SDDRC

san diego digestive diseases research center

CCBB | CENTER FOR COMPUTATIONAL BIOLOGY & BIOINFORMATICS

RNA-Seq Analysis & Interpretation

Kathleen Fisch, Ph.D.

Executive Director, Center for Computational Biology & Bioinformatics,
University of California, San Diego, La Jolla, CA, USA

Email: Kfisch@ucsd.edu
Website: compbio.ucsd.edu

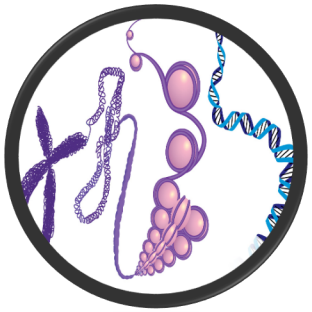
Outline

- **RNA-Seq Background**
  - Overview
  - Rationale & analysis goals
  - Library prep
  - Experimental design

- **RNA-Seq Analysis**
  - Overview
  - QC
  - Alignment
  - Gene & Transcript quantification
  - Normalization
  - Differential expression

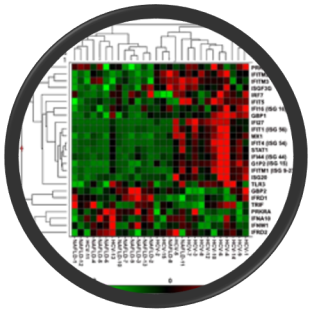
- **Downstream Analysis & Interpretation**
  - Hypergeometric test & Overrepresentation analysis
  - Functional enrichment analysis
  - Pathway analysis
  - Visualization with IGV
  - Network Analysis

UC San Diego
Center for Computational Biology & Bioinformatics

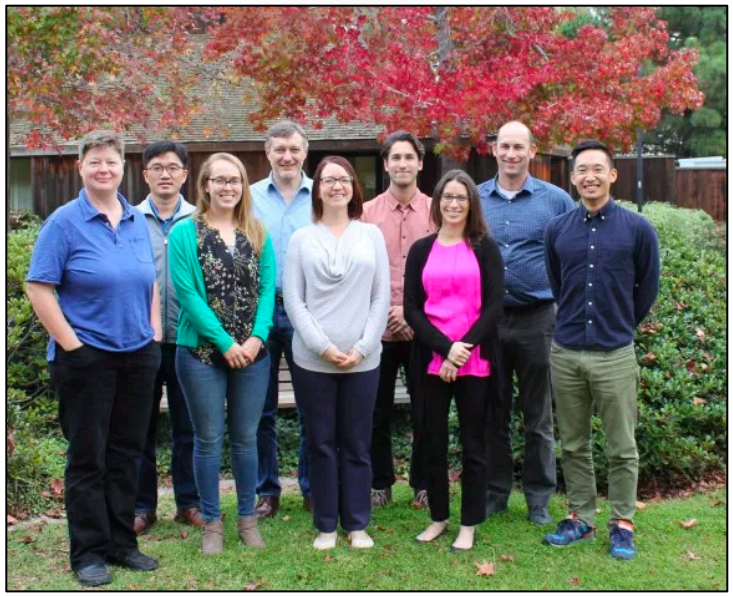


Epigenomics

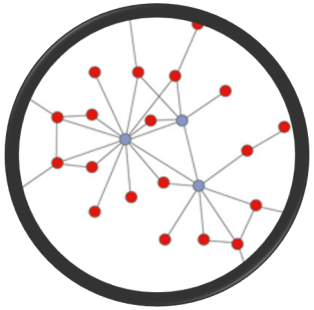
Whole Genome Analysis



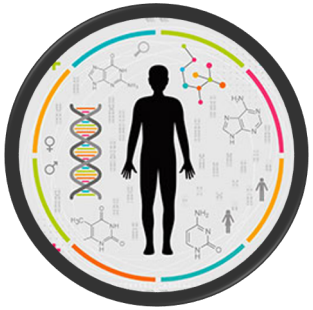
Gene Expression
& Regulation



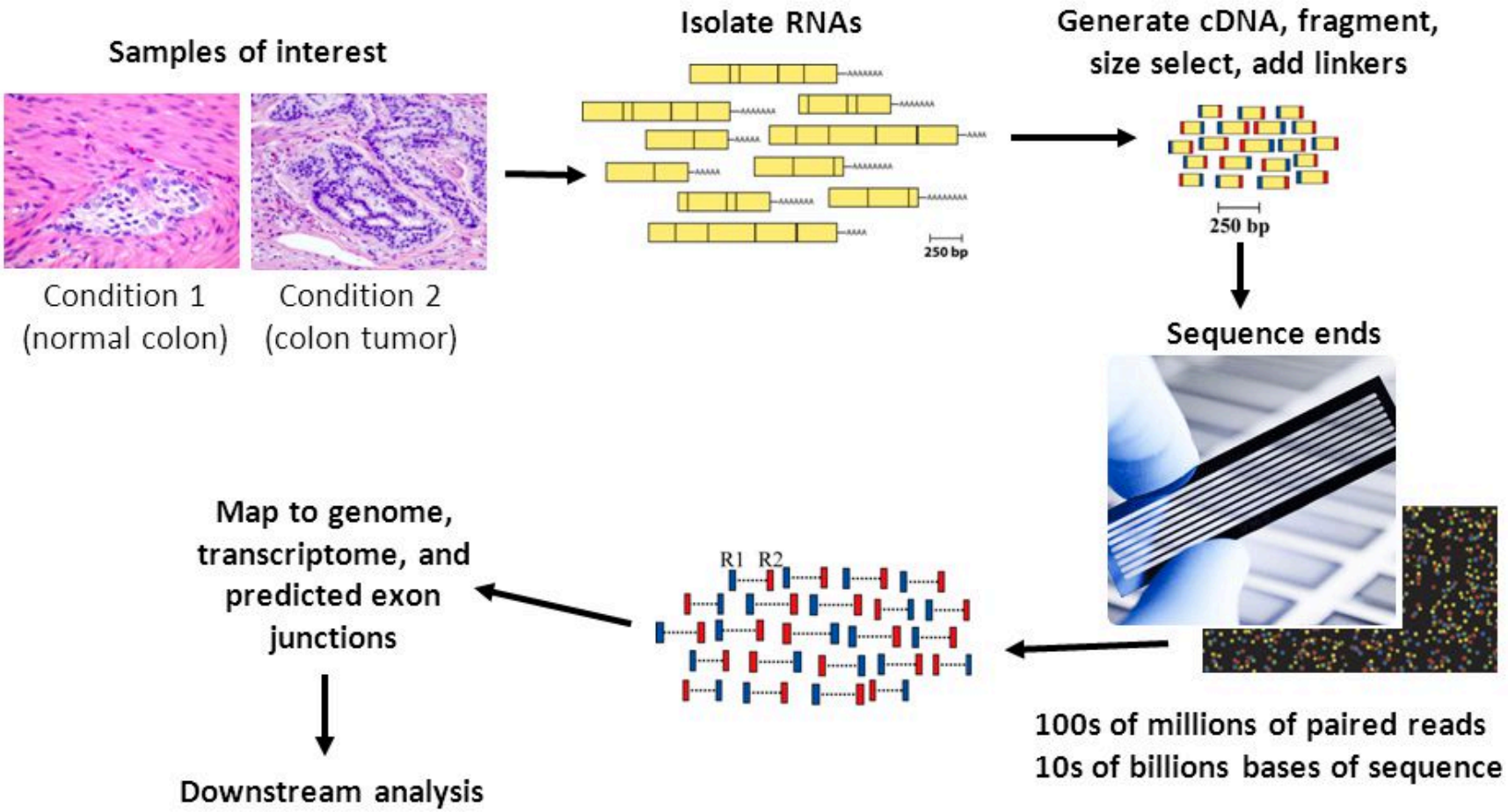
Networks &
Systems Biology



Biomarkers &
Therapeutic Targets

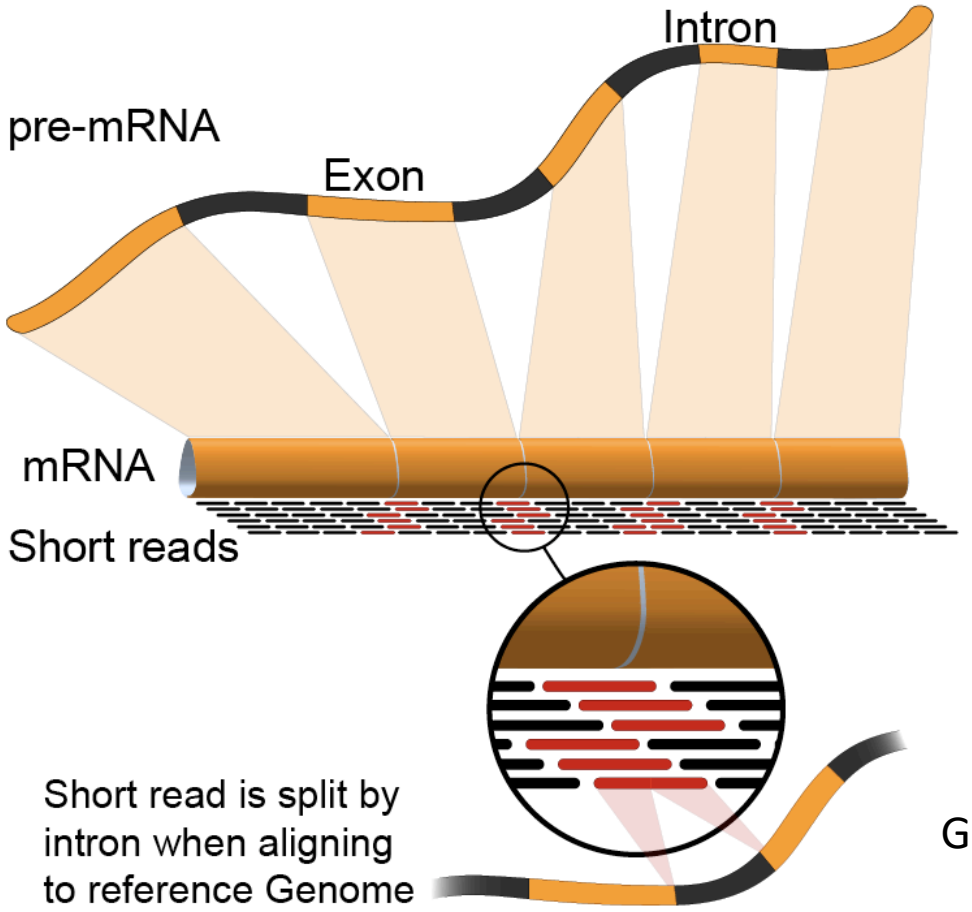


RNA-Seq Overview



Goldsby 2015 NESCent Academy

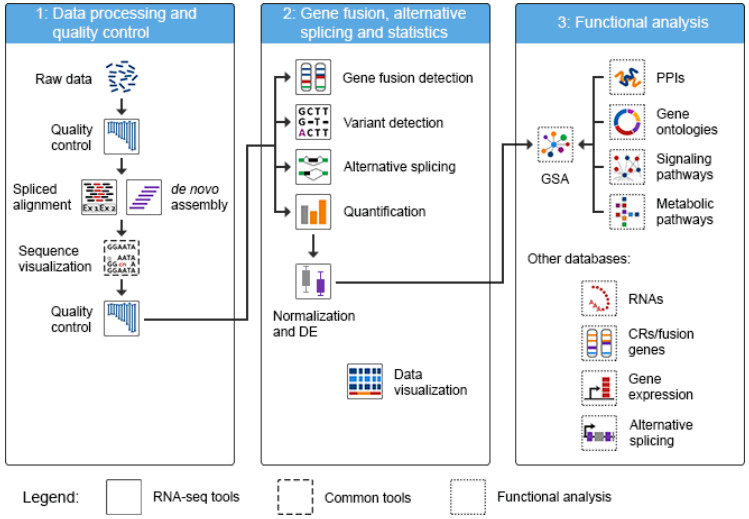
RNA sequencing Rationale



Griffith & Griffith 2013

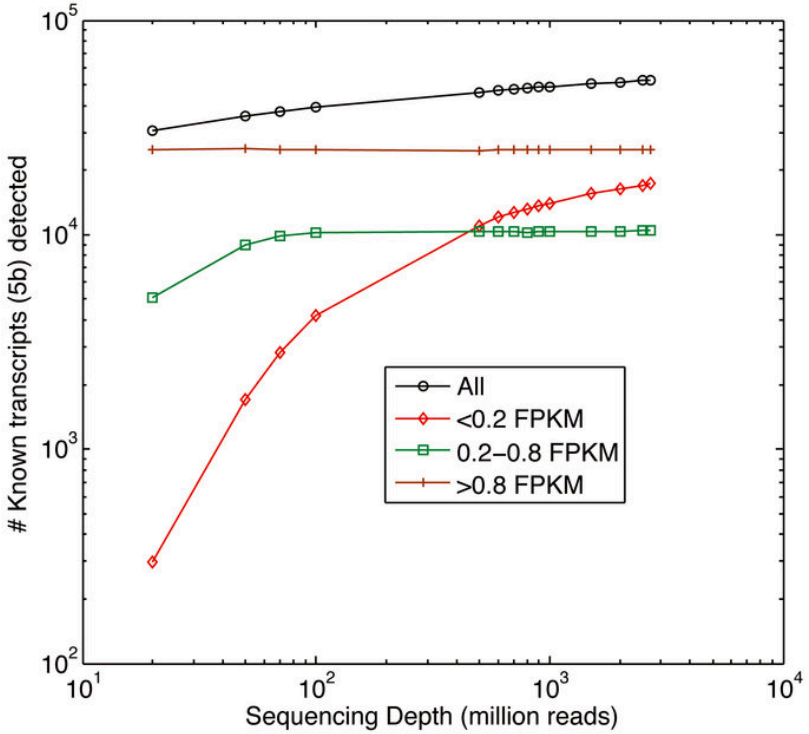
RNA-Seq Analysis Goals

- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- lncRNA
- Allele specific expression
- Mutation discovery
- Fusion detection
- RNA editing
- miRNAseq
- Single cell



