RNA-seq differential expression analysis

SciLifeLab RNA-seq workshop

April 13, 2016

Mikael Huss, SciLifeLab / Stockholm University, Sweden

SciLifeLab

Karolinska Institutet

KTH VETENSKAP OCH KONST

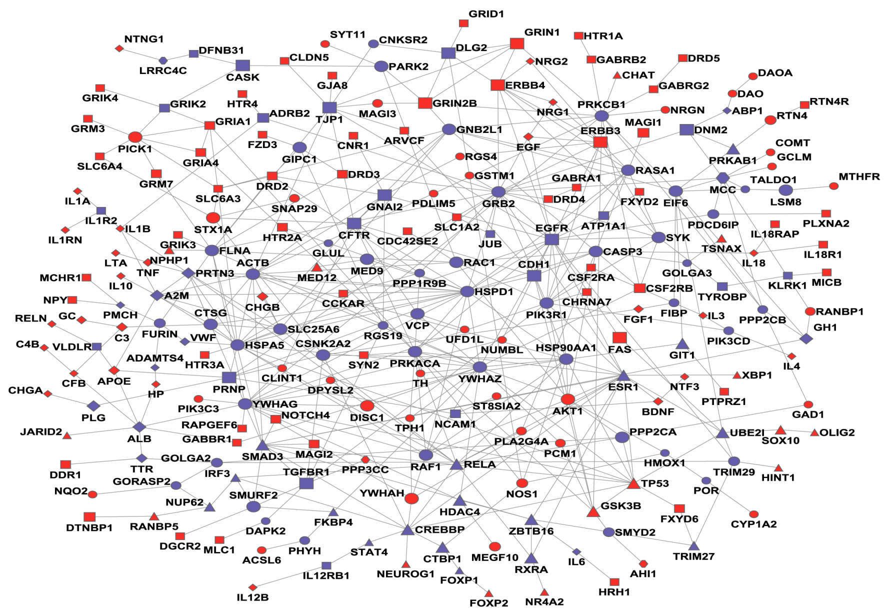
Stockholms universitet

UPPSALA UNIVERSITET

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.

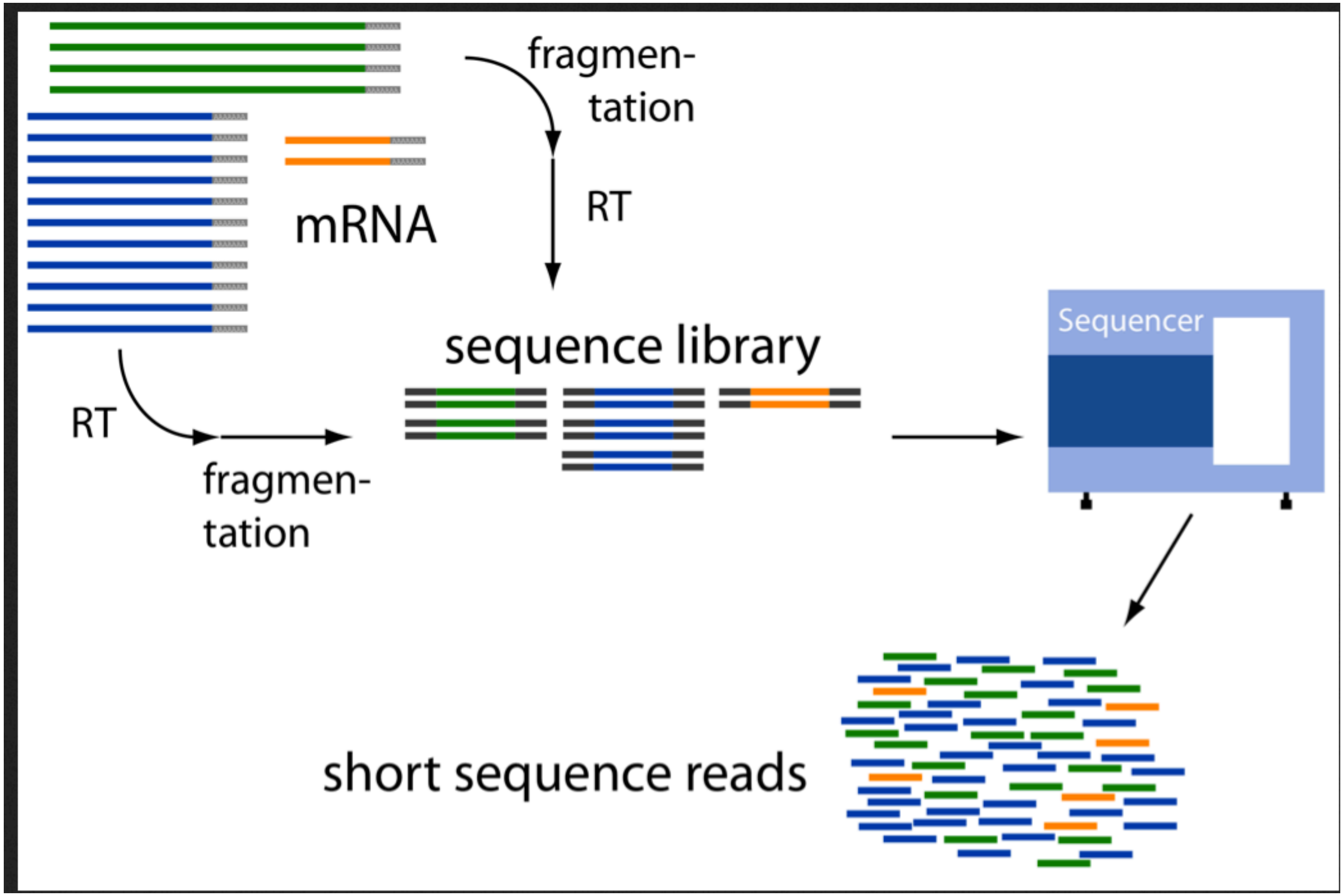
Typically **univariate** analysis (one gene at a time) – even though we know that genes are not independent



Some statistical aspects

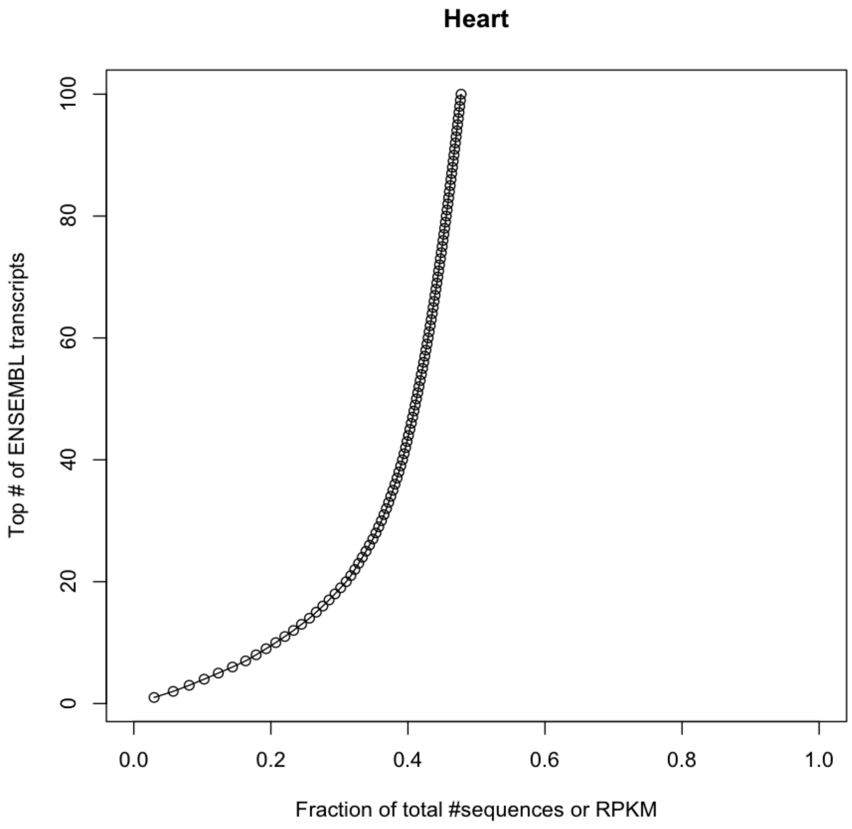
- Properties of RNA-seq data
- Replicates

