RNA-seq differential expression analysis

SciLifeLab RNA-seq & proteomics workshop

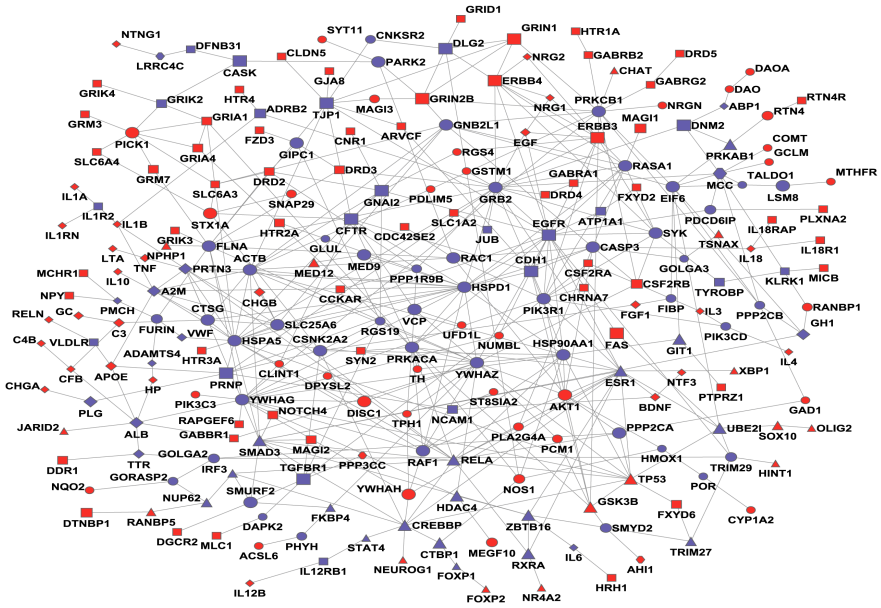
Oct 30, 2014

Estelle Proux-Wéra, SciLifeLab / Stockholm University, Sweden

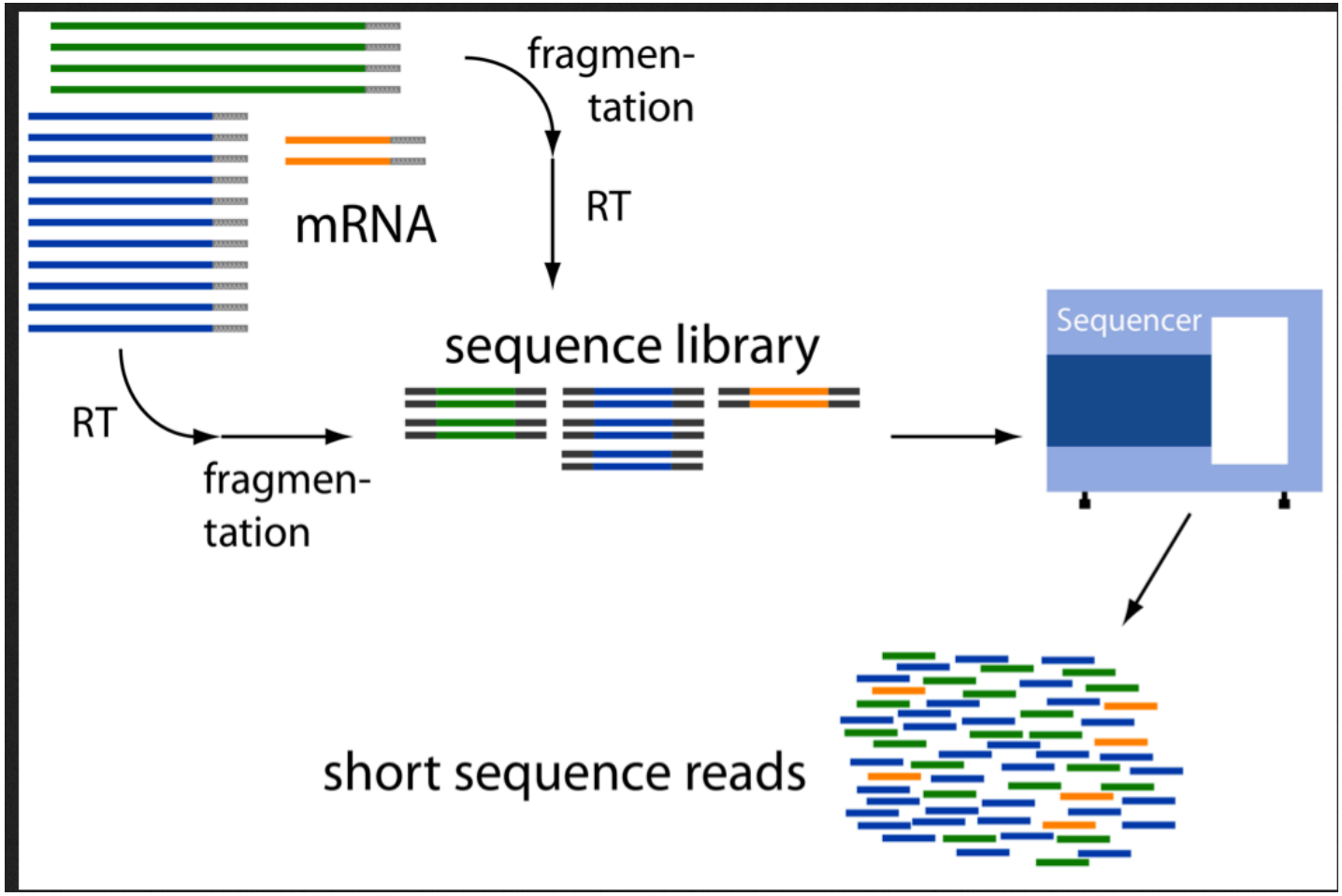
Differential expression analysis

The identification of genes (or other types of genomic features, such as transcripts or exons) that are expressed in significantly different quantities in distinct groups of samples, be it biological conditions (drug-treated vs. controls), diseased vs. healthy individuals, different tissues, different stages of development, or something else.

Typically **univariate** analysis (one gene at a time) – even though we know that genes are not independent