RNA-Seq Experimental Design

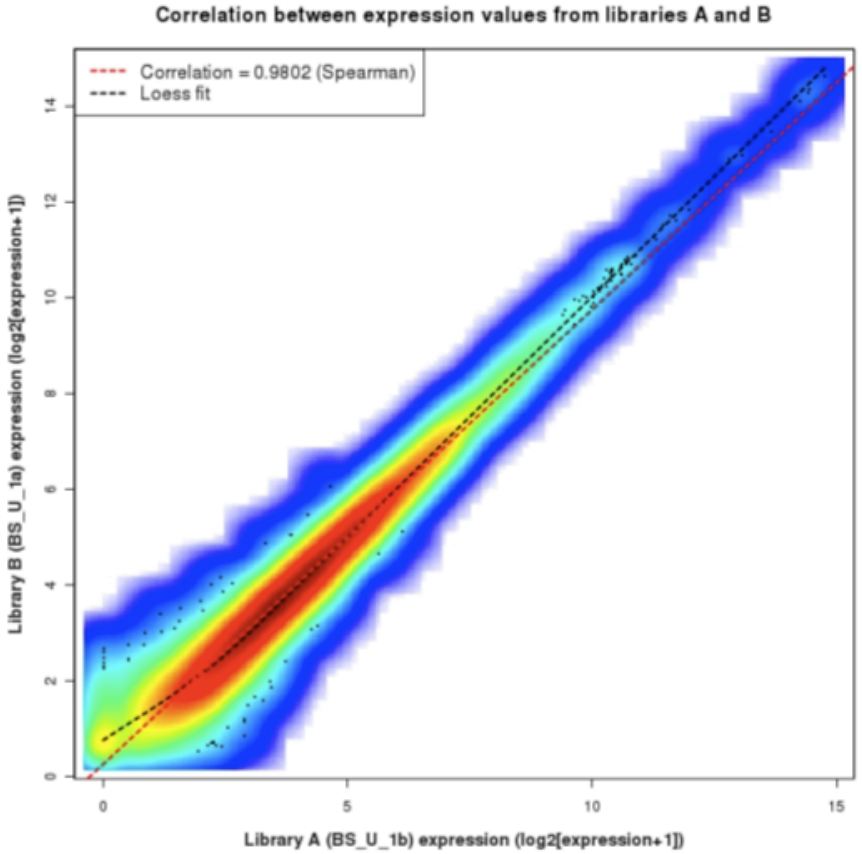
1. RNA extraction protocol
   - Poly(A) selection vs deplete rRNA
2. Stranded protocols
3. Single-end (SE) vs paired-end (PE) reads
4. Sequencing Depth aka library size
5. **Number of replicates**



Martin *et al*. 2014. *Scientific Reports*

RNA-Seq Experimental Design: Types of Replicates

- **Technical Replicate**
  - Multiple instances of sequence generation
    - Flow Cells, Lanes, Indexes
- **Biological Replicate**
  - Multiple isolations of cells showing the same phenotype, stage or other experimental condition
  - Some example concerns/challenges:
    - Environmental Factors, Growth Conditions, Time



Griffith & Griffith 2013

RNA-Seq Experimental Design: Number of Replicates

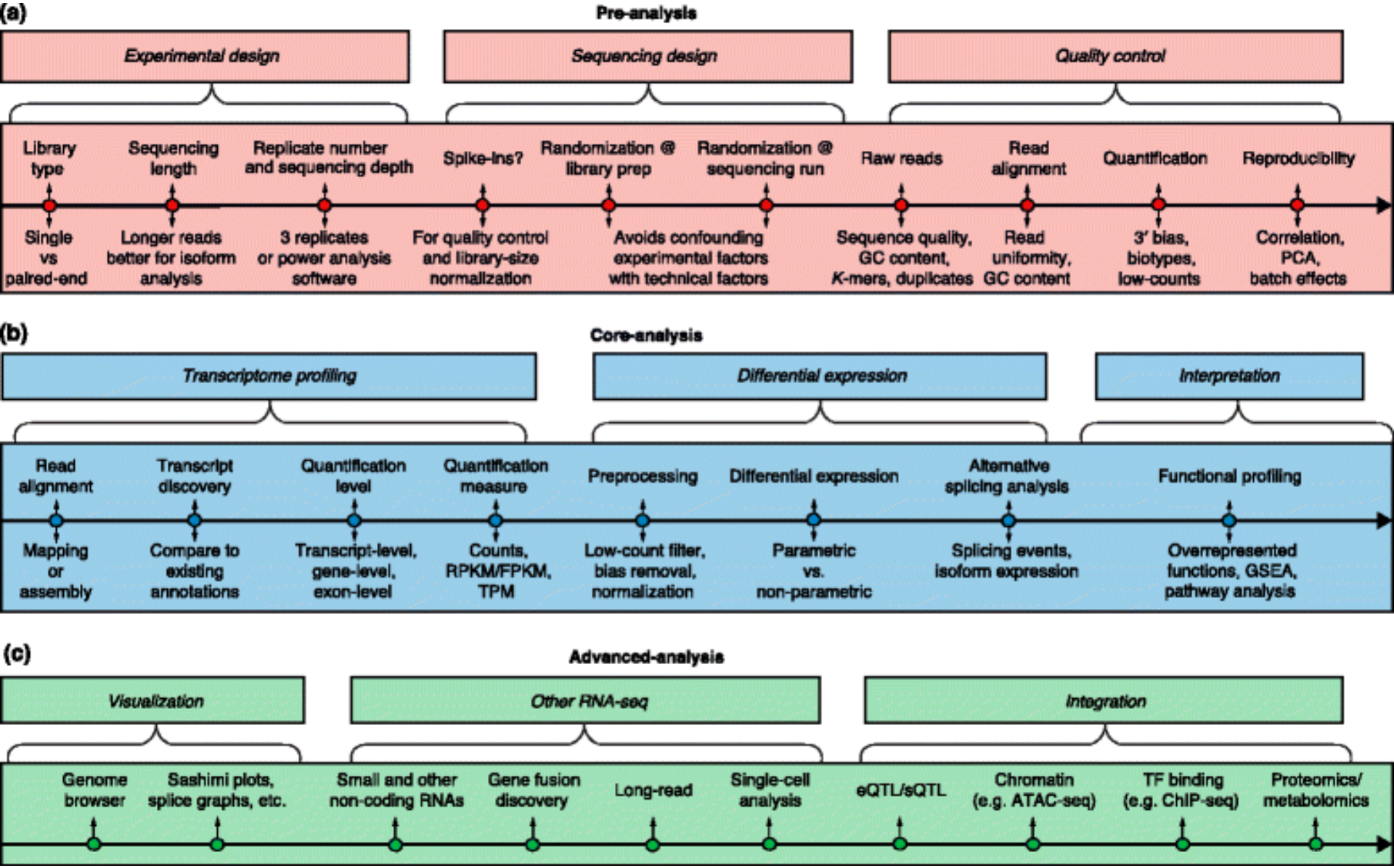
Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

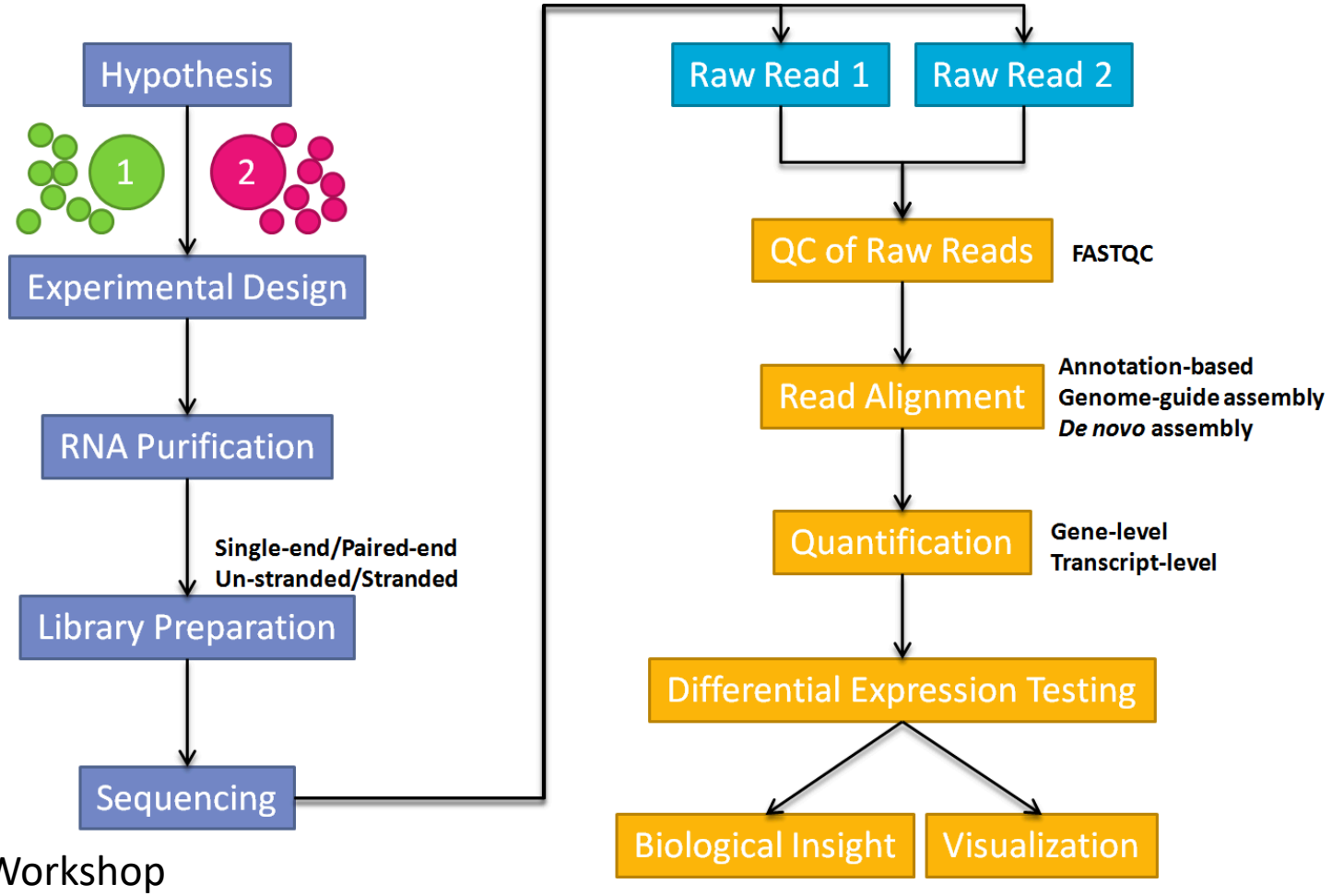
Example of calculations for the probability of detecting differential expression in a single test at a significance level of 5 %, for a two-group comparison using a Negative Binomial model, as computed by the RNASeqPower package of Hart et al. [190]. For a fixed within-group variance (package default value), the statistical power increases with the difference between the two groups (effect size), the sequencing depth, and the number of replicates per group. This table shows the statistical power for a gene with 70 aligned reads, which was the median coverage for a protein-coding gene for one whole-blood RNA-seq sample with 30 million aligned reads from the GTEx Project [214]

