11/06/2014

RNASeq
Differential Expression

Le Corguillé

v1.00

Introduction

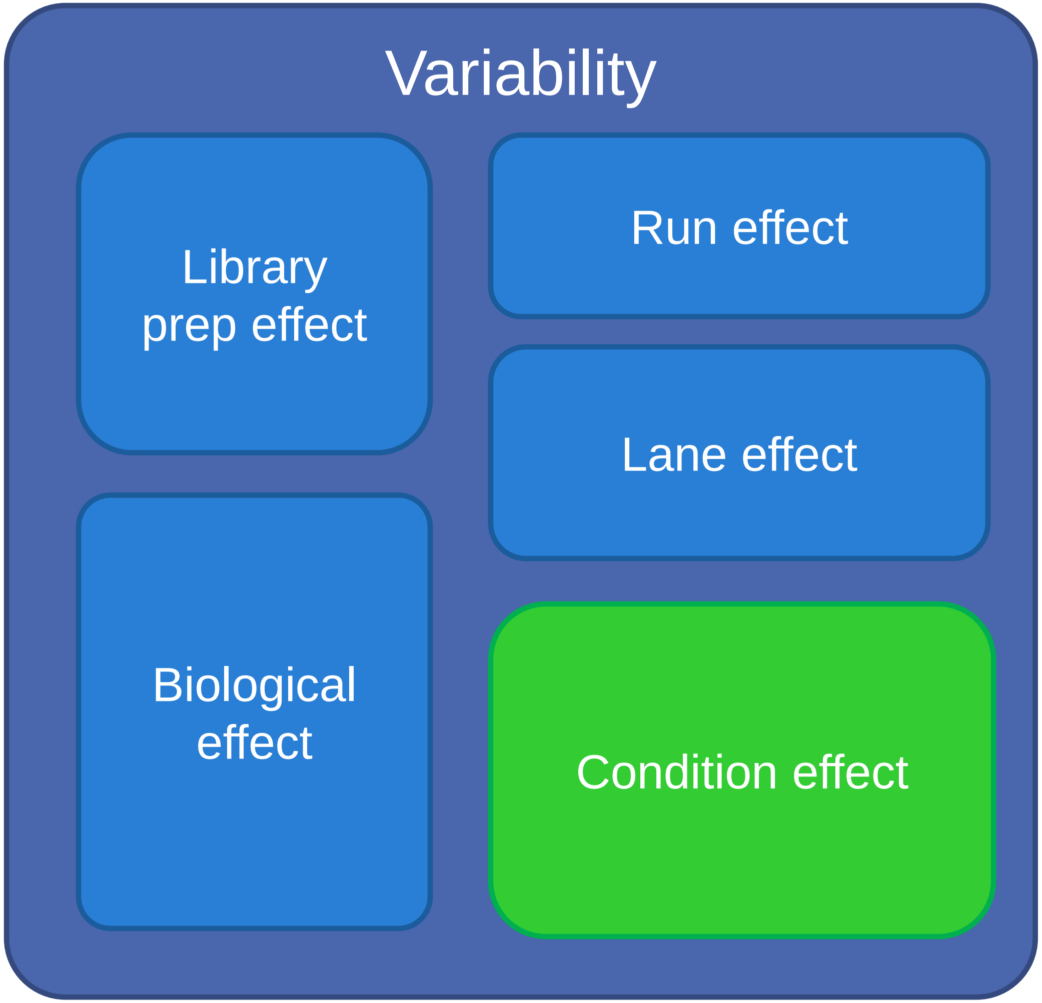
RNASeq

- No previous genomic sequence information is needed
  - In RNA-seq the expression signal of a transcript is limited by the sequencing depth and is dependent on the expression levels of other transcripts.
  - Discreet distributions.

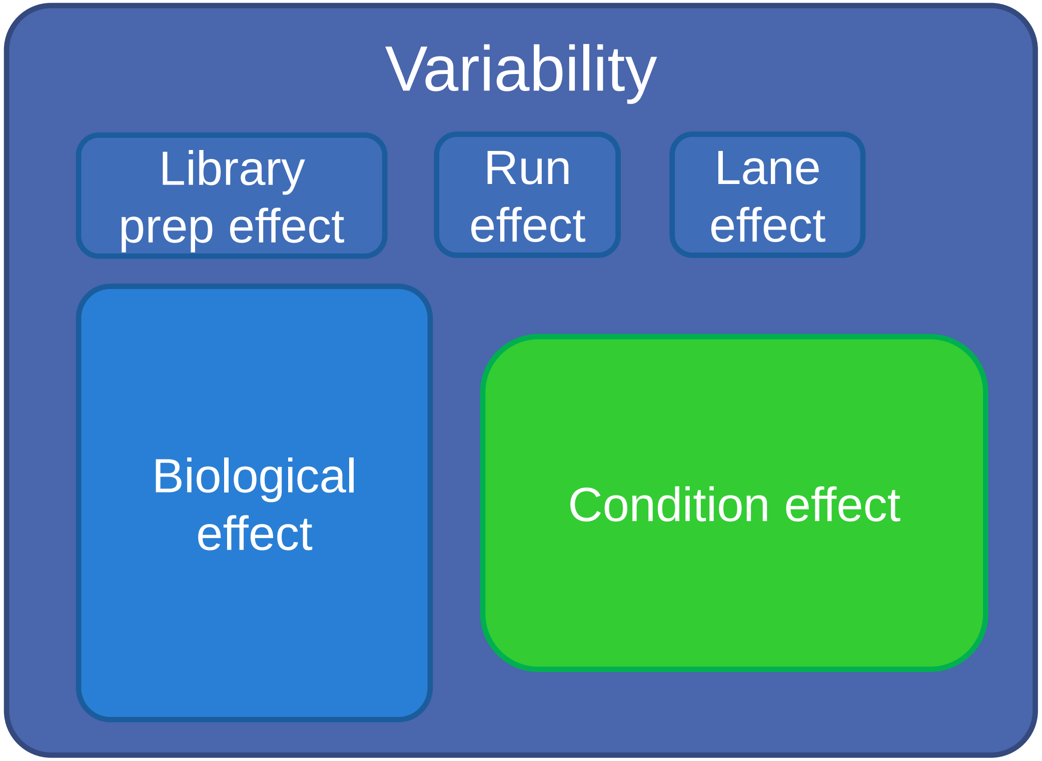
Microarray

- An existing library of expressed sequence tags is required
  - In array-based methods probe intensities are independent of each other such as microarrays.
- Continuous distributions

Introduction

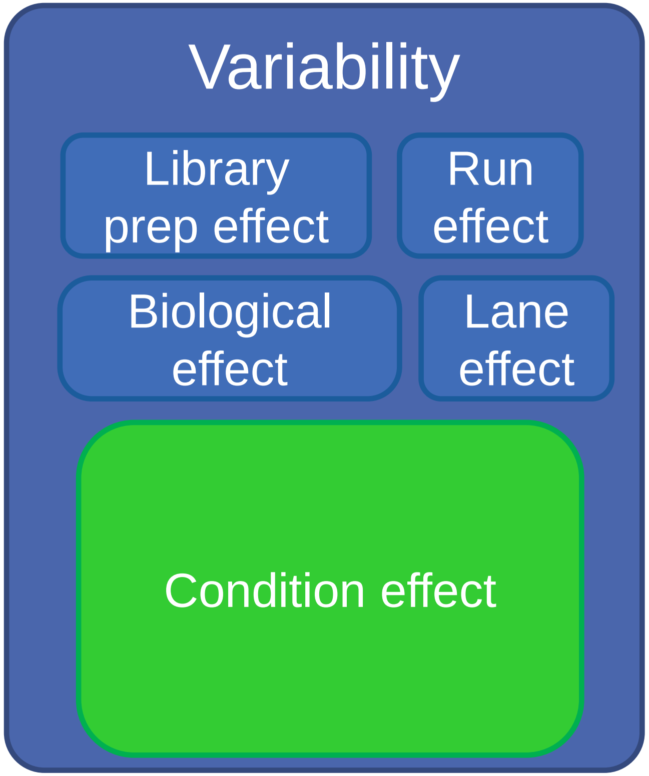


Introduction



Technical replicates
+ normalization
+ statistics

Introduction



Technical replicates
+ normalization
+ statistics

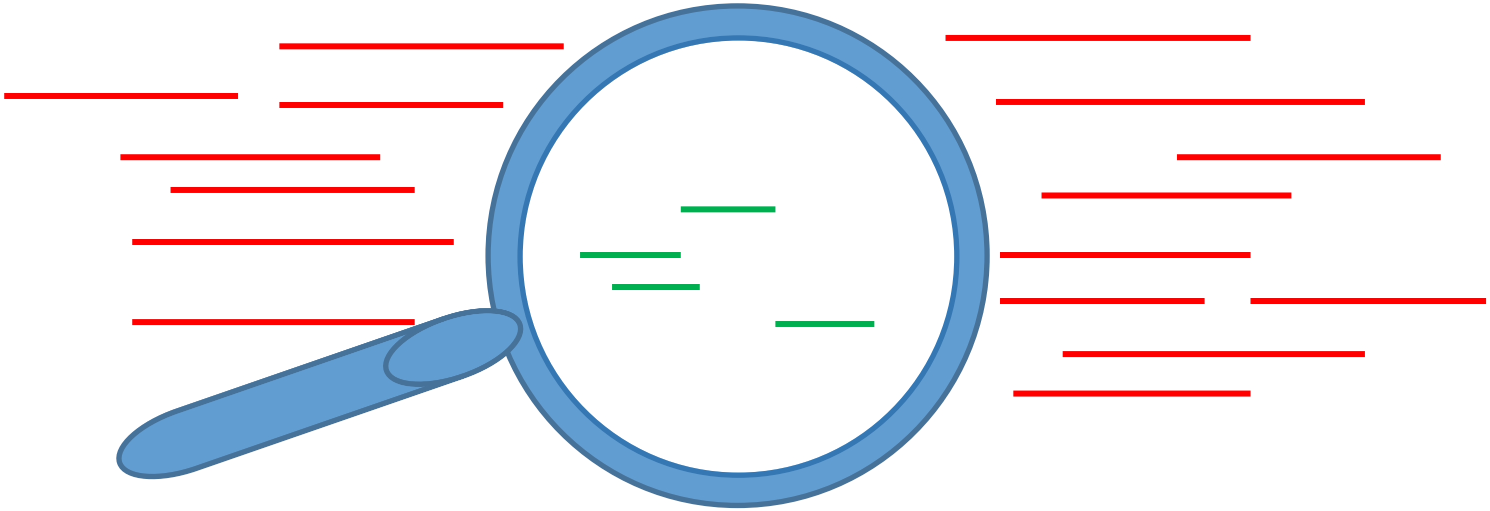
+

Biological replicates
+ statitics

Introduction

- Replicates

↗ Accuracy



INPUTS

Inputs

- Raw count table

id	LL06_1	LL06_2	LL09_1	LL09_2
comp3130_seq1	12	6	9	15
comp3131_seq2	167	233	987	856
comp4523_seq1	685	785	648	458
comp6984_seq3	87	68	354	591

Inputs

- Samples metadata / Samples info

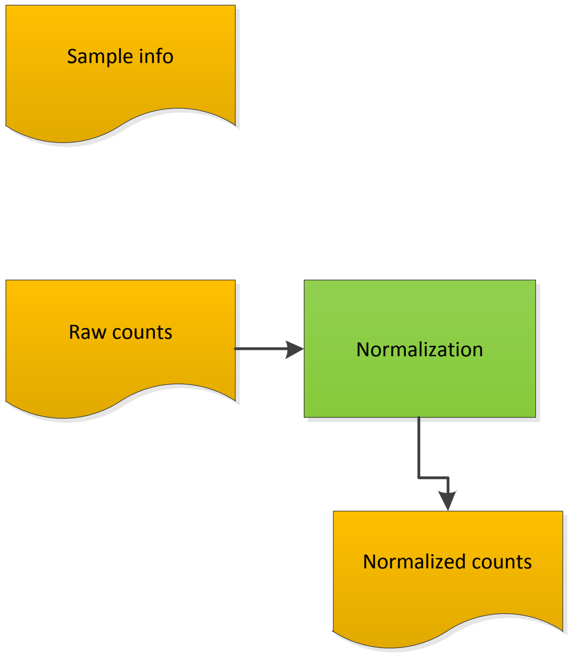
samplename	batch	light	hour	…
LL06_1	1	LL	06	
HL06_1	1	HL	06	
LL09_1	1	LL	09	
HL09_1	1	HL	09	
LL12_1	1	LL	12	
HL12_1	1	HL	12	
LL06_2	2	LL	06	
HL06_2	2	HL	06	
LL09_2	2	LL	09	

Inputs

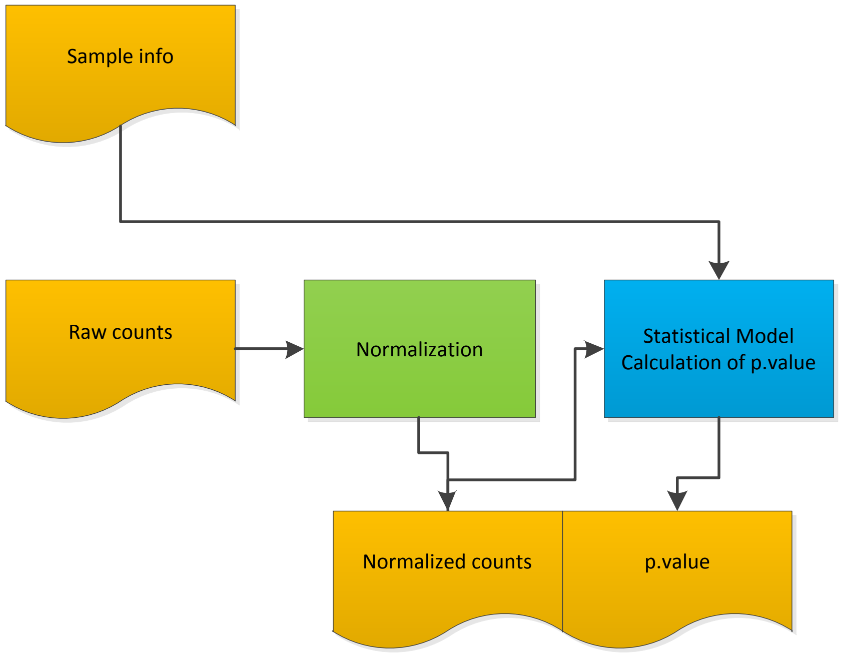
- Scale
  - Exon level -> DEXSeq
  - Gene level
  - Isoform level