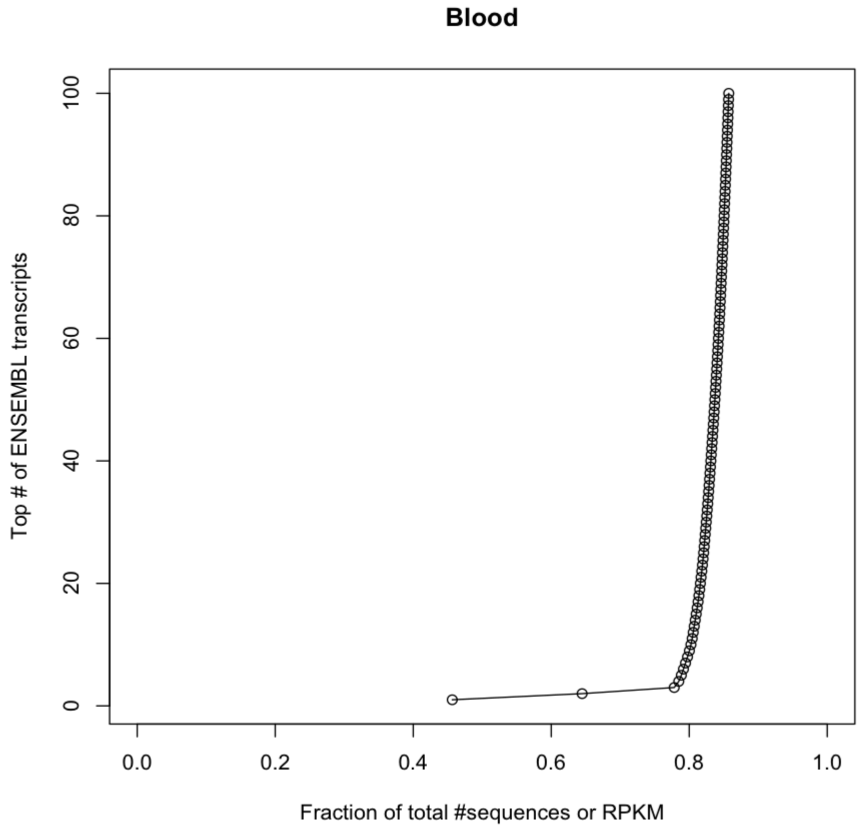
How are RNA-seq data generated?



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.

Thus, many DE analysis tools demand tables of integer read counts as input, rather than RPKM/FPKM/TPM.

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

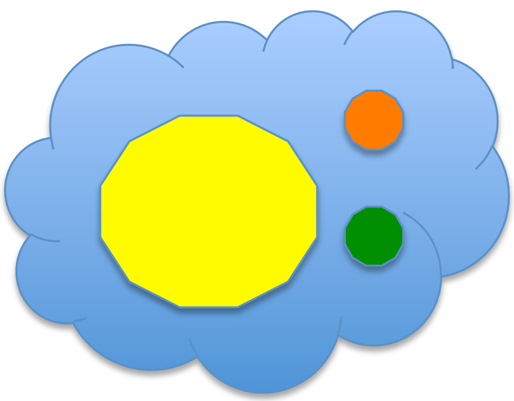
Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

TMM – Trimmed Mean of M values

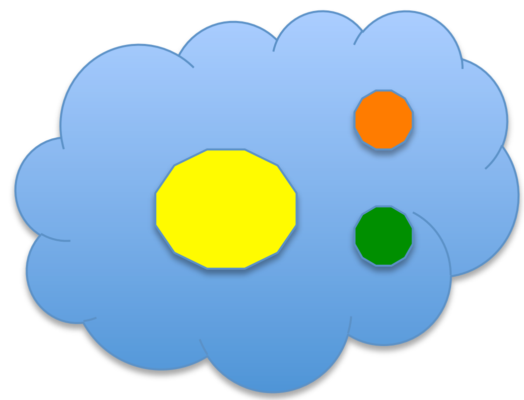
Attempts to correct for differences in RNA *composition* between samples

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 inboth 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

Experimental design

Copyright © 2010 by the Genetics Society of America
DOI: 10.1534/genetics.110.114983

Statistical Design and Analysis of RNA Sequencing Data

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

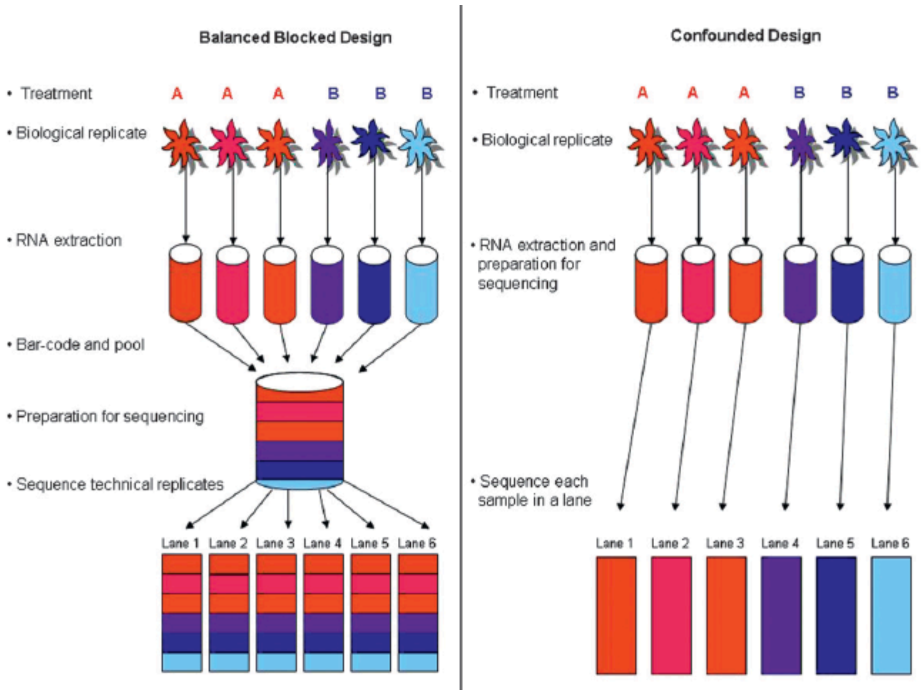
Department of Statistics, Purdue University, West Lafayette, Indiana 47907

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 ….

