



RNA-seq differential expression analysis

SciLifeLab RNA-seq workshop

November 15, 2018

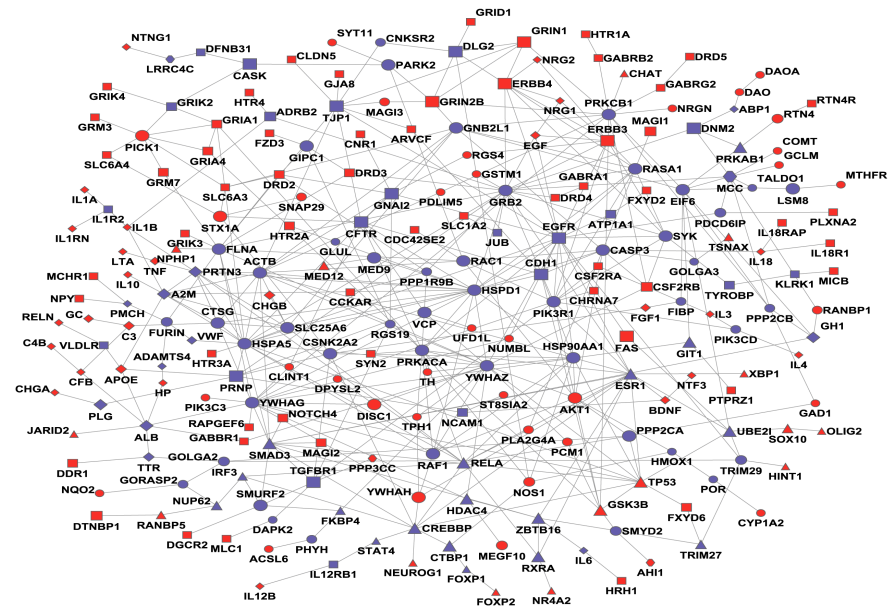
Dag Ahrén, NBIS / Lund University, Sweden

Differential expression analysis

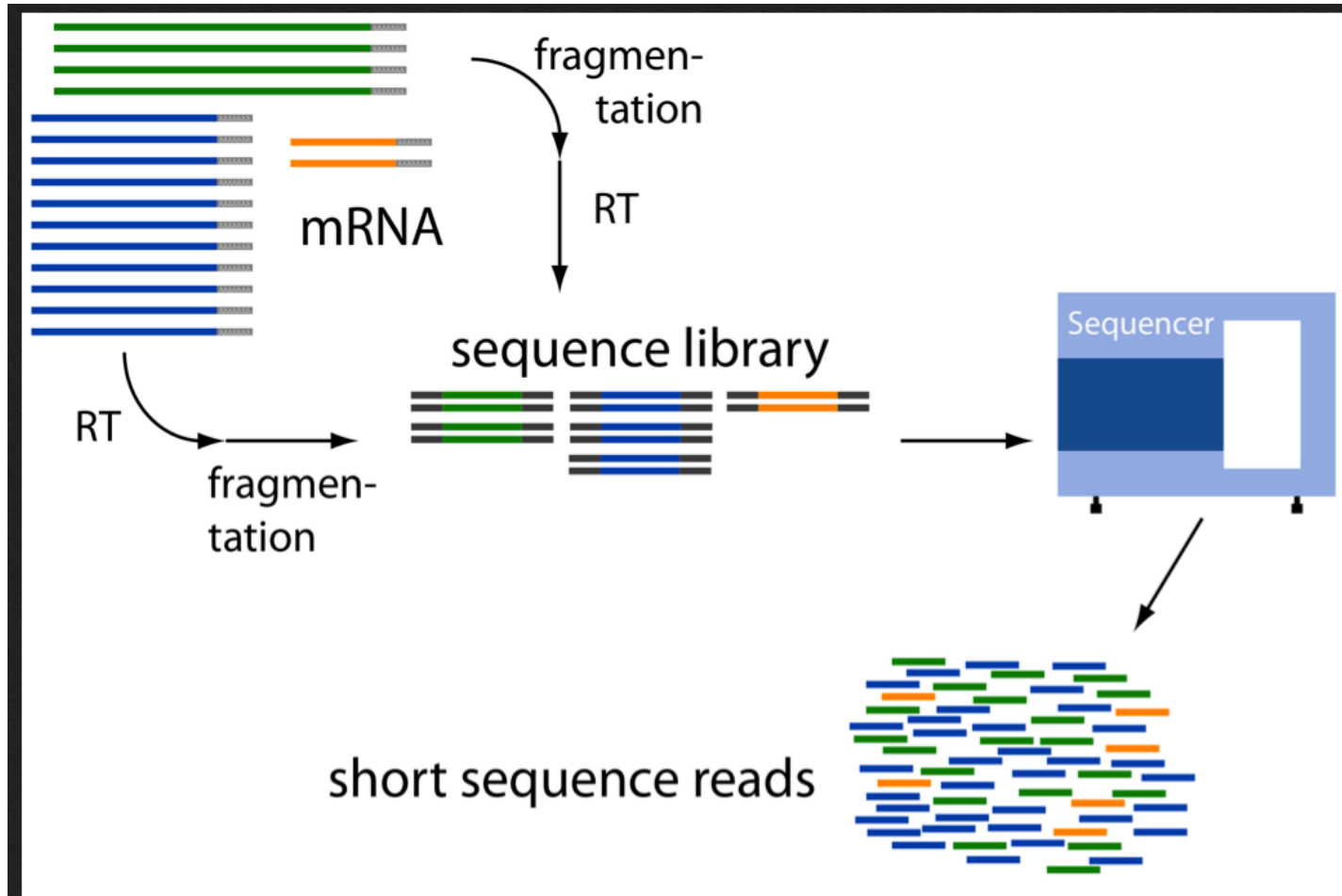
Goal: identify significantly differentially expressed genes/exons/transcripts

examples: 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

Count-based statistics

Researchers 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.