THE WORKFLOW

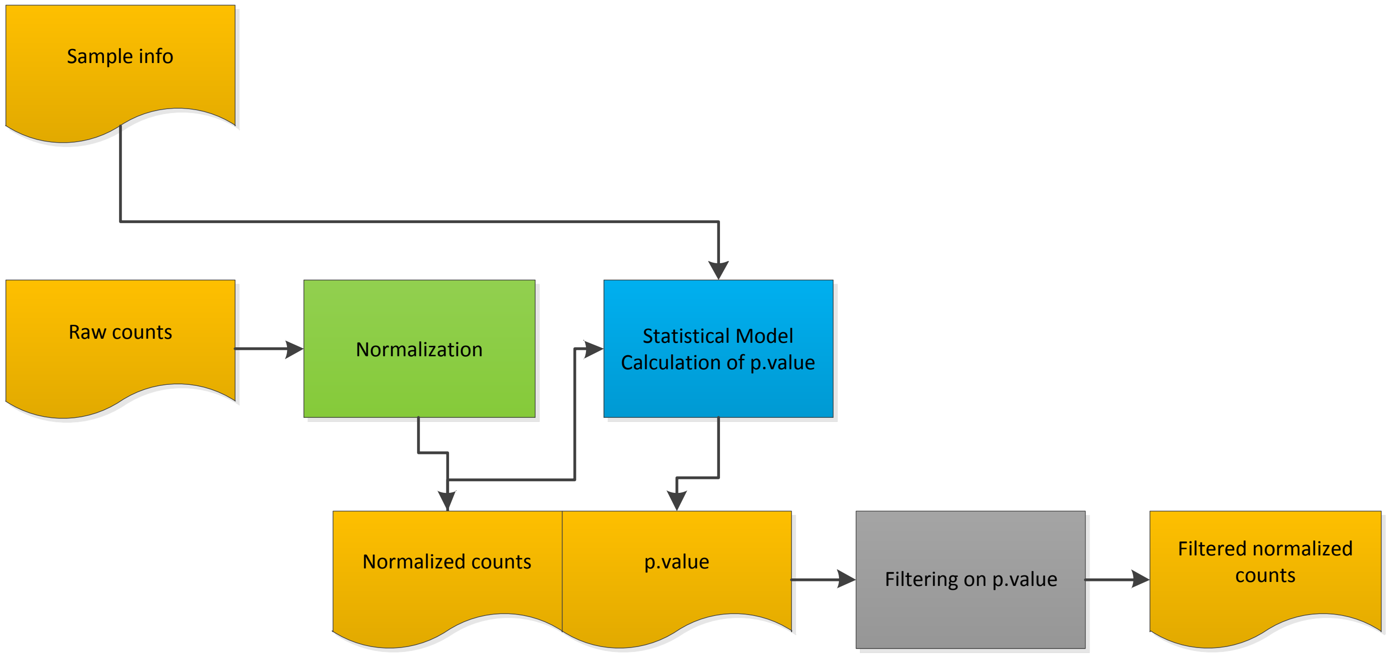
The workflow



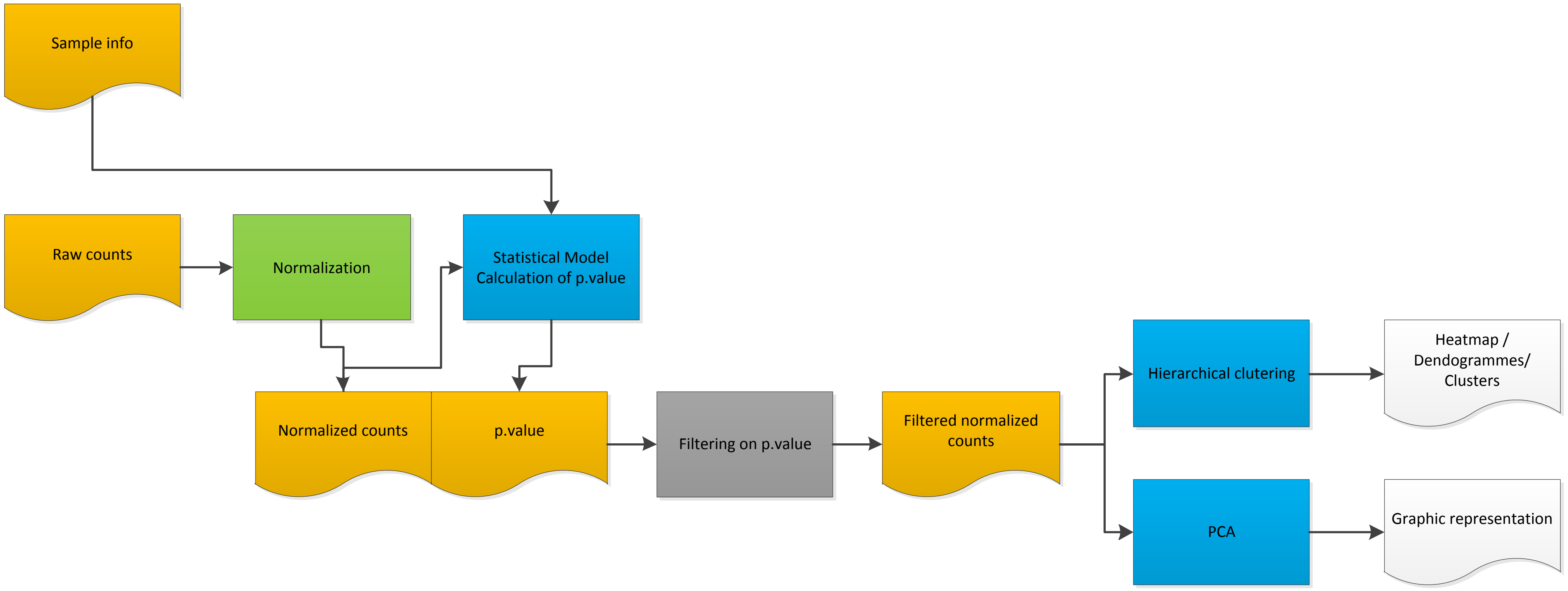
The workflow



The workflow

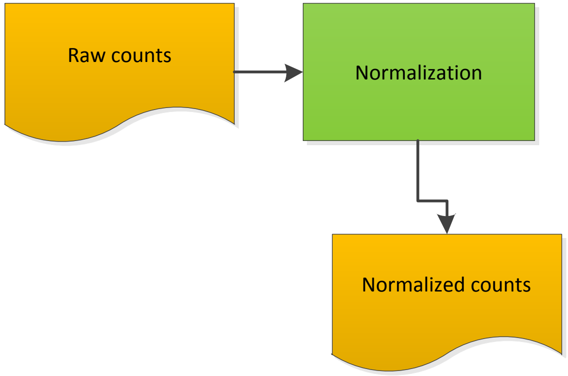


The workflow



NORMALIZATION

Normalization



Normalization

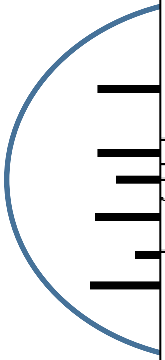
- Why ?
  - Between-sample → compare a gene in different sample
    - Depth of sequencing == library size
    - Sampling bias during the libaries construction == batch effect
    - Presence of majority fragments == saturation
    - Sequence composition du to PCR-amplification step (GC content)
  - Within-sample → compare genes in a sample
    - Gene length
    - Sequence composition (GC content)

Normalization

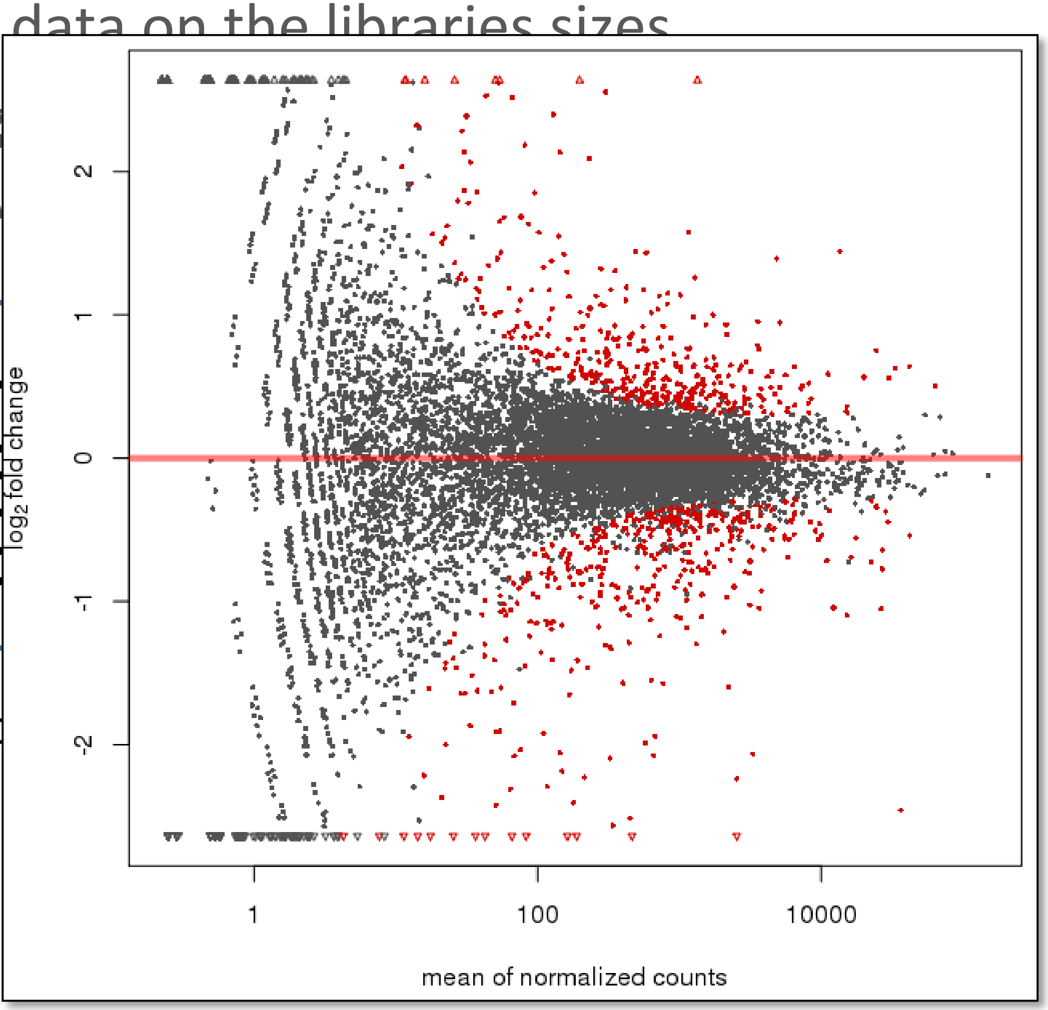
- How ?
  - Between-lane → compare a gene in different sample
    - Scale data on the libraries sizes
    - Using housekeeping genes
  - Within-lane → compare genes in a sample
    - Normalize on gene lengths

Normalization

- How ?
  - Between-lane → compare a gene in different sample
    - Scale data on the libraries sizes
    - Using
      - W

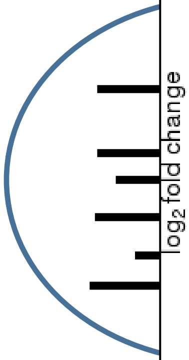


Co

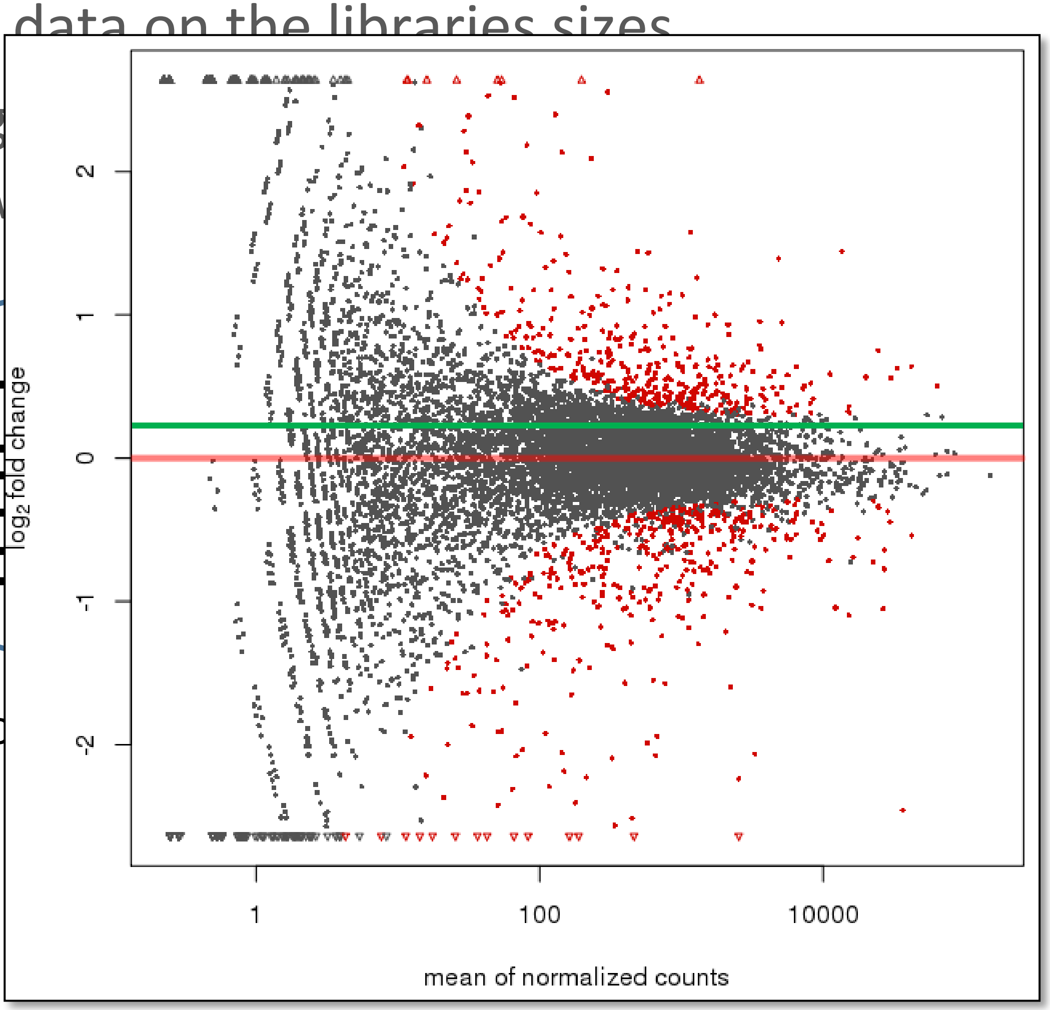


Normalization

- How ?
  - Between-lane → compare a gene in different sample
    - Scale data on the libraries sizes
    - Using
      - W