Conesa *et al.* 2016 *Genome Biology*

RNA-Seq Analysis Overview



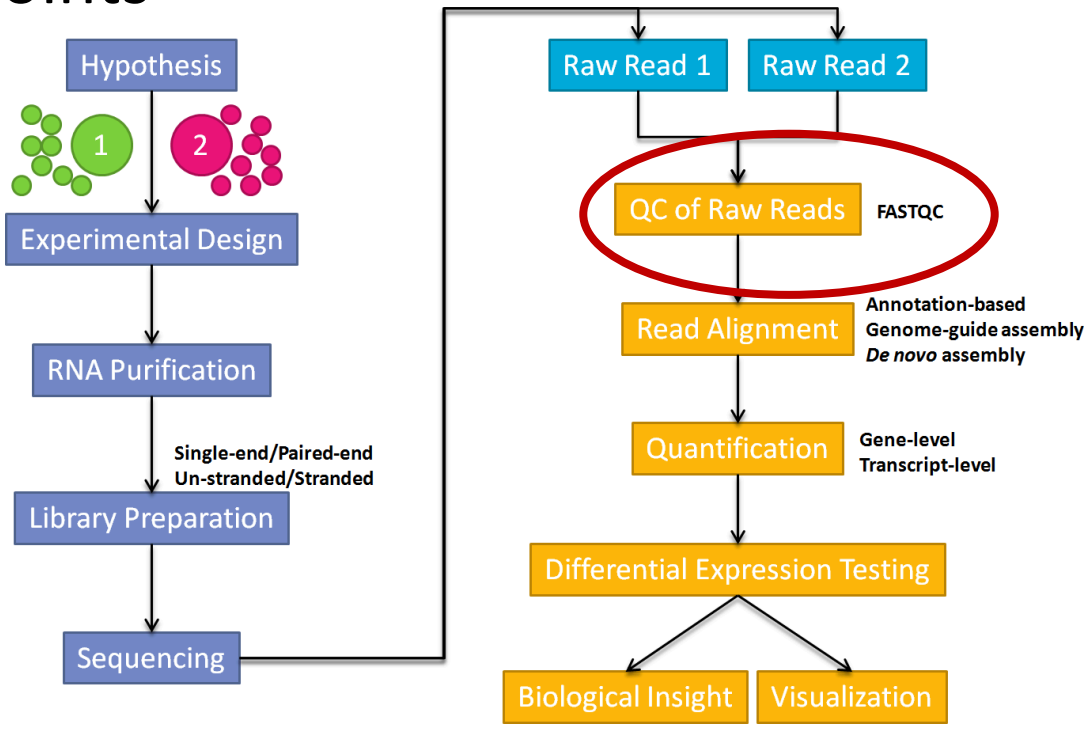
RNA-Seq Analysis Overview



Lin 2015 NGS Workshop

RNA-Seq Analysis QC

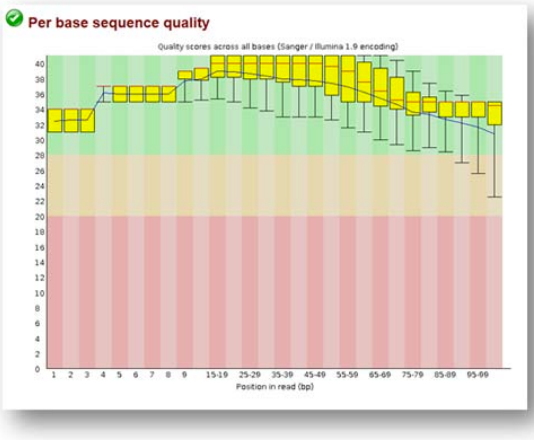
- Quality-control checkpoints
  - Raw reads
  - Read alignment
  - Quantification
  - Reproducibility



RNA-Seq Analysis QC

- Quality-control checkpoints
  - **Raw reads**
    - Sequence quality, GC content, presence of adaptors, overrepresented k-mer and duplicated reads
    - Tools: FASTQC
    - Goals
      - Detect sequencing errors
      - PCR artifacts or contaminations

Good Sequence Quality

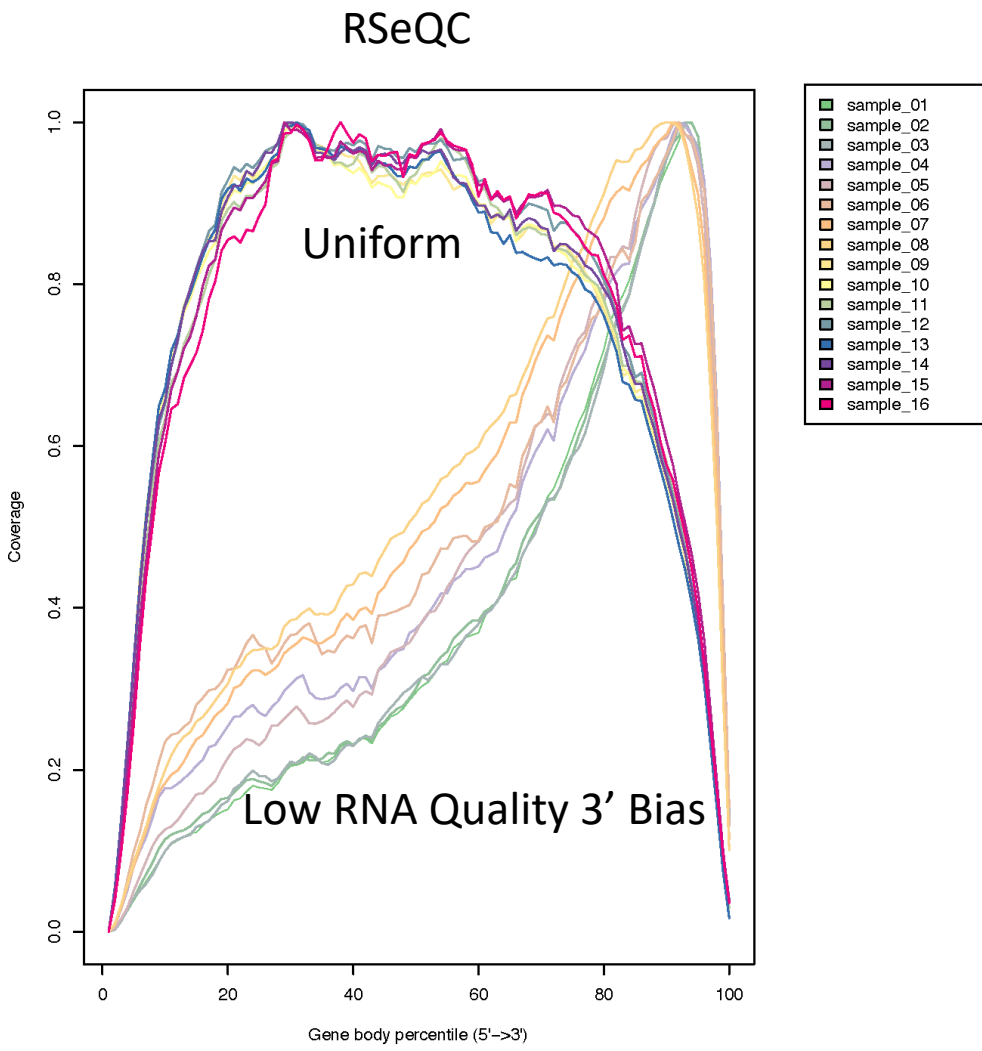


Poor Sequence Quality at 3' Ends

FASTQC Report

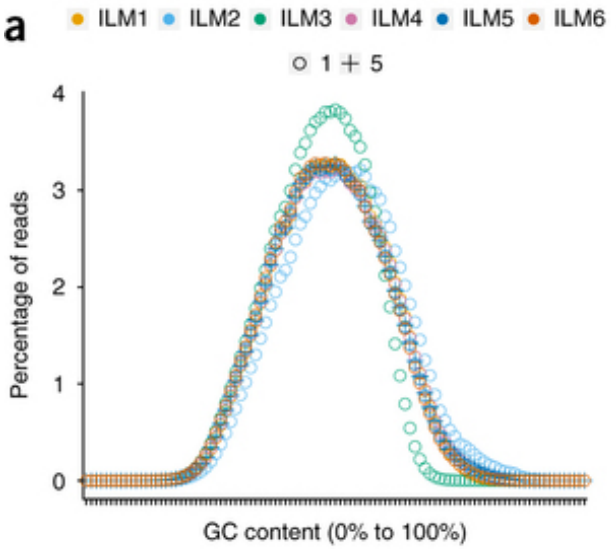
RNA-Seq Analysis QC

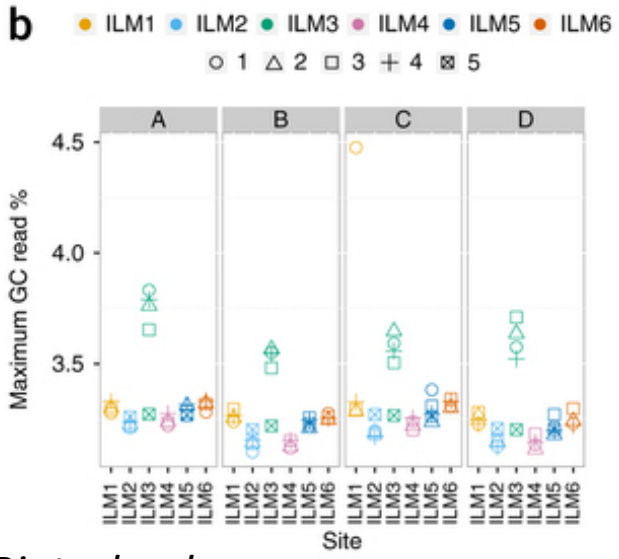
- Quality-control checkpoints
  - **Read alignment**
    - % Mapped reads
    - Uniformity of read coverage on exons and mapped strand
    - Ideal: 70-90% mapped to human genome
    - Tools: Picard, RSeQC, Qualimap
    - Goals:
      - Global indicator of overall sequencing accuracy and presence of contaminating DNA
      - Non-uniformity may indicate low RNA quality in starting material



RNA-Seq Analysis QC

- Quality-control checkpoints
  - **Quantification**
    - Check GC content and gene length bias
    - Tools: Bioconductor packages – NOISeq or EDA-Seq
    - Goal:
      - Apply correcting normalization methods, if necessary

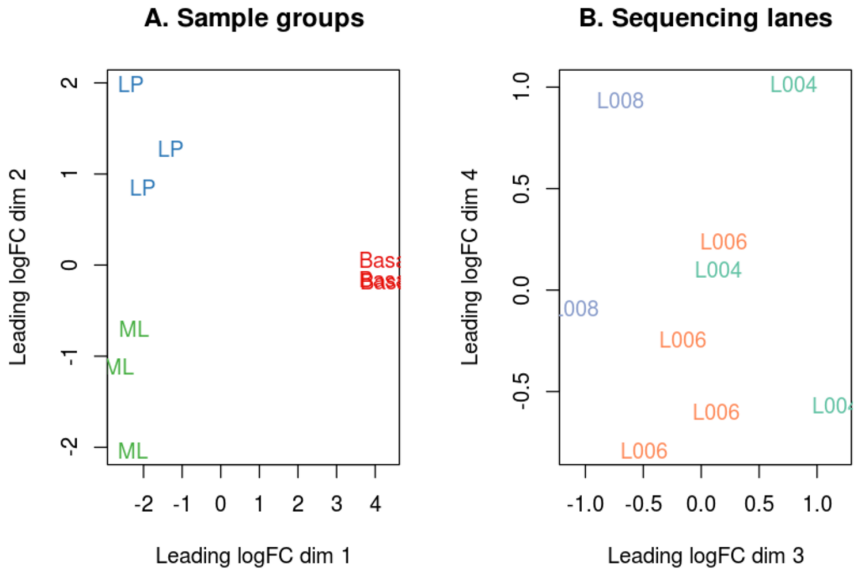




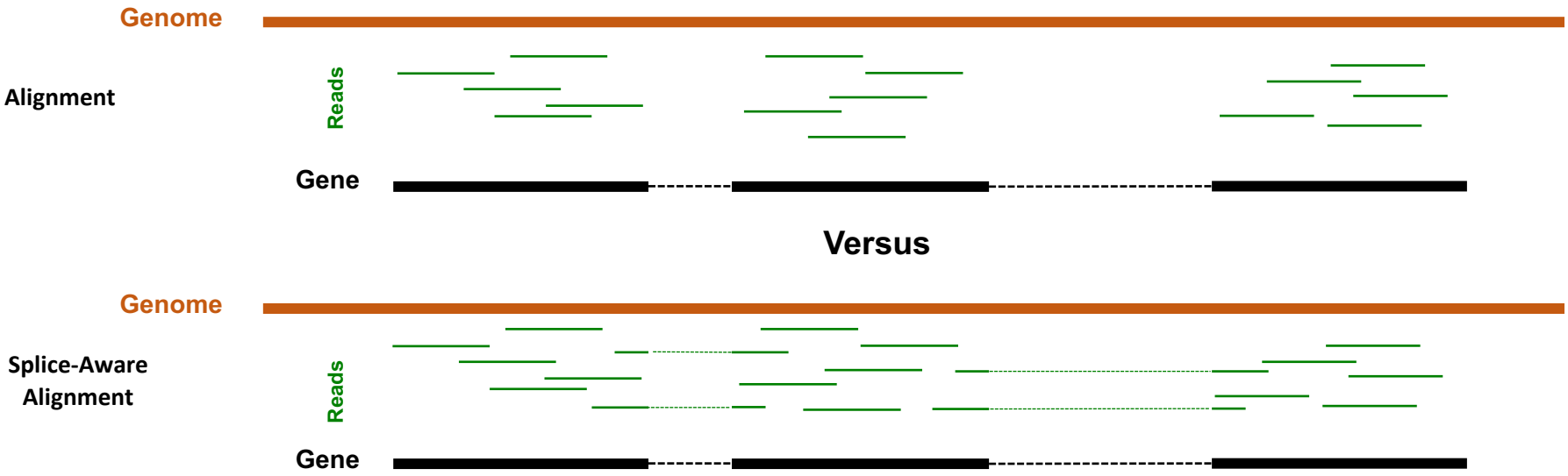
Li *et al*. 2014. *Nature Biotechnology*