RPKM= Reads Per Kilobase Million

FPKM= Fragments Per Kilobase Million

TPM= Transcripts Per Million

Count nature of RNA-seq data

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

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.

What would happen with the RPKM?

Which one would you consider more reliable and why?

Think-Pair-Share

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 1, we can be more confident that there is a true difference in the expression level than in scenario 2 (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

Technical vs biological replicates

Technical replicates:

- Assess variability of measurement technique
- Typically low for bulk RNA-seq (not necessarily true in 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

Ideal case: Large variation between groups & low variation within groups

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

Depends:

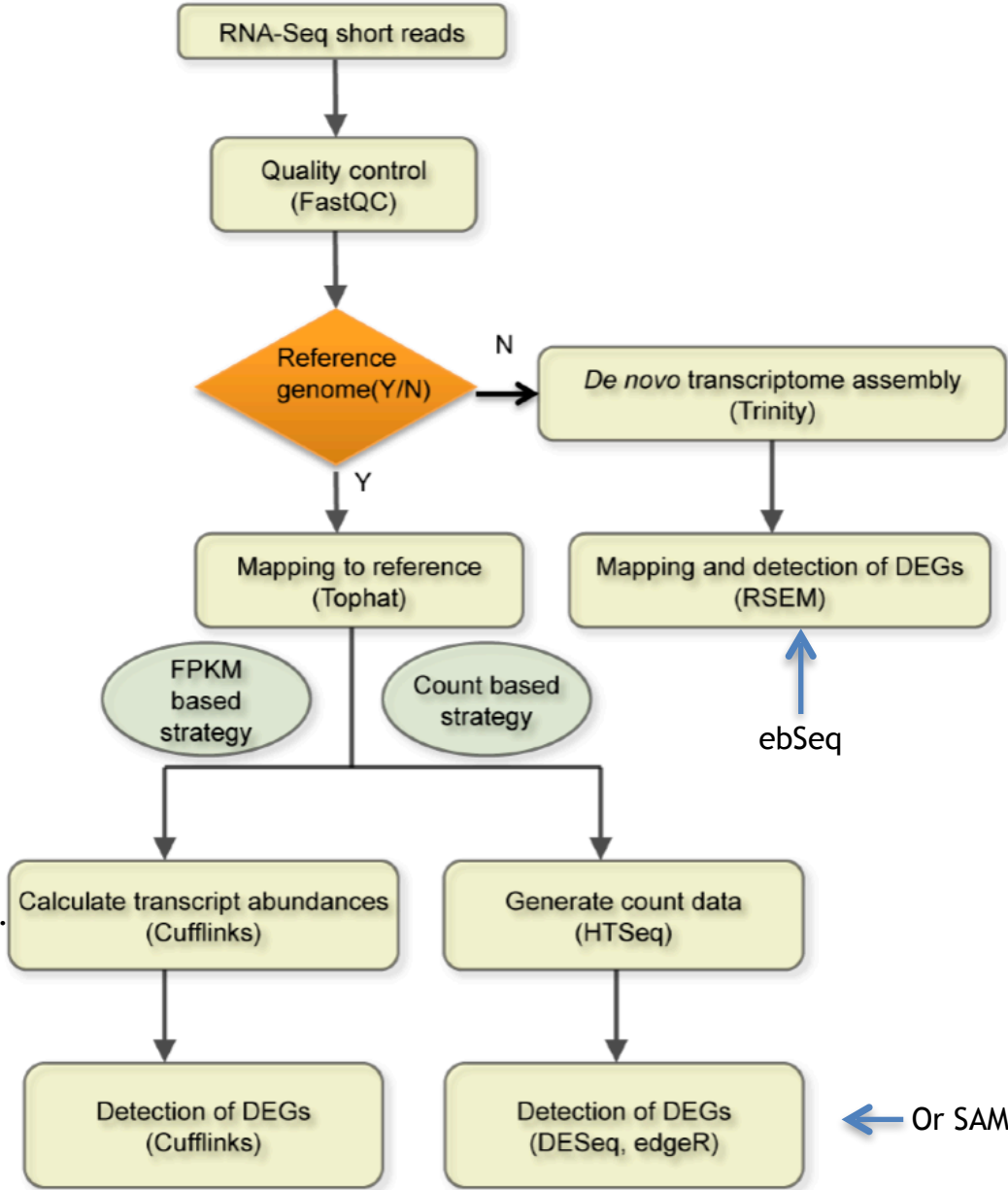
Homogeneous cell lines, inbred mice etc: maybe 3 samples / group enough.
Clinical case-control studies on patients: can need a dozen, hundreds or thousands, depending on the specifics

Also depends on your research question...

Different software packages and choices

- Many different options at each stage of the analysis:
 - Mapping software (alignment vs pseudo alignment)
 - Differential expression analysis (parametric vs non-parametric and complexity of design)