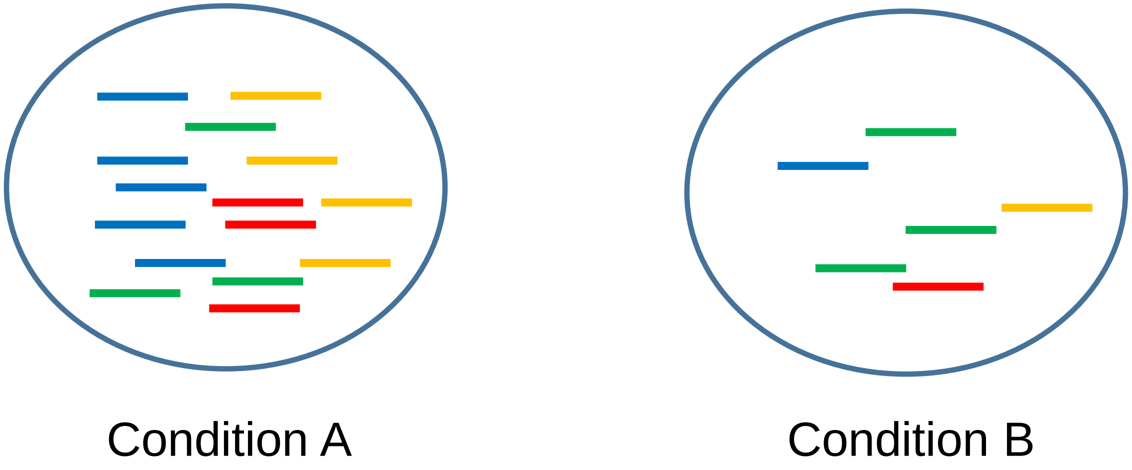


Co



Normalization

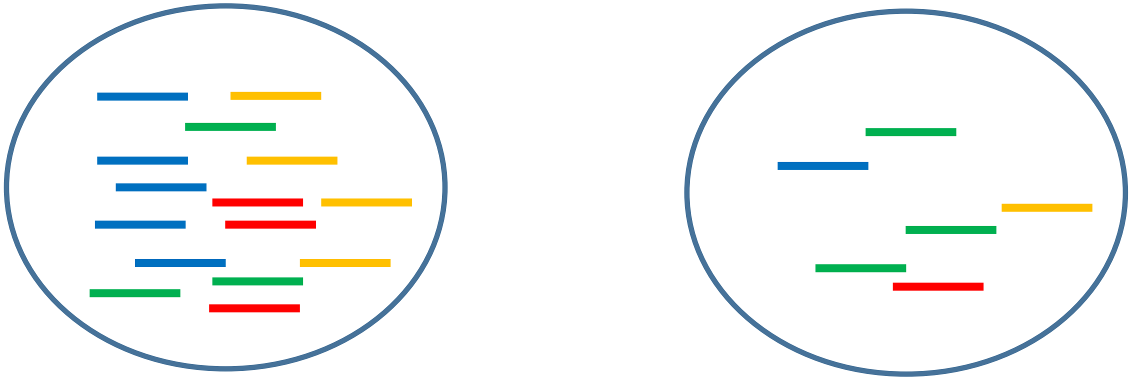
- How ?
  - Between-lane → compare a gene in different sample
    - Scale data on the libraries sizes
    - Using housekeeping genes
      - When :



- Examples : actin, GAPDH, ubiquitin, HSP90, Histone, rRNA, tRNA …

Normalization

- How ?
  - Between-lane → compare a gene in different sample
    - Scale data on the libraries sizes
    - Using housekeeping genes
      - When :



Condition A

Condition B

- Examples : actin, GA 0, Histone, rRNA, tRNA …

It depends

Normalization

- Normalization methods

Total Counts (TC)

- Motivation: greater lane sequencing depth => greater counts
- Assumption: read counts are proportional to expression level and sequencing depth (same RNAs in equal proportion)
- Method: divide transcript read count by total number of reads

- Problem: Sensitive to the presence of majority genes

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/Julie_AUBERT.pdf

Normalization

- Normalization methods

Upper Quartile normalization (UQ) or Median (Med)

- Motivation: total read count is strongly dependent on a few highly expressed transcripts
- Assumption: read counts are proportional to expression level and sequencing depth
- Method: divide transcript read count by, e.g., upper quartile

- Problem: Sensitive to the presence of majority genes

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/Julie_AUBERT.pdf

Normalization

- Normalization methods

Full quantile normalization (FQ)

- Motivation: total read count is strongly dependent on a few highly expressed transcripts
- Assumption: read counts have identical distribution across lanes
- Method: all quantiles of the count distributions are matched between lanes

- Problem: Can increase between group variance
  Is based on an very (too) strong assumption (similar distributions)

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/Julie_AUBERT.pdf

Normalization

- Normalization methods

Reads Per Kilobase per Million mapped reads (RPKM / FPKM)

- Motivation: greater lane sequencing depth and gene length => greater counts whatever the expression level
  Allow comparaison of expression of different genes in a sample
- Assumption: read counts are proportional to expression level, gene length and sequencing depth (same RNAs in equal proportion)
- Method: divide gene read count by total number of reads (in million) and gene length (in kb)

- Problem: Sensitive to the presence of majority genes
  Implies a similarity between RNA repertoires expressed

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/Julie_AUBERT.pdf

Normalization

- RPFM / FPFM

  http://blog.nextgenetics.net/?e=51

  – Pro
    - Simple, easy to understand
    - Comparable between different genes within the same dataset

  – Cons
    - Small changes in highly expressed genes (especially differences in rRNA contamination) cause a global shift in all other values
    - Small changes across lowly expressed genes (especially differences in DNA contamination) cause differences across a wide number of genes.

    - Mixing of noise levels
    - Noise is generally linked to the number of observations
    - The same RPKM value could come from
      - A small lowly observed gene with high noise
      - A large well observed gene with low noise

Normalization

- Library size VS RPFM
  - If we only normalized the individual tag counts by total library tag count, we would get a constant average normalized abundance. The numerator and denominater in this normalization are both in the same unit - tag counts.
  - For RPKMs, we are normalizing the tags by gene length first, and then normalizing by library size. The first normalization by length produces the unit of tag count/kilobase. The second normalization by library size divides tags/kilobase by tag count. This improper unit of normalization in the denominater is what is causing the inconsistent average RPKMs.
  - The average number of read across samples differe after RPKM normalization (NDR: The author propose to fix this issue)

http://blog.nextgenetics.net/?e=51

http://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/

Normalization

- Library size VS RPFM
  - If … ry tag count, w… e numerator a… unit - tag c…
  - F… and then n… produces the u… ary size d… malization in t… RPKMs.
  - T… KM normalization (NDR: The author propose to fix this issue)

http://blog.nextgenetics.net/?e=5

http://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/

Normalization

- Normalization methods

Reads Per Kilobase per Million mapped reads (RPKM / FPKM)

- Motivation: greater lane sequencing depth and gene length => greater counts whatever the expression level
  Allow comparaison of expression of different genes in a sample
- Assumption: read counts are proportional to expression level, gene length and sequencing depth (same RNAs in equal proportion)
- Method: divide gene read count by total number of reads (in million) and gene length (in kb)

- Problem: Sensitive to the presence of majority genes
  Implies a similarity between RNA repertoires expressed

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/Julie_AUBERT.pdf

Normalization

- Normalization methods

The Effective Library Size concept : TMM (edgeR) and DESeq

- Motivation: Different biological conditions express different RNA repertoires, leading to different total amounts of RNA
- Assumption: A majority of transcripts is not differentially expressed
- Method: Minimizing effect of (very) majority sequences

- Problem: ?

implies a similarity between RNA repertoires expressed

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/Julie_AUBERT.pdf

Normalization

- Normalization methods

The Effective Library Size concept : TMM (edgeR) and DESeq

- Motivation: Different biological conditions express different RNA repertoires, leading to different total amounts of RNA
- Assumption: A majority of transcripts is not differentially expressed
- Method: Minimizing effect of (very) majority sequences

- Problem: ?

implies a similarity between RNA repertoires expressed

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/Julie_AUBERT.pdf

Normalization

- **The Effective Library Size**
  - TMM / edgeR
    - uses the number of mapped reads (i. e., count table column sums) and estimates an additional normalization factor to account for sample-specific effects (e. g., diversity); these two factors are combined and used as an offset in the NB model.
  - DESeq
    - defines a virtual reference sample by taking the median of each gene’s values across samples, and then computes size factors as the median of ratios of each sample to the reference sample.

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor
Simon Anders, Davis J. McCarthy, Yunshen Chen, Michal Okoniewski, Gordon K.Smyth, Wolfgang Huber & Mark D. Robinson

Normalization

WARNING

- It is important to recognize that the number of reads which overlap a gene is not a direct measure of the gene’s expression.

=> Genes length bias

=> One effect of this bias is to reduce the ability to detect differential expression among shorter genes simply from the a lack of coverage since the power of statistical tests involving count data decreases with lower number of count

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data
Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E. Mason, Nicholas D. Socci and Doron Betel

Normalization

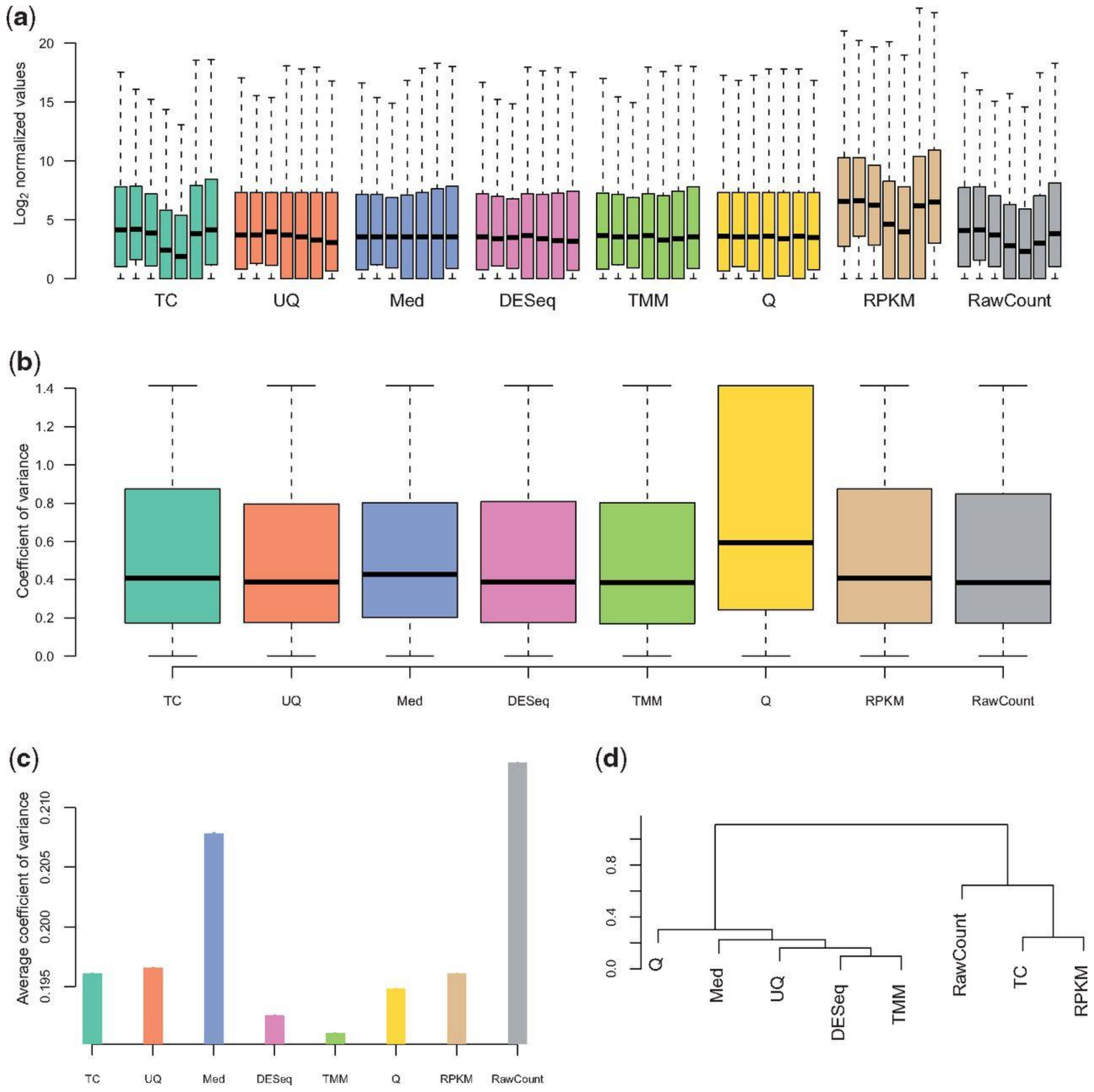


Figure 1:

Comparison of normalization methods for real data.

(A) Boxplots of log2(counts + 1) for all conditions and replicates in the M. musculus data, by normalization method.

(B) Boxplots of intra-group variance for one of the conditions (labeled ‘B’ in the corresponding data found in Supplementary Data) in the M. musculus data, by normalization method.

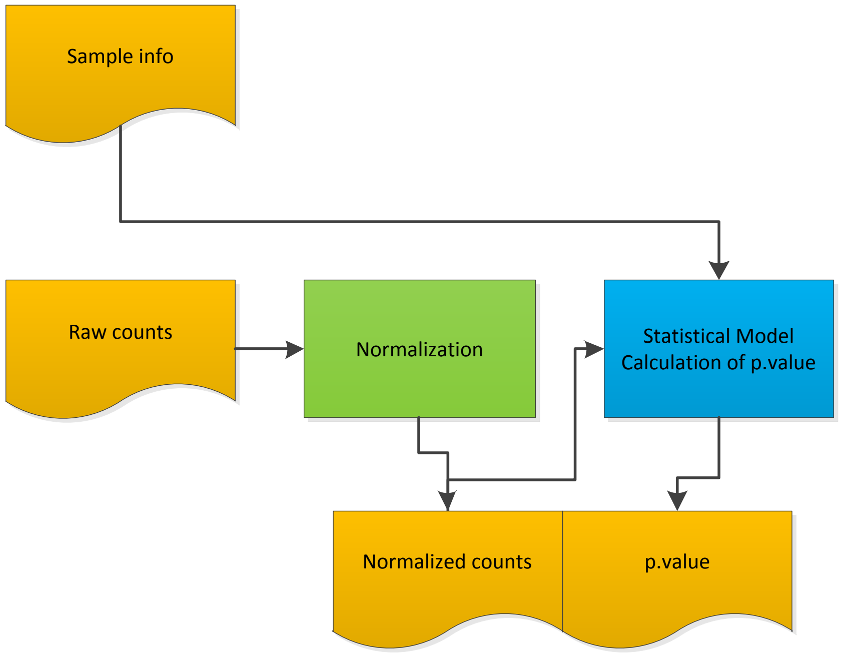
(C) Analysis of housekeeping genes for the H. sapiens data.

(D) Consensus dendrogram of differential analysis results, using the DESeq Bioconductor package, for all normalization methods across the four datasets under consideration.