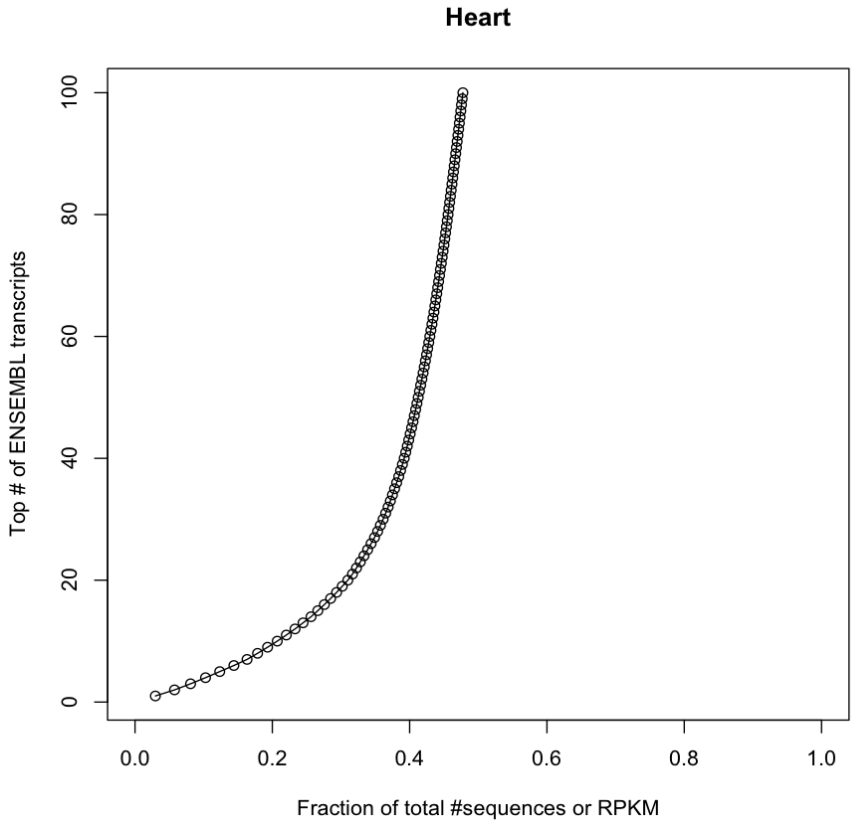


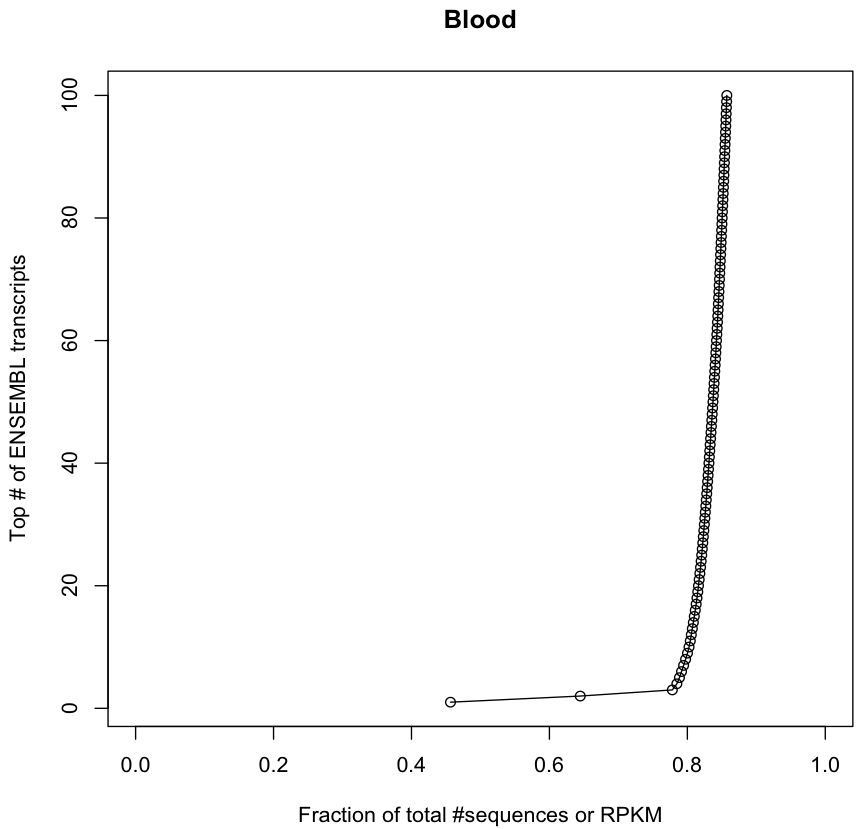
How are RNA-seq data generated?



Sampling process

”Transcriptional real estate”





Count-based statistics

People often use discrete distributions (Poisson, negative binomial etc.) rather than continuous (e g normal) distributions for modeling RNA-seq data.

This is natural when you consider the way data are generated.

Experimental design

DOI: 10.1534/genetics.110.114983

Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge[1]

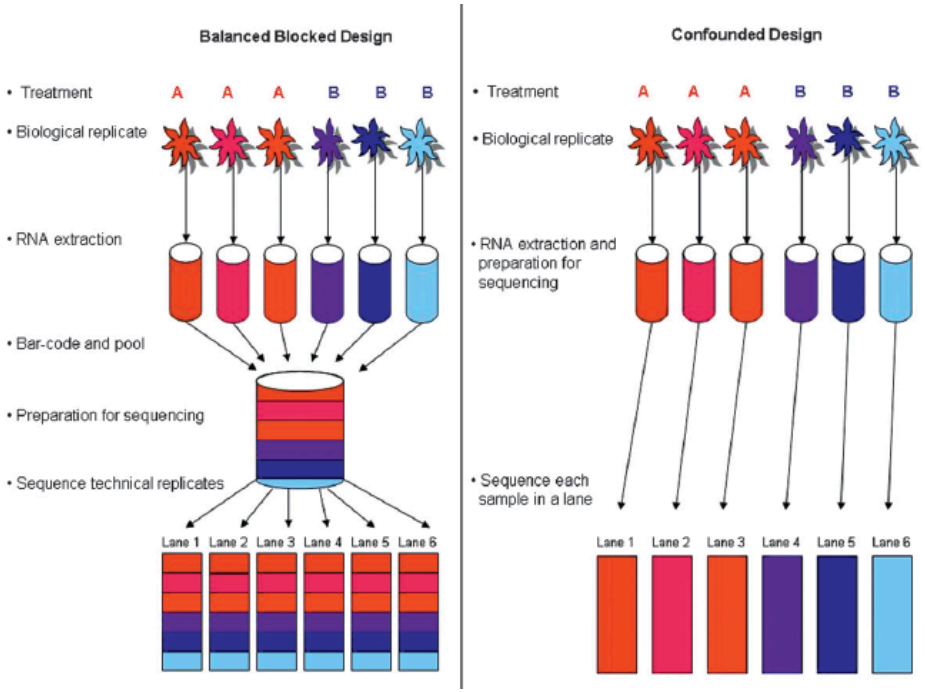
Department of Statistics, Purdue University, West Lafayette, Indiana 47907

Manuscript received January 31, 2010
Accepted for publication March 15, 2010

http://www.genetics.org/content/185/2/405

Important for subsequent DE analysis!

Replication
Randomization
Blocking



Technical vs biological replicates

Technical replicates:

Assess variability of measurement technique
Typically low for bulk RNA-seq (not necessarily single-cell RNA-seq)
Poisson distribution can model variability between RNA-seq technical replicates rather well

Biological replicates:

Assess variability between individuals / “normal” biological variation
Necessary for drawing conclusions about biology
Variability across RNA-seq biological replicates not well modelled by Poisson – usually negative binomial (“overdispersed Poisson”) is used

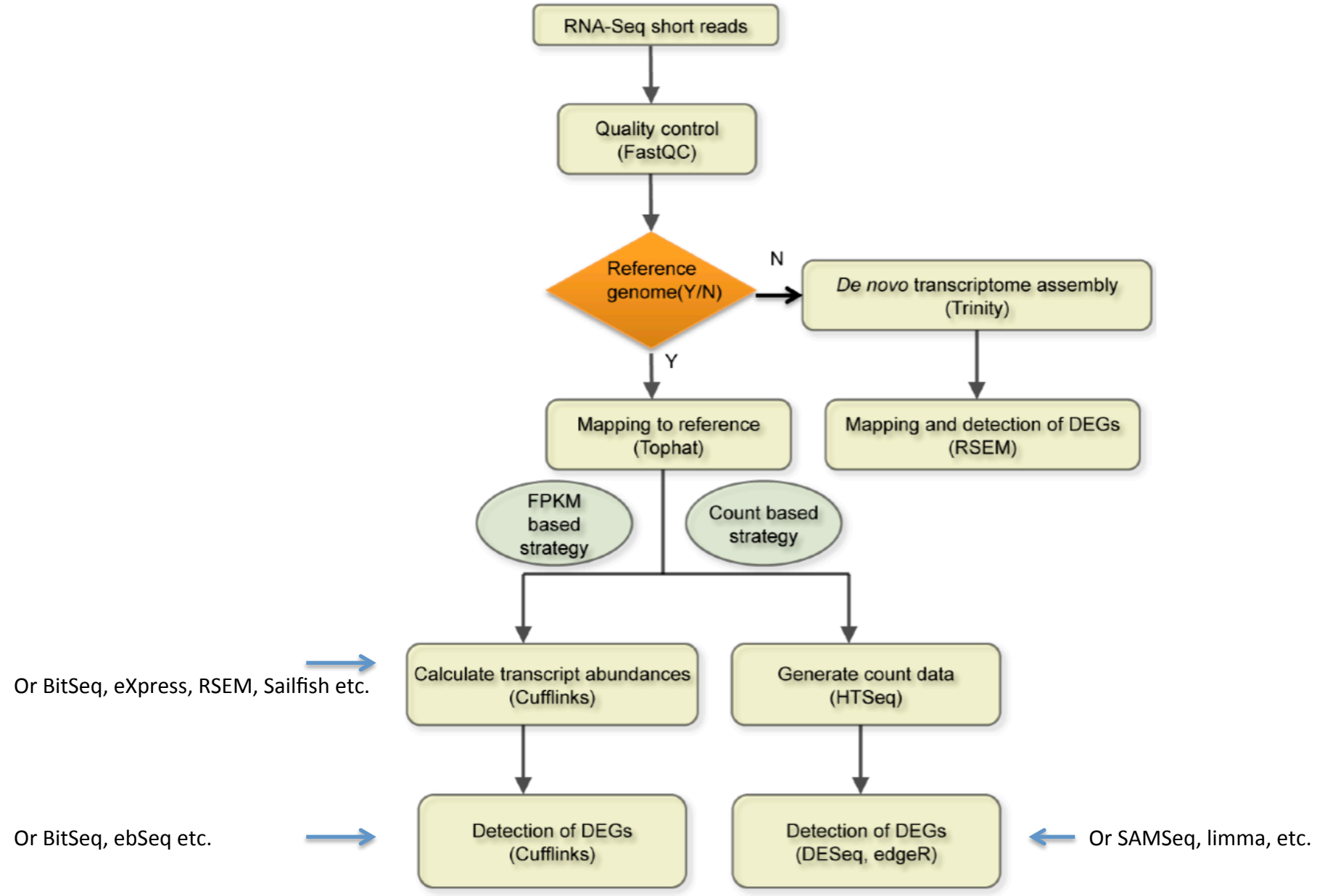
Replicates and differential expression

Intuitively, the variation **between** the groups that you want to compare should be large compared to the variation **within** each group to be able to say that we have differential expression.