How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

RNA 22:1–13, 2016

NICHOLAS J. SCHURCH,[1,6] PIETÁ SCHOFIELD,[1,2,6] MAREK GIERLIŃSKI,[1,2,6] CHRISTIAN COLE,[1,6] ALEXANDER SHERSTNEV,[1,6] VIJENDER SINGH,[2] NICOLA WROBEL,[3] KARIM GHARBI,[3] GORDON G. SIMPSON,[4] TOM OWEN-HUGHES,[2] MARK BLAXTER,[3] and GEOFFREY J. BARTON[1,2,5]

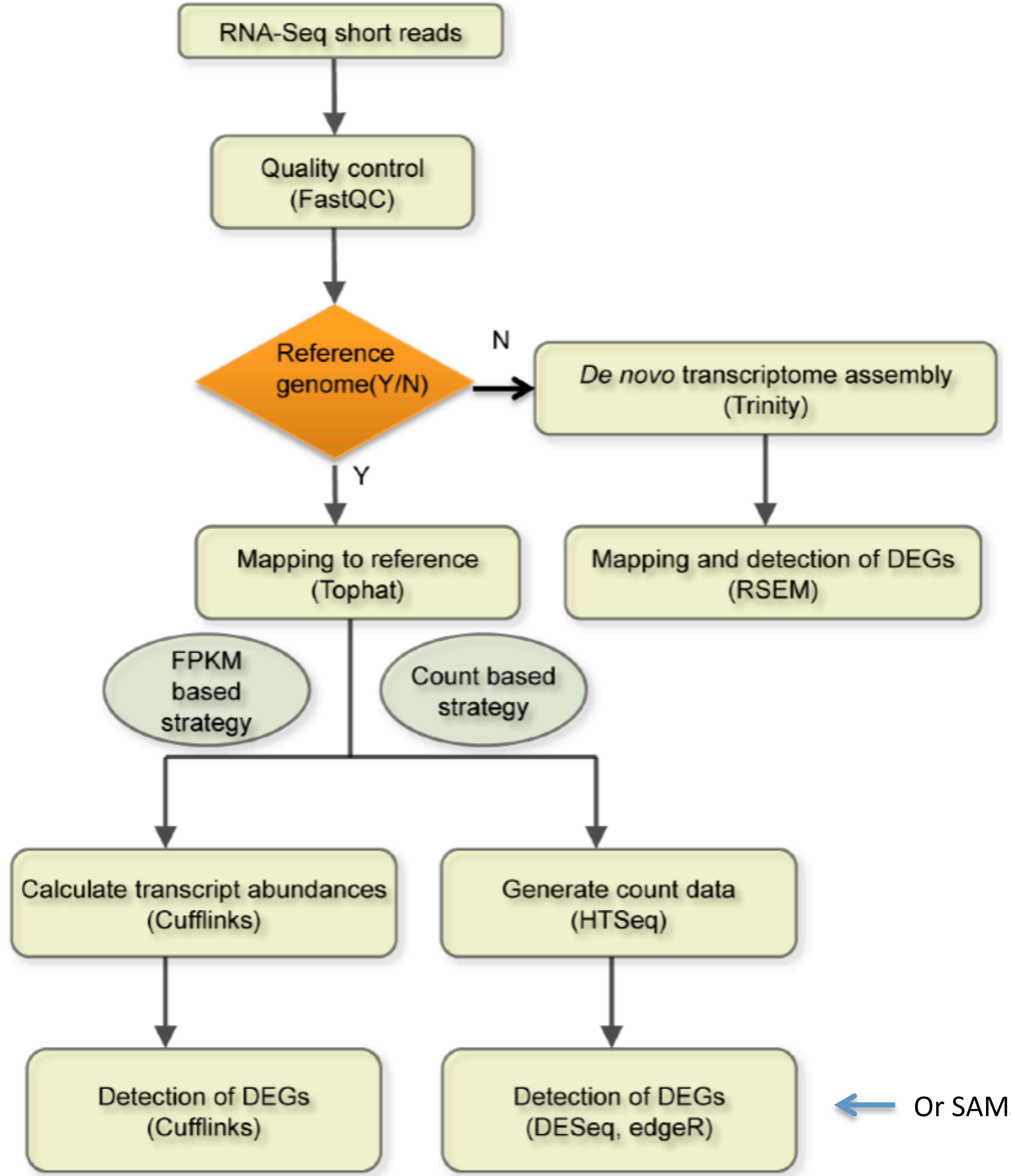
48 wild-type and 48 mutant (snf2 deletion) biological replicates in yeast
(well studied, relatively small genome, few multi-exonic genes => should be a relatively "simple" case)

Recommendation:

At least six replicates per condition for all experiments.
At least 12 replicates per condition for experiments where identifying the majority of all DE genes is important.

Different software packages and choices

- Mapping vs pseudo-alignment
- Parametric vs non-parametric
- Isoform-level vs gene-level
- Complex vs simple comparisons



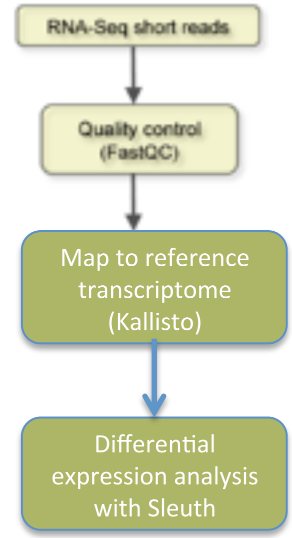
Or BitSeq, eXpress, RSEM, Sailfish etc.

Or BitSeq, ebSeq etc.

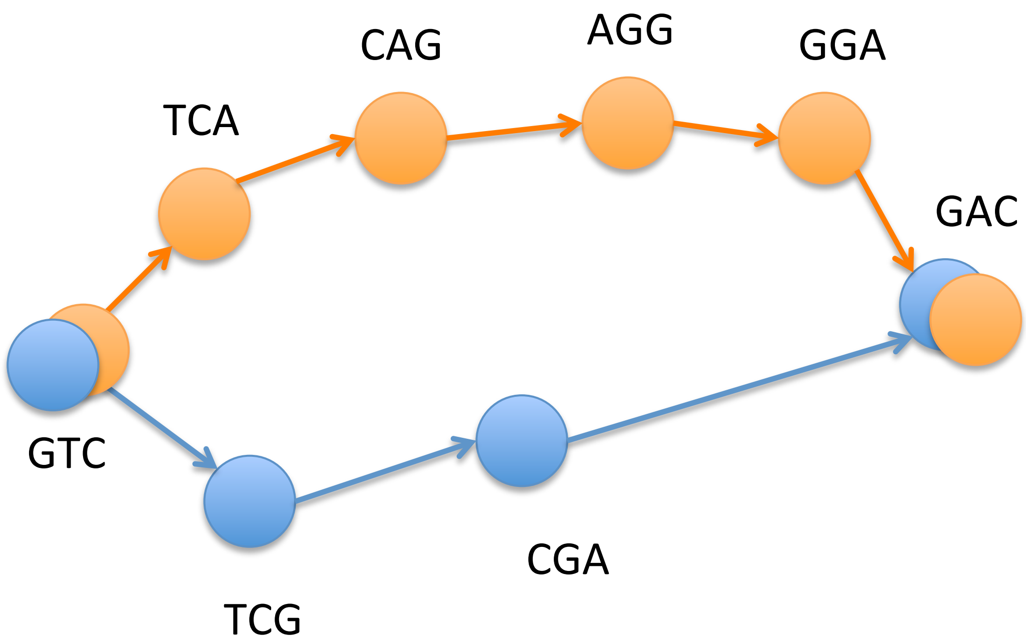
Or SAMSeq, limma, etc.