MA Dillies, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform (2013) 14 (6): 671-683 :480

STATISTICS

Statistics


Sample info
Raw counts
Normalization
Statistical Model
Calculation of p.value
Normalized counts
p.value

Statistics Models

- Rappel

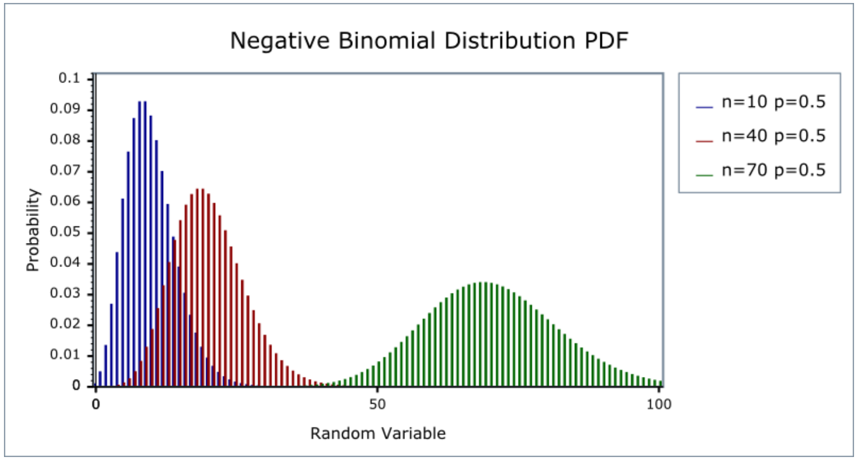
$$\mathbb{P}(X=k)=\int_0^{+\infty}\frac{\lambda^k e^{-\lambda}}{k!}\frac{\lambda^{r-1}e^{-\lambda/\theta}}{\Gamma(r)\theta^r}\mathrm{d}\lambda$$

$$\begin{aligned}\mathbb{P}(X_n\le k)&=I_p(n,k+1)\\&=1-I_{1-p}(k+1,n)\\&=1-I_{1-p}((k+n)-(n-1),(n-1)+1)\\&=1-\mathbb{P}(Y_{k+n}\le n-1)\\&=\mathbb{P}(Y_{k+n}\ge n)\end{aligned}$$

$$\mathbb{P}(X=k)=\left(\frac{\theta}{\theta+1}\right)^{r+k}\int_0^{+\infty}\frac{1}{\Gamma(r)k!\theta^r}t^{k+r-1}e^{-t}dt$$

$$\begin{aligned}f(k;r,p)&=\int_0^\infty f_{\mathrm{Poisson}(\lambda)}(k)\cdot f_{\mathrm{Gamma}\left(r,\frac{p}{1-p}\right)}(\lambda)\,\mathrm{d}\lambda\\&=\int_0^\infty\frac{\lambda^k}{k!}e^{-\lambda}\cdot\lambda^{r-1}\frac{e^{-\lambda(1-p)/p}}{\left(\frac{p}{1-p}\right)^r\Gamma(r)}\,\mathrm{d}\lambda\\&=\frac{(1-p)^r p^{-r}}{k!\,\Gamma(r)}\int_0^\infty\lambda^{r+k-1}e^{-\lambda/p}\,\mathrm{d}\lambda\\&=\frac{(1-p)^r p^{-r}}{k!\,\Gamma(r)}\,p^{r+k}\,\Gamma(r+k)\\&=\frac{\Gamma(r+k)}{k!\,\Gamma(r)}\,(1-p)^r p^k.\end{aligned}$$

$$\begin{aligned}\binom{n}{k}+\binom{n}{k+1}&=\frac{n!}{k!(n-k)!}+\frac{n!}{(k+1)!(n-(k+1))!}\\&=\frac{n!(k+1)}{k!(k+1)(n-k)!}+\frac{n!(n-k)}{(k+1)!(n-k-1)!(n-k)}\\&=\frac{n!(k+1)}{(k+1)!(n-k)!}+\frac{n!(n-k)}{(k+1)!(n-k)!}\\&=\frac{n!((k+1)+(n-k))}{(k+1)!(n-k)!}\\&=\frac{n!(n+1)}{(k+1)!(n-k)!}\\&=\frac{(n+1)!}{(k+1)!((n+1)-(k+1))!}\\&=\binom{n+1}{k+1}.\end{aligned}$$



- The model

Poisson distribution

- Motivation: Poisson distribution appears when things are counted
- Assumption: mean and variance are the same
- Method: Poisson distribution has only one parameter λ (expected number of reads)

- Problem:

  Good distribution for technical replicates
  But biological variability of RNA-seq count data cannot be capture using the Poisson distribution because data present overdispersion
  (i.e., variance of counts larger than mean)

http://smpgd2012.univ-lyon1.fr/IMG/pdf/NGS/gonzalez-smpgd2012.pdf

Statistics Models

- The model

Poisson distribution

- M appears when

- As are the same

- M only one para reads)

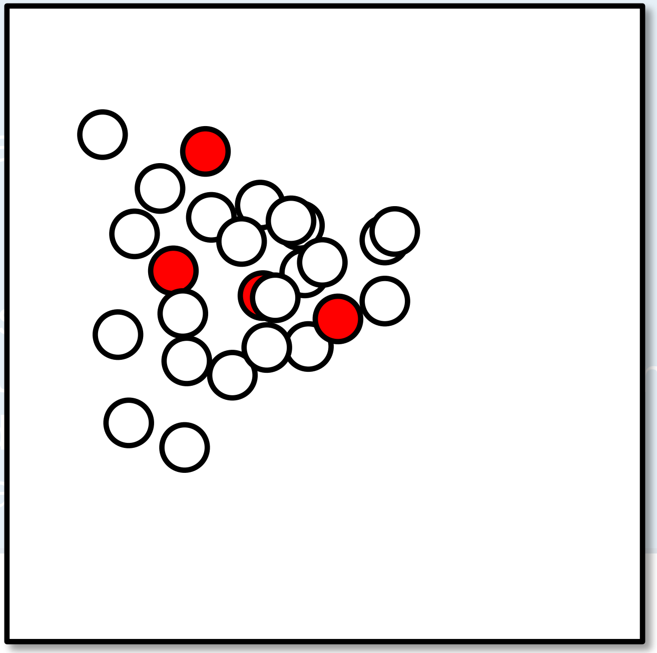
- Pr

nical replicates

f RNA-seq cou ng the

Pois present overd

rger than mea



- The model

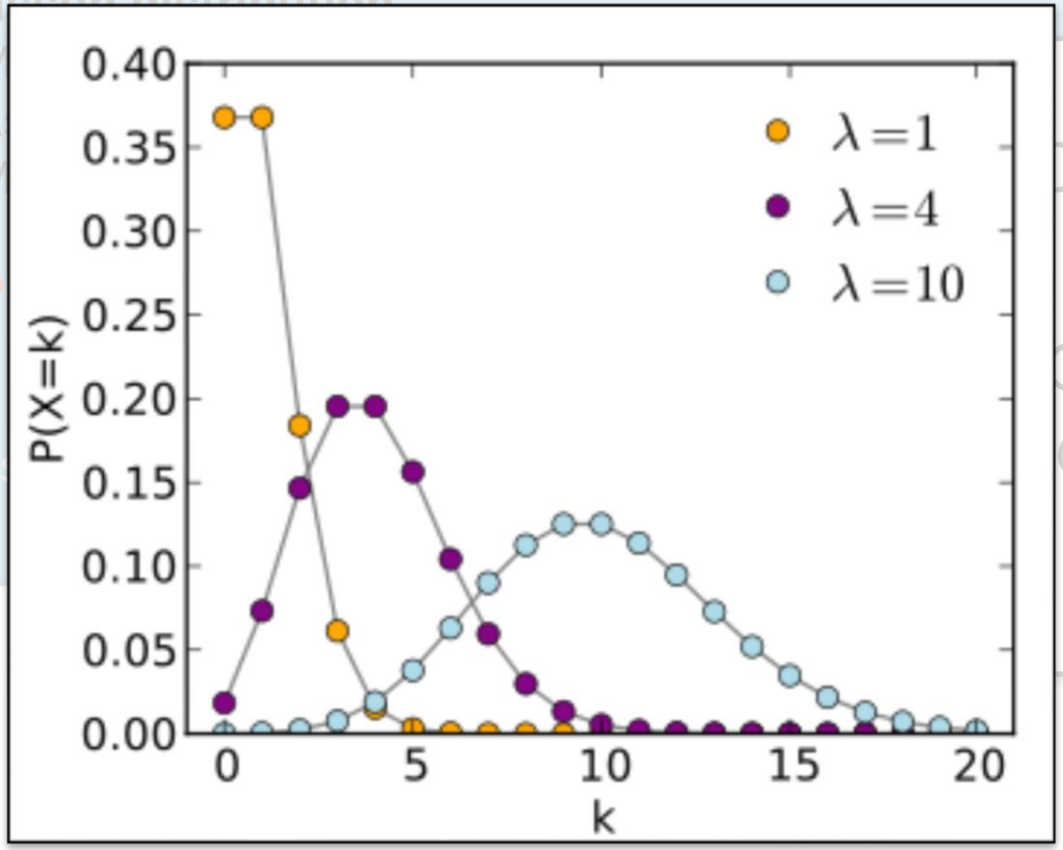
Poisson distribution

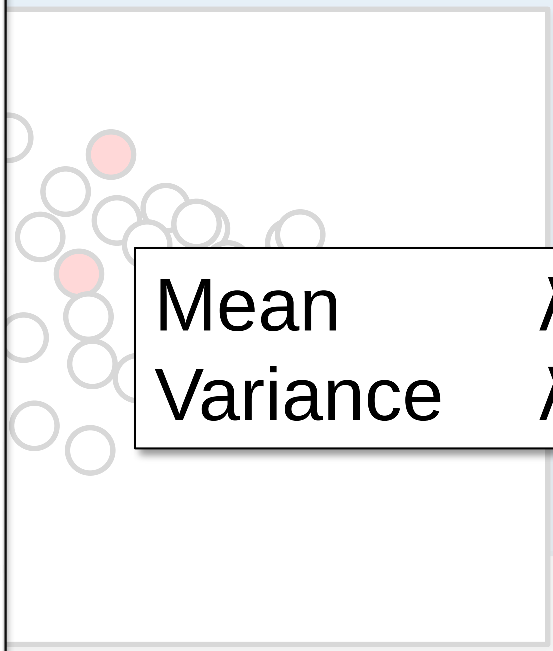
- M
- A
- M

reads)

- P

Poi



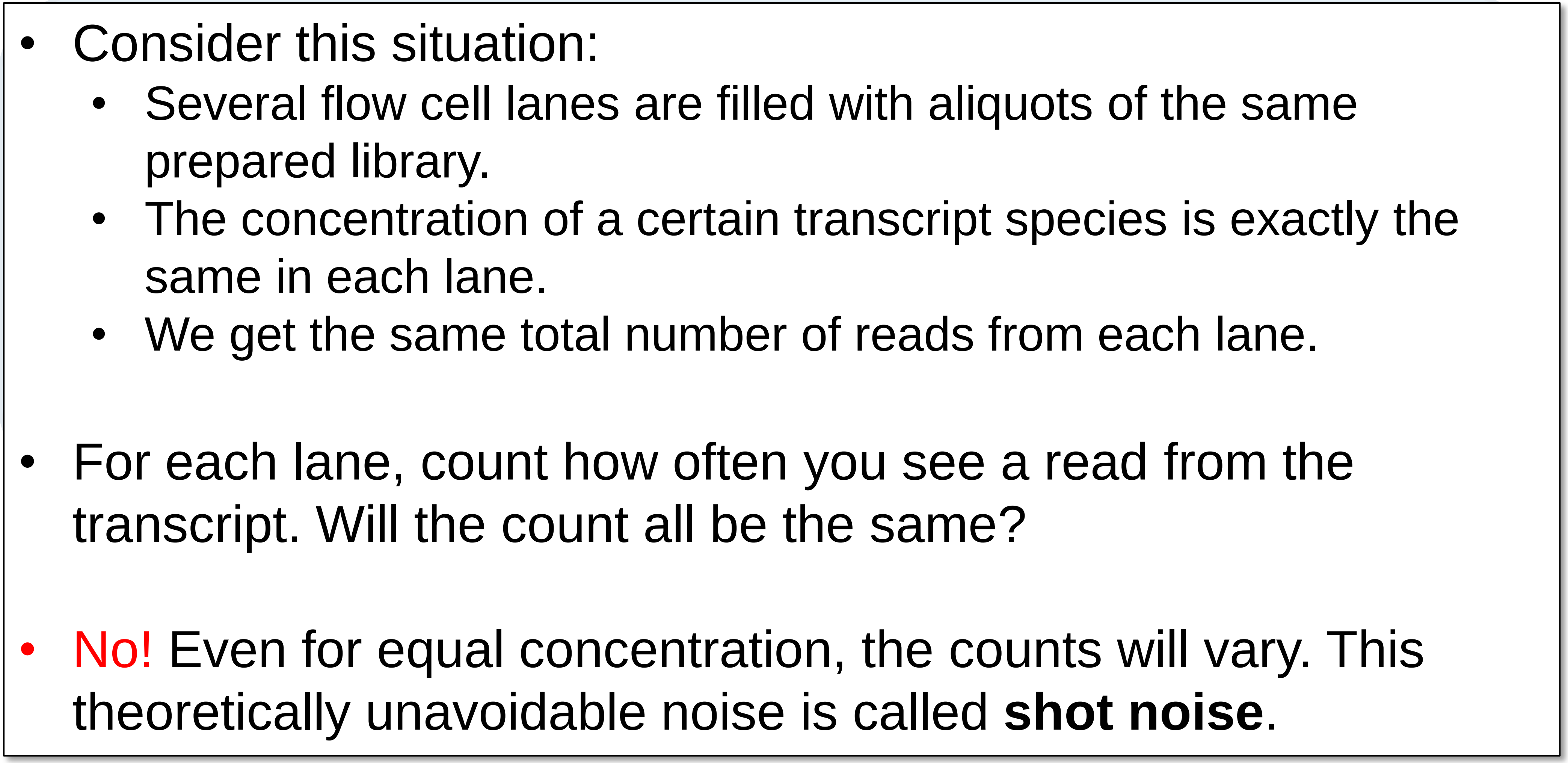


Mean	λ
Variance	λ

Wikipedia

Statistics Models

- The model



- Consider this situation:
  - Several flow cell lanes are filled with aliquots of the same prepared library.
  - The concentration of a certain transcript species is exactly the same in each lane.
  - We get the same total number of reads from each lane.
- For each lane, count how often you see a read from the transcript. Will the count all be the same?
- No! Even for equal concentration, the counts will vary. This theoretically unavoidable noise is called **shot noise**.