Possible workflows

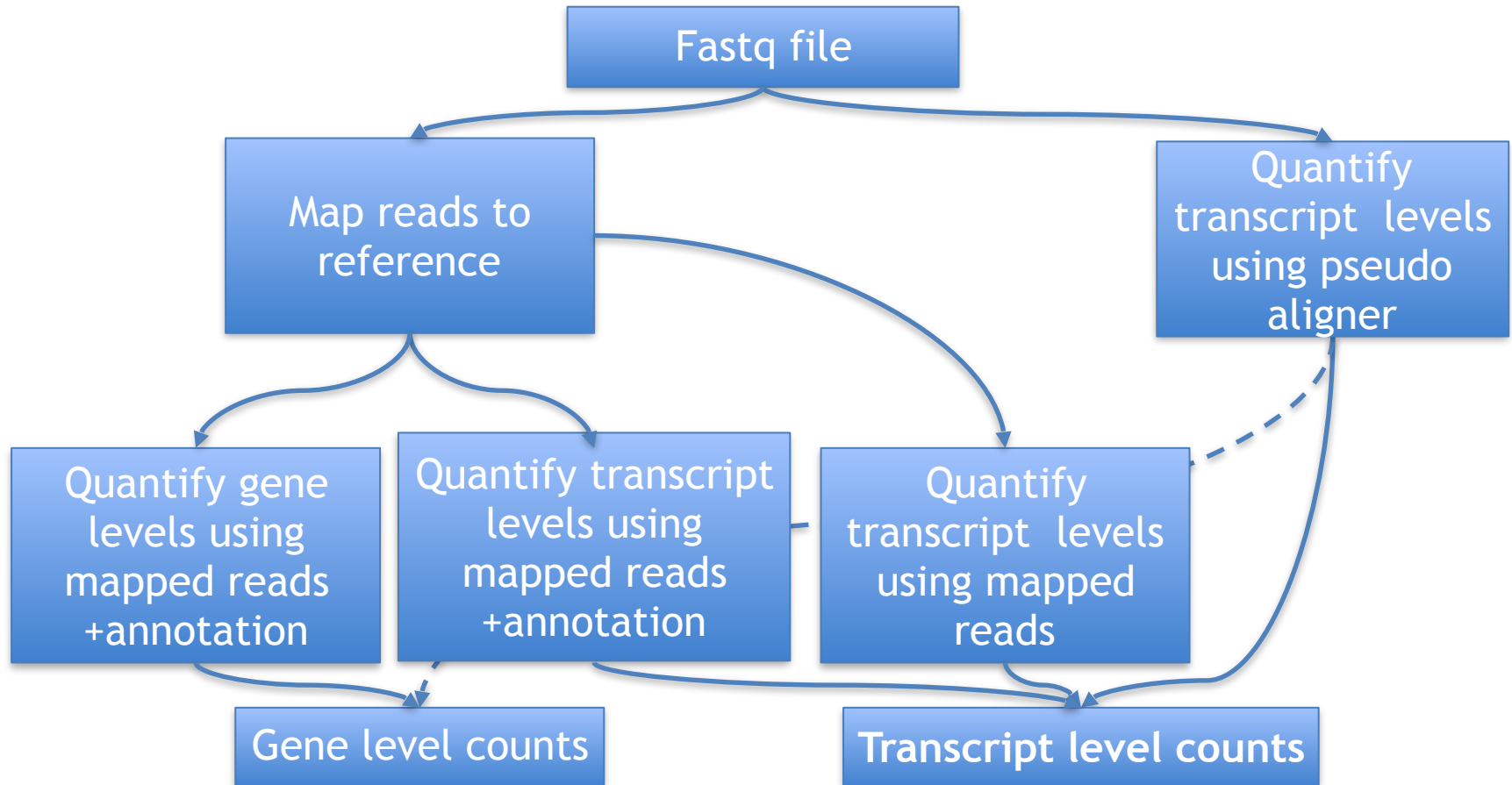


Or BitSeq, eXpress, RSEM, Sailfish etc. →

Or BitSeq, ebSeq etc. →

← Or SAMSeq, limma, etc.

Read alignment pipelines and gene expression estimates



Transcript level analysis

Zhang *et al.* *BMC Genomics* (2017) 18:583
DOI 10.1186/s12864-017-4002-1

BMC Genomics

RESEARCH ARTICLE

Open Access



Evaluation and comparison of computational tools for RNA-seq isoform quantification

Chi Zhang¹, Baohong Zhang¹, Lih-Ling Lin² and Shanrong Zhao^{1*}

Methods used in paper

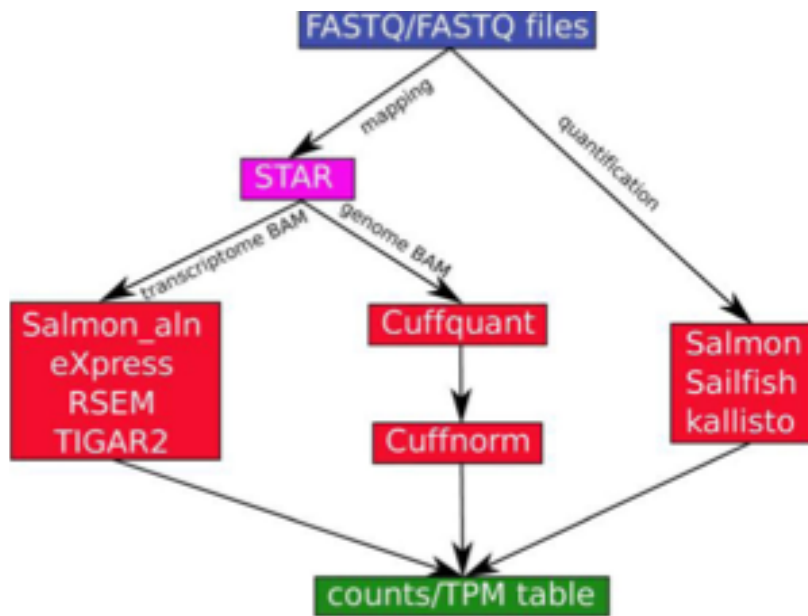


Table 1 Run time metrics of each method on 50 million paired-end reads of length 76 bp in an high performance computing cluster

	Memory (Gb)	Run time (min)	Algorithm	Multi-thread
Cufflinks	3.5	117	ML	Yes
RSEM	5.6	154	ML	Yes
eXpress	<u>0.55</u>	30	ML	No
TIGAR2	28.3	1045	VB	Yes
kallisto	3.8	7	ML	Yes
Salmon	6.6	6	VB/ML	Yes
Salmon_aln	3	7	VB/ML	Yes
Sailfish	6.3	<u>5</u>	VB/ML	Yes

For methods that support multi-threading, eight threads were used. For alignment-free methods (Kallisto, Salmon and Sailfish), a mapping step was included. The best performer in each category is underlined and the worst performer is in bold
ML Maximum Likelihood, *VB* Variational Bayes