Alternative - Kallisto/Sleuth “pipeline”:

Pseudo-alignment + transcript-centric quantification and DE analysis



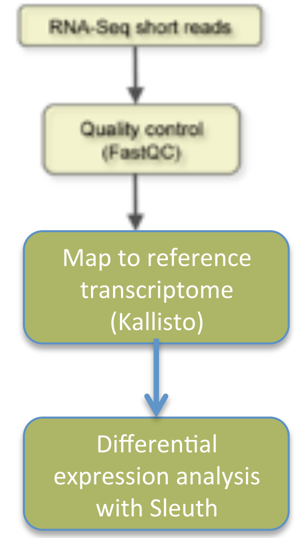
- Build an index by chopping ref transcriptomes into k-mers and putting them into “colored” graphs

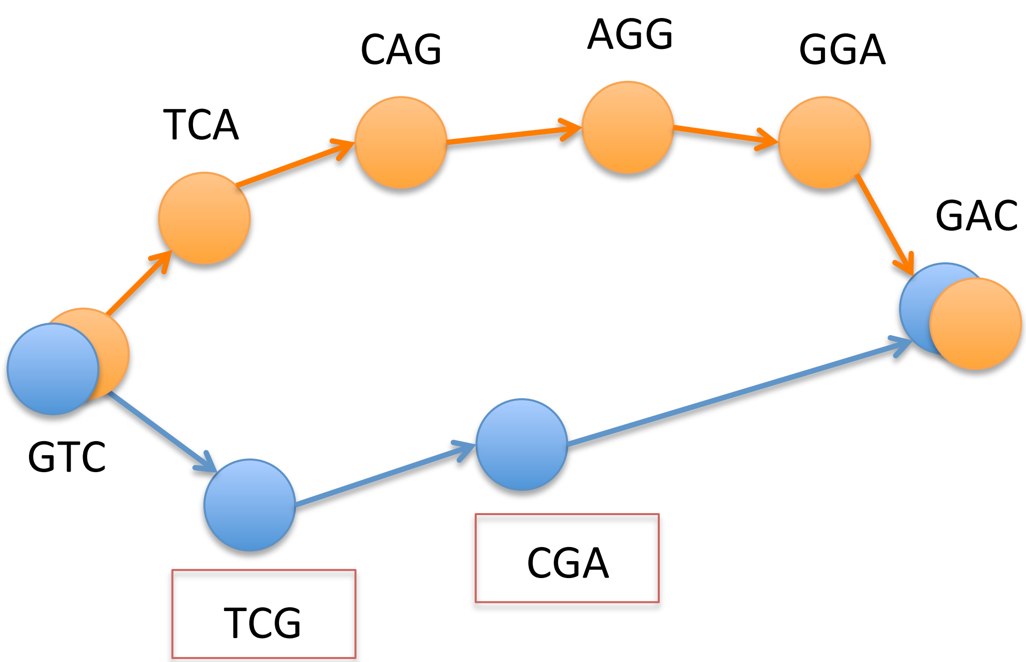


SciLifeLab
Karolinska Institutet
KTH
Stockholms universitet
UPPSALA UNIVERSITET

Kallisto/Sleuth "pipeline":

Pseudo-alignment + transcript-centric quantification and DE analysis



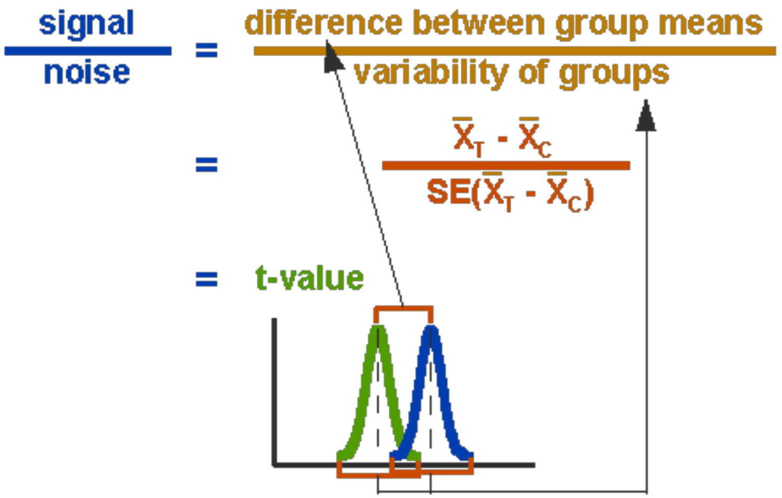


When a new read is observed, chop it into k-mers and see what it is *compatible* with. E g observe TCGA:

TCGA
T**CGA**

Differential expression analysis?

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.

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

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

… but ...

But: Revisiting the 48-replicate benchmark paper

TABLE 1. RNA-seq differential gene expression tools and statistical tests

Name	Assumed distribution	Normalization	Description
t-test	Normal	DEseq[a]	Two-sample *t*-test for equal variances
log *t*-test	Log-normal	DEseq[a]	Log-ratio *t*-test
Mann-Whitney	None	DEseq[a]	Mann-Whitney test
Permutation	None	DEseq[a]	Permutation test
Bootstrap	Normal	DEseq[a]	Bootstrap test
baySeq[c]	Negative binomial	Internal	Empirical Bayesian estimate of posterior likelihood
Cuffdiff	Negative binomial	Internal	Unknown
DEGseq[c]	Binomial	None	Random sampling model using Fisher's exact test and the likelihood ratio test
DESeq[c]	Negative binomial	DEseq[a]	Shrinkage variance
DESeq2[c]	Negative binomial	DEseq[a]	Shrinkage variance
EBSeq[c]	Negative binomial	DEseq[a] (median)	Empirical Bayesian estimate of posterior likelihood
edgeR[c]	Negative binomial	TMM[b]	Empirical Bayes estimation and either an exact test analogous to Fisher's exact test but adapted to over-dispersed data or a generalized linear model
Limma[c]	Log-normal	TMM[b]	Generalized linear model
NOISeq[c]	None	RPKM	Nonparametric test based on signal-to-noise ratio
PoissonSeq[c]	Poisson log-linear model	Internal	Score statistic
SAMSeq[c]	None	Internal	Mann-Whitney test with Poisson resampling

For experiments with <12 replicates per condition; use *edgeR (exact)*.
For experiments with >12 replicates per condition; use *DESeq*.

Parametric methods apparently working better …

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 q values or some other type of 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”

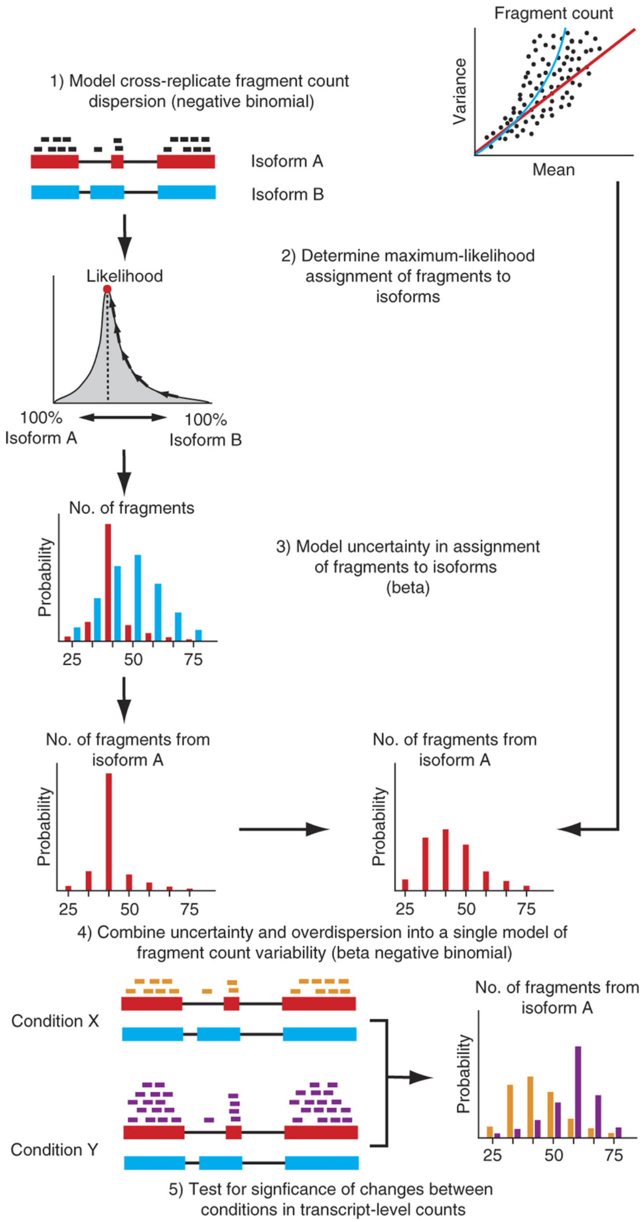
Distributional issue: Solved by variance stabilizing transform in limma – voom() function