Simon Anders – EMBL – Differential expression analysis for sequencing count data

Statistics Models

- The model

Negative Binomial (NB): edgeR and DESeq

- Motivation: distribution takes into account Overdispersion
- Assumption:
- Method: NB is a two-parameter distribution

  Origin: Y ~ NB (p, m)

  Y ... number of successes in a sequence of Bernoulli trials with probability p before r failures occur

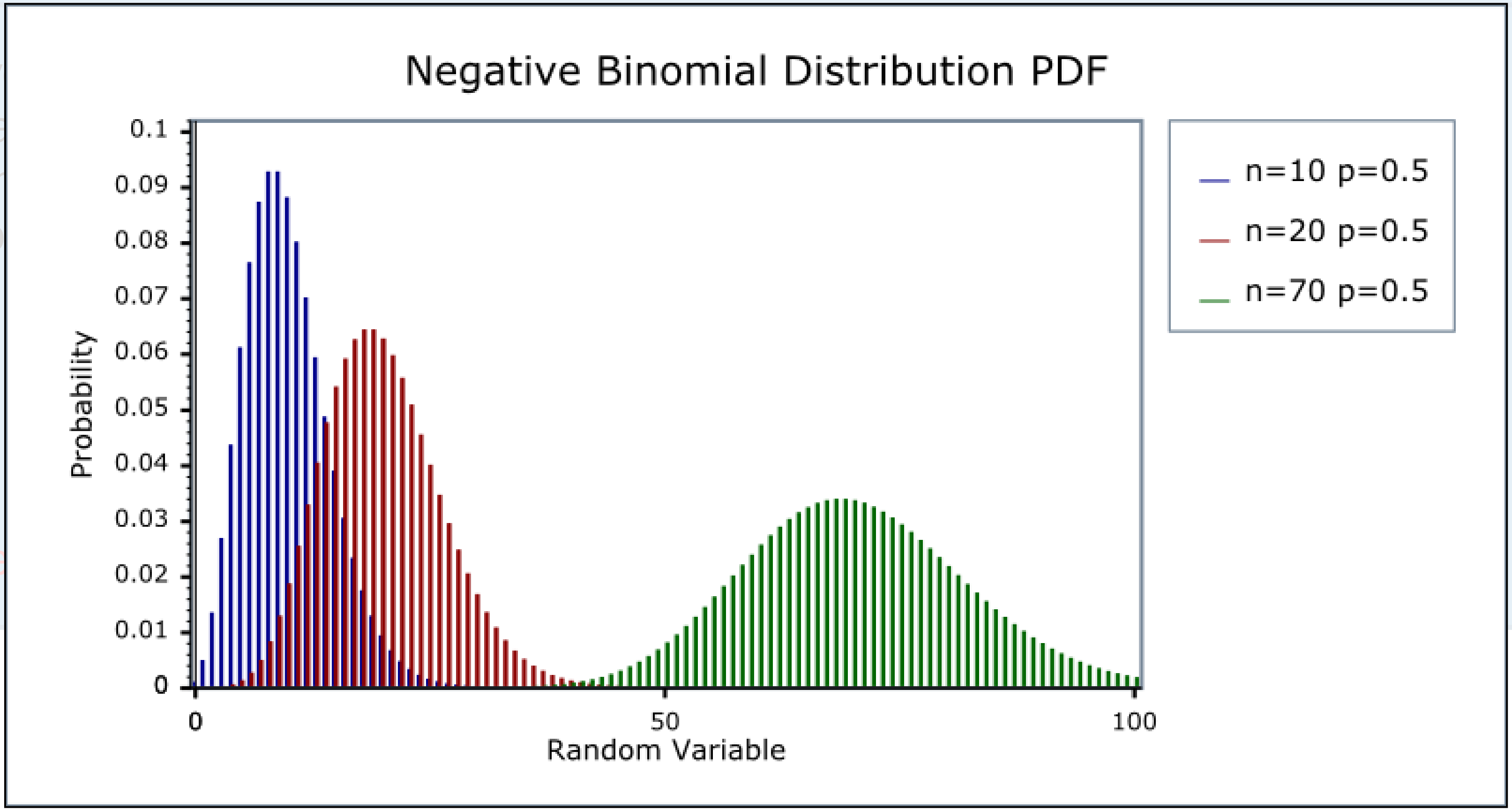
  RNASeq case: λ (mean) and φ (overdispersion)

- Problem: φ_i / gene cannot be estimated due to the small number of individuals

http://smpgd2012.univ-lyon1.fr/IMG/pdf/NGS/gonzalez-smpgd2012.pdf

- The model

Negativ
- Motiva
- Assum
- Metho
- Proble

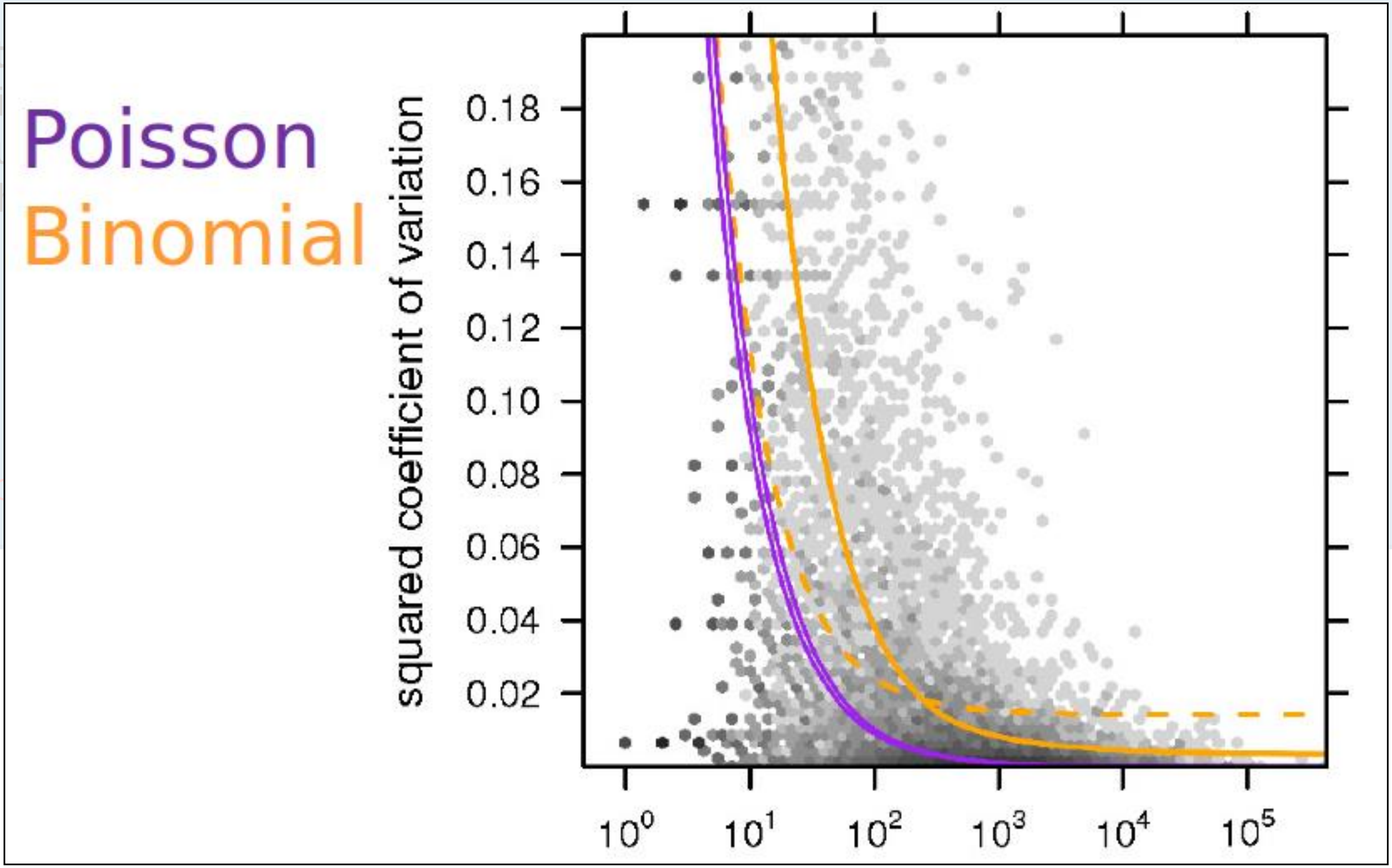


Statistics Models

- The model

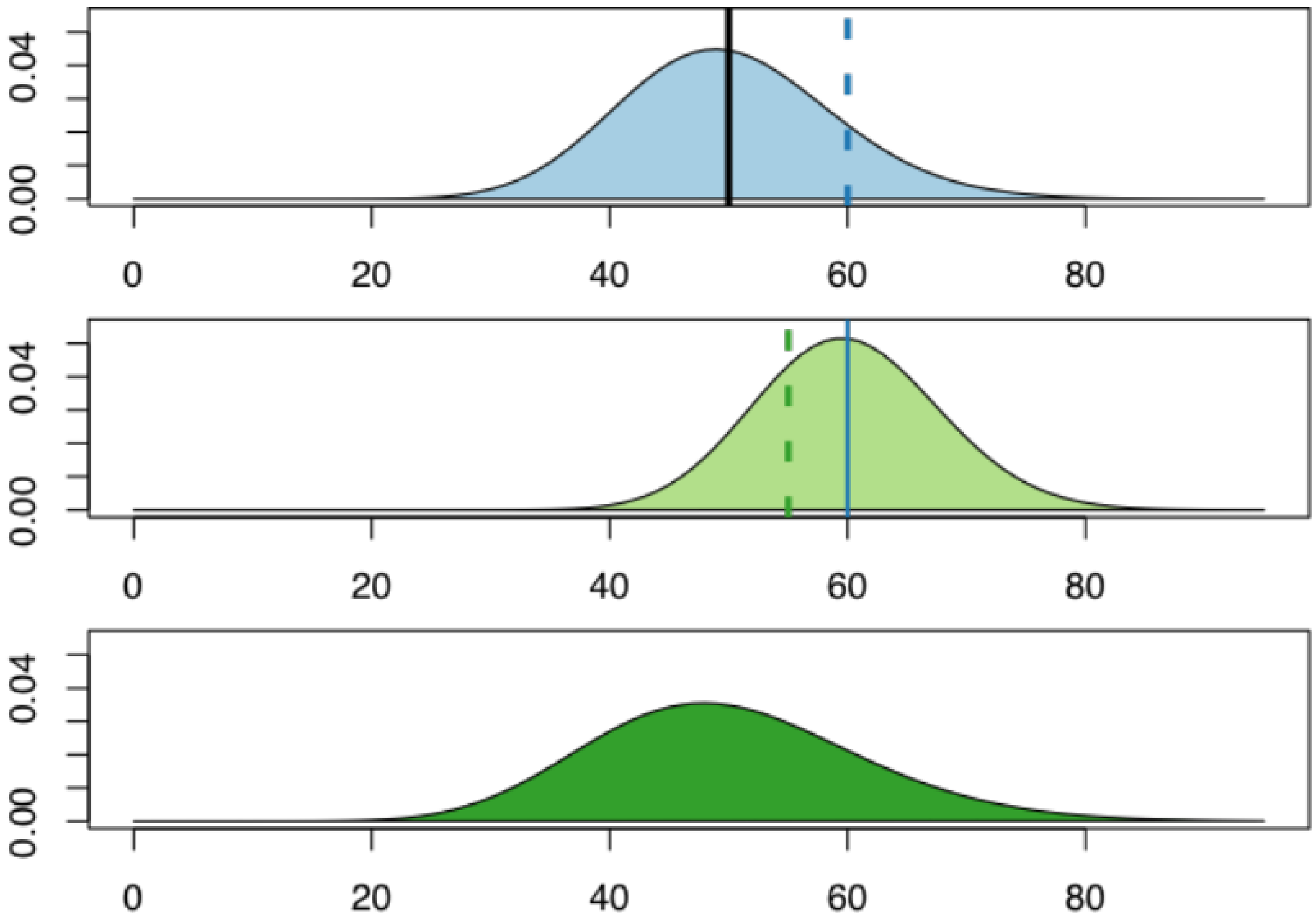
Negat

- Moti
- Assu
- Metl
- Prob



Statistics Models

- The model



Biological sample with mean μ and variance *v*

Poisson distribution with mean *q* and variance *q*.

Negative binomial with mean *μ* and variance *q*+*v*.

Statistics Models

- The model

Negative Binomial (NB): edgeR and DESeq

- Motivation: distribution takes into account Overdispersion
- Assumption:
- Method: NB is a two-parameter distribution

    Origin: $Y \sim NB(p, m)$

    Y ... number of successes in a sequence of Bernoulli trials with probability p before r failures occur

    RNASeq case: λ (mean) and φ (overdispersion)

- Problem: φ_i / gene cannot be estimated due to the small number of individuals

http://smpgd2012.univ-lyon1.fr/IMG/pdf/NGS/gonzalez-smpgd2012.pdf

Statistics Softw

Unaffected by outiliers
New
Limma+Voom

Because
Bayesien
ROCKS
EBSeq

TOP PERFORMING METHODS FOR DATA SETS WITH LARGE SAMPLE SIZES
SAMSEQ

NBPSeq
2000

FOR EXON
DEXSEQ

100%
Garantee
DESeq

For more and more
genes
edgeR

ALWAYS TRUST THE ORIGINAL
BAYSEQ

SAMSeq
The number

Statistics Software

- In this paper, we have evaluated and compared eleven methods for differential expression analysis of RNA-seq data. Table 2 summarizes the main findings and observations. No single method among those evaluated here is optimal under all circumstances, and hence the method of choice in a particular situation depends on the experimental conditions. Among the methods evaluated in this paper, those based on a variance-stabilizing transformation combined with **limma** (i.e., voom+limma and vst+limma) performed well under many conditions, were relatively unaffected by outliers and were computationally fast, but they required at least 3 samples per condition to have sufficient power to detect any differentially expressed genes. As shown in the supplementary material (Additional file 1), they also performed worse when the dispersion differed between the two conditions. The non-parametric **SAMseq**, which was among the top performing methods for data sets with large sample sizes, required at least 4-5 samples per condition to have sufficient power to find DE genes. For highly expressed genes, the fold change required for statistical significance by **SAMseq** was lower than for many other methods, which can potentially compromise the biological significance of some of the statistically significantly DE genes. The same was true for **ShrinkSeq**, which however has an option for imposing a fold change requirement in the inference procedure.
- Small sample sizes (2 samples per condition) imposed problems also for the methods that were indeed able to find differentially expressed genes, there leading to false discovery rates sometimes widely exceeding the desired threshold implied by the FDR cutoff. For the parametric methods this may be due to inaccuracies in the estimation of the mean and dispersion parameters. In our study, **TSPM** stood out as the method being most affected by the sample size, potentially due to the use of asymptotic statistics. Even though the development goes towards large sample sizes, and barcoding and multiplexing create opportunities to analyze more samples at a fixed cost, as of today RNA-seq experiments are often too expensive to allow extensive replication. The results conveyed in this study strongly suggest that the differentially expressed genes found between small collections of samples need to be interpreted with caution and that the true FDR may be several times higher than the selected FDR threshold.
- **DESeq, edgeR and NBPSeq** are based on similar principles and showed, overall, relatively similar accuracy with respect to gene ranking. However, the sets of significantly differentially expressed genes at a pre-specified FDR threshold varied considerably between the methods, due to the different ways of estimating the dispersion parameters. With default settings and for reasonably large sample sizes, **DESeq** was often overly conservative, while **edgeR** and in particular **NBPSeq** often were too liberal and called a larger number of false (and true) DE genes. In the supplementary material (Additional file 1) we show that varying the parameters of **edgeR** and **DESeq** can have large effects on the results of the differential expression analysis, both in terms of the ability to control type I error rates and false discovery rates and in terms of the ability to detect the truly DE genes. These results also show that the recommended parameters (that are used in the main paper) are indeed well chosen and often provide the best results.
- **EBSeq, baySeq and ShrinkSeq** use a different inferential approach, and estimate the posterior probability of being differentially expressed, for each gene. **baySeq** performed well under some conditions but the results were highly variable, especially when all DE genes were upregulated in one condition compared to the other. In the presence of outliers, **EBSeq** found a lower fraction of false positives than **baySeq** for large sample sizes, while the opposite was true for small sample sizes.