The more biological replicates, the better you can estimate the variation. But how many replicates are needed?

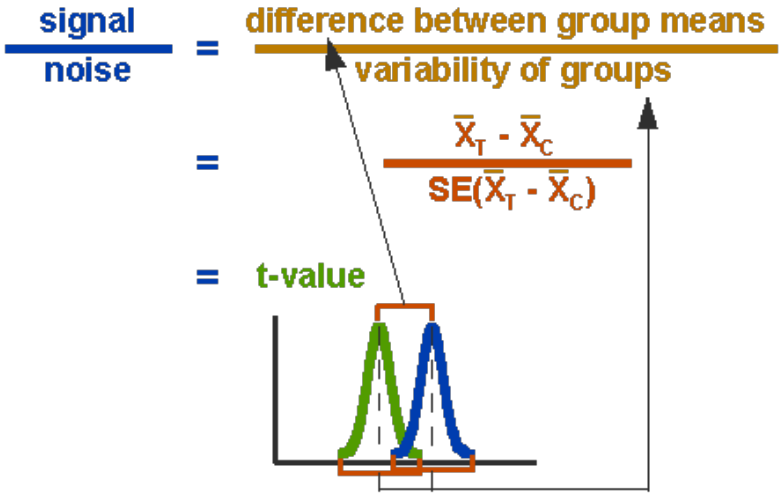
Depends:

Homogeneous cell lines, inbred mice etc: maybe 3 samples / group enough.
Clinical case-control studies on patients: can need a dozen, hundreds or thousands, depending on the specifics ….


RNA-Seq short reads
Quality control (FastQC)
Reference genome(Y/N)
N
De novo transcriptome assembly (Trinity)
Y
Mapping to reference (Tophat)
Mapping and detection of DEGs (RSEM)
FPKM based strategy
Count based strategy
Calculate transcript abundances (Cufflinks)
Generate count data (HTSeq)
Detection of DEGs (Cufflinks)
Detection of DEGs (DESeq, edgeR)
Or BitSeq, eXpress, RSEM, Sailfish etc.
Or BitSeq, ebSeq etc.
Or SAMSeq, limma, etc.

Problems associated with a t test

Couldn't we just use a Student's t test for each gene?



http://www.socialresearchmethods.net/kb/stat_t.php

Problems with this approach:

- May have **few replicates**

- Distribution is **not normal**

- **Multiple testing** issues

TABLE 8.1 List of (some) Software Tools for Differential Expression Analysis

Software Tool	Type of Software	Analysis Approach	Comment
DESeq	R/Bioconductor package	Count-based (negative binomial)	Considered conservative (low false-positive rate)
edgeR	R/Bioconductor package	Count-based (negative binomial)	Similar to DESeq in philosophy
tweeDESeq	R/Bioconductor package	Count-based (Tweedie distribution family)	More general than DESeq/edgeR, but new and not widely tested
Limma	R/Bioconductor package	Linear models on continuous data	Originally developed for microarray analysis, very thoroughly tested. Need to preprocess counts to continuous values
SAMSeq (samr)	R package	Nonparametric test	Adapted from the SAM microarray DE analysis approach. Works better with more replicates
NOISeq	R/Bioconductor package	Nonparametric test	
CuffDiff	Linux command line tool	Isoform deconvolution + count-based tests	Can give differentially expressed isoforms as well as genes (also differential usage of TSS, splice sites)
BitSeq	Linux command line tool and R package	Isoform deconvolution in a Bayesian framework	Can give differentially expressed isoforms. Also calculates (gene and isoform) expression estimates
ebSeq	R/BioConductor package	Isoform deconvolution in a Bayesian framework	Can give differentially expressed isoforms. Can be used in a pipeline preceded by RSEM expression estimation

Parametric vs. non-parametric methods

It would be nice to not have to assume anything about the expression value distributions but only use rank-order statistics. -> methods like SAM (Significance Analysis of Microarrays) or SAM-seq (equivalent for RNA-seq data)

However, it is (typically) harder to show statistical significance with non-parametric methods with few replicates.

My rule of thumb:

-Many replicates (~ >10) in each group -> use SAM(Seq)
-Otherwise use DESeq or other parametric method

Note that according to Simon Anders (creator of DESeq) says that non-parametric methods are definitely better with 12 replicates and maybe already at five

http://seqanswers.com/forums/showpost.php?p=74264&postcount=3

Dealing with the “t test issues”

Distributional issue: Solved by variance stabilizing transform in limma - voom() function

edgeR and DESeq model the count data using a *negative binomial distribution* and use their own modified statistical tests based on that.

Dealing with the “t test issues”

Distributional issue: Solved by variance stabilizing transform in limma – voom() function

edgeR and DESeq model the count data using a *negative binomial distribution* and use their own modified statistical tests based on that.

Multiple testing issue: All of these packages report false discovery rate (corrected p values). For SAMseq based on resampling, for others usually Benjamini-Hochberg corrected p values.

Dealing with the “t test issues”

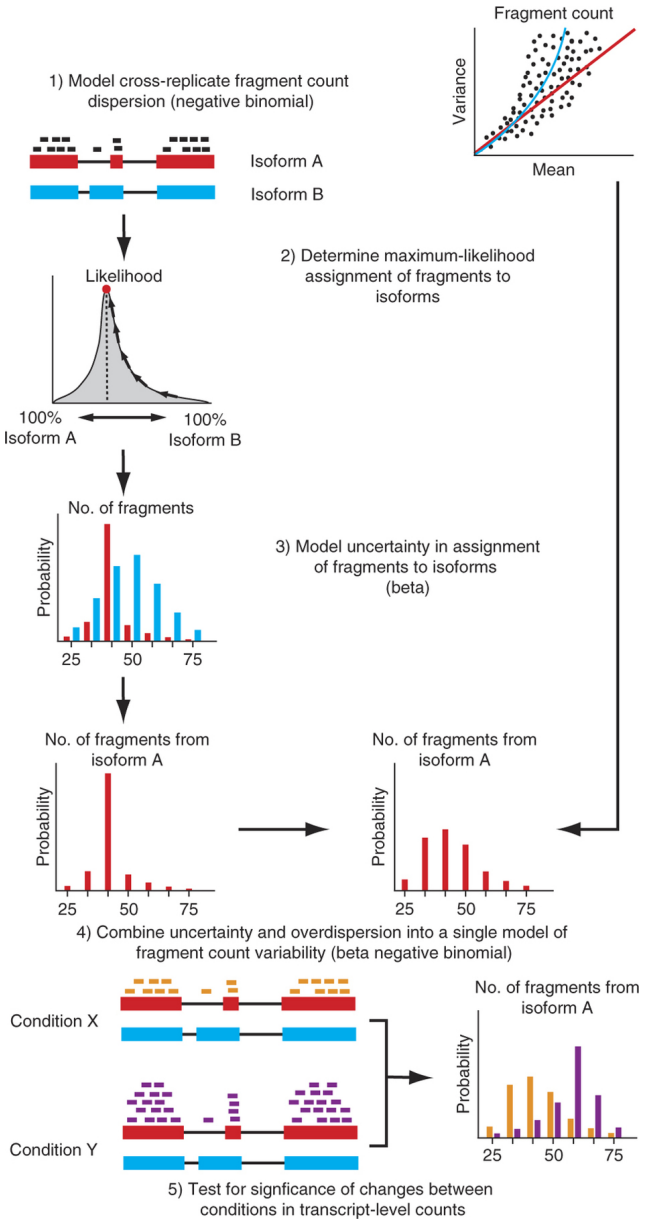
Distributional issue: Solved by variance stabilizing transform in limma – voom() function

edgeR and DESeq model the count data using a *negative binomial distribution* and use their own modified statistical tests based on that.

Multiple testing issue: All of these packages report false discovery rate (corrected p values). For SAMseq based on resampling, for others usually Benjamini-Hochberg corrected p values.

Variance estimation issue: edgeR, DESeq2 and limma (in slightly different ways) “borrow” information across genes to get a better variance estimate. One says that the estimates “shrink” from gene-specific estimates towards a common mean value.

CuffDiff2



Integrates isoform quantification + differential expression analysis.

