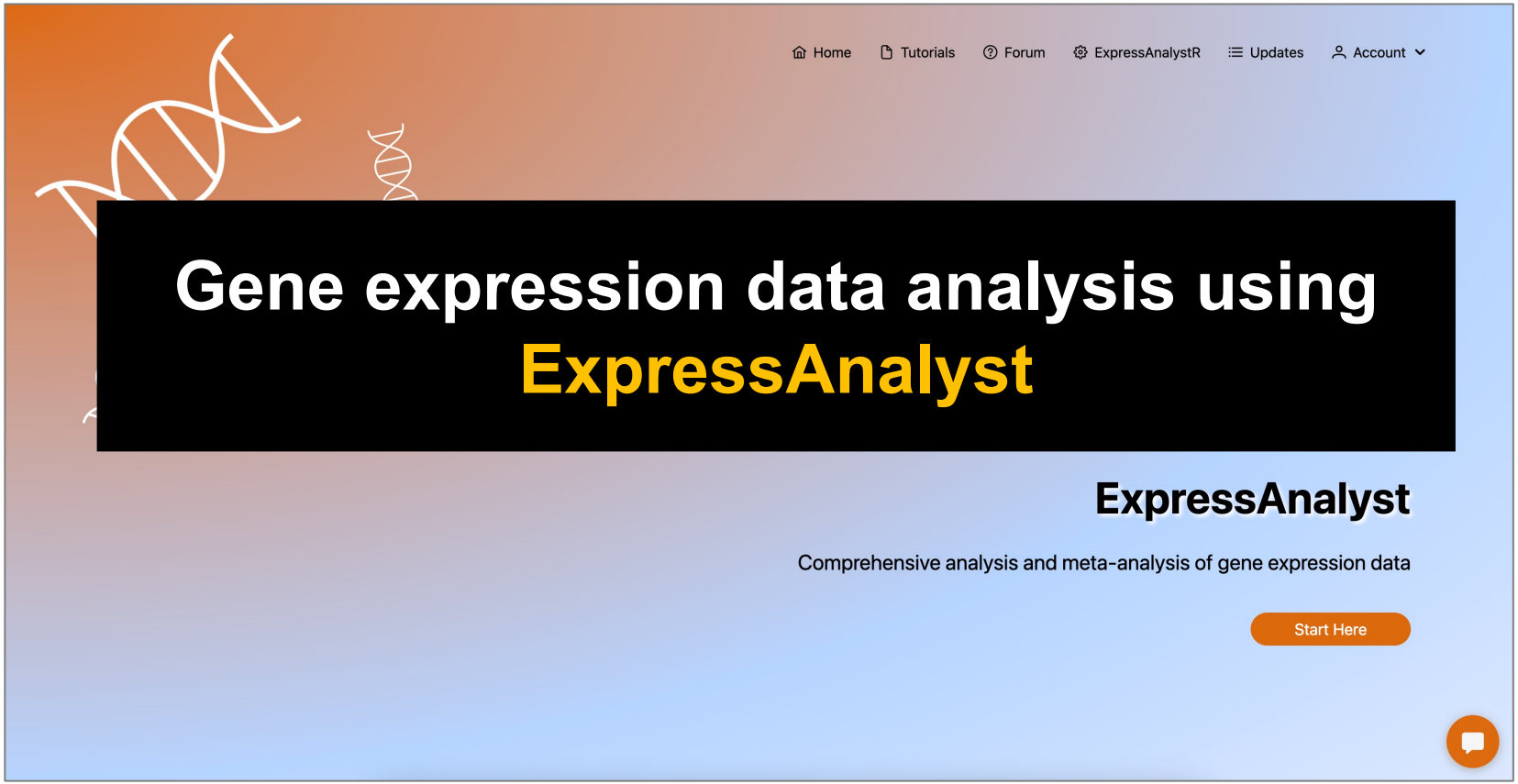
Pathways will converge on the global network to reveal common themes



Schedule for today

Time	Topics
9:00 – 9:10	Introduction
9:10 – 10:10	Gene expression data analysis workflow
10:15 – 11:25	Live demo & Hands-on practice
	Summary and discussion




Home
Tutorials
Forum
ExpressAnalystR
Updates
Account
Gene expression data analysis using
ExpressAnalyst
ExpressAnalyst
Comprehensive analysis and meta-analysis of gene expression data
Start Here

You are encouraged to follow the demo

Community version:

- https://www.expressanalyst.ca
- https://new.expressanalyst.ca

Pro version:

- https://pro.expressanalyst.ca

* You will be automatically assigned to one of the nodes

ExpressAnalyst

nature communications

Article

https://doi.org/10.1038/s41467-023-38785-y

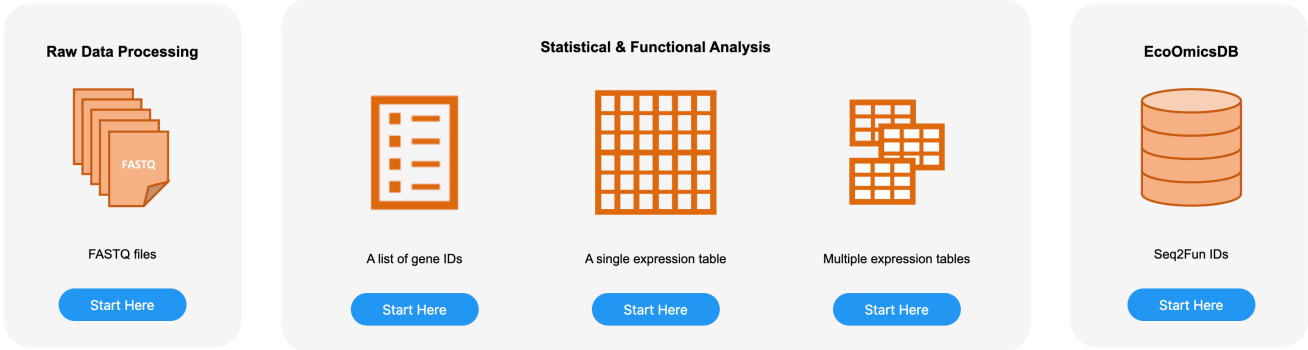
ExpressAnalyst: A unified platform for RNA-sequencing analysis in non-model species

Received: 20 October 2022

Accepted: 16 May 2023

Published online: 24 May 2023

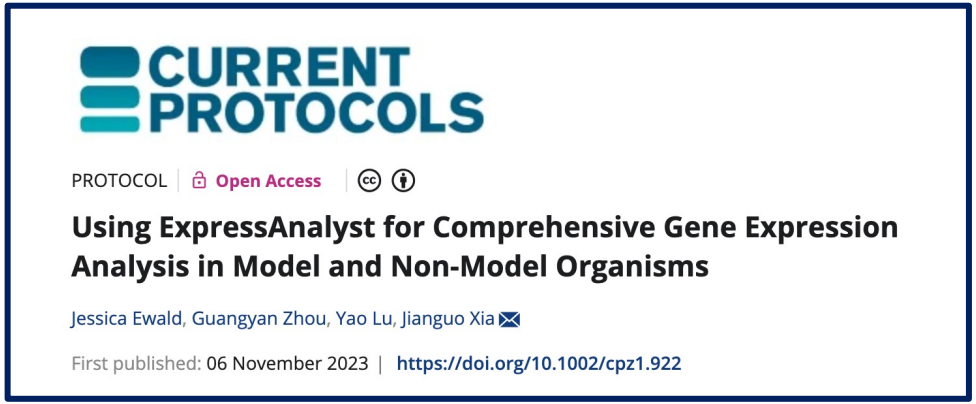
Check for updates



ExpressAnalyst

- RNAseq read mapping and quantification
  - Kallisto (model species)
  - Seq2Fun (non-model species)
- Differential gene expression analysis
  - Limma
  - edgeR
  - DEseq2
    - Sample size below 100
- Visual analytics
  - Volcano Plots
  - Heatmaps clustering
  - Ridgeline plots
  - Enrichment Networks
  - Upset Diagram (meta-analysis)
- Integrated with enrichment analysis
  - ORA
  - GSEA

Our Tasks



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

First published: 06 November 2023 | https://doi.org/10.1002/cpz1.922

- **Basic Protocol 1**: RNA-seq count table uploading, processing, and normalization
- **Basic Protocol 2**: Differential expression analysis with linear models
- **Basic Protocol 3**: Functional analysis with volcano plot, enrichment network, and ridgeline visualization
- **Basic Protocol 4**: Hierarchical clustering analysis of transcriptomics data using interactive heatmaps
- **Basic Protocol 9**: Functional analysis of transcriptomics signatures

Demo Overview

- Dataset: RNA-seq data collected from mouse liver to study the effect of Bisphenol-A exposure during pregnancy on offspring.
  - 16 samples
  - 8 male and 8 female offsprings
- Objectives:
  - Format data and metadata file for ExpressAnalyst
  - Data processing
    - Quality check
    - Filtering and normalization
  - Differential expression analysis
  - Functional analysis
- **Pro features:**
  - Project saving
  - Report generation