Isoform quantification problematic for genes with many isoforms

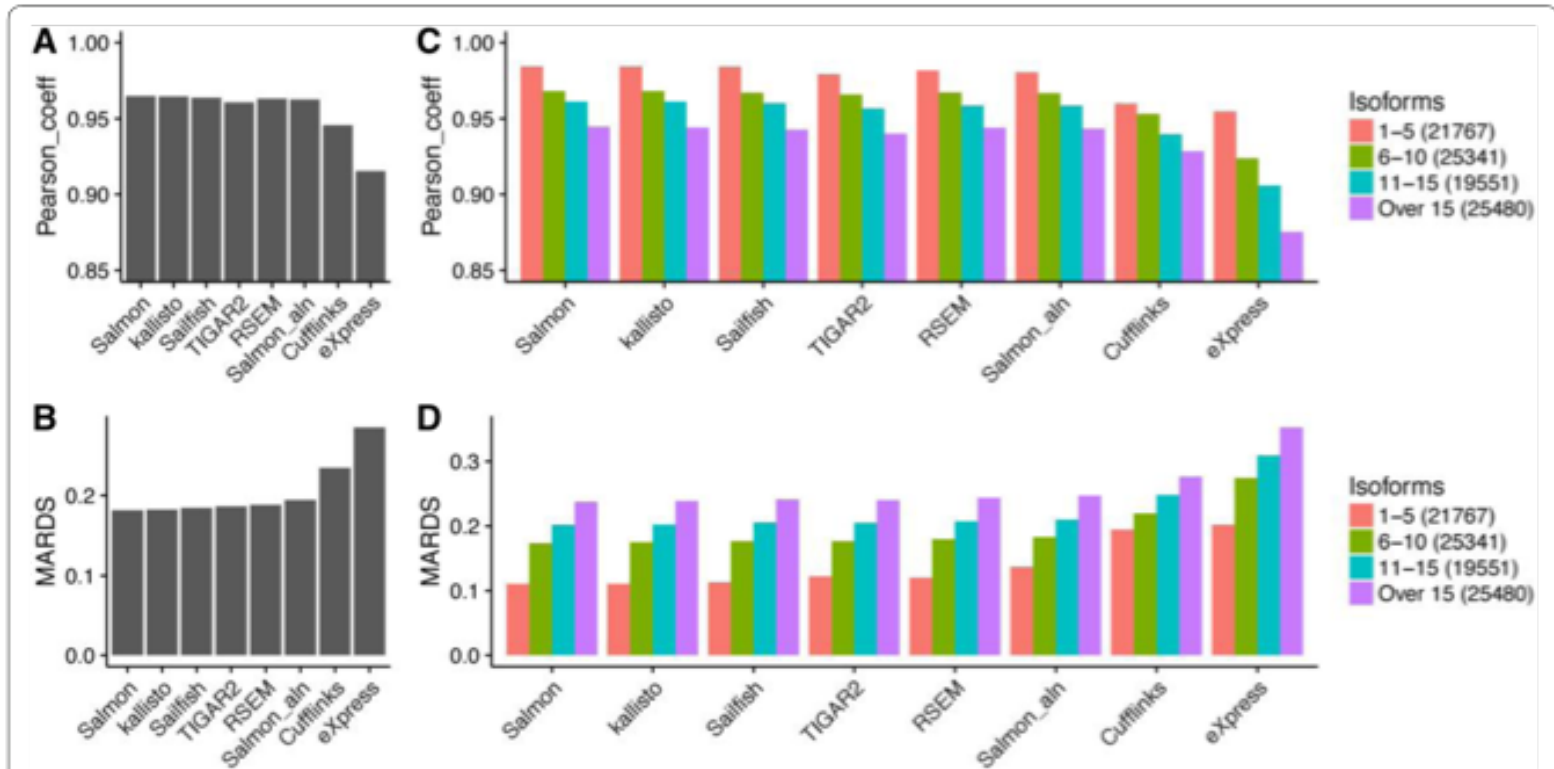



Fig. 2 Comparisons of the overall performance among different methods and the impact of the number of transcripts on the accuracy of isoform quantification. **a** Pearson correlation coefficient **b** mean absolute relative differences and **c-d**) The above metrics were broken into separate groups according to the number of annotated transcript isoforms for each gene. The number of transcripts in each group is shown in figure legends. The accuracy metrics were calculated by comparing the estimated counts with the “ground truths” in simulated dataset

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

Differential expression analysis?

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

$$\begin{aligned} \frac{\text{signal}}{\text{noise}} &= \frac{\text{difference between group means}}{\text{variability of groups}} \\ &= \frac{\bar{X}_T - \bar{X}_C}{\text{SE}(\bar{X}_T - \bar{X}_C)} \\ &= \text{t-value} \end{aligned}$$


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

Problems with this approach:

- May have **few replicates**
- **Multiple testing** issues
- Distribution is **not normal**

Dealing with the “t test issues”

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.

Dealing with the “t test issues”

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.

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”

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.

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.

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.