Also: **BitSeq**

Complex designs

The simplest case is when you just want to compare two groups against each other.

But what if you have several factors that you want to control for?

E.g. you have taken tumor samples at two different time points from six patients, cultured the samples and treated them with two different anticancer drugs and a mock control treatment. -> 2x6x3 = 36 samples.

Now you want to assess the differential expression in response to one of the anticancer drugs, drug X. You could just compare all “drug X” samples to all control samples but the inter-subject variability might be larger than the specific drug effect.

→Enter limma / DESeq / edgeR which can work with factorial designs

(SAMSeq cannot, which is another reason one might not want to use it)

DESeq and factorial design

Count matrix: countdata

	S1	S2	S3	S4	S5	S6	S7	S8
Gene1	679	448	873	408	1138	1047	770	572
Gene2	467	515	621	365	587	799	417	508
Gene3	260	211	263	164	245	331	233	229
Gene4	60	55	40	35	78	63	76	60
…								

From http://www.bioconductor.org/help/workflows/rnaseqGene/#de

DESeq and factorial design

Metadata table: coldata

	SampleName	cell	dex
S1	GSM1275862	N61311	untrt
S2	GSM1275863	N61311	trt
S3	GSM1275866	N052611	untrt
S4	GSM1275867	N052611	trt
S5	GSM1275870	N080611	untrt
S6	GSM1275871	N080611	trt
S7	GSM1275874	N061011	untrt
S8	GSM1275875	N061011	trt

DESeq and factorial design

In R: create a DESeq object

```
dds <- DESeqDataSetFromMatrix(
    countData = countdata,
    colData = coldata,
    design = ~ cell + dex))
```

=> we want to test for the effect of dexamethasone (the last factor), controlling for the effect of different cell line (the first factor).

Limma and factorial designs

limma stands for “linear models for microarray analysis” – but it can be used for RNA-seq after applying voom() to a count matrix
Essentially, the expression of each gene is modeled with a linear relation

Linear Models

- In general, need to specify:
  - Dependent variable
  - Explanatory variables (experimental design, covariates, etc.)
- More generally:

$$y = X\beta + \epsilon$$

vector of observed data — design matrix — Vector of parameters to estimate

http://www.math.ku.dk/~richard/courses/bioconductor2009/handout/19_08_Wednesday/KU-August2009-LIMMA/PPT-PDF/Robinson-limma-linear-models-ku-2009.6up.pdf

The design matrix describes all the conditions, e g treatment, patient, time etc
y = a + b*treatment + c*time + d*patient + e*batch + f

Baseline/average

Error term/noise

Decision tree for software selection

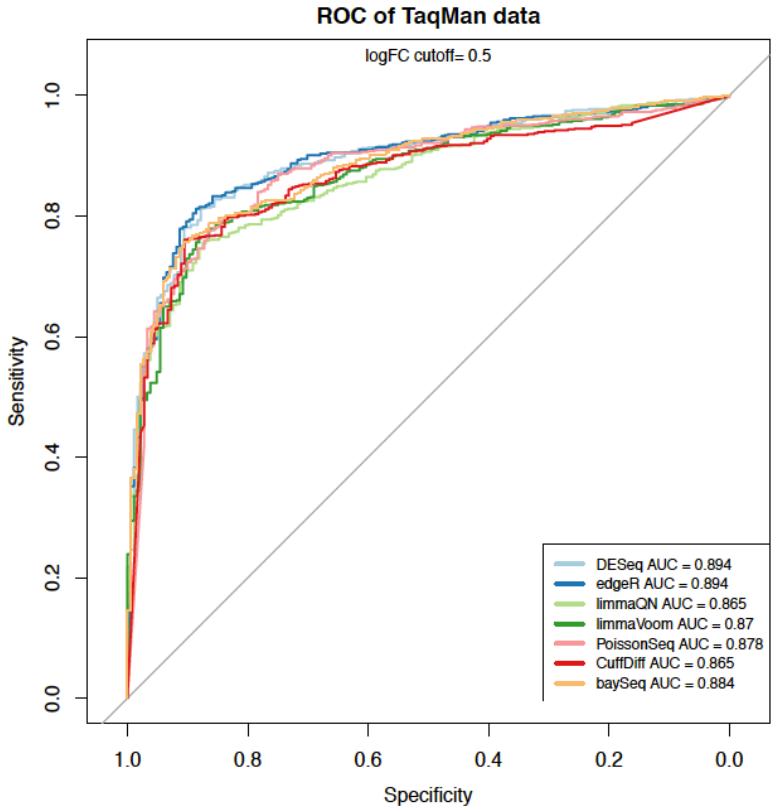
Differentially expressed **exons** => *DEXSeq*

Differentially expressed **isoforms** => *BitSeq, Cuffdiff or ebSeq*

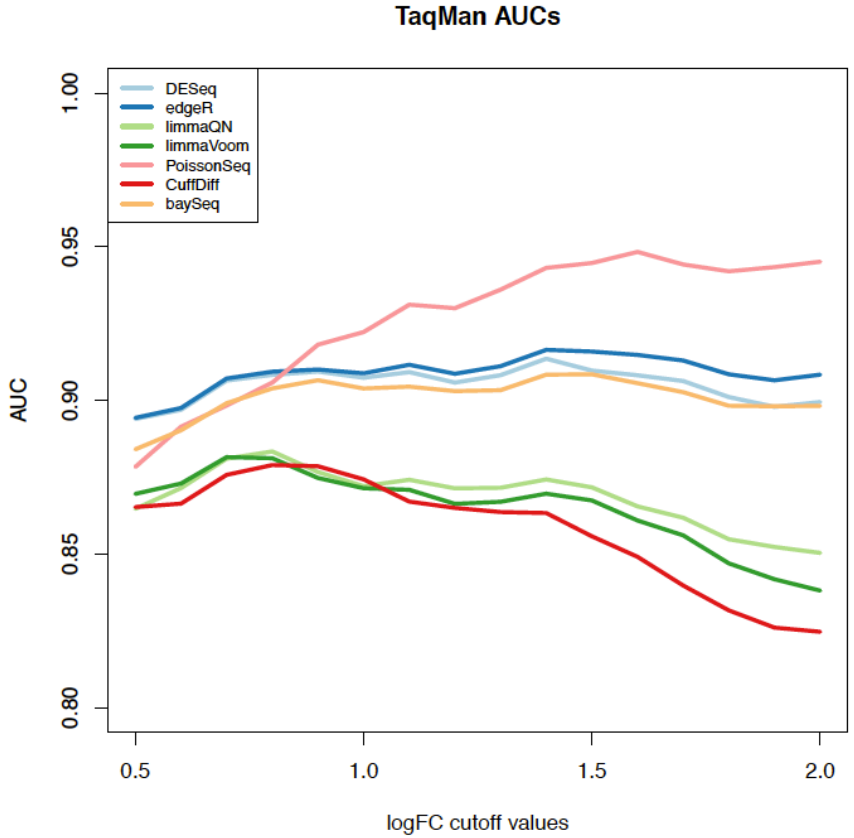
Differentially expressed genes => **Select type of experimental design**

- Complex design (more than one varying factor) => *DESeq, edgeR, limma*
- Simple comparison of groups => **How many biological replicates?**
  - More than about 5 biological replicates per group => *SAMSeq*
  - Less than 5 biological replicates per group => *DESeq, edgeR, limma*

Recent DE software comparisons (1)



(a)



(b)