RNA-Seq Analysis QC

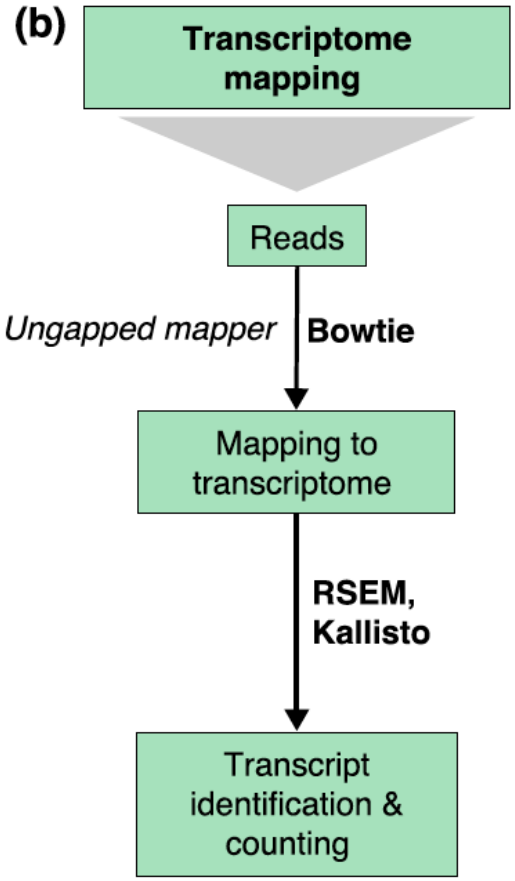
- Quality-control checkpoints
  - **Reproducibility**
    - Checking on reproducibility among replicates and for possible batch effects
    - Tool: Principal component analysis (PCA)/Multidimensional Scaling
    - Goal: Assess global quality of RNA-seq dataset



RNA-Seq Analysis -- Alignment

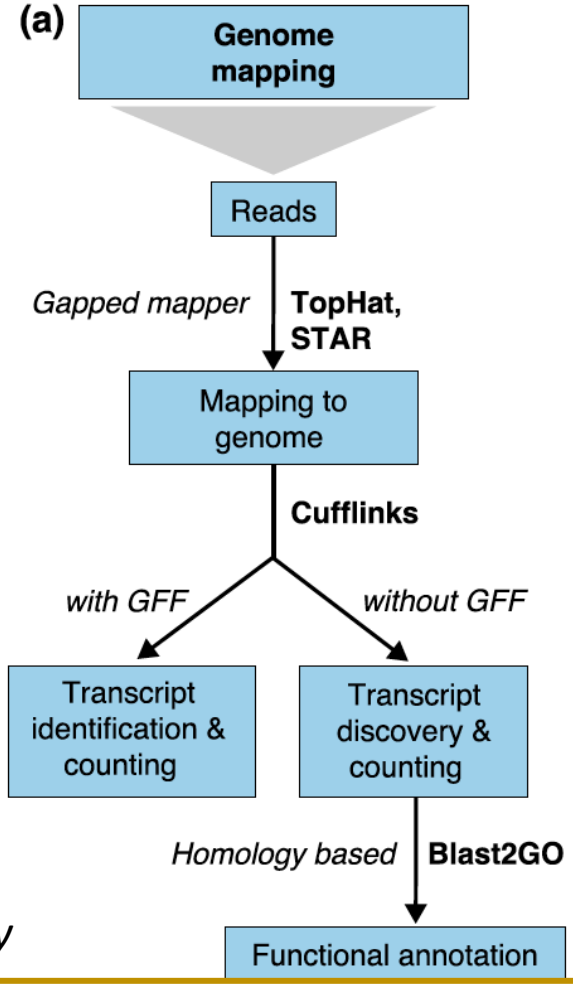


RNA-Seq Analysis
Alignment – Transcriptome Mapping



Conesa *et al.* 2016 *Genome Biology*

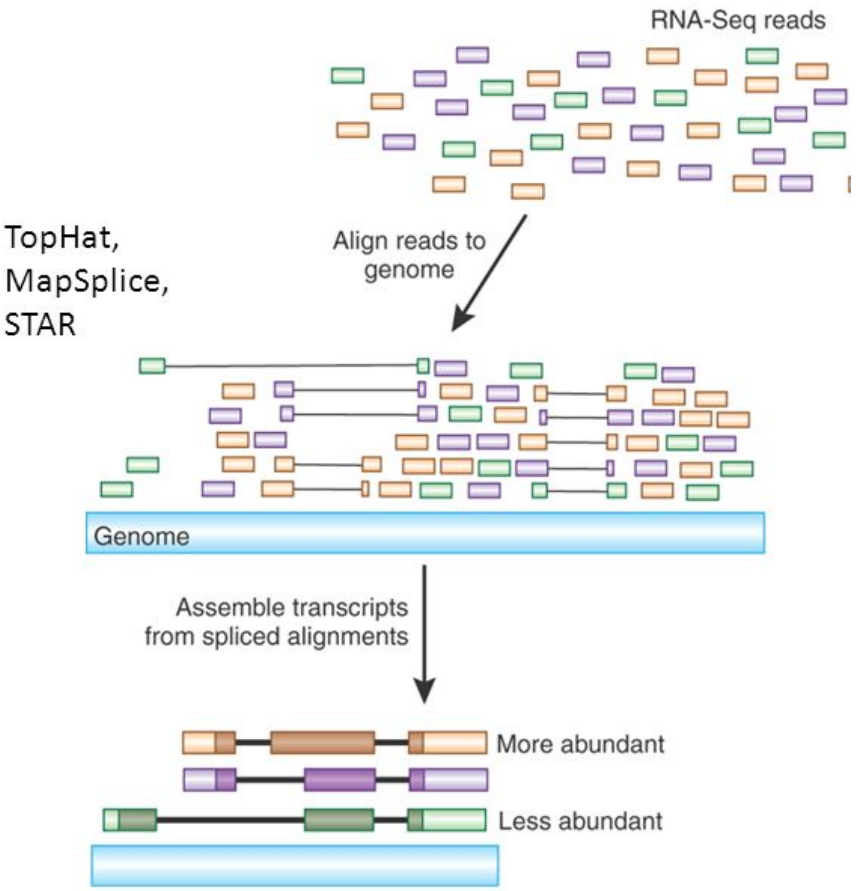
RNA-Seq Analysis
Alignment – Genome Mapping



Conesa *et al.* 2016 *Genome Biology*

RNA-Seq Analysis
Alignment -- Important Parameters

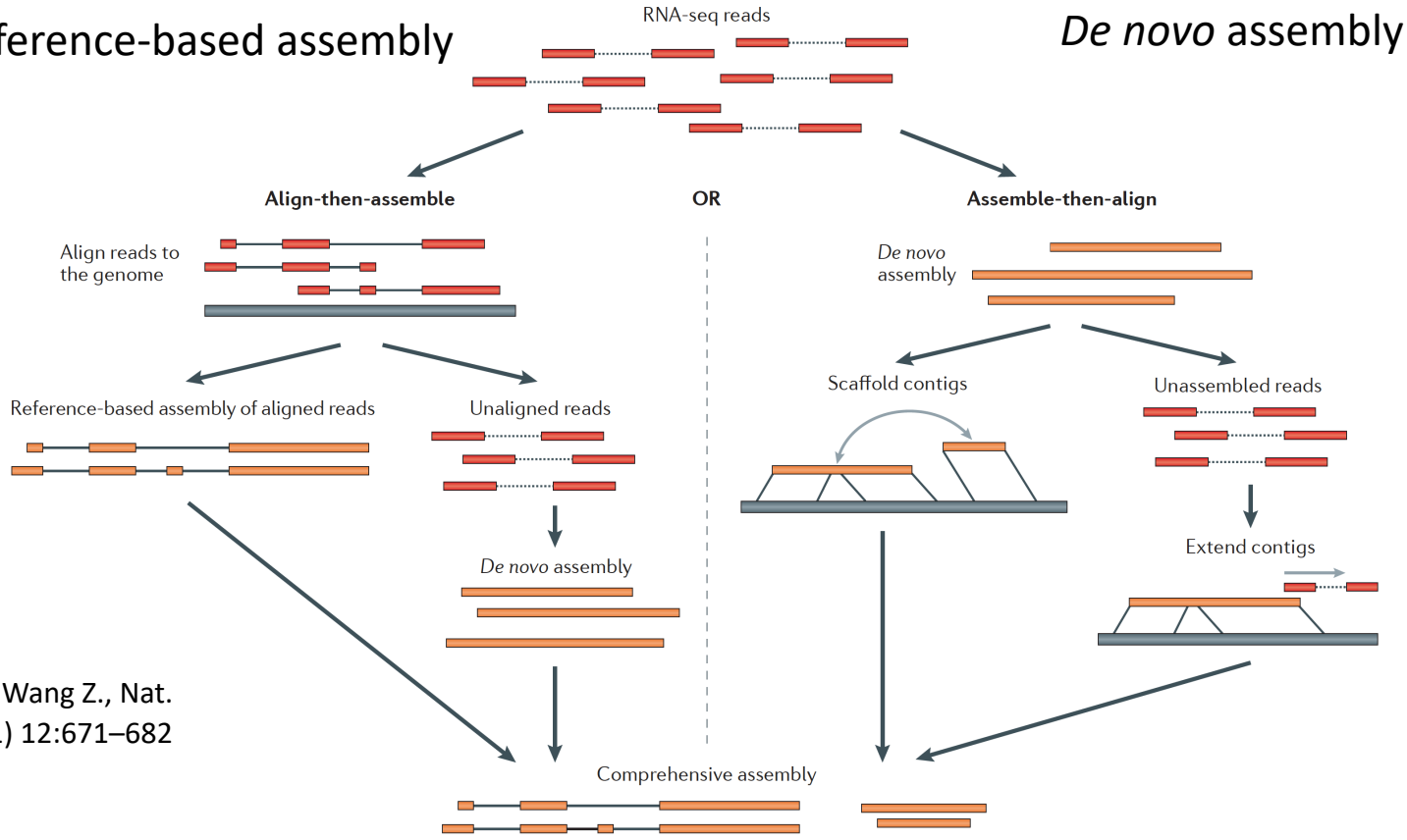
- Strandedness of the RNA-seq library
- # of mismatches to accept
- Read length
- Type of reads (SE or PE)
- Fragment length



RNA-Seq Analysis – Transcript Discovery

Reference-based assembly

De novo assembly

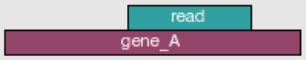
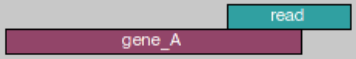
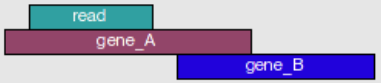
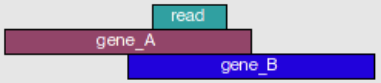


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

RNA-Seq Analysis
Gene-level Quantification

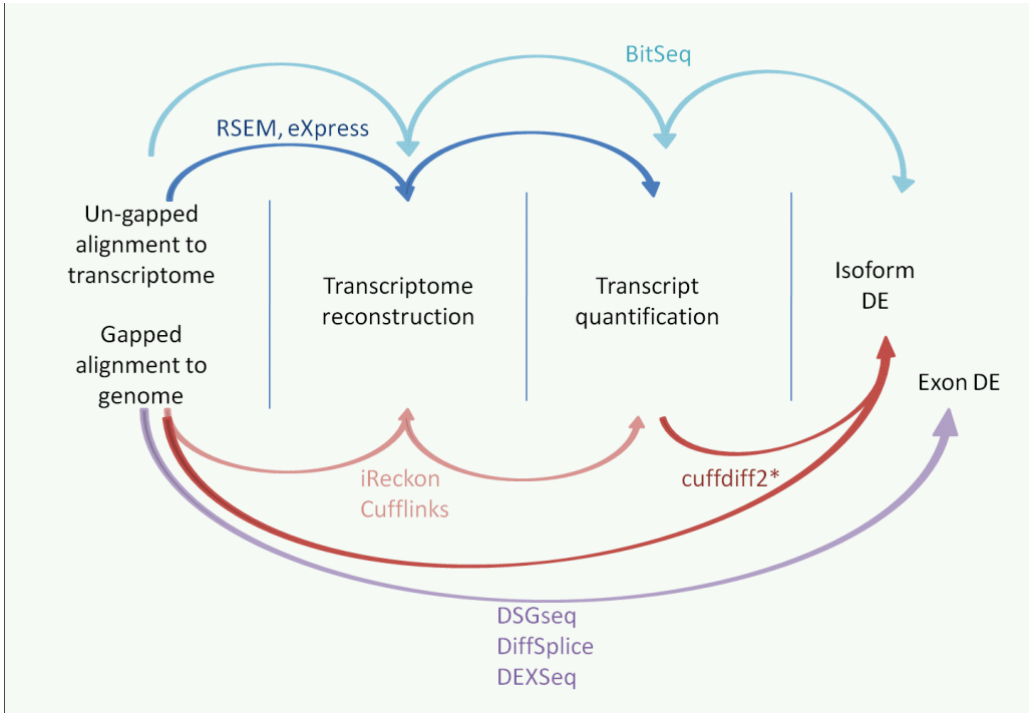
- Aggregation of raw counts of mapped reads
  - HTSeq-count or featureCounts
  - Gene-level approach based on GTF gene coordinates
  - Discard multimappers