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 q values or some other type of 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**

Sleuth

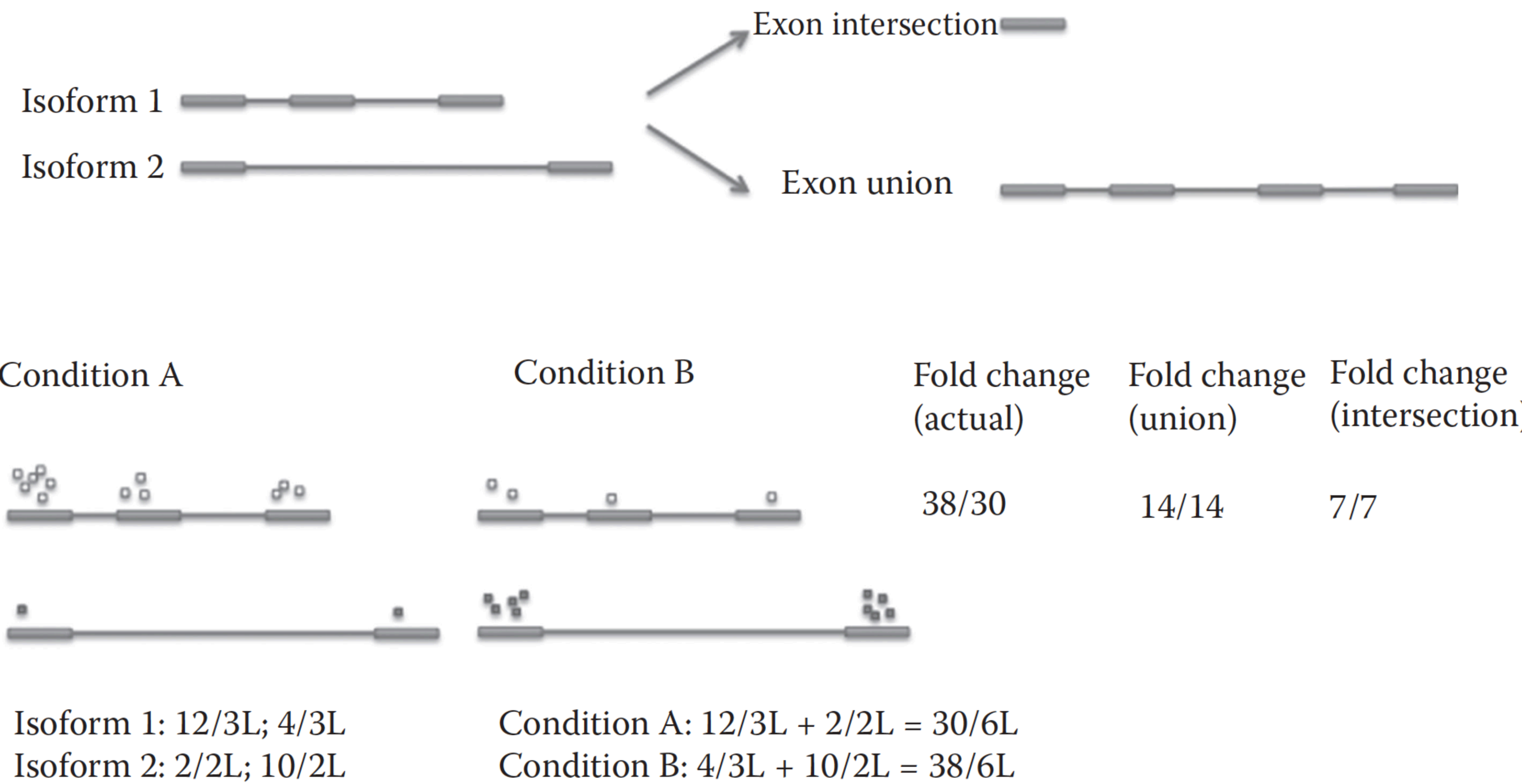
Developed by the same team as CuffDiff, and superior to it according to them. Based on Kallisto.

Transcript-oriented (like CuffDiff)

Includes uncertainty coming from “quantification noise” (like CuffDiff)

Supports modelling multiple experimental factors (unlike CuffDiff)

Reason to use transcript-level analysis



Assembly-based DE: Ballgown

Ballgown bridges the gap between transcriptome assembly and expression analysis

Alyssa C Frazee, Geo Pertea, Andrew E Jaffe, Ben Langmead, Steven L Salzberg & Jeffrey T Leek

Affiliations | Corresponding author

Nature Biotechnology **33**, 243–246 (2015) | doi:10.1038/nbt.3172

Relatively untested, but more general than most existing tools:

- Does isoform-level expression
- Can test DE for novel transcripts
- Supports complex designs
- More sensitive than CuffDiff
- Provides a database back-end for handling transcript assemblies

Set of BAM files

↓ Reference guided assembly (Cufflinks)

Set of assemblies

↓ Merge assemblies (CuffMerge)

Consensus assembly

↓ Generate FPKM/counts for each assembled transcript in each sample (tablemaker)

Contig expression table

↓ Process in Ballgown

Differentially expressed contigs

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?

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.

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.

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.

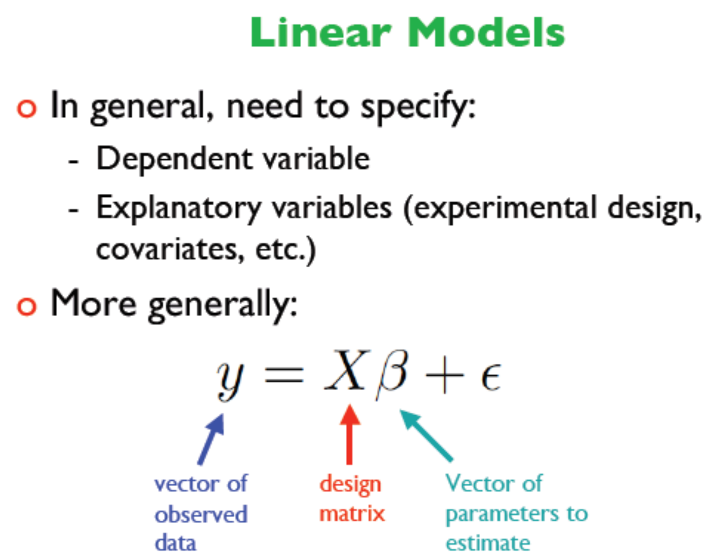
→limma / DESeq / edgeR / Sleuth which can work with factorial designs

(but not e g CuffDiff2, SAMSeq)

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



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 (a)

Error term/noise (f)

Which software to choose?