A comparison of methods for differential expression analysis of RNA-seq data. Charlotte Soneson1 and Mauro Delorenzi
BMC Bioinformatics 2013, 14:91 doi:10.1186/1471-2105-14-91

Statistics Software

- **limma** (i.e., voom+limma and vst+limma)
  - unaffected by outliers
  - but they required at least 3 samples per condition
- **SAMseq, ShrinkSeq** (The non-parametric)
  - top performing methods for data sets with large sample sizes
  - required at least 4-5 samples per condition
  - fold change required for statistical significance was lower → compromise the biological significance
  - Small sample sizes inaccuracies in the estimation of the mean and dispersion parameters
- **TSPM**
  - most affected by the sample size
- **DESeq, edgeR and NBPSeq**
  - showed, overall, relatively similar accuracy with respect to gene ranking
  - recommended parameters well chosen and often provide the best results
  - pre-specified FDR threshold varied considerably between the methods
  - **DESeq** : overly conservative
  - **edgeR**, **NBPSeq** : too liberal and called a larger number of false (and true) DE genes.
  - **edgeR**, **DESeq :** varying the parameters of can have large effects on the results
- **EBSeq, baySeq and ShrinkSeq** (posterior probability)
  - **baySeq** performed well under some conditions ; results were highly variable, especially when all DE genes were upregulated in one condition
  - **EBSeq** In the presence of outliers, found a lower fraction of false positives for large sample sizes not fot small sample sizes
  - **baySeq** In the presence of outliers, found a lower fraction of false positives true for small sample sizes not fot large sample sizes

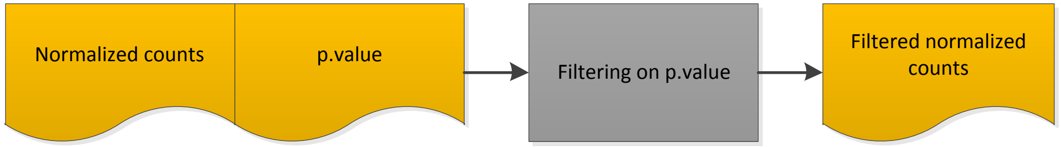
A comparison of methods for differential expression analysis of RNA-seq data. Charlotte Soneson1 and Mauro Delorenzi
BMC Bioinformatics 2013, 14:91 doi:10.1186/1471-2105-14-91

Statistics Model

- Modes
  - 2 conditions:
    - between two (t-test like)
    - between more groups: (pairwise.t-test like)
  - N conditions – Multivariate analysis: generalized linear models (GLMs) (Anova-like)
    - 1 factor
    - 2 factors
    - N factors

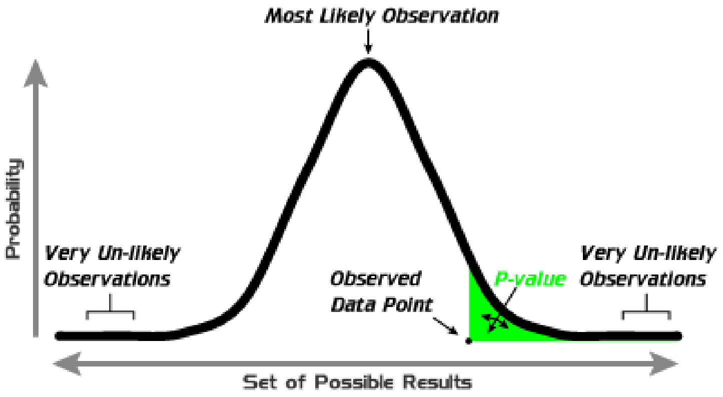
    Ex: ~Treatment+Time+Batch

Statistics Selection



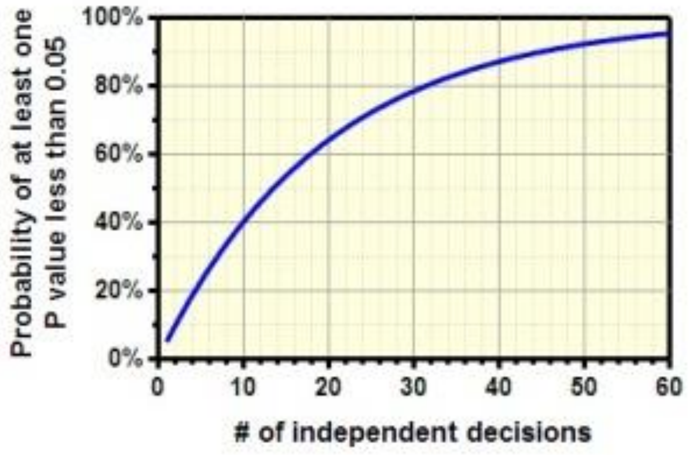
Statistics Selection

- The results
  - p.value
    - The *p*-value of the test statistic is a way of saying how extreme that statistic is for our sample data. The smaller the *p*-value, the more unlikely the observed sample.



A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

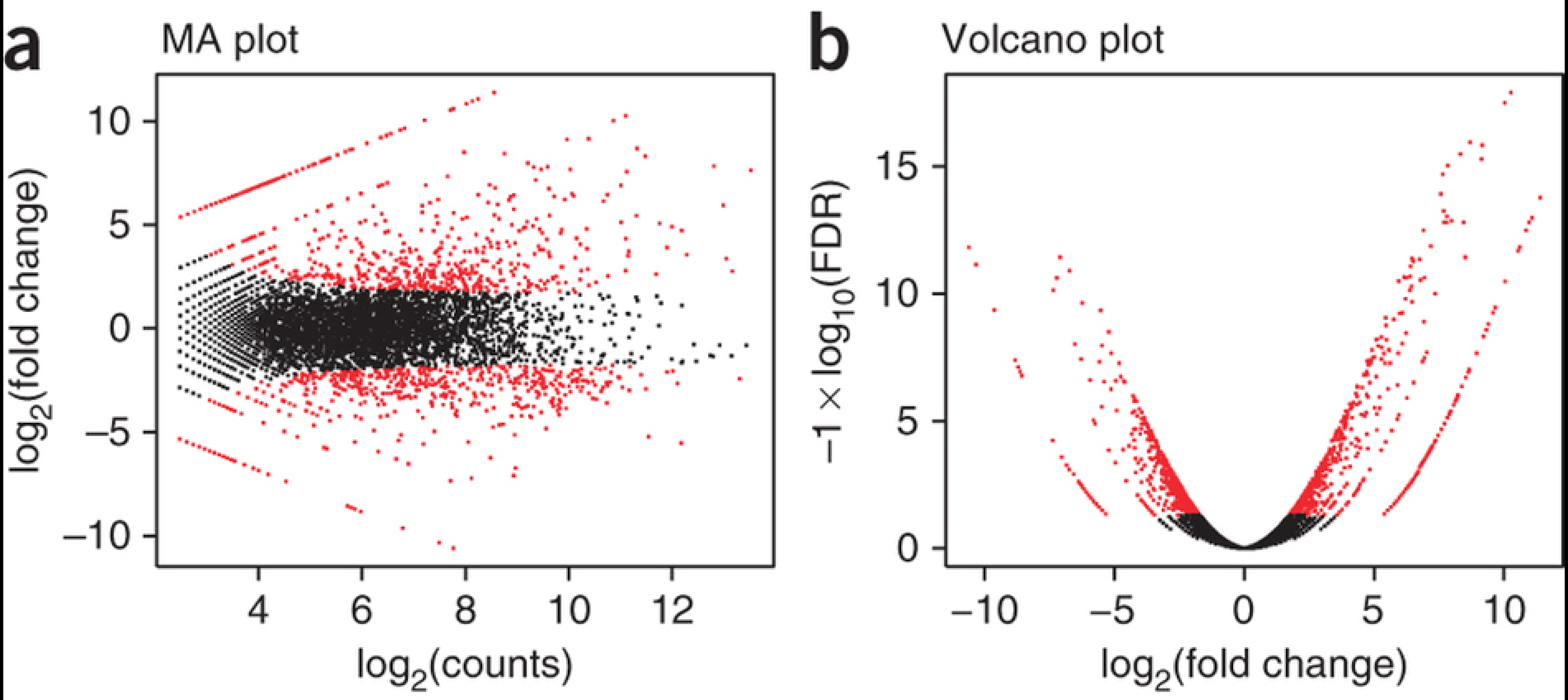
  - adjusted p.value / False Discovery Rate
    - Used in multiple hypothesis testing
    - Corrections
      - Bonferroni
      - Benjamini-Hochberg (BH)



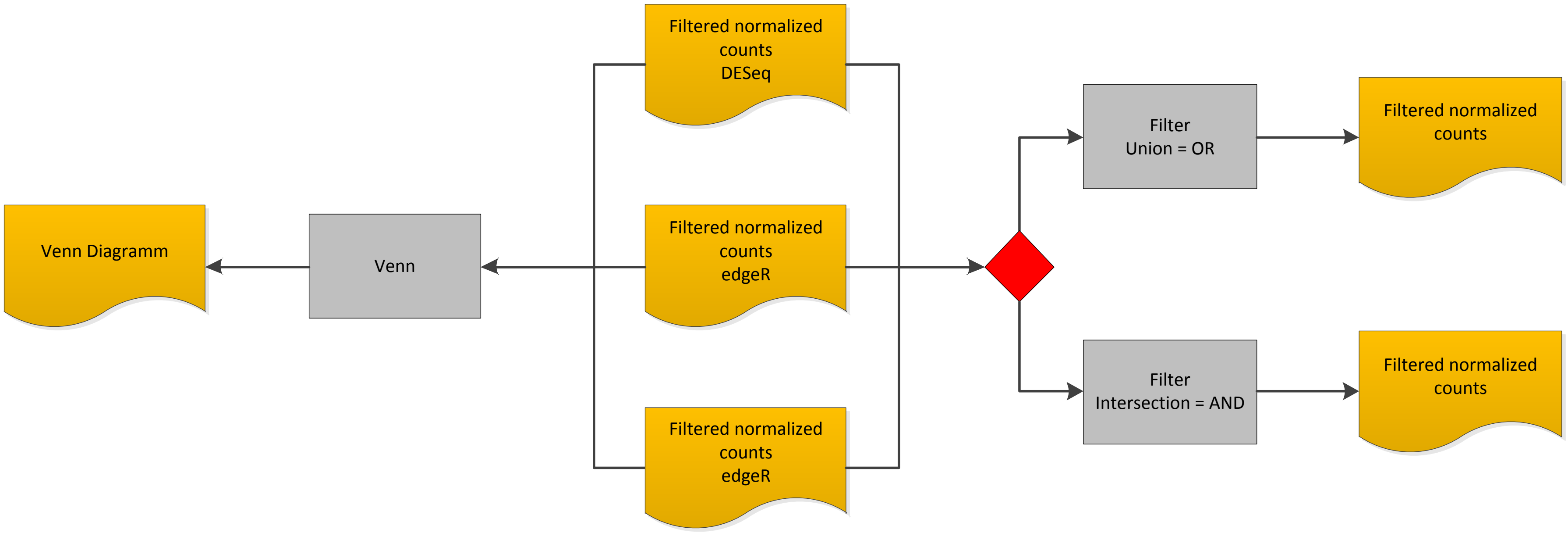
Statistics Selection

- Filtering
  - alpha threshold
    - The number alpha is the threshold value that we measure *p*-values against. It tells us how extreme observed results must be in order to reject the null hypothesis of a significance test.
    - Must be set in advance !

    - Ex:
      - For results with a 90% level of confidence, the value of alpha is 1 - 0.90 = 0.10.
      - For results with a 95% level of confidence, the value of alpha is 1 - 0.95 = 0.05.
      - For results with a 99% level of confidence, the value of alpha is 1 - 0.99 = 0.01.
    - So:
      - alpha > pvalue → H0 is rejected → 👍

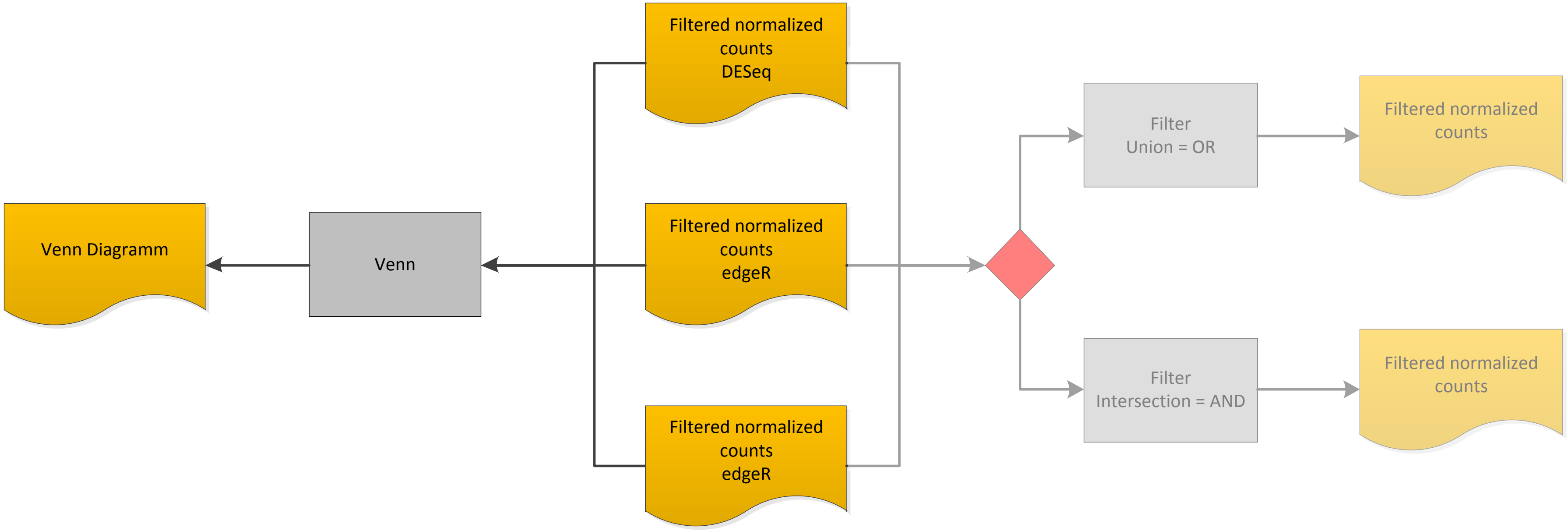

a
MA plot
log$_2$(fold change)
10
5
0
–5
–10
4
6
8
10
12
log$_2$(counts)
b
Volcano plot
–1 × log$_{10}$(FDR)
15
10
5
0
–10
–5
0
5
10
log$_2$(fold change)