HTSeq-count

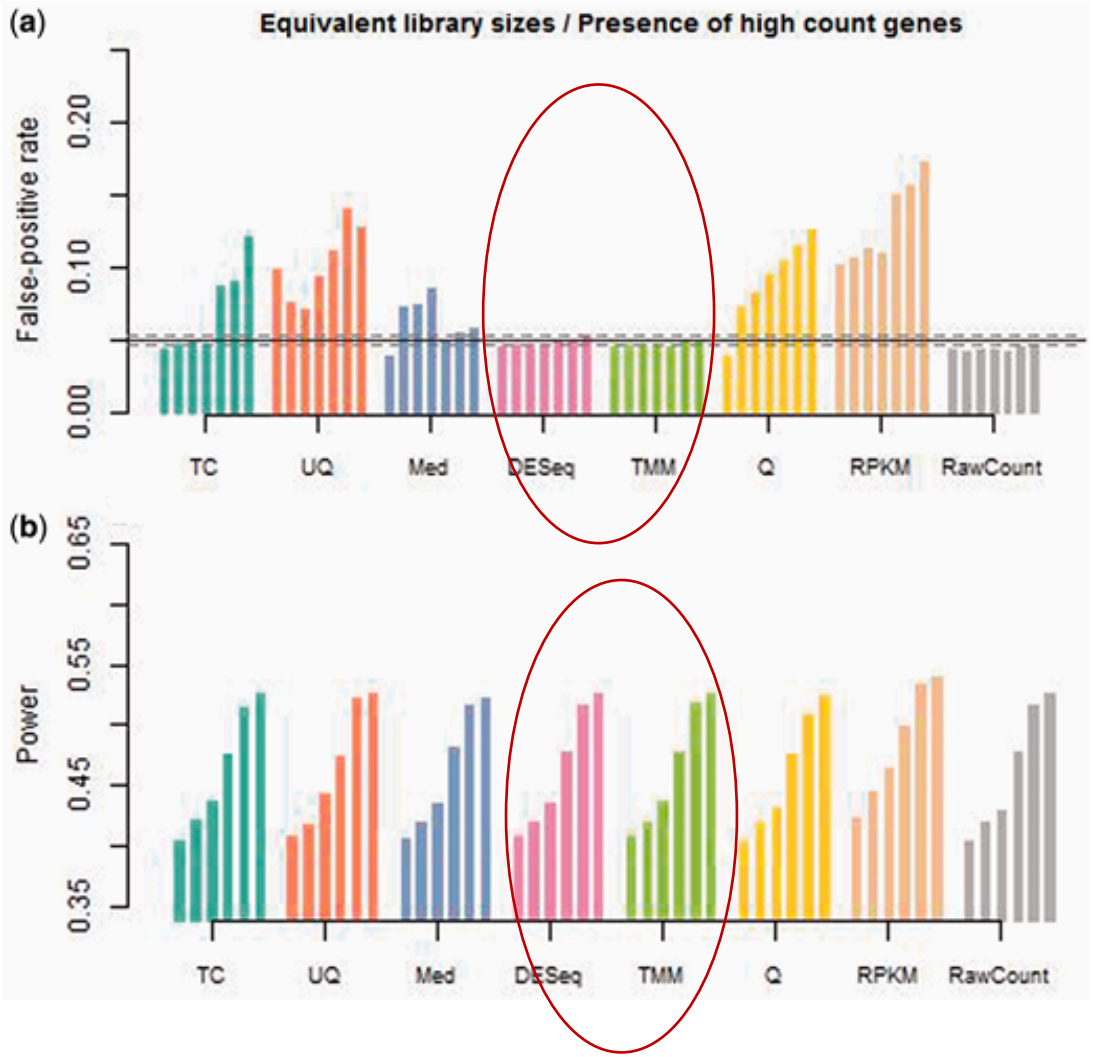
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

RNA-Seq Analysis
Transcript-level Quantification

- Transcript-level expression algorithms
  - Allocate multi-mapping reads among transcript and output within-sample normalized values corrected for sequencing biases.
  - RSEM (RNA-Seq by Expectation Maximization)
  - Cufflinks, eXpress, Kallisto



RNA-Seq Analysis Normalization



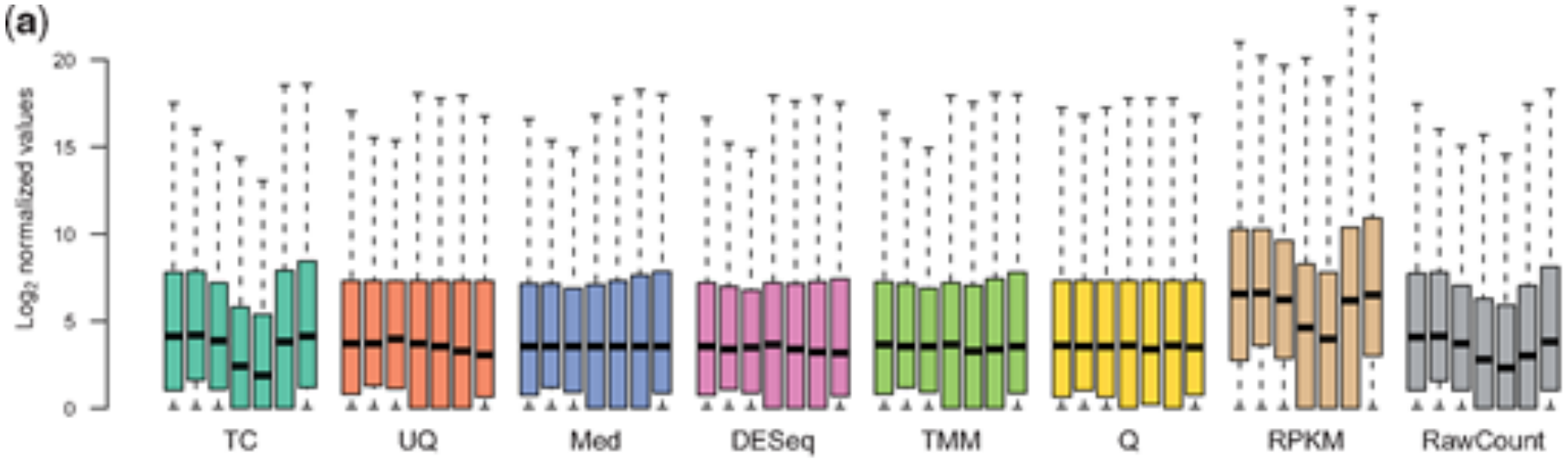
Dillies et al. 2013
Briefings in Bioinformatics

RNA-Seq Analysis Normalization

Summary of comparison results for the seven normalization methods under consideration

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	–	+	+	–	–
UQ	++	++	+	++	–
Med	++	++	–	++	–
DESeq	**++**	**++**	**++**	**++**	**++**
TMM	**++**	**++**	**++**	**++**	**++**
Q	++	–	+	++	–
RPKM	–	+	+	–	–

A '–' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.