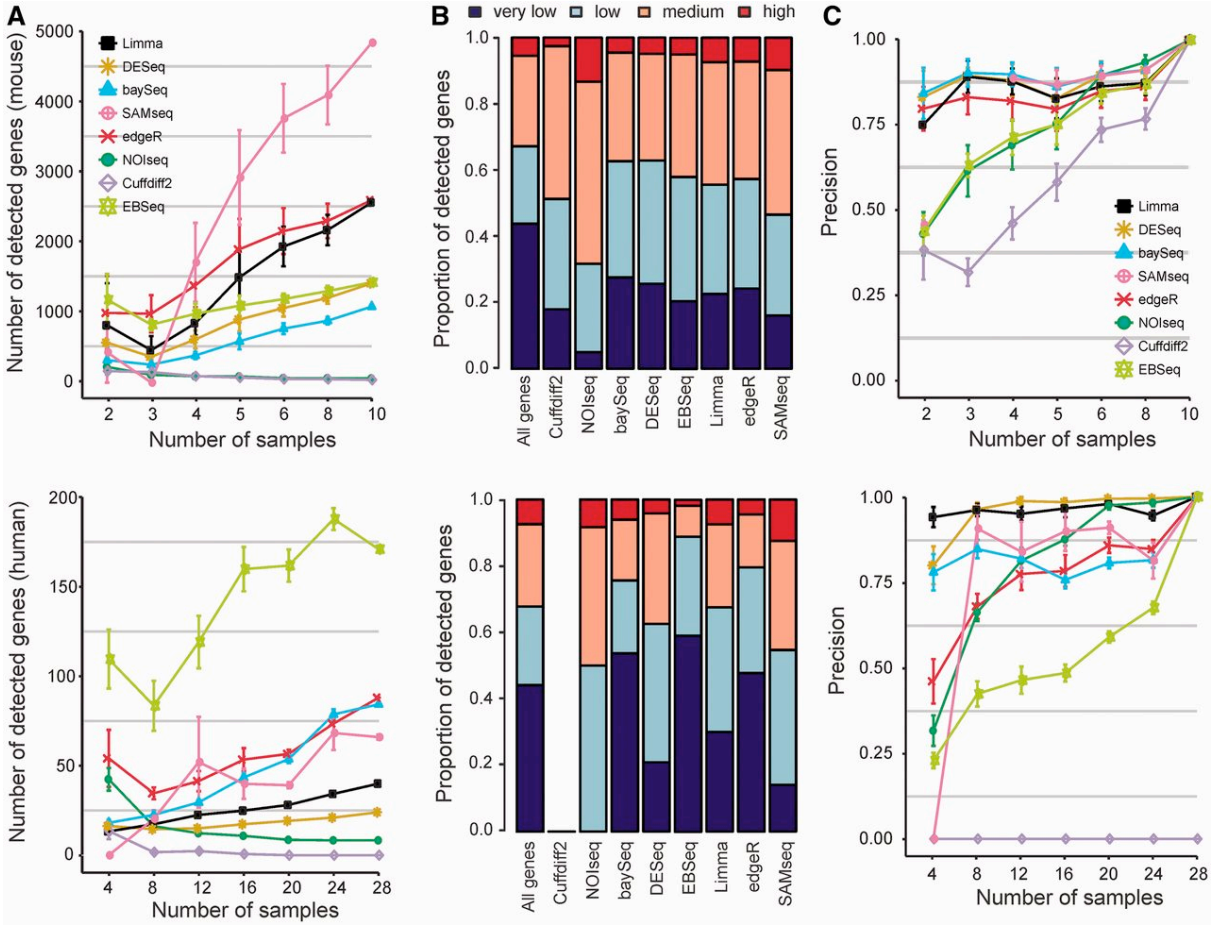
Comprehensive evaluation of differential expression analysis methods for RNA-seq data

Franck Rapaport [1], Raya Khanin [1], Yupu Liang [1], Azra Krek [1], Paul Zumbo [2,4], Christopher E. Mason [2,4], Nicholas D. Socci [1], Doron Betel [3,4]

[1]Bioinformatics Core, Memorial Sloan-Kettering Cancer Center, New York
[2]Department of Physiology and Biophysics, Weill Cornell Medical College, New York
[3] Division of Hematology/Oncology, Department of Medicine, Weill Cornell Medical College, New York
[4] Institute for Computational Biomedicine, Weill Cornell Medical College, New York

January 24, 2013

Recent DE software comparisons (2)



Briefings in Bioinformatics Advance Access published December 2, 2013
BRIEFINGS IN BIOINFORMATICS. page 1 of 12
doi:10.1093/bib/bbt086

Comparison of software packages for detecting differential expression in RNA-seq studies

Fatemeh Seyednasrollah, Asta Laiho and Laura L. Elo

Submitted: 20th August 2013; Received (in revised form): 9th October 2013

Recent DE software comparisons (3-4)

Research article

Highly accessed | Open Access

A comparison of methods for differential expression analysis of RNA-seq data

Charlotte Soneson[1]* and **Mauro Delorenzi**[1,2]

* Corresponding author: Charlotte Soneson Charlotte.Soneson@isb-sib.ch ▾ Author Affiliations

[1] Bioinformatics Core Facility, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[2] Département de formation et recherche, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland

For all author emails, please log on.

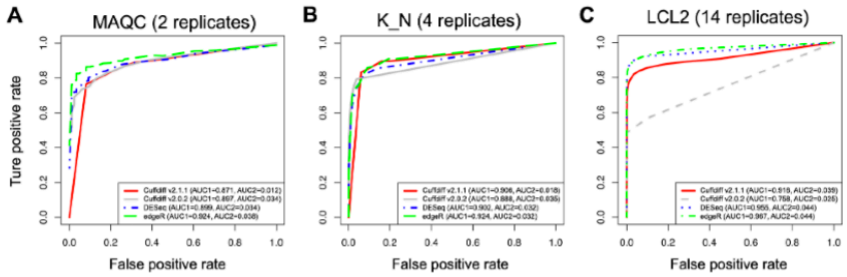
BMC Bioinformatics 2013, **14**:91 doi:10.1186/1471-2105-14-91

Nice code examples in supplementary material: R code for all tested packages

A comparative study of techniques for differential expression analysis on RNA-Seq data

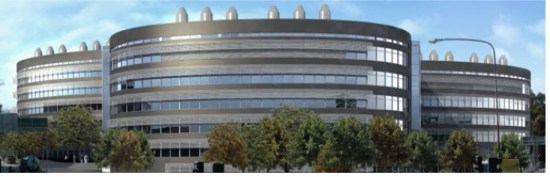
Zong Hong Zhang, Dhanisha J. Jhaveri, Vikki M. Marshall, et al.

bioRxiv posted online May 28, 2014
Access the most recent version at doi: http://dx.doi.org/10.1101/005611



SciLifeLab

Karolinska Institutet | KTH Vetenskap och Konst | Stockholms universitet | Uppsala Universitet



Take-away messages from DE tool comparison

-CuffDiff2, which should theoretically be better, seems to work worse, perhaps due to the increased “statistical burden” from isoform expression estimation. Two studies also report it has an optimum number of replicates at ~4-5

-The HTSeq quantification which is theoretically “wrong” seems to give good results with downstream software

-Limma, which does not use the negative binomial distribution seems to work well

-It is practically always better to sequence more biological replicates than to sequence the same samples deeper

Omitted from these comparisons:

- gains from ability to do complex designs
- isoform-level DE analysis (hard to establish ground truth)
- some packages like BitSeq

Normalization/scaling/transformation: different goals

- **R/FPKM:** (Mortazavi et al. 2008)
  - Correct for: differences in sequencing depth and transcript length
  - Aiming to: compare a gene across samples and diff genes within sample

- **TMM**: (Robinson and Oshlack 2010)
  - Correct for: differences in transcript pool composition; extreme outliers
  - Aiming to: provide better across-sample comparability

- **TPM**: (Li et al 2010, Wagner et al 2012)
  - Correct for: transcript length distribution in RNA pool
  - Aiming to: provide better across-sample comparability

- **Limma voom (logCPM)**: (Lawet al 2013)
  - Aiming to: stabilize variance; remove dependence of variance on the mean

Optimal Scaling of Digital Transcriptomes

Gustavo Glusman, Juan Caballero, Max Robinson, Burak Kutlu, Leroy Hood

Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

TMM – Trimmed Mean of M values

Attempts to correct for differences in RNA *composition* between samples

E g if certain genes are very highly expressed in one tissue but not another, there will be less "sequencing real estate" left for the less expressed genes in that tissue and RPKM normalization (or similar) will give biased expression values for them compared to the other sample

RNA population 1

RNA population 2

Equal sequencing depth -> orange and red will get lower RPKM in RNA population 1 although the expression levels are actually the same in populations 1 and 2

Robinson and Oshlack Genome Biology 2010, 11:R25, http://genomebiology.com/2010/11/3/R25

Normalization in DE analysis

edgeR, DESeq2 and some others want to keep the (integer) read counts in the DE testing because they

- Use a discrete statistical model
- Want to retain statistical power (see next slide)

… but they **implicitly** normalize (by TMM in edgeR and RLE in DESeq2) as part of the DE analysis.

Programs like SAMSeq and limma are fine with continuous values (like FPKM), the former because it has a **rank based model** and the latter because it cares more about the **mean-variance relationship** being weak. They also apply their own types of normalization as part of the DE testing.

Count nature of RNA-seq data

Programs like edgeR and DESeq2 want to make use of the count nature of RNA-seq data to increase statistical power. The reasoning goes something like this:

(simplified toy example!)

Scenario 1: A 30000-bp transcript has 1000 counts in sample A and 700 counts in sample B.

Scenario 2: A 300-bp transcript has 10 counts in sample A and 7 counts in sample B.