Data format

- .csv or .txt format
- Samples in columns.
- Genes in rows.
- Important that first row starts with **#NAME** followed by sample names

	A	B	C	D	E	F	G	H	I	J	K	L
1	#NAME	BPA5F1	BPA1F1	BPA4F2	BPA4F3	CON12F1	CON12F2	CON8F1	CON9F2	BPA4M2	BPA5M1	BPA1M2
2	Plekhg2	109	174	80	76	83	88	52	76	99	76	76
3	Plekhg3	1281	1227	1263	990	546	687	561	661	866	799	895
4	Plekhg1	485	462	467	562	347	296	458	487	261	349	338
5	Plekhg6	126	172	206	198	137	164	110	152	247	193	223
6	4930592I03Rik	7	5	9	7	2	7	0	7	4	3	9
7	Plekhg4	0	0	0	0	0	0	0	0	0	0	0
8	Plekhg5	117	150	138	95	83	109	105	91	77	120	78
9	Nsa2	1540	1472	1577	1792	1419	1580	1288	1670	1171	1260	1275
10	Nampt	2882	3103	3487	3090	2400	3148	2343	2569	2242	3137	2349
11	Vmo1	43	68	51	62	24	30	22	26	30	54	51
12	Man2a1	11979	13268	15988	14592	7416	9708	7285	10071	11742	13363	12304
13	F930015N05Rik	15	17	17	17	2	9	18	9	14	5	11
14	4931406P16Rik	955	1038	971	934	514	810	559	874	526	572	708
15	Vwf	289	329	362	284	208	211	170	181	151	246	244
16	Mir302d	0	0	0	0	0	0	0	0	0	0	0
17	Mier1	1262	1667	1510	1597	534	1086	558	1135	976	1084	1154
18	Mier2	152	245	184	166	114	141	104	126	180	142	180
19	Mier3	949	1066	1220	1214	780	1177	808	1179	829	937	834
20	Aqp11	361	357	420	431	184	307	238	247	343	439	384
21	Aqp12	0	0	0	0	0	0	0	0	0	0	0
22	1810010H24Rik	71	74	92	87	97	135	89	132	68	105	74
23	Rpl31-ps12	239	223	216	238	75	119	85	112	151	161	193
24	Dag1	589	3324	3078	743	659	841	984	728	1910	2736	3035
25	Myadm	585	562	600	734	525	727	432	641	485	477	341
26	Gm15008	0	0	0	0	0	0	0	0	0	0	0
27	Sacs	26	22	24	28	11	16	17	23	12	17	14
28	Zfp712	82	79	77	89	71	80	71	55	33	60	54
29	Zfp710	447	531	451	448	329	393	405	298	328	397	432
30	Zfp711	0	2	0	0	0	0	0	0	0	2	0
31	Zfp715	539	653	632	624	377	495	357	520	328	432	472

Metadata file

- .csv or .txt
- Samples in rows
  - Make sure sample names match with the ones in data file.
- Metadata variables in columns
- Important that first row starts with **#NAME** followed by metadata group name

	A	B	C	D	E	F	G
1	#NAME	Treatment	Sex	liver_TG	liver_TC	liver_UC	liver_PC
2	BPA5F1	BPA	Female	77.5	2.82	2.35	15.26
3	BPA1F1	BPA	Female	77.5	2.82	2.35	15.26
4	BPA4F2	BPA	Female	81.65	3.22	2.47	14.85
5	BPA4F3	BPA	Female	98.92	3.24	2.27	16.38
6	CON12F1	Control	Female	180.19	3.45	2.68	18.4
7	CON12F2	Control	Female	134	3	2.4	15.6
8	CON8F1	Control	Female	137.74	3.55	2.71	18.78
9	CON9F2	Control	Female	160.54	3.27	2.72	15.51
10	BPA4M2	BPA	Male	88.97	2.74	2.28	15.29
11	BPA5M1	BPA	Male	63.03	2.91	2.42	16.24
12	BPA1M2	BPA	Male	99.23	2.77	2.31	14.77
13	BPA1M3	BPA	Male	123.62	3.84	3.2	24.72
14	CON11M2	Control	Male	84.46	2.29	2.11	14.6
15	CON8M1	Control	Male	73.19	2.12	1.93	12.71
16	CON8M2	Control	Male	91.94	2.83	2.36	16.27
17	CON9M2	Control	Male	110.68	2.72	2.14	15.15

Next Lecture

1. Network analysis from a gene list
   - ➢PPI network
   - ➢Tissue-specific, signaling networks
2. Meta-analysis
3. RNAseq in non-model species

We would like to hear your comments & feedback

contact@xialab.ca

See you next week!