Dillies et al. 2013 *Briefings in Bioinformatics*

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

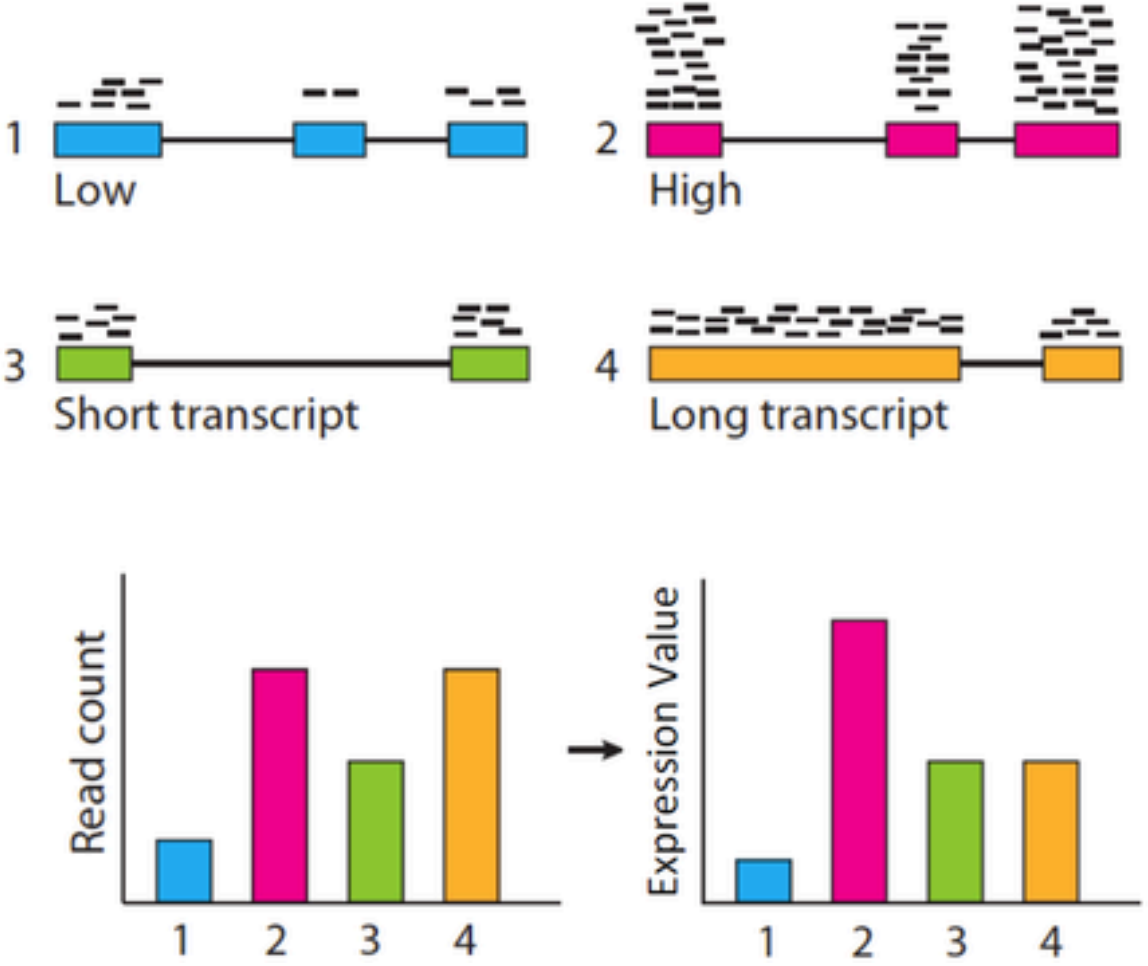
RNA population 1

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

RNA-Seq – Differential Expression Analysis Overview



http://sourceforge.net/projects/trinityrnaseq/files/misc/RNASEQ_WORKSHOP/r

RNA-Seq – Differential Expression Analysis Methods

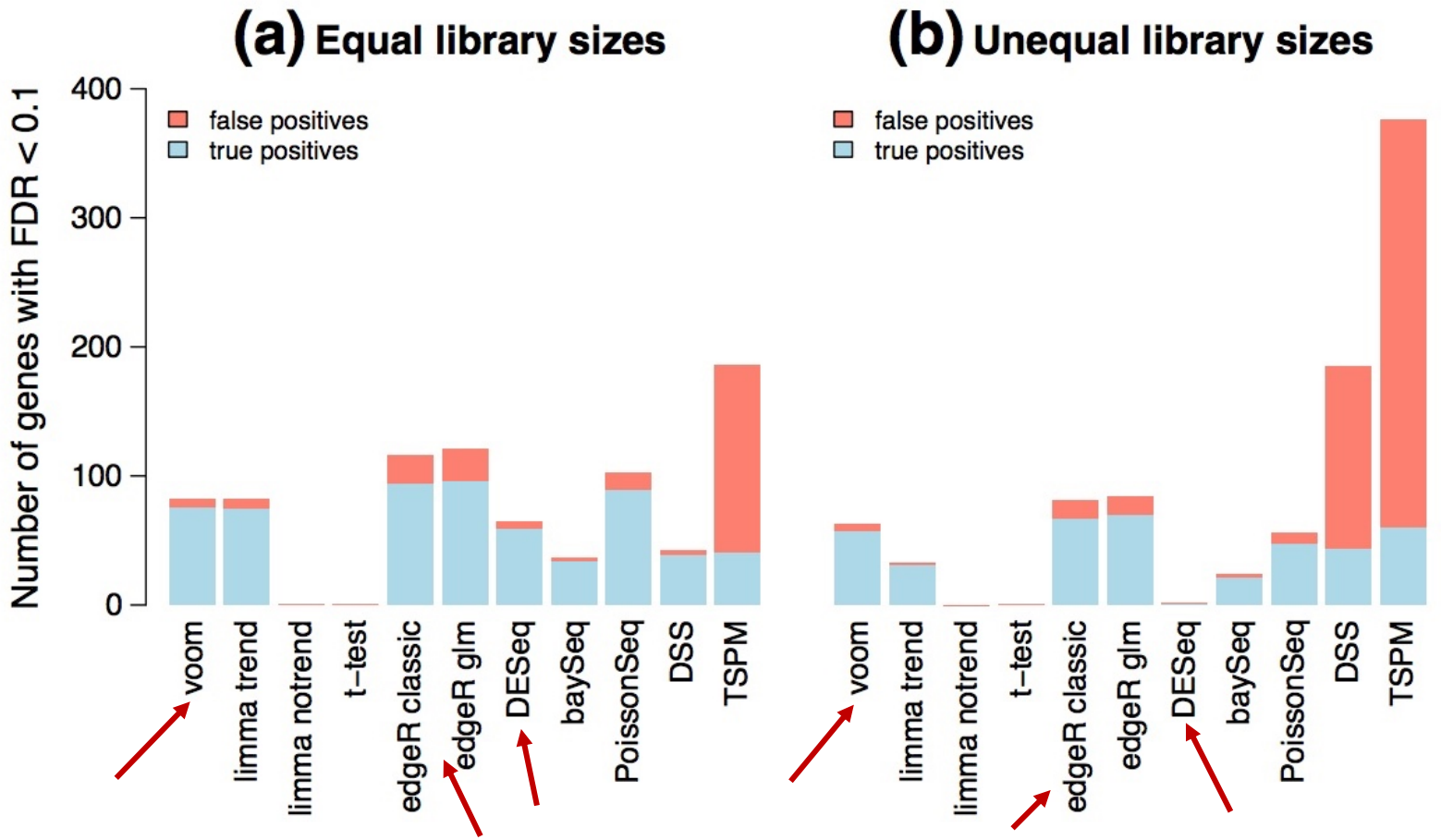


Figure 4 Power to detect true differential expression. Bars show the total number of genes that are detected as statistically significant (FDR < 0.1) **(a)** with equal library sizes and **(b)** with unequal library sizes. The blue segments show the number of true positives while the red segments show false positives. 200 genes are genuinely differentially expressed. Results are averaged over 100 simulations. Height of the blue bars shows empirical power. The ratio of the red to blue segments shows empirical FDR. FDR, false discovery rate.

Law et al. 2014 *Genome Biology*

RNA-Seq – Differential Expression Analysis – Bioconductor RNAseq123

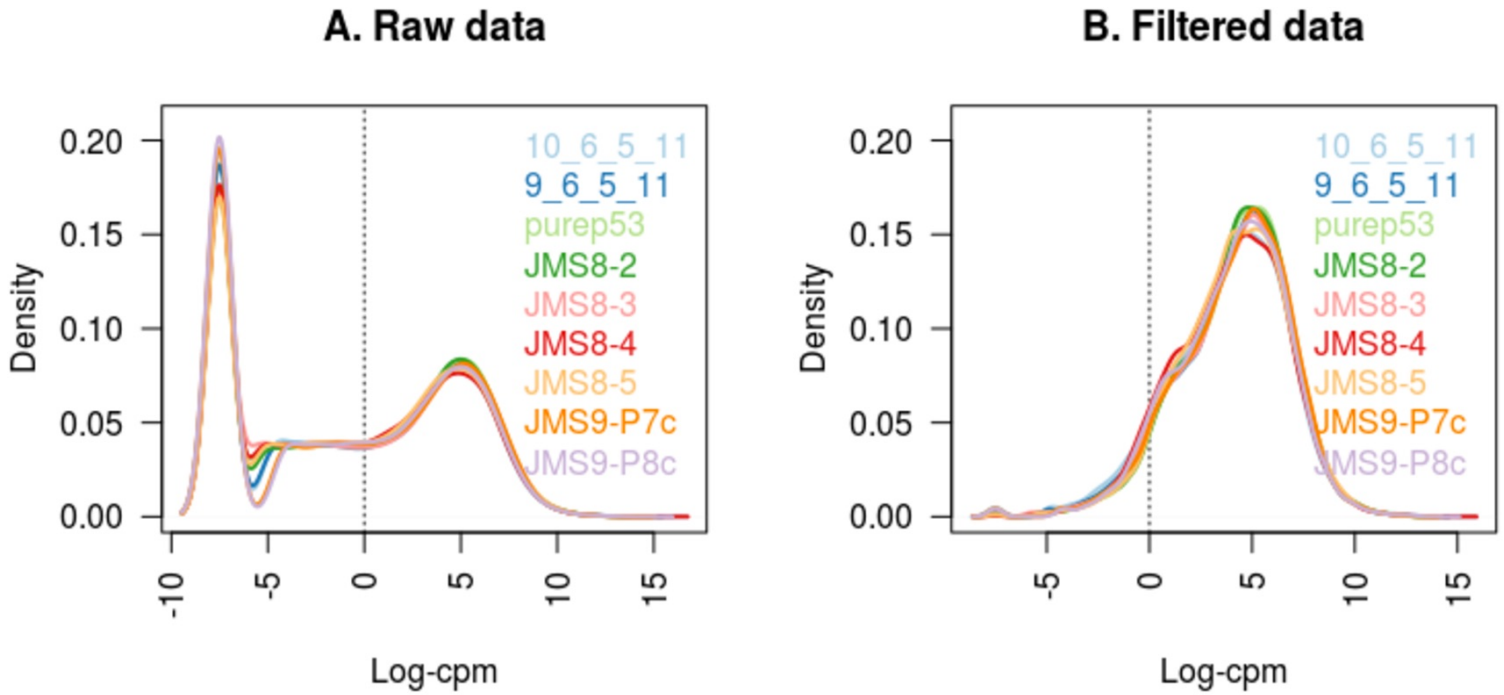
Data pre-processing
- Transformations from the raw-scale
- Removing genes that are lowly expressed
- Normalising gene expression distributions
- Unsupervised clustering of samples

Differential expression analysis
- Creating a design matrix and contrasts
- Removing heteroscedascity from count data
- Fitting linear models for comparisons of interest
- Examining the number of DE genes
- Examining individual DE genes from top to bottom
- Useful graphical representations of differential expression results

https://www.bioconductor.org/help/workflows/RNAseq123/#normalising-gene-expression-distributions

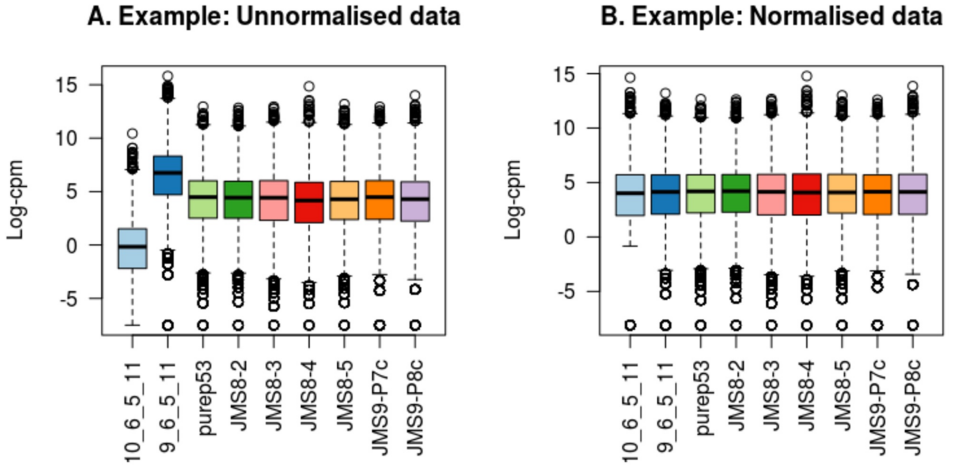
RNA-Seq – Differential Expression Analysis: Data Pre-processing

1. Transform raw counts into counts per million (CPM) or log2-counts per million (log-CPM)
2. Remove genes that are lowly expressed (CPM > 1)

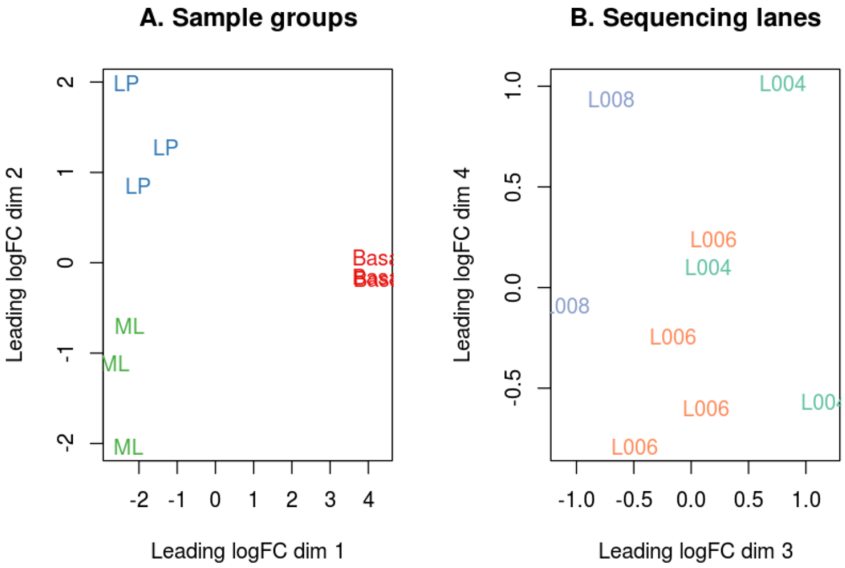


RNA-Seq – Differential Expression Analysis: Data Pre-processing

3. Normalize gene expression distributions (TMM)
4. Unsupervised clustering of samples



Example data: Boxplots of log-CPM values showing expression distributions for unnormalised data (A) and normalised data (B) for each sample in the modified dataset where the counts in samples 1 and 2 have been scaled to 5% and 500% of their original values respectively.