Statistics Selection



Filtered normalized counts DESeq
Filtered normalized counts edgeR
Filtered normalized counts edgeR
Venn
Venn Diagramm
Filter Union = OR
Filter Intersection = AND
Filtered normalized counts
Filtered normalized counts

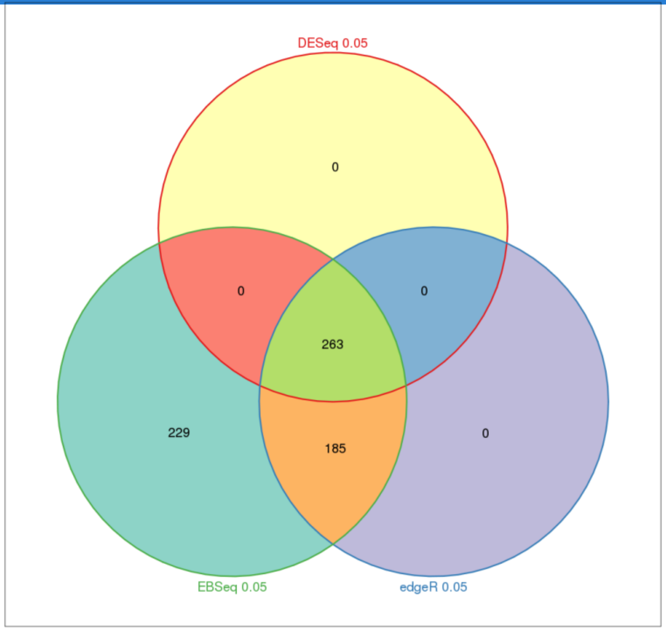
Statistics Selection

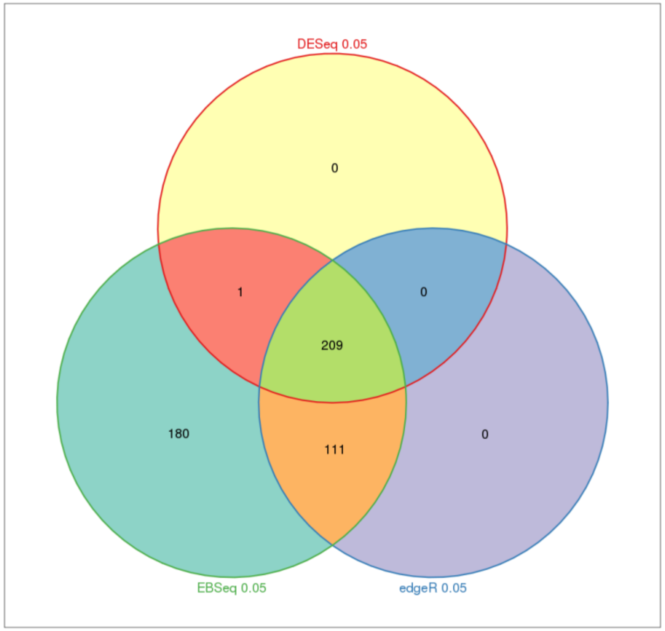


Venn Diagramm
Venn
Filtered normalized counts DESeq
Filtered normalized counts edgeR
Filtered normalized counts edgeR
Filter Union = OR
Filter Intersection = AND
Filtered normalized counts
Filtered normalized counts

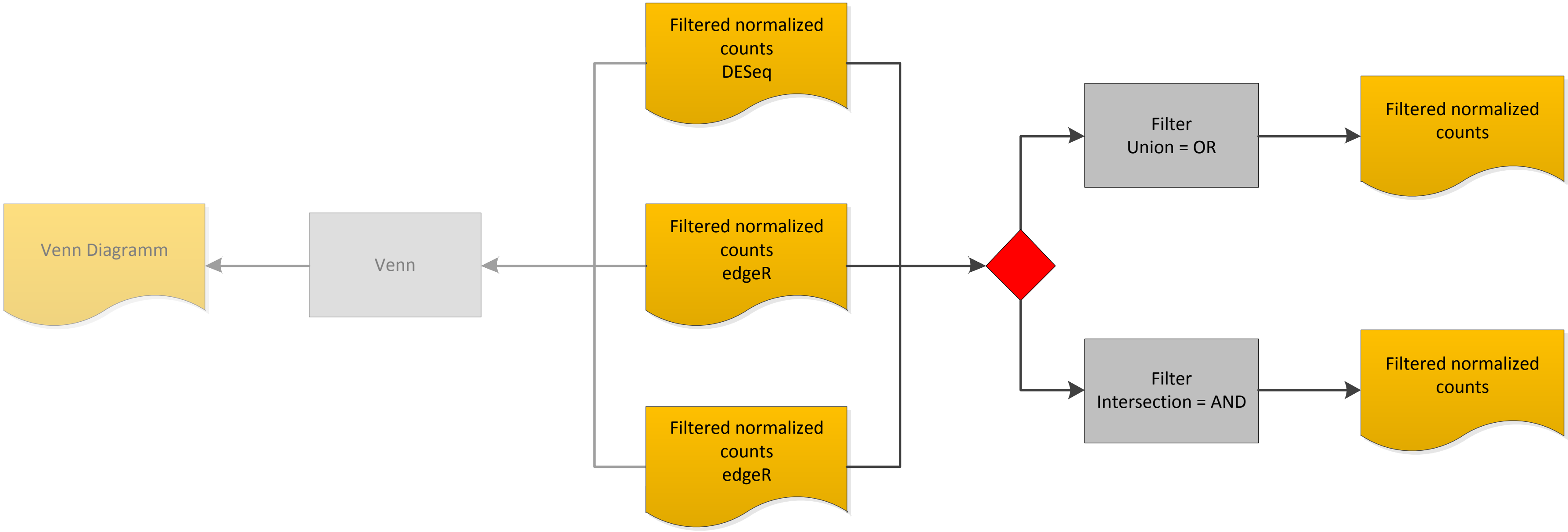
Statistics Selection

- Conciliation
  - Venn
    - Stringency or Liberal ?
    - Intersection or Union ?





Statistics Selection


Filtered normalized counts DESeq
Filtered normalized counts edgeR
Filtered normalized counts edgeR
Venn
Venn Diagramm
Filter Union = OR
Filter Intersection = AND
Filtered normalized counts
Filtered normalized counts

Statistics Selection

- Filtering
  - Intersection

    (DESeq <= 0.005 and edgeR <= 0.005)

    =

    (c25 <= 0.005 and c25 <= 0.005)

  - Union

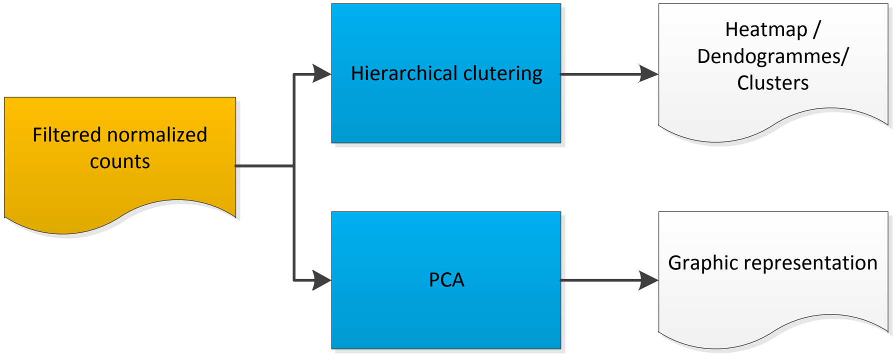
    (DESeq <= 0.005 or edgeR <= 0.005)

    =

    (c25 <= 0.005 or c25 <= 0.005)

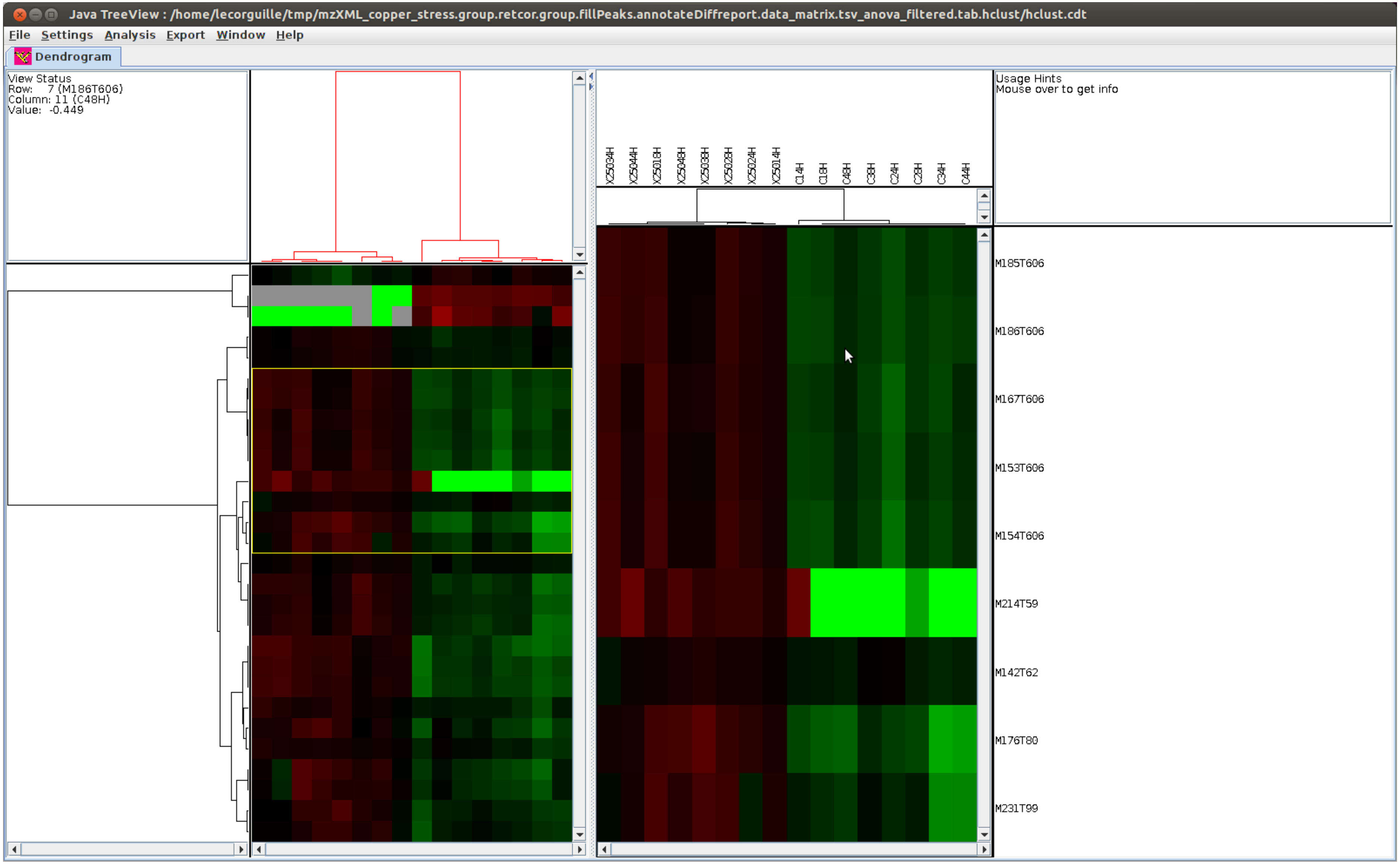
POST-ANALYSIS

Post-analysis



Post-analysis

- Hierarchical Ascendant Clustering (HAC)



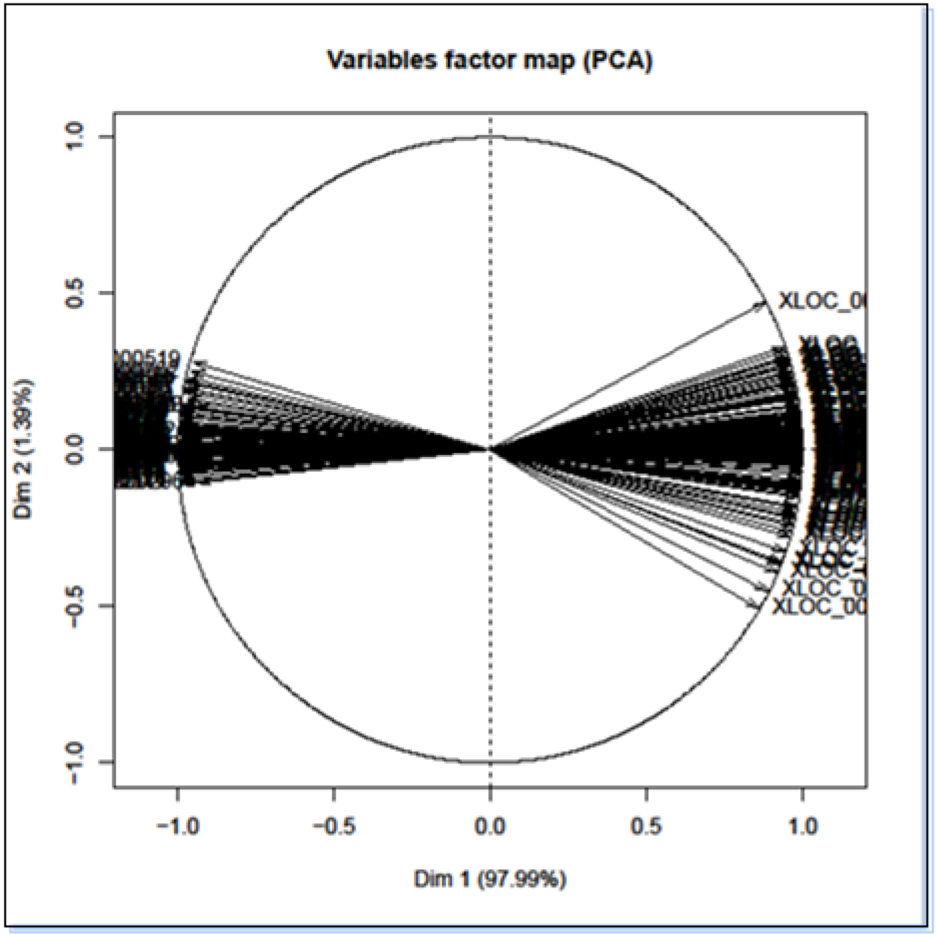
Java TreeView : /home/lecorguille/tmp/mzXML_copper_stress.group.retcor.group.fillPeaks.annotateDiffreport.data_matrix.tsv_anova_filtered.tab.hclust/hclust.cdt