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

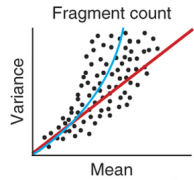
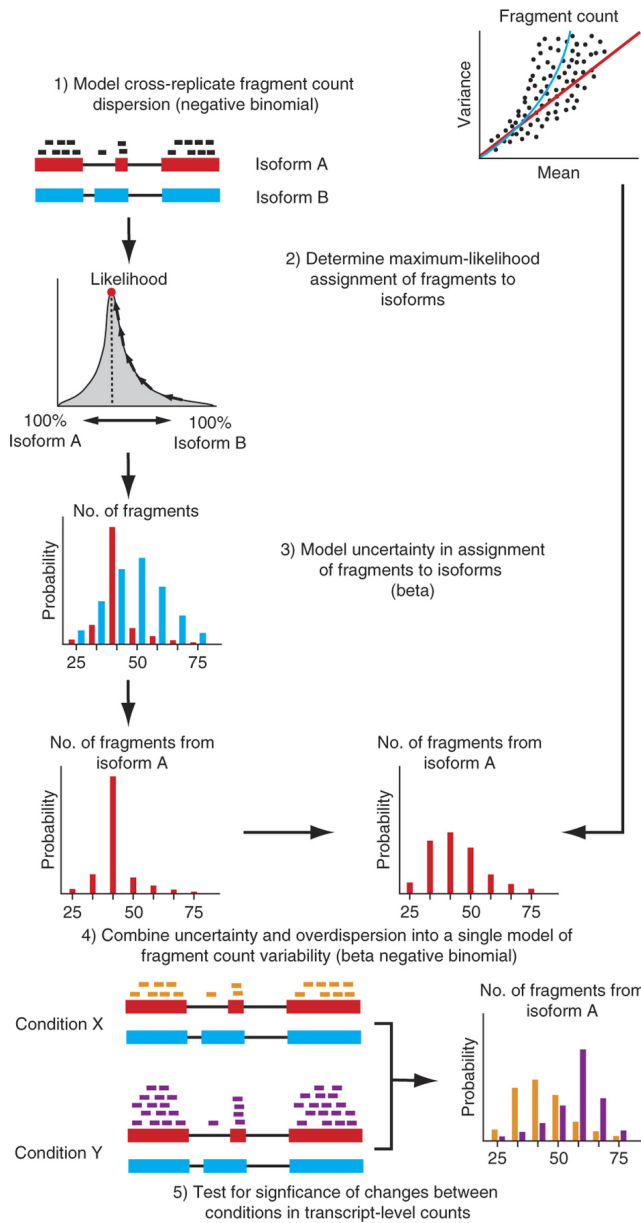
Name	Assumed distribution	Normalization	Description
t-test	Normal	DESeq ^a	Two-sample t-test for equal variances
log t-test	Log-normal	DESeq ^a	Log-ratio t-test
Mann-Whitney	None	DESeq ^a	Mann-Whitney test
Permutation	None	DESeq ^a	Permutation test
Bootstrap	Normal	DESeq ^a	Bootstrap test
baySeq ^c	Negative 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 ...

CuffDiff2



Integrates isoform quantification + differential expression analysis.

Also: **BitSeq**

Sleuth

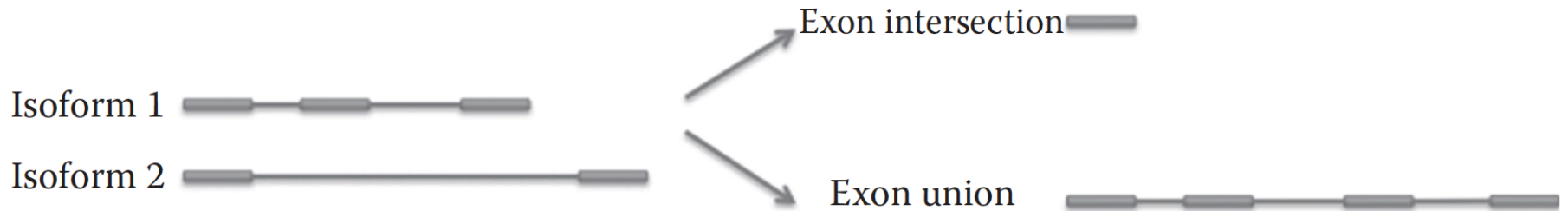
Developed by the same team as CuffDiff, and superior to it according to them. Based on Kallisto.

Transcript-oriented (like CuffDiff)

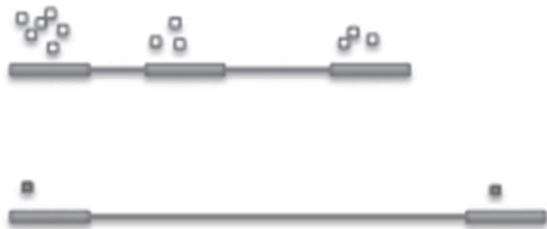
Includes uncertainty coming from “quantification noise” (like CuffDiff)

Supports modelling multiple experimental factors (unlike CuffDiff)

Reason to use transcript-level analysis counting can hide DE



Condition A



Condition B



Fold change
(actual)

38/30

Fold change
(union)

14/14

Fold change
(intersection)

7/7

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. -> $2 \times 6 \times 3 = 36$ 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. $\rightarrow 2 \times 6 \times 3 = 36$ 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. $\rightarrow 2 \times 6 \times 3 = 36$ 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.

\rightarrow DESeq2 / edgeR / Sleuth which can work with factorial designs

(but not e.g. CuffDiff2, SAMSeq)

Decision tree for software selection (2016)

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

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

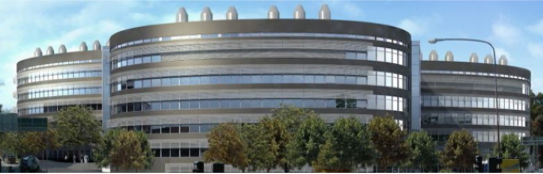
Differentially expressed genes => **Select type of experimental design**

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



Take-away messages from DE tool comparison

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

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.

Gene level analysis

SCIENTIFIC REPORTS



OPEN

Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data

Received: 18 July 2016

Accepted: 3 April 2017

Published online: 08 May 2017

Celine Everaert^{1,2,3}, Manuel Luypaert⁴, Jesper L. V. Maag⁵, Quek Xiu Cheng⁵, Marcel E. Dinger⁵, Jan Hellemans⁴ & Pieter Mestdagh^{1,2,3}