RNA-Seq – Differential Expression Analysis

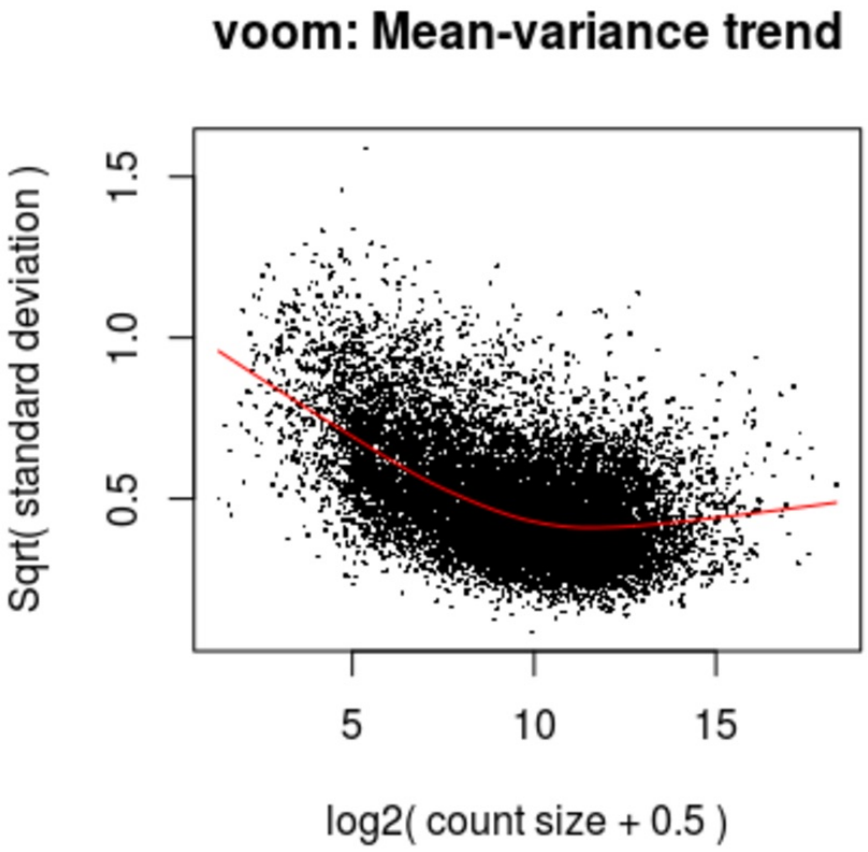
1. Create design matrix and contrasts

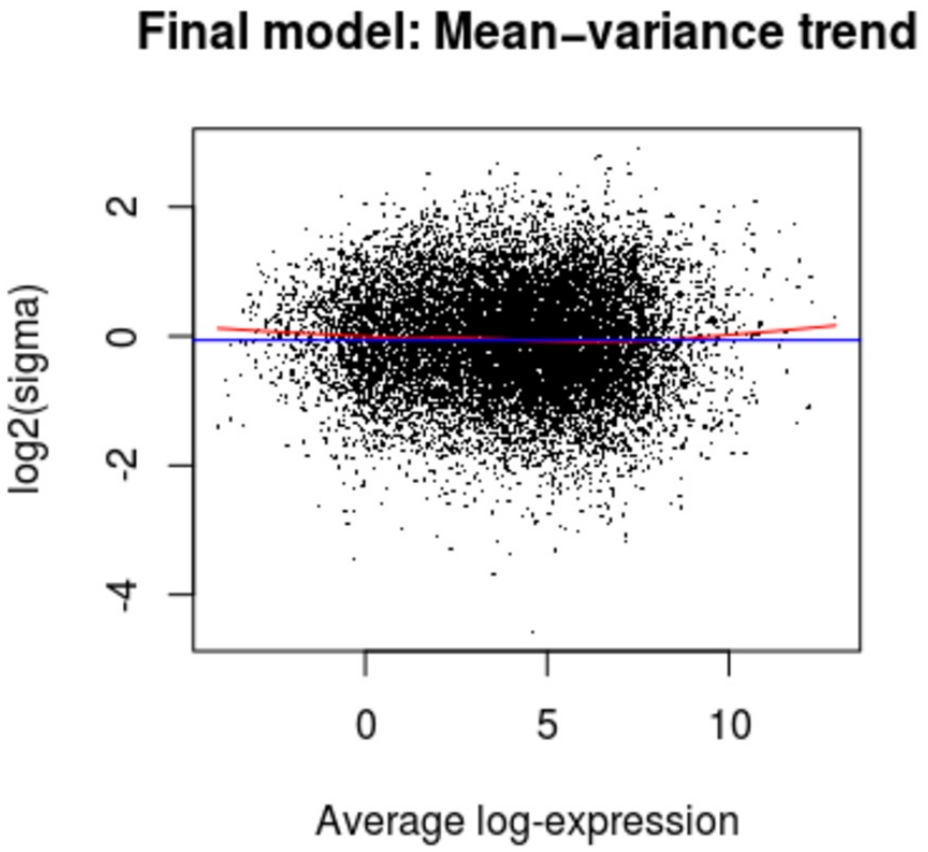
```
design <- model.matrix(~0+group+lane)
colnames(design) <- gsub("group", "", colnames(design))
design

##   Basal LP ML laneL006 laneL008
## 1     0  1  0        0        0
## 2     0  0  1        0        0
## 3     1  0  0        0        0
## 4     1  0  0        1        0
## 5     0  0  1        1        0
## 6     0  1  0        1        0
## 7     1  0  0        1        0
## 8     0  0  1        0        1
## 9     0  1  0        0        1
```

```
contr.matrix <- makeContrasts(
   BasalvsLP = Basal-LP,
   BasalvsML = Basal - ML,
   LPvsML = LP - ML,
   levels = colnames(design))
contr.matrix

##           Contrasts
## Levels     BasalvsLP BasalvsML LPvsML
##   Basal            1         1      0
##   LP              -1         0      1
##   ML               0        -1     -1
##   laneL006         0         0      0
##   laneL008         0         0      0
```

RNA-Seq – Differential Expression Analysis

2. Remove heteroscedasticity from count data

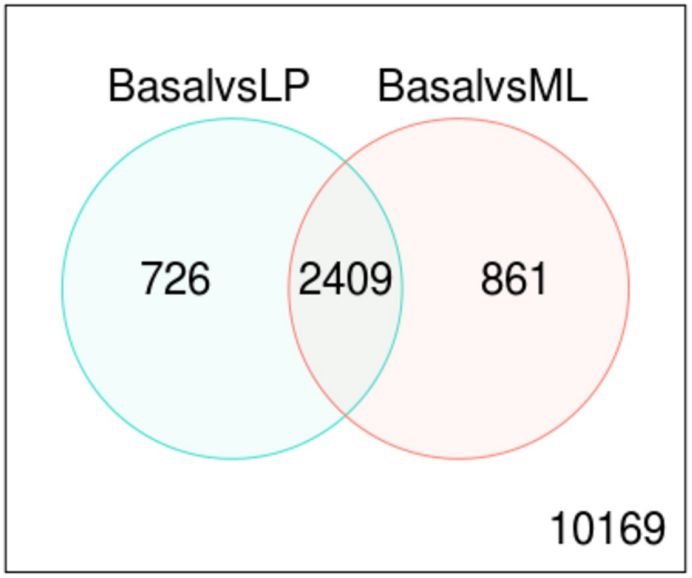




RNA-Seq – Differential Expression Analysis

3. Fitting linear models for comparisons of interest – limma
4. Examining the number of DE genes

```
summary(decideTests(efit))

##    BasalvsLP BasalvsML LPvsML
## -1      4127      4338   2895
## 0       5740      5655   8825
## 1       4298      4172   2445
```



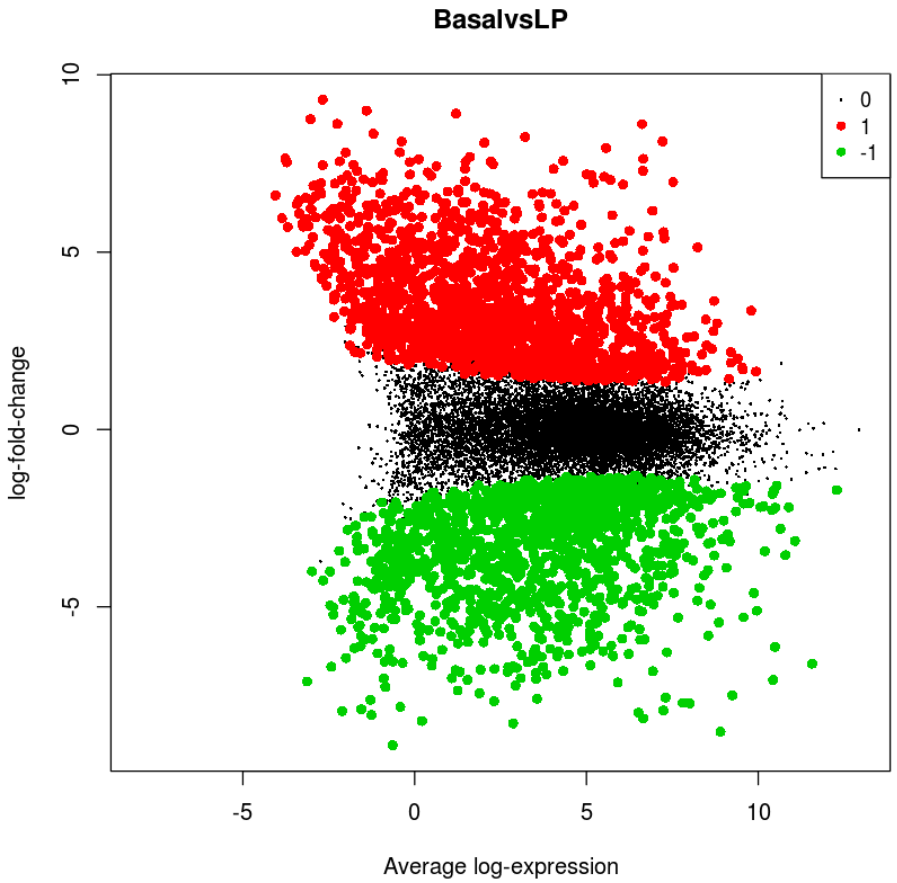
RNA-Seq – Differential Expression Analysis

5. Examining individual DE genes from top to bottom

```
basal.vs.lp <- topTreat(tfit, coef=1, n=Inf)
basal.vs.ml <- topTreat(tfit, coef=2, n=Inf)
head(basal.vs.lp)

##        ENTREZID SYMBOL TXCHROM logFC AveExpr     t  P.Value adj.P.Val
## 12759     12759    Clu   chr14 -5.44    8.86 -33.4 3.99e-10   2.7e-06
## 53624     53624  Cldn7   chr11 -5.51    6.30 -32.9 4.50e-10   2.7e-06
## 242505   242505  Rasef    chr4 -5.92    5.12 -31.8 6.06e-10   2.7e-06
## 67451     67451   Pkp2   chr16 -5.72    4.42 -30.7 8.01e-10   2.7e-06
## 228543   228543   Rhov    chr2 -6.25    5.49 -29.5 1.11e-09   2.7e-06
## 70350     70350  Basp1   chr15 -6.07    5.25 -28.6 1.38e-09   2.7e-06

head(basal.vs.ml)

##        ENTREZID  SYMBOL TXCHROM logFC AveExpr     t  P.Value adj.P.Val
## 242505   242505   Rasef    chr4 -6.51    5.12 -35.5 2.57e-10  1.92e-06
## 53624     53624   Cldn7   chr11 -5.47    6.30 -32.5 4.98e-10  1.92e-06
## 12521     12521    Cd82    chr2 -4.67    7.07 -31.8 5.80e-10  1.92e-06
## 71740     71740 Nectin4    chr1 -5.56    5.17 -31.3 6.76e-10  1.92e-06
## 20661     20661   Sort1    chr3 -4.91    6.71 -31.2 6.76e-10  1.92e-06
## 15375     15375   Foxa1   chr12 -5.75    5.63 -28.3 1.49e-09  2.28e-06
```

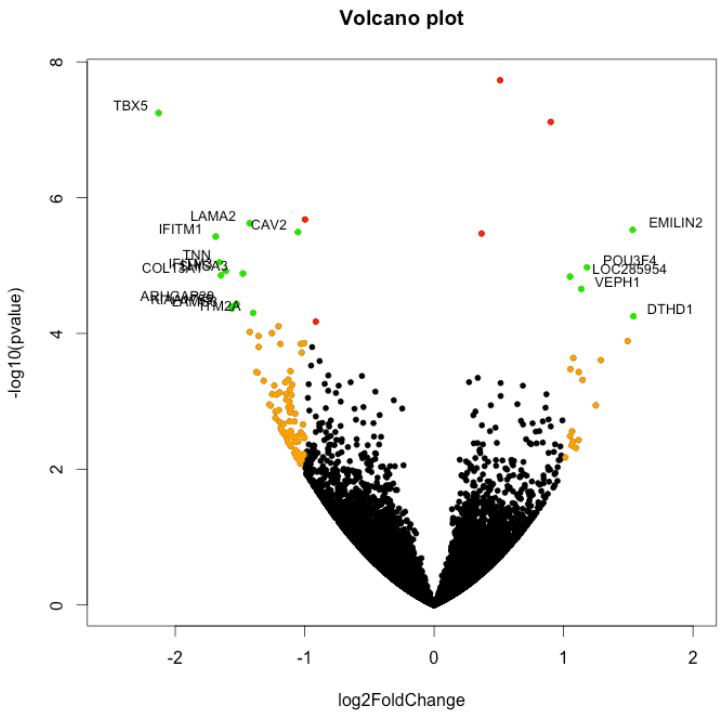
RNA-Seq – Differential Expression Analysis

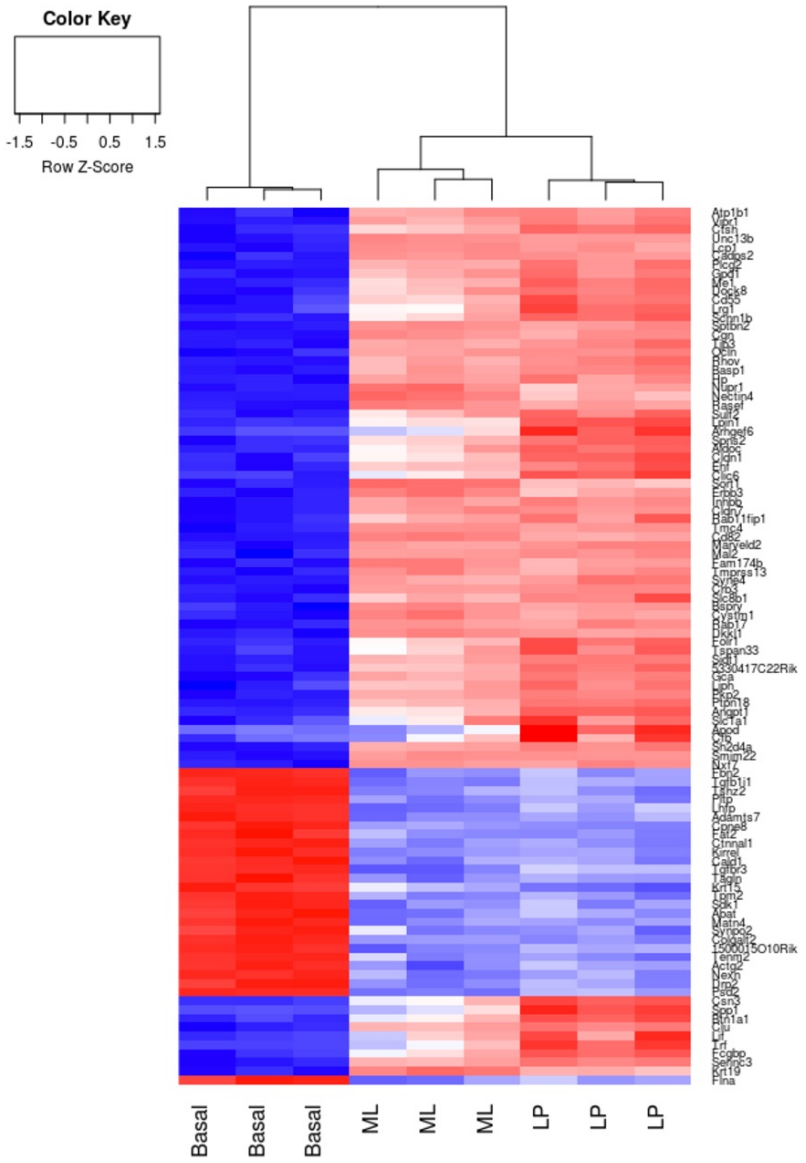
6. Useful graphical representations of differential expression