Assume that the sequencing depths are the same in both samples and both scenarios. Then **the RPKM is the same** in sample A in both scenarios, and in sample B in both scenarios.

In scenario A, we can be more confident that there is a true difference in the expression level than in scenario B (although we would want replicates of course!) by analogy to a coin flip – 600 heads out of 1000 trials gives much more confidence that a coin is biased than 6 heads out of 10 trials

Batch normalization

Often, putting the experimental batch as a **factor** in the **design matrix** is enough.

If you wish to explicitly normalize away the batch effects (to get a new, batch-normalized expression matrix with continuous values), you can use a method such as ComBat.

(Designed for microarrays, should use voom():ed values for RNA-seq)

Johnson, WE, Rabinovic, A, and Li, C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics 8(1):118-127.

Independent filtering

Pre-filtering a count table using criteria such as (for example):

- Keep only genes with an average count per million > 1
- Keep only genes where 50% of the samples have > 0 counts
- Keep only genes where the sum of counts across samples > 10

(etc)

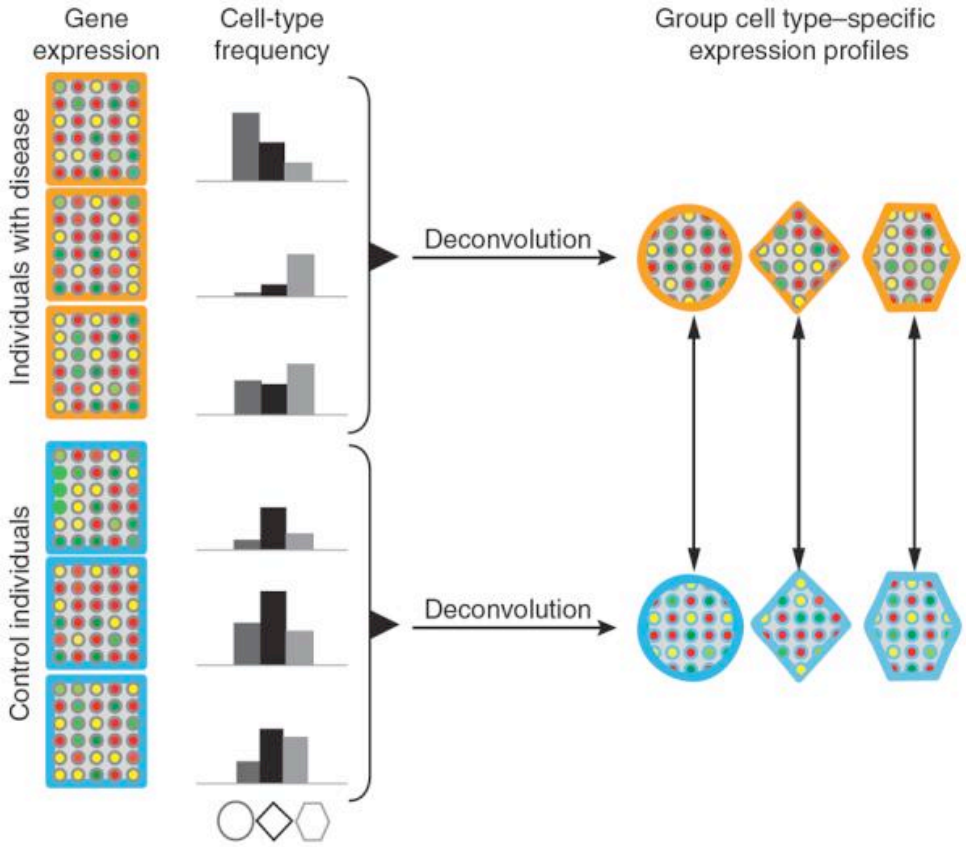
Usually helps achieve better adjusted p values, probably due to both smaller multiple-testing penalty and irrelevant “noisy” genes messing with parameter fitting

More sophisticated data-driven methods exist, e g

HTSFilter: An independent data-based filter for replicated high-throughput transcriptome sequencing experiments

A. Rau[a], **M. Gallopin**[a], **G. Celeux**[b] and **F. Jaffrézic**[a]

DE analysis in mixtures of cell types



CellMix, R package implementing several deconvolution methods (most for microarray)

Gaujoux R, Seoighe C. CellMix: a comprehensive toolbox for gene expression deconvolution. Bioinformatics. 2013 Sep 1;29(17):2211-2. doi: 10.1093/bioinformatics/btt351.

Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ. Cell type-specific gene expression differences in complex tissues. Nat Methods. 2010 Apr;7(4):287-9.

Differential expression analysis output

Top 10 differentially expressed genes tables for each contrast

Top differentially expressed genes: full_table_E16.5wt-E16.5ko.txt

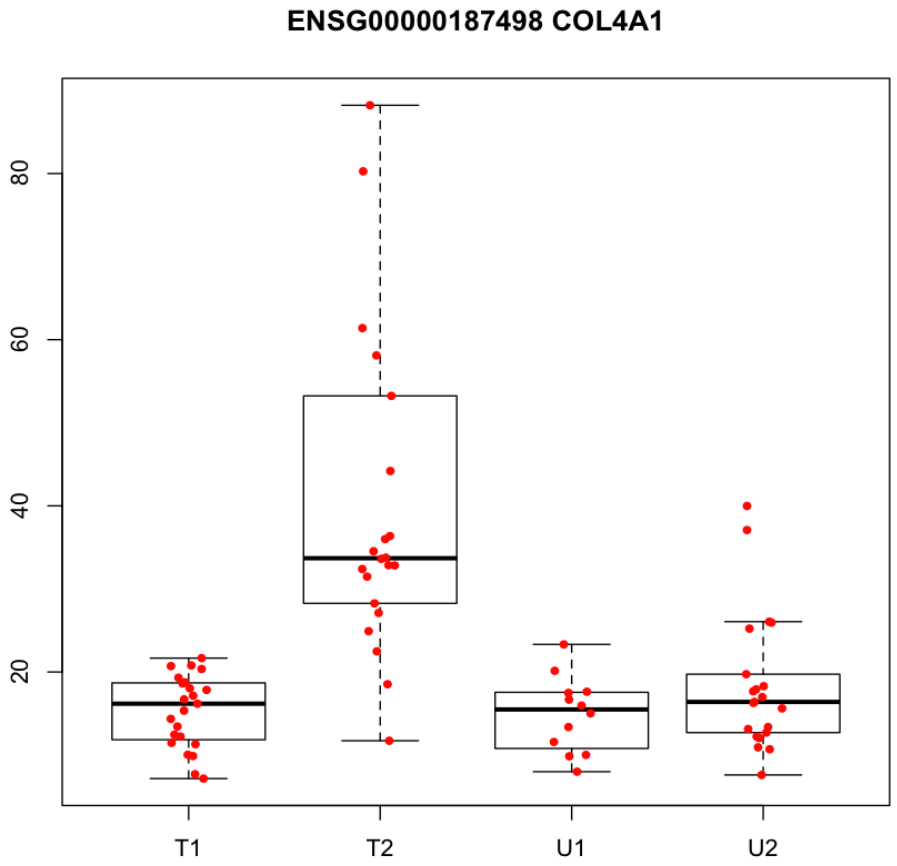
Identifier	logFC	logCPM	LR	PValue	FDR
ENSMUSG00000046623	-5.461022655078558	0.6874706484171422	130.8203992586715	2.710534641577855e-30	1.029732110335425e-25
ENSMUSG00000046623	-5.461022655078558	0.6874706484171422	130.8203992586715	2.710534641577855e-30	1.029732110335425e-25

(and so on …)