Expression levels are similar between RT-qPCR and RNA-seq data

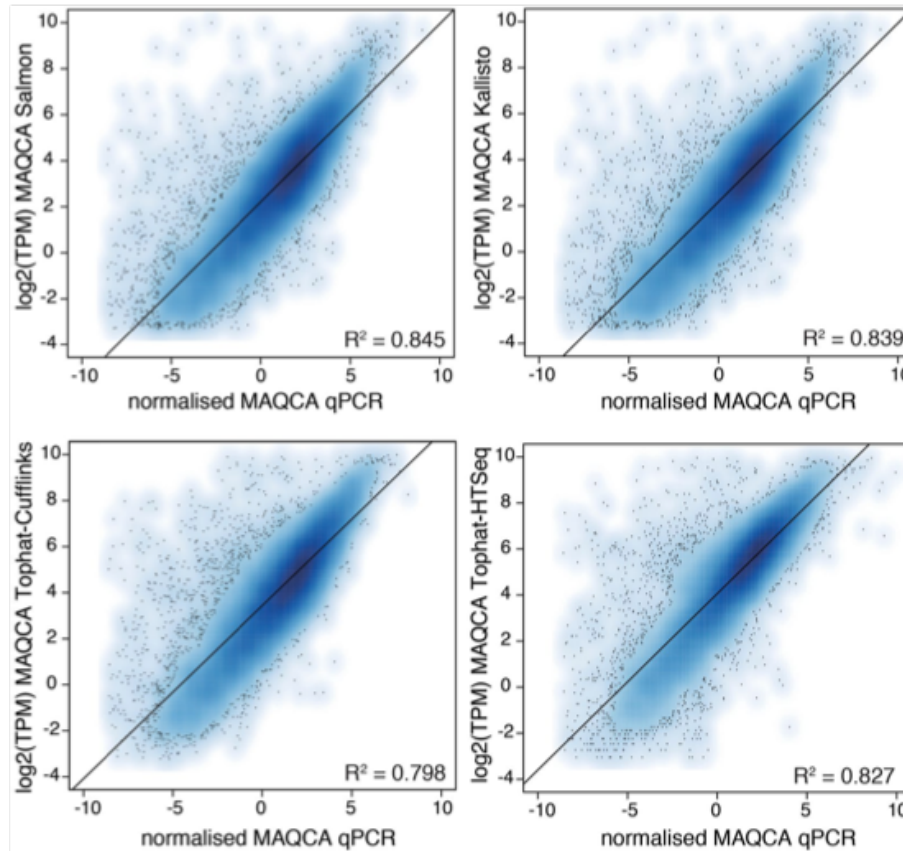
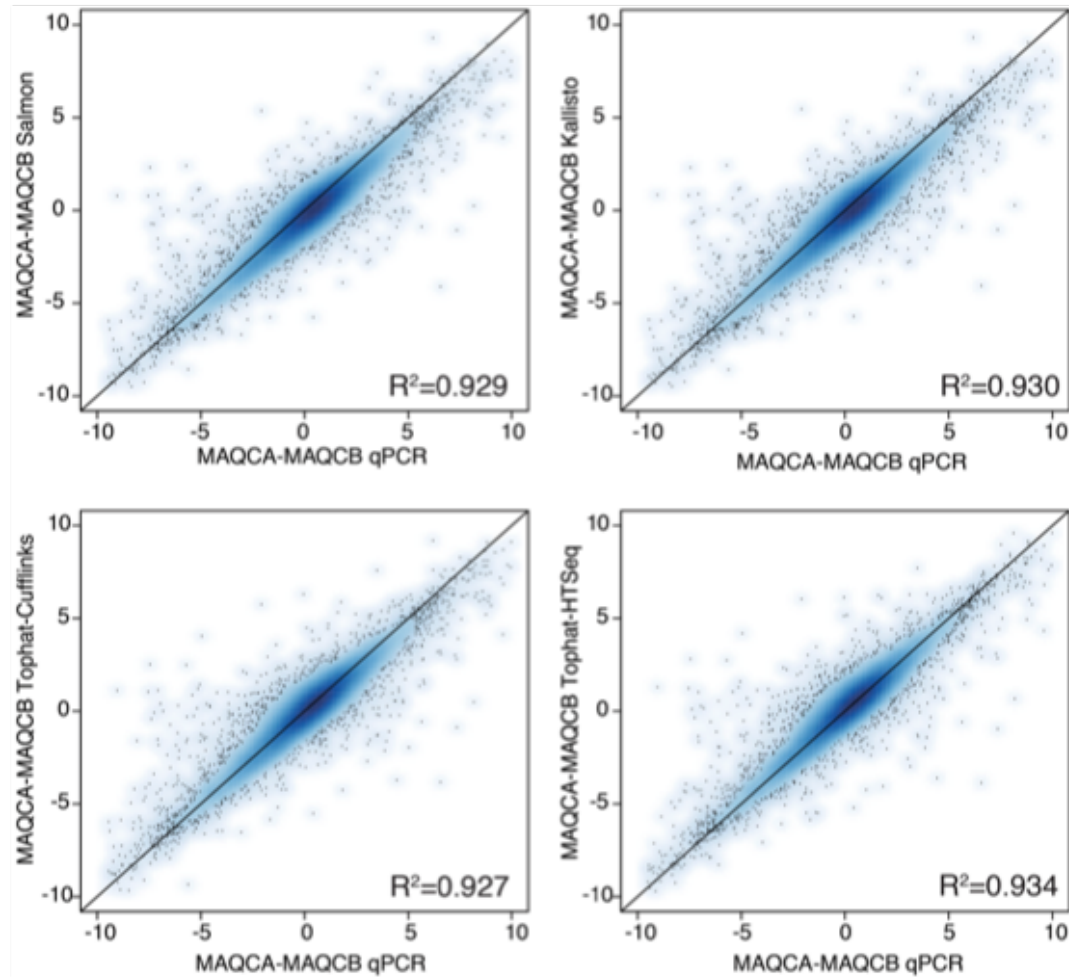


Figure 1. Gene expression correlation between RT-qPCR and RNA-seq data. The Pearson correlation coefficients and linear regression line are indicated. Results are based on RNA-seq data from dataset 1.