- Based on need
- Benchmarks

Decision tree for software selection (2015)

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*

Decision tree for software selection (2016)

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

Sleuth

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*, Sleuth

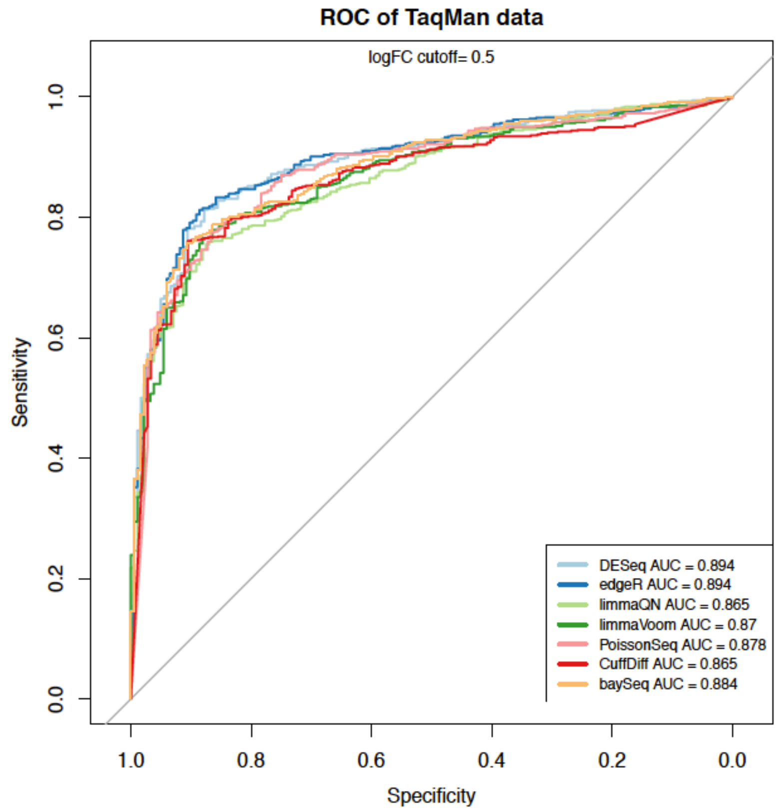
Simple comparison of groups => **How many biological replicates?**

More than about 5 biological replicates per group => *~~SAMSeq~~*

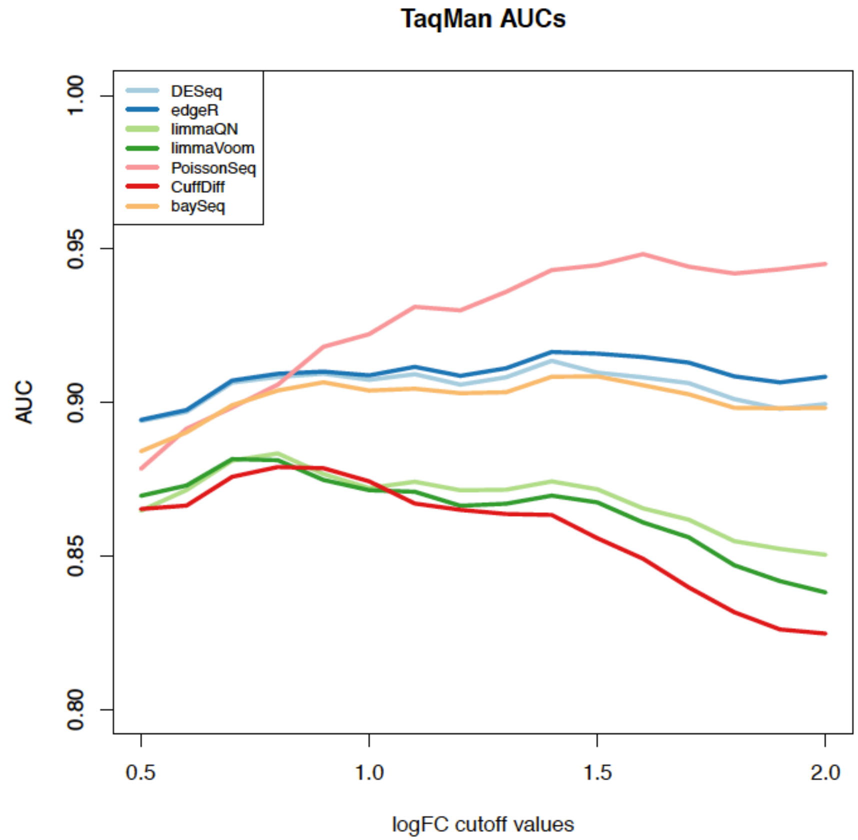
Less than 5 biological replicates per group => *DESeq, edgeR, limma*

?

Other DE software comparisons (1)



(a)



(b)

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

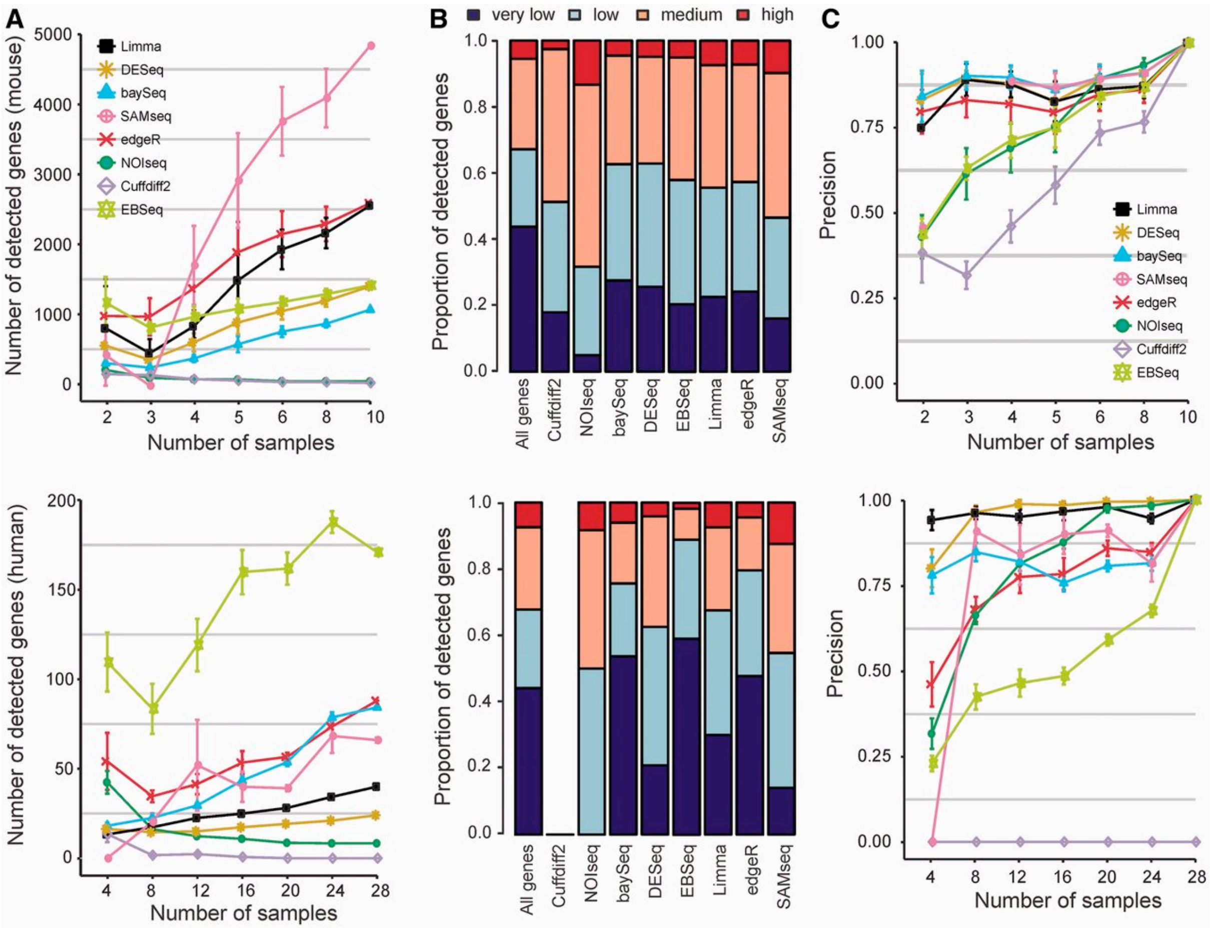
Franck Rapaport [1], Raya Khanin [1], Yupu Liang [1], Azra Krek [1], Paul Zumbo [2,4], Christopher E. Mason [2,4], Nicholas D. Socci [1], Doron Betel [3,4]

[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

DESeq, edgeR, PoissonSeq come out well

Other 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

Limma, DESeq, baySeq

Other 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.