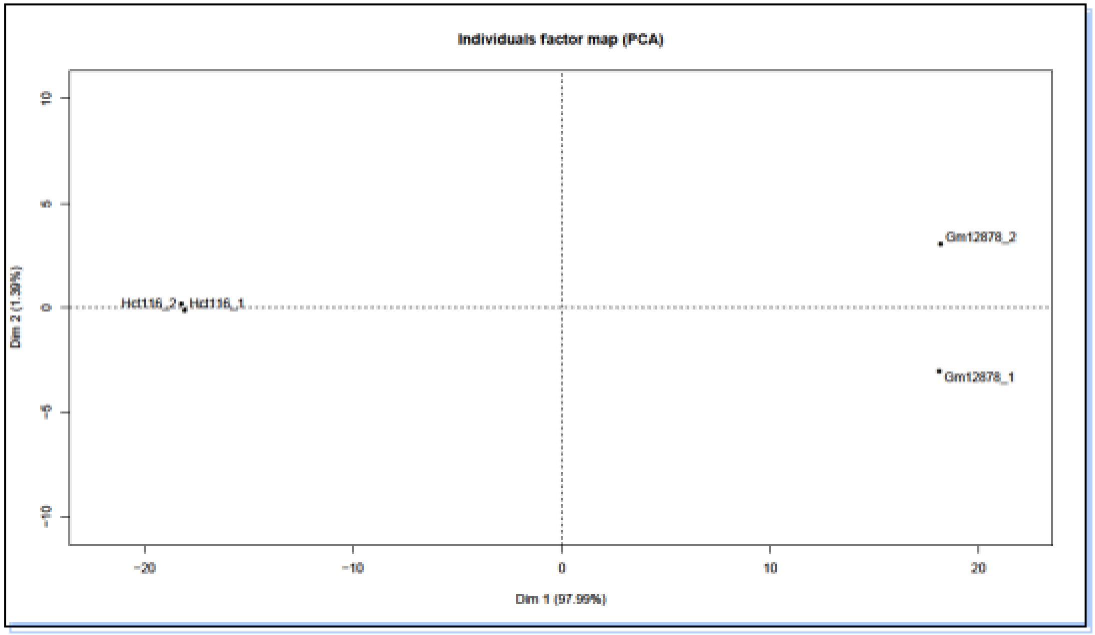
View Status
Row: 7 (M186T606)
Column: 11 (C48H)
Value: -0.449

Usage Hints
Mouse over to get info

Post-analysis

- Principal component analysis (PCA)





TP

TP

- Inputs

	Gm12878_1	Gm12878_2	Hct116_1	Hct116_2
NM_001003891	86	98	140	139
NM_033200	1379	1639	3499	3583
NM_152513	523	589	36	33
NM_015330	7	8	17	19
NR_046423	0	0	2	1

Tabular Merge (Count table)

Gm12878	Gm12878_1
Gm12878	Gm12878_2
Hct116	Hct116_1
Hct116	Hct116_2

Sample info

TP

- Step 1a: Run DE analysis

Merge output file	count_table.tab
Method	DESeq
Replicates	Yes
Sample file	sample_info.tab

- Step 1b: Run DE analysis

Merge output file	count_table.tab
Method	edgeR
Replicates	Yes
Sample file	sample_info.tab

TP

- Step 2a: Filter
  - **Filter** DESeq: Normalized counts used by DE method
  - **With following condition** c12<0.001

  → **renaming** DESeq_filtered_0.001

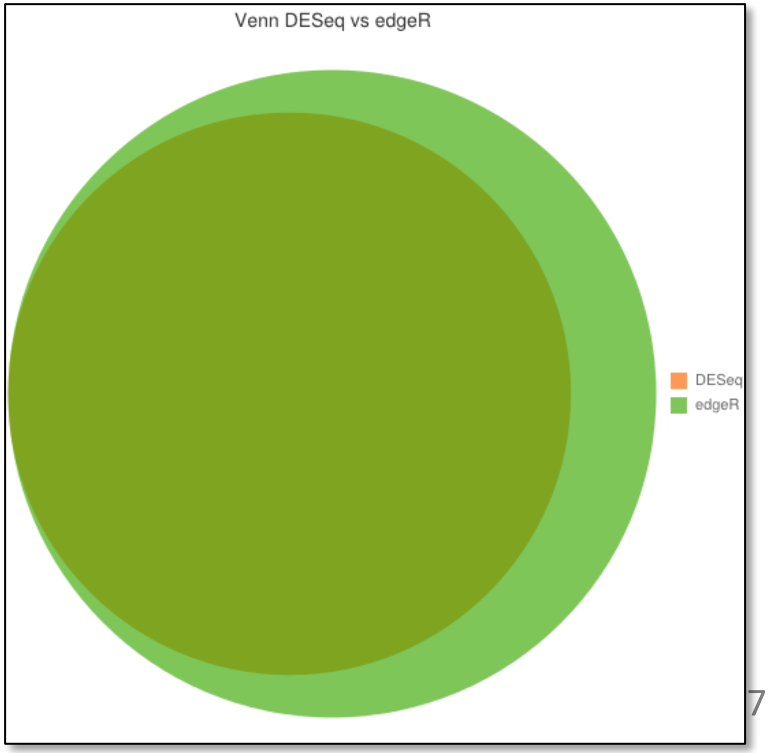
- Step 2b: Filter
  - **Filter** edgeR: Normalized counts used by DE method
  - **With following condition** c9<0.001

  → **renaming** edgeR_filtered_0.001

TP

- Step 3: proportional venn

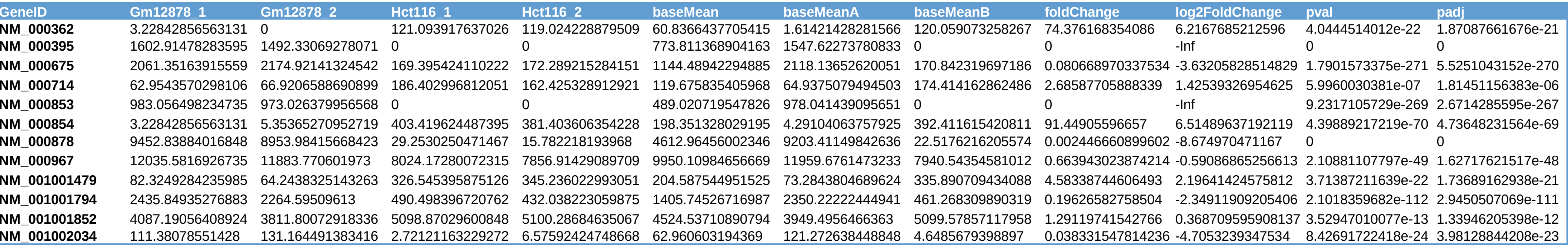
title	Venn DESeq vs edgeR
size	540
input file 1	DESeq_filtered_0.001
column index	0
as name	DESeq
input file 2	edgeR_filtered_0.001
column index file 2	0
as name file 2	edgeR



TP

- Step 4: Compare two Datasets

• **Header**	Yes
• **Compare**	DESeq_filtered_0.001
• **Using column**	c1
• **again**	edgeR_filtered_0.001
• **and column**	c1
→ **renaming**	DESeq_edgeR_0.001_intersect

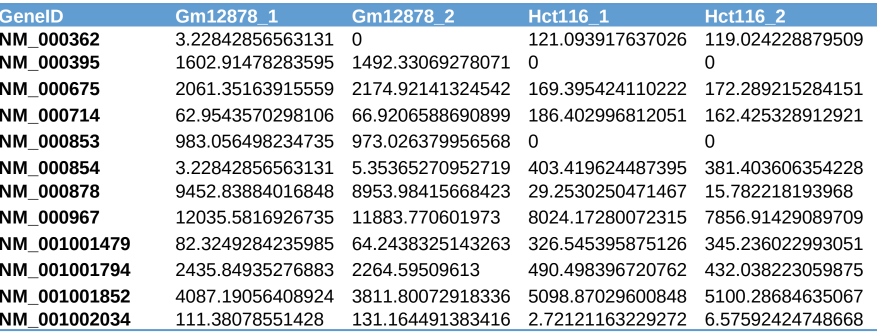


GeneID	Gm12878_1	Gm12878_2	Hct116_1	Hct116_2	baseMean	baseMeanA	baseMeanB	foldChange	log2FoldChange	pval	padj
NM_000362	3.22842856563131	0	121.093917637026	119.024228879509	60.8366437705415	1.61421428281566	120.059073258267	74.376168354086	6.2167685212596	4.0444514012e-22	1.87087661676e-21
NM_000395	1602.91478283595	1492.33069278071	0	0	773.811368904163	1547.62273780833	0	0	-Inf	0	0
NM_000675	2061.35163915559	2174.92141324542	169.395424110222	172.289215284151	1144.48942294885	2118.13652620051	170.842319697186	0.080668970337534	-3.63205828514829	1.7901573375e-271	5.5251043152e-270
NM_000714	62.9543570298106	66.9206588690899	186.402996812051	162.425328912921	119.675835405968	64.9375079494503	174.414162862486	2.68587705888339	1.42539326954625	5.9960030381e-07	1.81451156383e-06
NM_000853	983.056498234735	973.026379956568	0	0	489.020719547826	978.041439095651	0	0	-Inf	9.2317105729e-269	2.6714285595e-267
NM_000854	3.22842856563131	5.35365270952719	403.419624487395	381.403606354228	198.351328029195	4.29104063757925	392.411615420811	91.44905596657	6.51489637192119	4.39889217219e-70	4.73648231564e-69
NM_000878	9452.83884016848	8953.98415668423	29.2530250471467	15.782218193968	4612.96456002346	9203.41149842636	22.5176216205574	0.002446660899602	-8.674970471167	0	0
NM_000967	12035.5816926735	11883.770601973	8024.17280072315	7856.91429089709	9950.10984656669	11959.6761473233	7940.54354581012	0.663943023874214	-0.59086865256613	2.10881107797e-49	1.62717621517e-48
NM_001001479	82.3249284235985	64.2438325143263	326.545395875126	345.236022993051	204.587544951525	73.2843804689624	335.890709434088	4.58338744606493	2.19641424575812	3.71387211639e-22	1.73689162938e-21
NM_001001794	2435.84935276883	2264.59509613	490.498396720762	432.038223059875	1405.74526716987	2350.22222444941	461.268309890319	0.19626582758504	-2.34911909205406	2.1018359682e-112	2.9450507069e-111
NM_001001852	4087.19056408924	3811.80072918336	5098.87029600848	5100.28684635067	4524.53710890794	3949.4956466363	5099.57857117958	1.29119741542766	0.368709595908137	3.52947010077e-13	1.33946205398e-12
NM_001002034	111.38078551428	131.164491383416	2.72121163229272	6.57592424748668	62.960603194369	121.272638448848	4.6485679398897	0.038331547814236	-4.7053239347534	8.42691722418e-24	3.98128844208e-23

TP

- Step 5: Cut

• **Cut columns**	Yes
• **Tab**	c1-c5
• **From**	DESeq_edgeR_0.001_intersect
→ **renaming**	DESeq_edgeR_0.001_intersect_dataMatrix

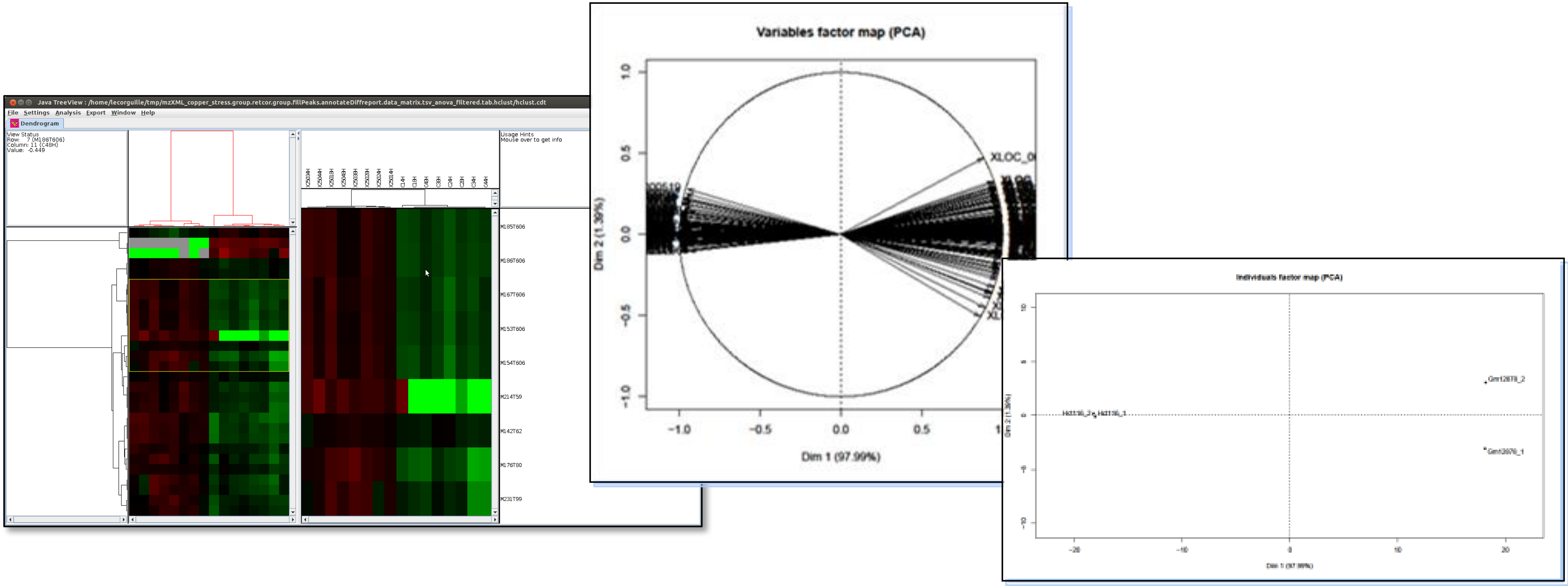


GeneID	Gm12878_1	Gm12878_2	Hct116_1	Hct116_2
NM_000362	3.22842856563131	0	121.093917637026	119.024228879509
NM_000395	1602.91478283595	1492.33069278071	0	0
NM_000675	2061.35163915559	2174.92141324542	169.395424110222	172.289215284151
NM_000714	62.9543570298106	66.9206588690899	186.402996812051	162.425328912921
NM_000853	983.056498234735	973.026379956568	0	0
NM_000854	3.22842856563131	5.35365270952719	403.419624487395	381.403606354228
NM_000878	9452.83884016848	8953.98415668423	29.2530250471467	15.782218193968
NM_000967	12035.5816926735	11883.770601973	8024.17280072315	7856.91429089709
NM_001001479	82.3249284235985	64.2438325143263	326.545395875126	345.236022993051
NM_001001794	2435.84935276883	2264.59509613	490.498396720762	432.038223059875
NM_001001852	4087.19056408924	3811.80072918336	5098.87029600848	5100.28684635067
NM_001002034	111.38078551428	131.164491383416	2.72121163229272	6.57592424748668

TP

- Step 6a: Hierarchical Clustering
  - **Data table file** DESeq_edgeR_0.001_intersect_dataMatrix

- Step 6b: PCA
  - **Data table file** DESeq_edgeR_0.001_intersect_dataMatrix



Java TreeView : /home/lecorguille/tmp/mzXML_copper_stress.group.retcor.group.fillPeaks.annotateDiffreport.data_matrix.tsv_anova_filtered.tab.hclust/hclust.cdt

View Status
Row: 7 (M186T606)
Column: 11 (C48H)
Value: -0.449

Usage Hints
Mouse over to get info