Most problems are consistent so they disappear when you do diff-exp analysis



Miscellaneous (if there is time)

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

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:

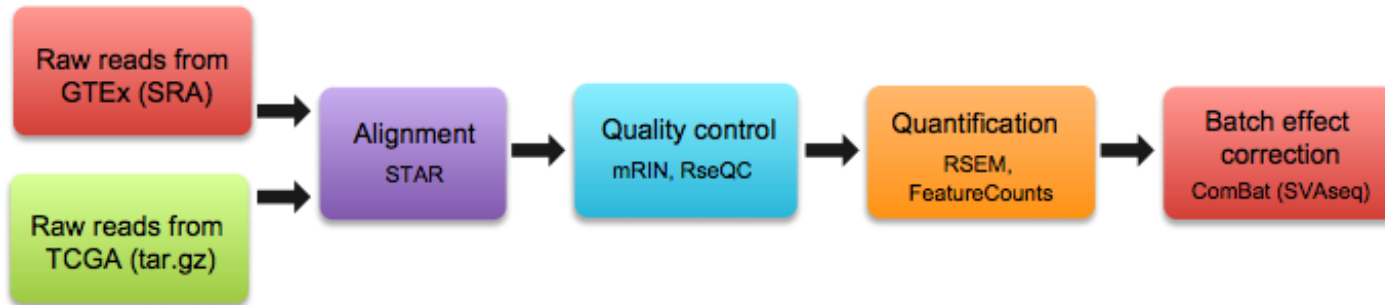
'COMBATTING' BATCH EFFECTS WHEN COMBINING
BATCHES OF GENE EXPRESSION MICROARRAY DATA

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

Enabling cross-study analysis of RNA-Sequencing data

Qingguo Wang^{1,2,3}, Joshua Armenia^{1,2}, Chao Zhang⁴, Alexander V. Penson^{1,2}, Ed Reznik^{1,2}, Liguo Zhang⁵, Angelica Ochoa^{1,2}, Benjamin E. Gross^{1,2}, Christine A. Iacobuzio-Donahue⁵, Doron Betel⁴, Barry S. Taylor^{1,2,6}, Jianjiong Gao^{1,2}, Nikolaus Schultz^{1,2,6}

Recent preprint
<http://biorxiv.org/content/early/2017/02/27/110734>



But see also 2015 paper

Assessing the consistency of public human tissue RNA-seq data sets

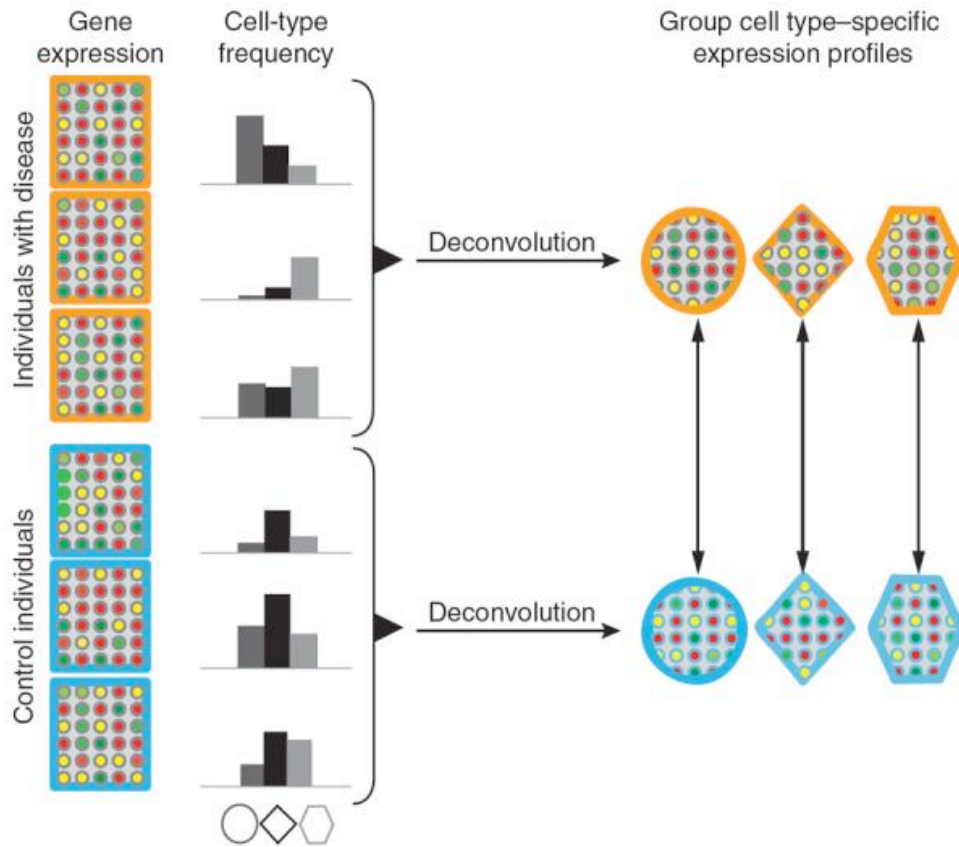
Frida Danielsson, Tojo James, David Gomez-Cabrero and Mikael Huss

Corresponding author. Mikael Huss, Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Box 1031, SE-171 21 Solna, Sweden. Tel.: +46735675775; Fax: +46852481425; E-mail: mikael.huss@scilifelab.se

Key Points

- Publicly available data sets with precomputed RNA expression levels are not comparable in their untransformed state in the sense that samples from the same tissues obtained in different experiments do not cluster by tissue.
- Logarithmic transformation improves clustering of samples in principal components 2 and 3, while principal component 1 still seems to be dominated by study-specific factors.
- RNA extraction method, read length and sequencing layout (single-end versus paired-end) contribute strongly to variation between samples.
- Removal of known batch effects is essential for clustering based on tissue type.
- Reprocessing raw data avoids loss of expression information because of gene identifier matching issues but does not serve to improve clustering.