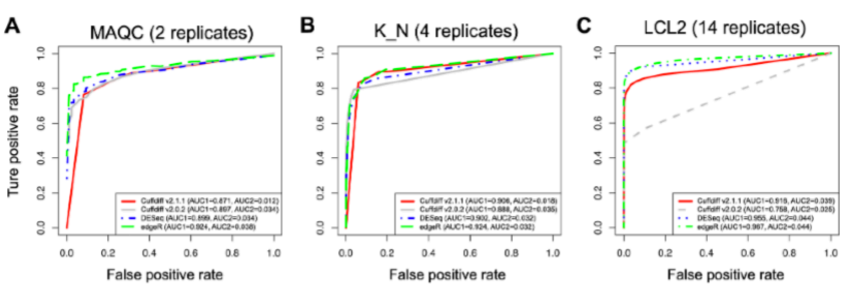
BMC Bioinformatics 2013, **14**:91 doi:10.1186/1471-2105-14-91

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

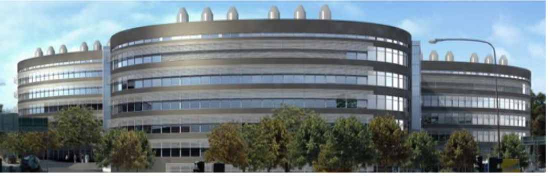
A comparative study of techniques for differential expression analysis on RNA-Seq data

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



edgeR, DESeq



Take-away messages from DE tool comparison

-edgeR, DESeq and limma (the latter of which does not use the negative binomial distribution) tend to to work well

-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 poor performance with >5 replicates

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

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

Not considered in these comparisons:

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

Miscellaneous (if there is time)

- Visualization of DE analysis results
- Normalization and scaling
- Batch normalization
- Mixtures of cell types
- Beyond univariate DE analysis

Differential expression analysis output

Top 10 differentially expressed genes tables for each contrast

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

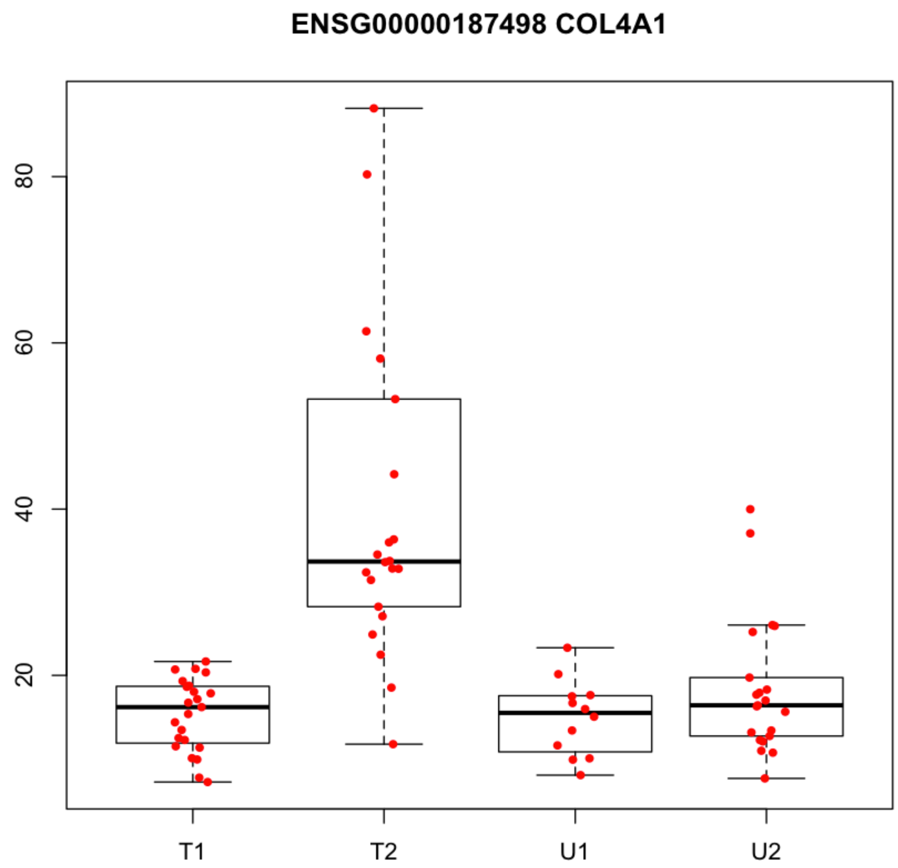
Identifier	logFC	logCPM	LR	PValue	FDR
ENSMUSG00000046623	-5.46102265507855	0.687470648417142	130.820399258671	2.71053464157785e-30	1.02973211033542e-25
ENSMUSG00000046623	-5.46102265507855	0.687470648417142	130.820399258671	2.71053464157785e-30	1.02973211033542e-25

(and so on …)

Log fold change, FDR

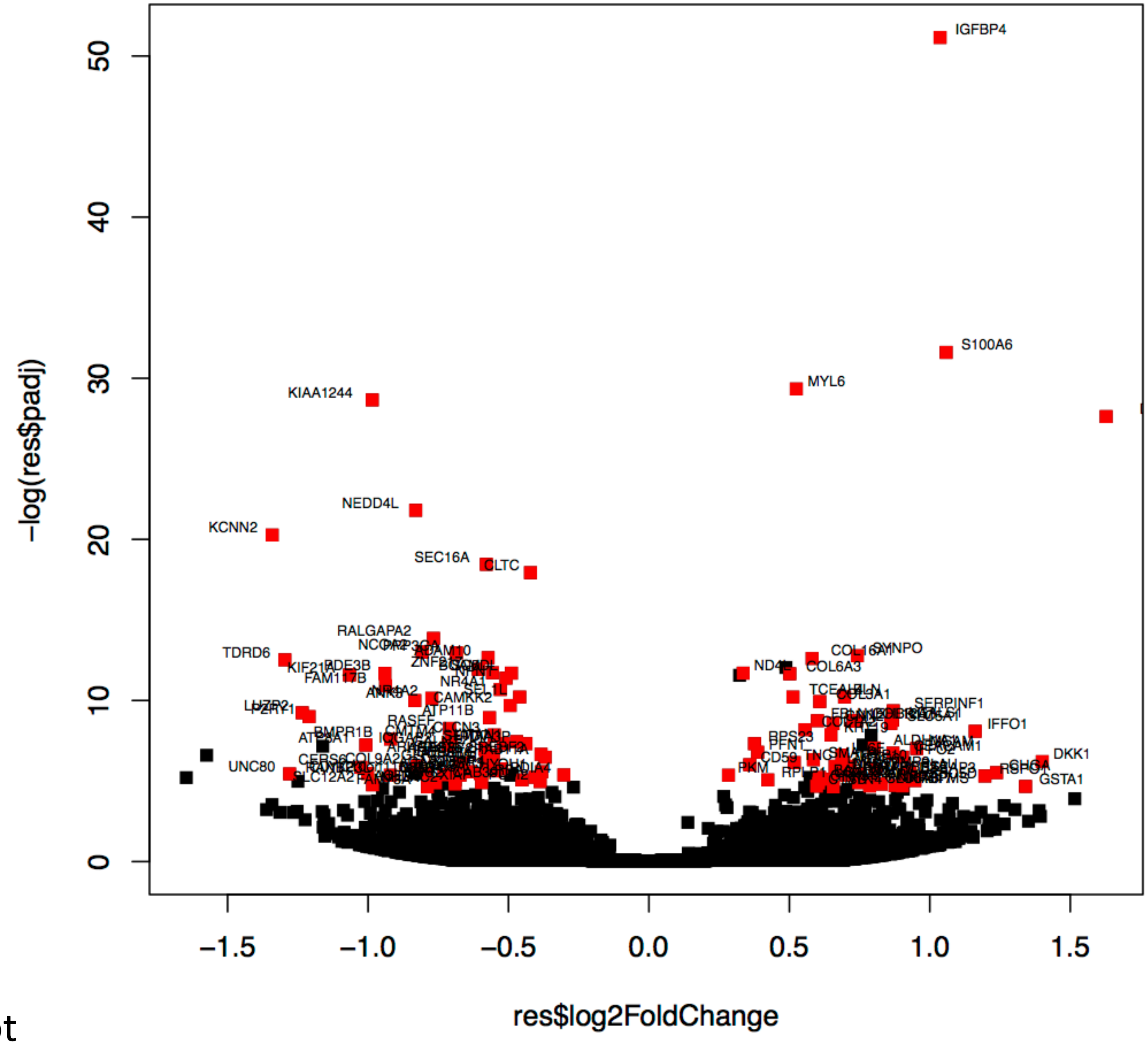
How to visualize?