Log fold change, FDR

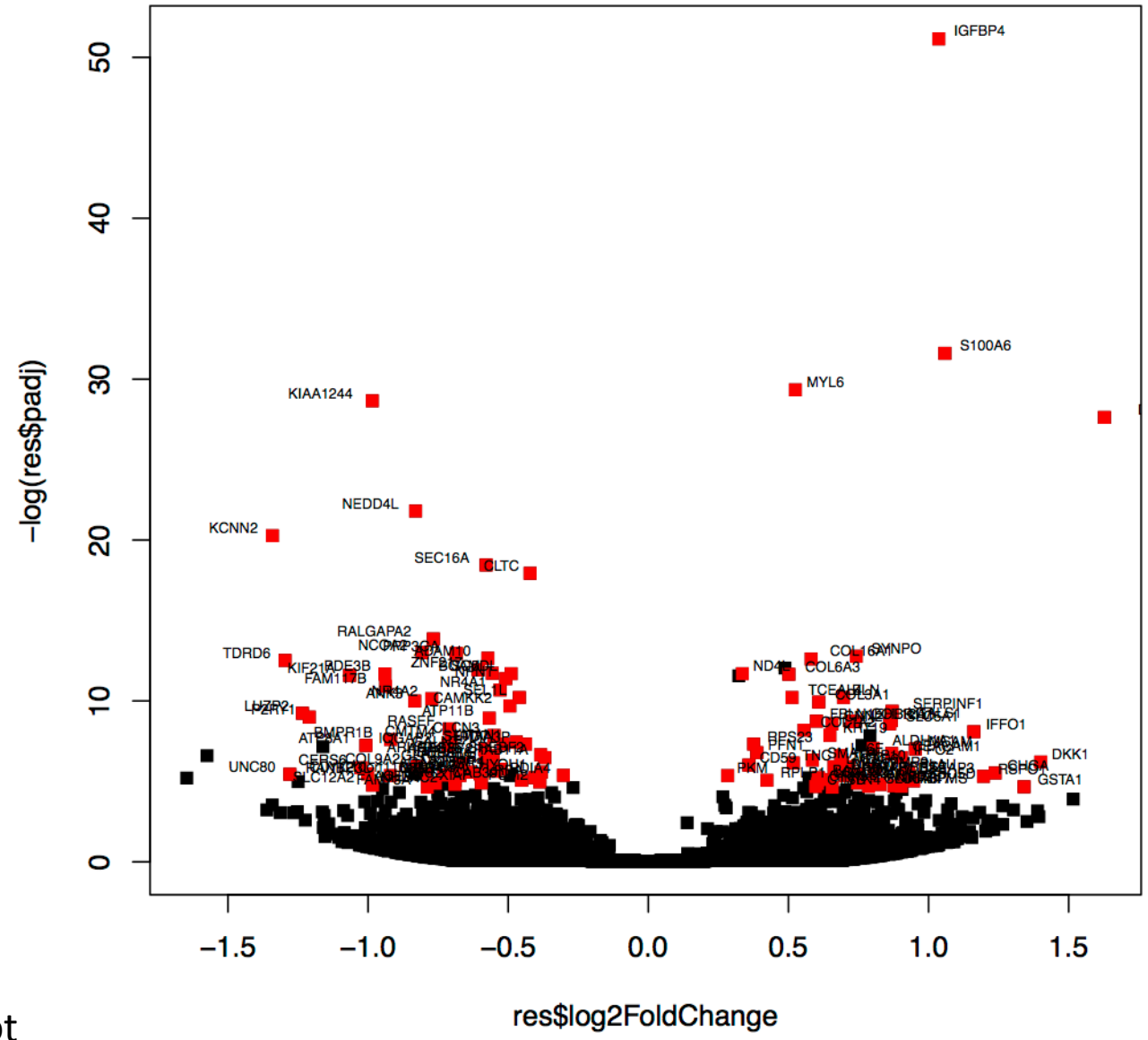
How to visualize?