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

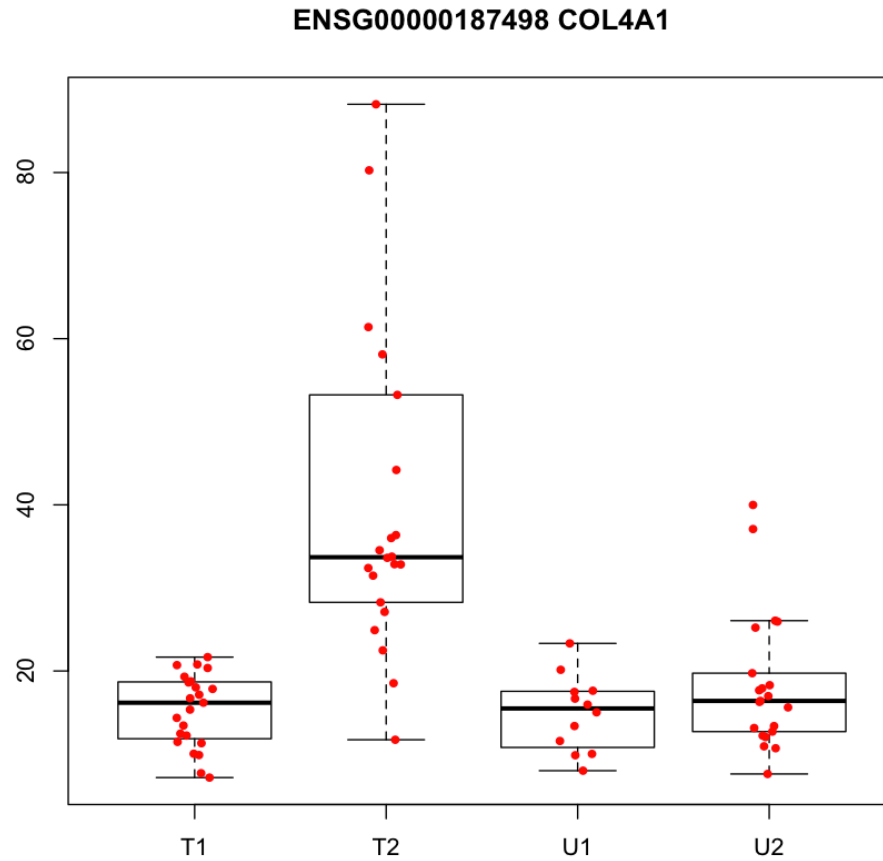
Identifier	logFC	logCPM	LR	PValue	FDR
ENSMUSG00000046623	-5.46102265507855	0.687470648417142	130.820399258671	2.71053464157785e-30	1.02973211033542e-25
ENSMUSG00000046623	-5.46102265507855	0.687470648417142	130.820399258671	2.71053464157785e-30	1.02973211033542e-25

(and so on ...)

Log fold change, FDR

How to visualize?

Looking at top genes one by one




Box plot

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

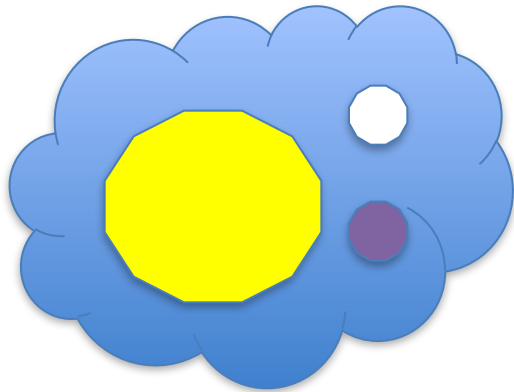
Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

TMM - Trimmed Mean of M values

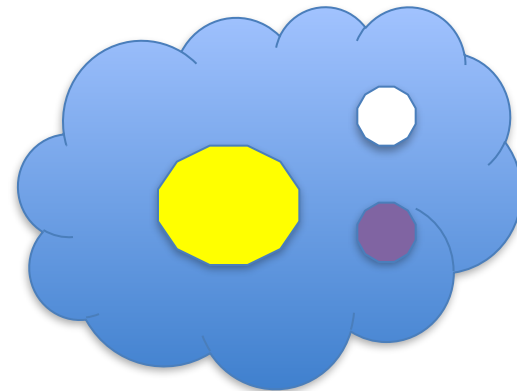
Attempts to correct for differences in RNA *composition* between samples

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 -> white and purple 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

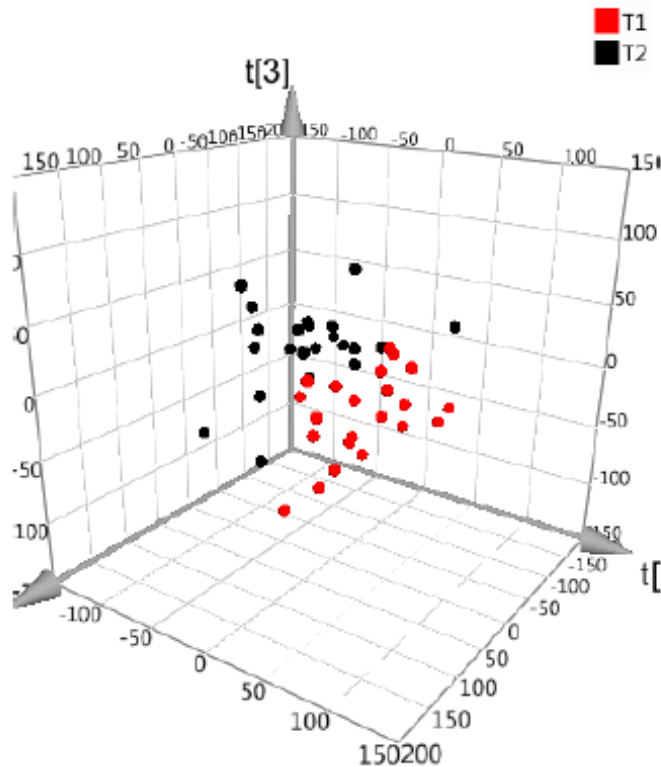
- Use a discrete statistical model
- Want to retain statistical power (see next slide)

... 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.

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).