Looking at top genes one by one



Box plot

More global view



Volcano plot

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

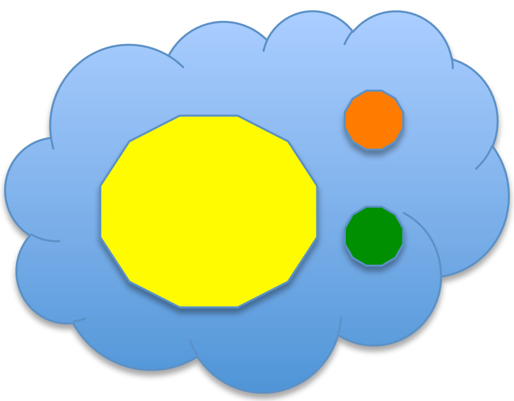
Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

TMM – Trimmed Mean of M values

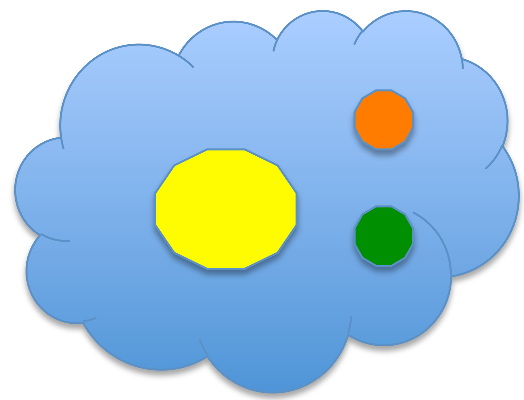
Attempts to correct for differences in RNA *composition* between samples

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

SciLifeLab
Karolinska Institutet
KTH
Stockholms universitet
UPPSALA UNIVERSITET

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.

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.

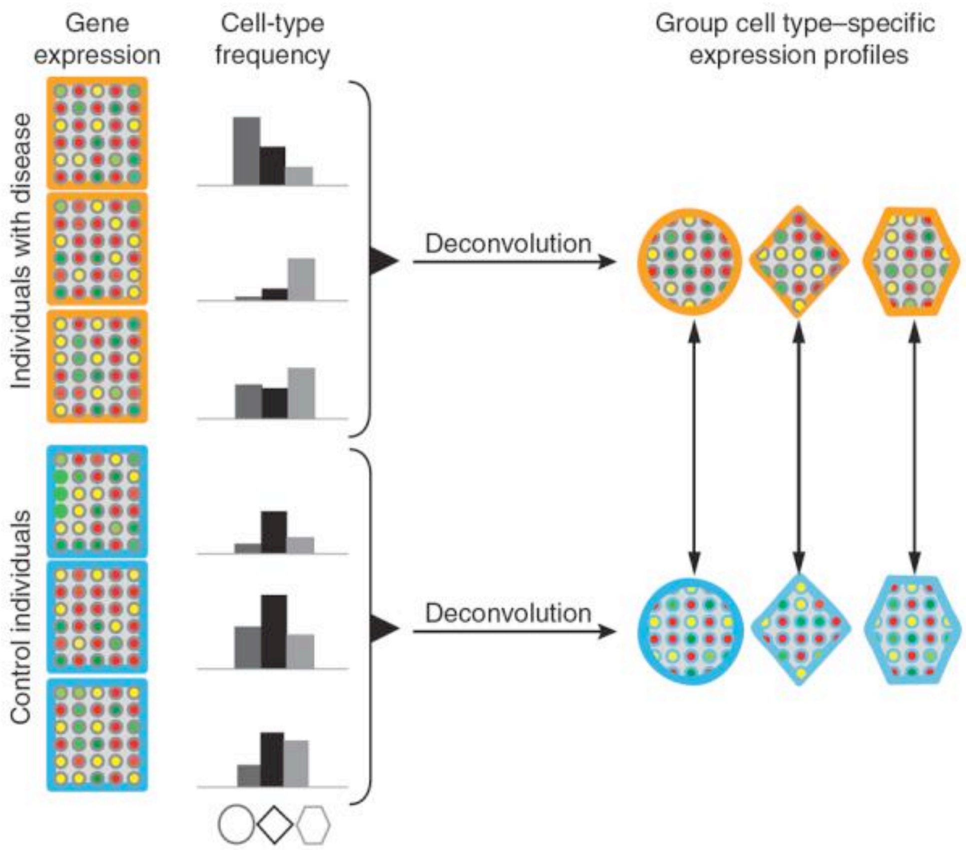
(Designed for microarrays, should use log scale values for RNA-seq)

COMBAT:

'COMBATTING' BATCH EFFECTS WHEN COMBINING BATCHES OF GENE EXPRESSION MICROARRAY DATA

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

DE analysis in mixtures of cell types



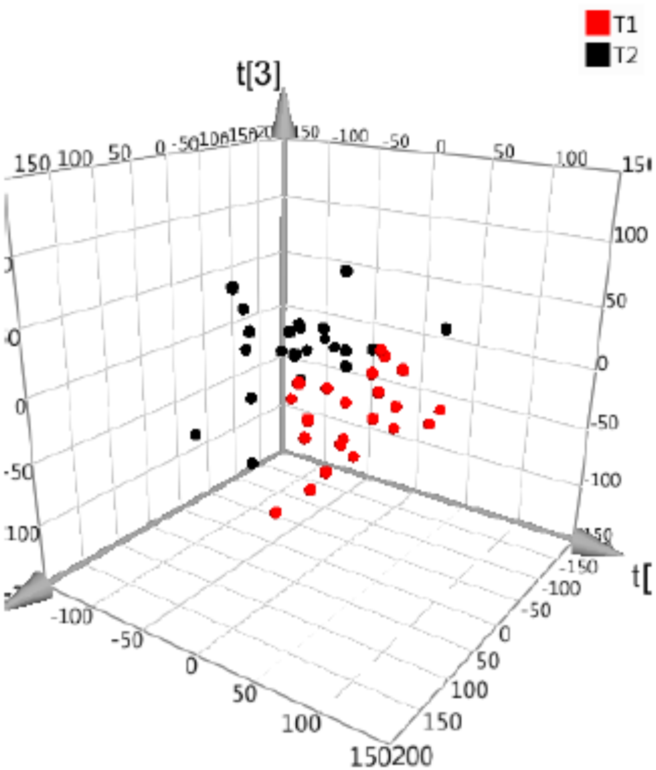
CellMix, R package implementing several deconvolution methods (most for microarray)

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.

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

Beyond univariate differential expression (2)

Statistical/machine learning approaches:

Can use gene or transcript expression levels as features in a statistical model when trying to predict some class (classification) or continuous variable (regression)

Feature selection methods frequently needed to reduce the number of genes/transcripts used in the model. E g lasso/elastic net or Boruta (random forest based feature selection).