Looking at top genes one by one



Box plot

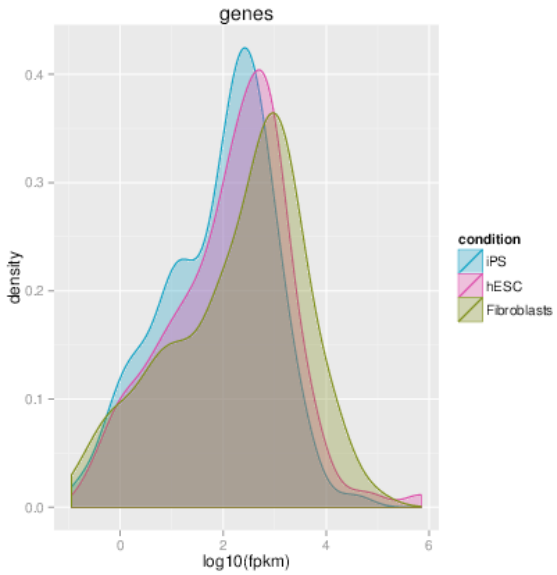
More global view



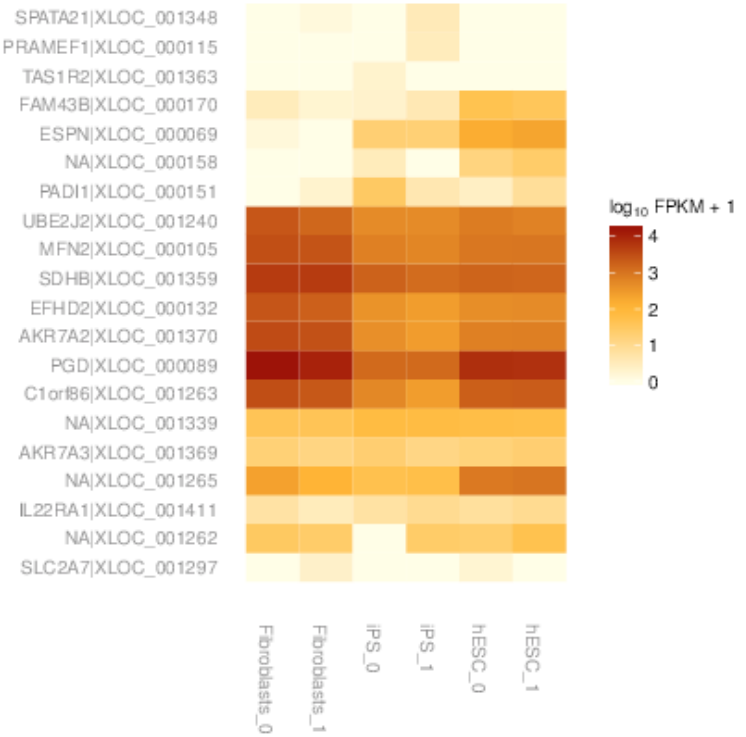
Volcano plot

cummeRbund

Works on output from CuffDiff

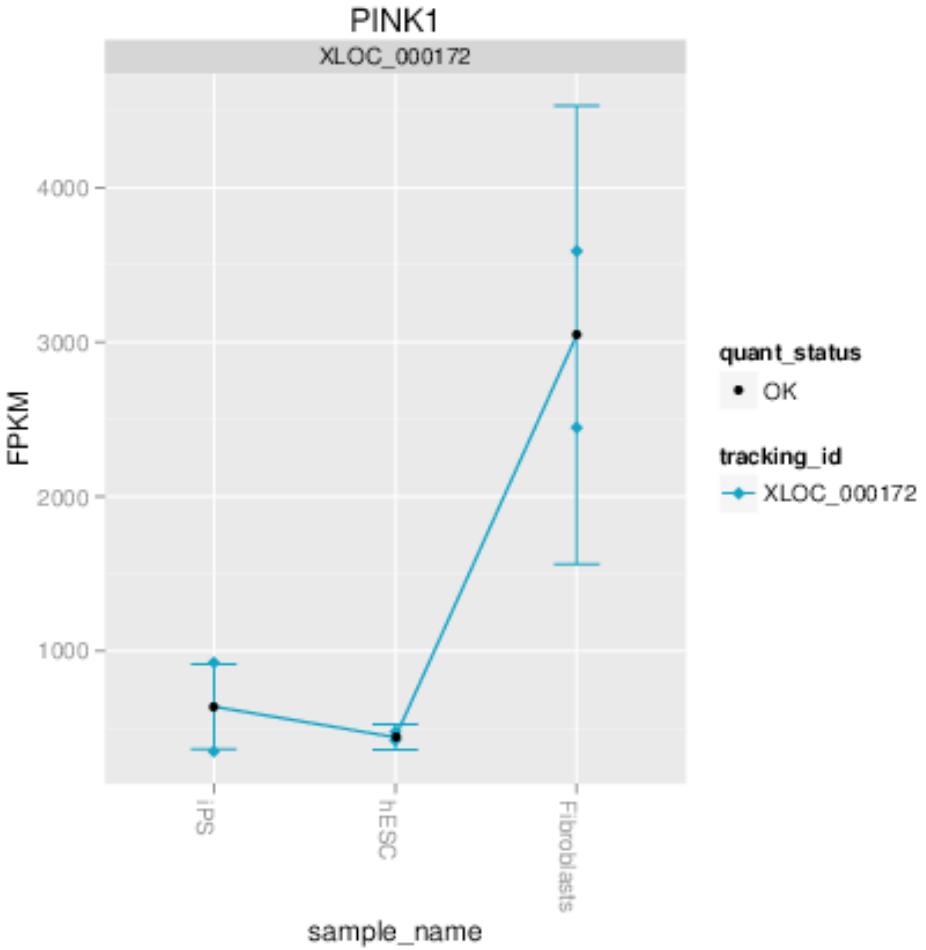


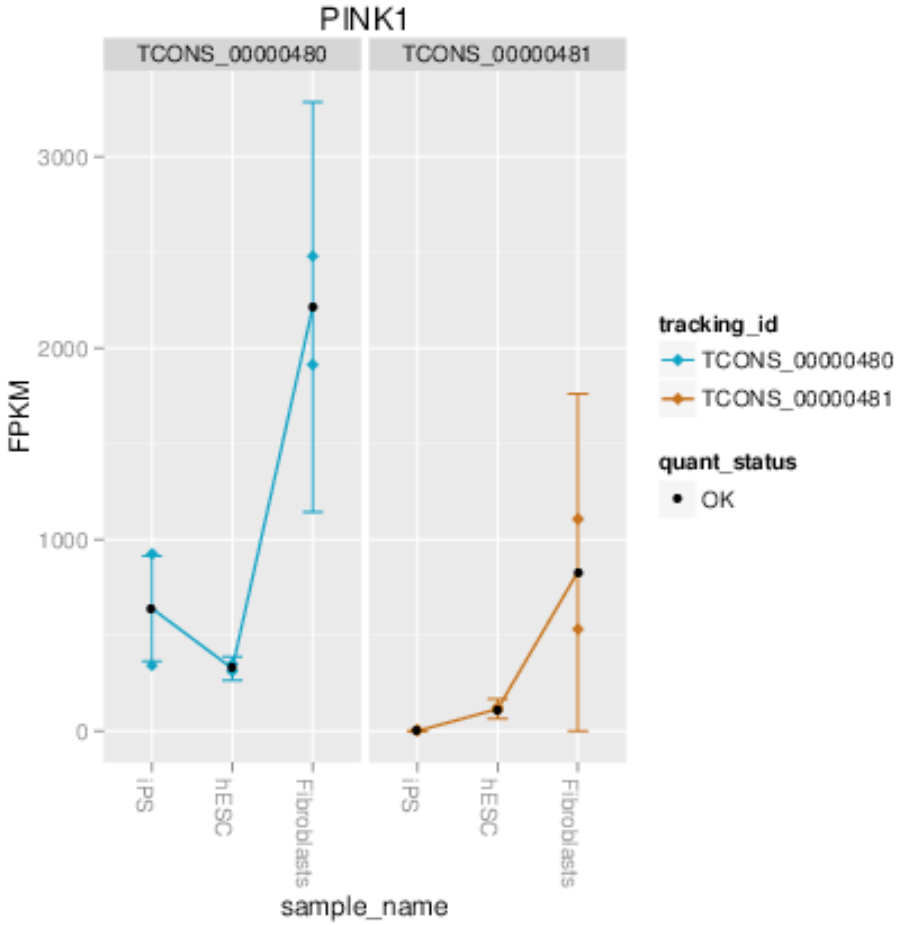
Distribution of FPKM values



Correlation heatmap for a gene set

cummeRbund

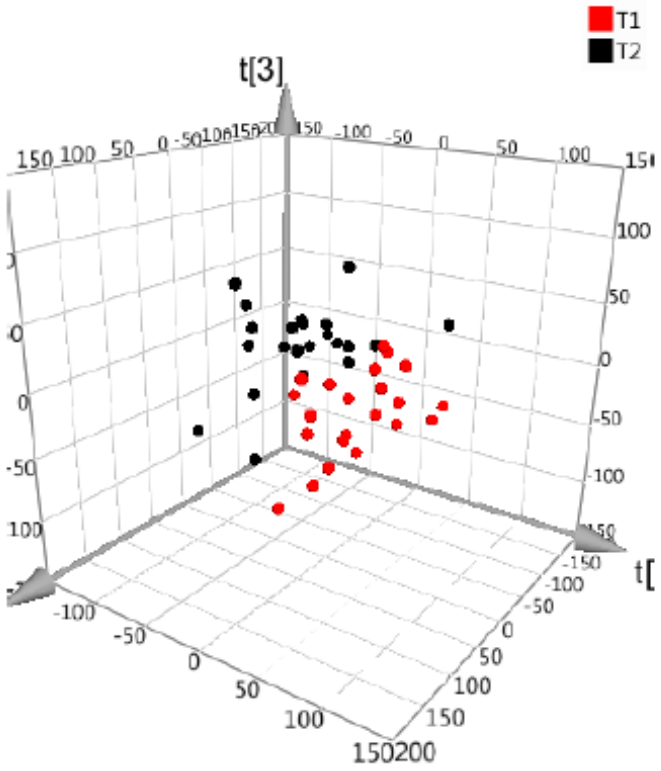




Gene-level plots

Beyond univariate differential expression (1)

Multivariate methods such as PCA (unsupervised) or PLS (supervised) can be used to obtain loadings for features (genes/transcripts/…) that contribute to separation of groups



The loading scores can be used as a different kind of measure of which genes are interesting

However – often similar to univariate results!

Beyond univariate differential expression (2)

Hudson *et al. BMC Genomics* 2012, **13**:356
http://www.biomedcentral.com/1471-2164/13/356

BMC Genomics

CORRESPONDENCE **Open Access**

Beyond differential expression: the quest for causal mutations and effector molecules

Nicholas J Hudson*, Brian P Dalrymple and Antonio Reverter

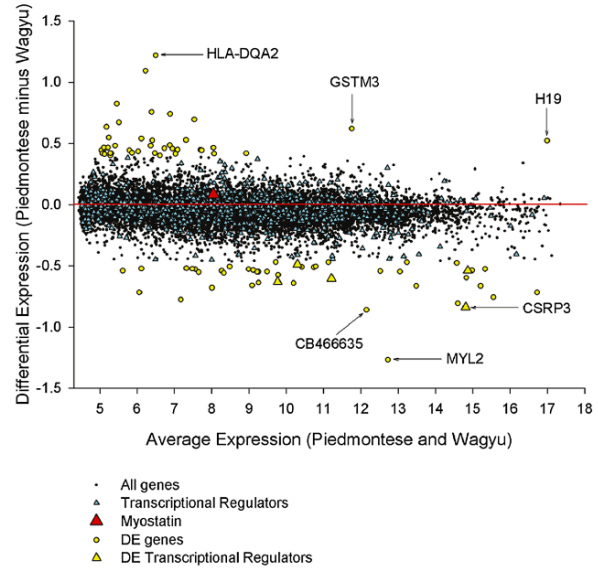


Figure 2 Needle in a numerical haystack. Despite being the causal effector molecule, *MSTN* is neither DE nor abundant when comparing *MSTN* mutant cattle versus *MSTN* wildtype cattle. Here DE is computed by subtracting the average expression in the Wagyu from the average in the Piedmontese, across the 10 time points. Figure from *PLoS Computational Biology*.

Table 1 Measures of gene expression in ascending order of complexity

Measure	Algebra formulae	Description
Expression	$E_{i,A} = \frac{1}{n}\sum_{k=1}^{n} x_{i,k}$	Average (normalized) expression of the *i*-th gene across the *n* samples (eg. biological replicates) of experimental condition *A* and where each $x_{i,k}$ corresponds to the expression of the *i*-th gene in the *k*-th sample ($k = 1, \ldots, n$).
Differential Expression	$dE_i = E_{i,A} - E_{i,B}$	Difference in the expression of the *i*-th gene in the two conditions under scrutiny, *A* and *B* (eg. healthy and diseased, two breeds, two diets, two time points, . . .). Note that it is not a requirement to have the same number of samples surveyed in the two conditions.
Co-Expression	$C_{i,j} = r_A(i,j) = \frac{Cov(i,j)}{\sigma_i \sigma_j}$	Similarity of expression profile (typically and shown here the Spearman correlation coefficient) between the *i*-th and the *j*-th genes across the *n* samples of condition *A*.
Differential Co-Expression	$dC_{i,j} = r_A(i,j) - r_B(i,j)$	Difference in the co-expression between the *i*-th and the *j*-th genes in the two conditions under scrutiny, *A* and *B*. Note that it is not a requirement to have the same number of samples surveyed in the two conditions.
Co-Differential Expression	$CdE_{i,j} = r(dE_i, dE_j)$	Similarity of the profile of differential expression of genes *i* and *j* across the levels of another experimental design effect such as time points. Two conditions, *A* and *B*, are being surveyed across a series of developmental time points.

RNA-seqlopedia

RNA-seq produces millions of sequences from complex RNA samples. With this powerful approach, you can:

1. Measure gene expression.
2. Discover and annotate complete transcripts.
3. Characterize alternative splicing and polyadenylation.

The RNA-seqlopedia provides an overview of RNA-seq and of the choices necessary to carry out a successful RNA-seq experiment.

rnaseq.uoregon.edu