RNA-Seq – Differential Expression Analysis

6. Useful graphical representations of differential expression





Downstream Analysis & Interpretation

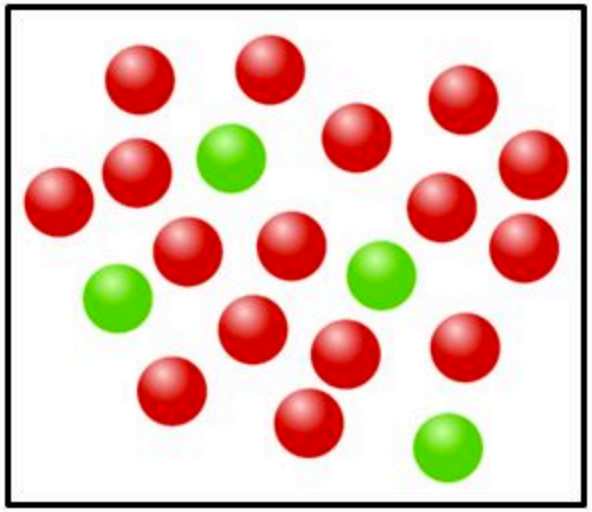
- Hypergeometric test and overrepresentation analysis
- Functional Gene Set Enrichment Analysis
- Pathway Analysis
- Visualize Alignments with IGV
- Network Analysis

ENTREZID	SYMBOL
242505	Rasef
53624	Cldn7
12521	Cd82
71740	Nectin4
20661	Sort1
15375	Foxa1

Hypergeometric test

- Uses hypergeometric distribution to measure the probability of having drawn a **specific number of successes** (out of a total number of draws) from a population

- Example:



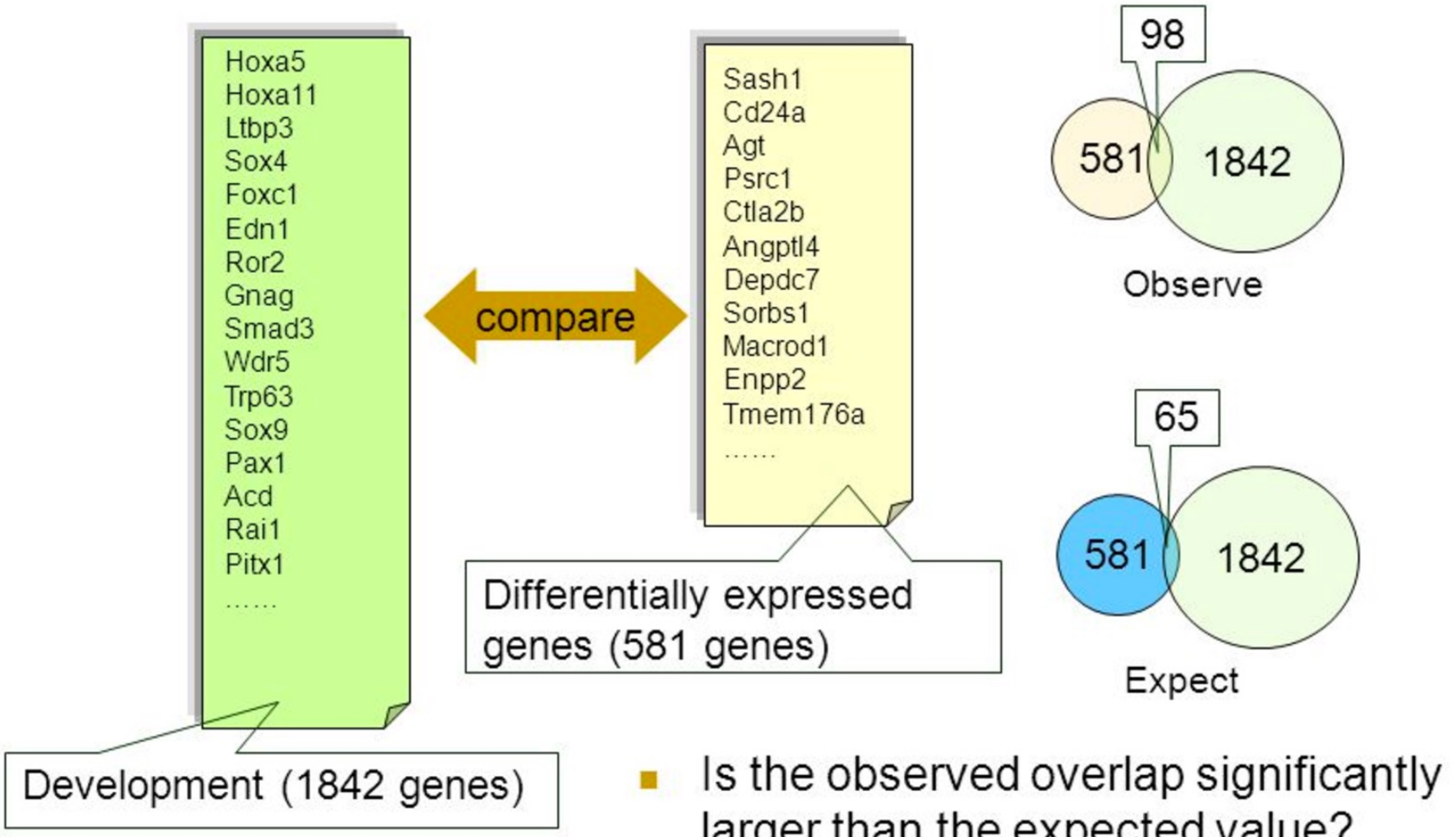
Imagine that there are 4 green and 16 red marbles in a box.

You close your eyes and draw 5 marbles **without replacement**

What is the probability that exactly 2 of the 5 are green?

Babraham Bioinformatics

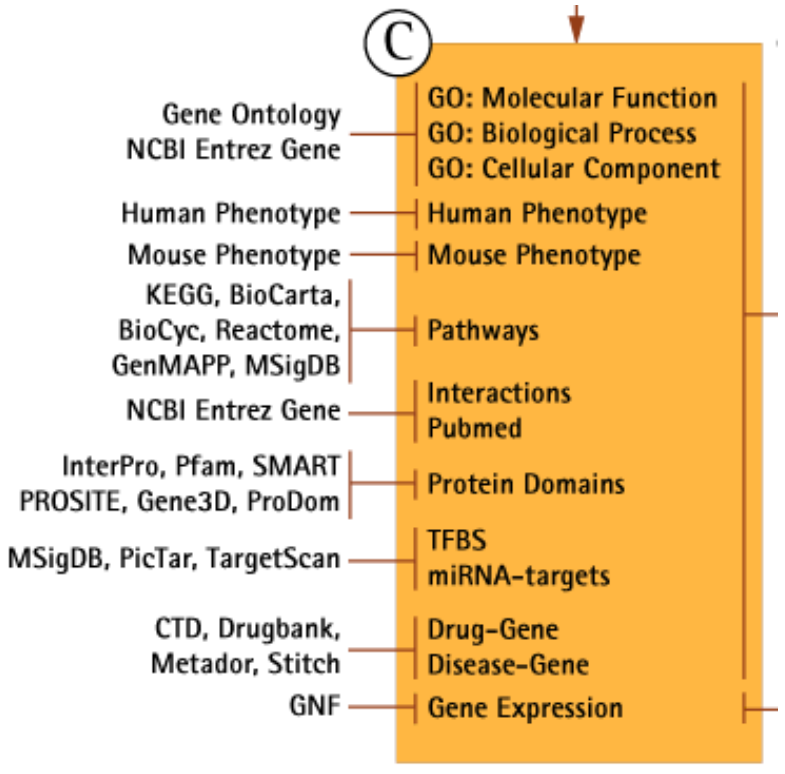
Hypergeometric Test for Overrepresentation Analysis



- Is the observed overlap significantly larger than the expected value?

Downstream Analysis & Interpretation: Functional Enrichment

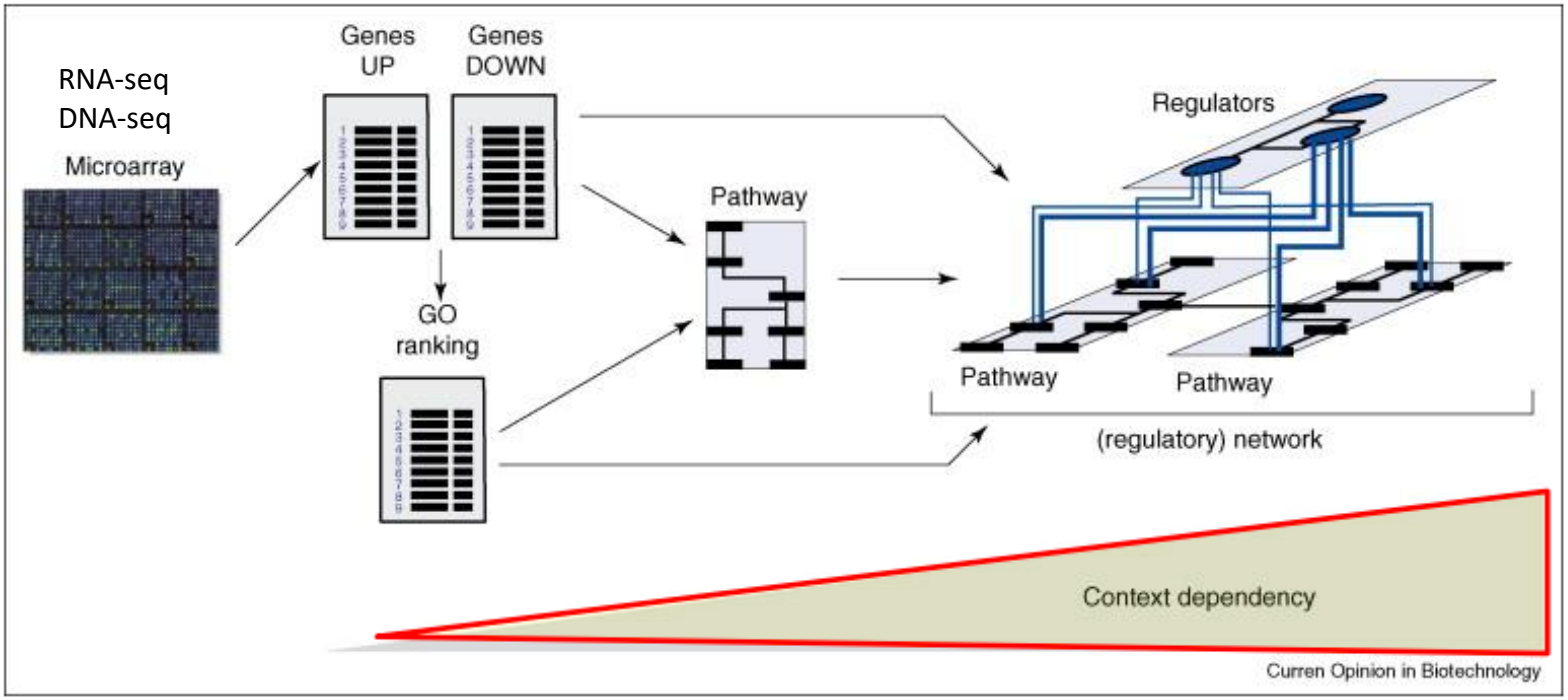
- Gene list enrichment analysis (Hypergeometric test) based on functional annotations
- Tools
  - ToppGene, GSEA, Webgestalt, DAVID



7: Pathway [Display Chart] 75 annotations before applied cutoff / 10916 genes in category

	ID	Name	Source	pValue	FDR B&H	FDR B&Y	Bonferroni	Genes from Input	Genes in Annotation
1	198802	Heart Development	BioSystems: WikiPathways	3.193E-14	2.394E-12	1.174E-11	2.394E-12	6	47
2	M2288	NFAT and Hypertrophy of the heart (Transcription in the broken heart)	MSigDB C2 BIOCARTA (v5.1)	1.851E-8	6.943E-7	3.403E-6	1.389E-6	4	54
3	672464	SRF and miRs in Smooth Muscle Differentiation and Proliferation	BioSystems: WikiPathways	1.558E-7	3.895E-6	1.909E-5	1.168E-5	3	19
4	712094	Cardiac Progenitor Differentiation	BioSystems: WikiPathways	3.731E-6	6.996E-5	3.429E-4	2.799E-4	3	53
5	198878	Serotonin Receptor 2 and ELK-SRF/GATA4 signaling	BioSystems: WikiPathways	4.772E-5	7.158E-4	3.508E-3	3.579E-3	2	17

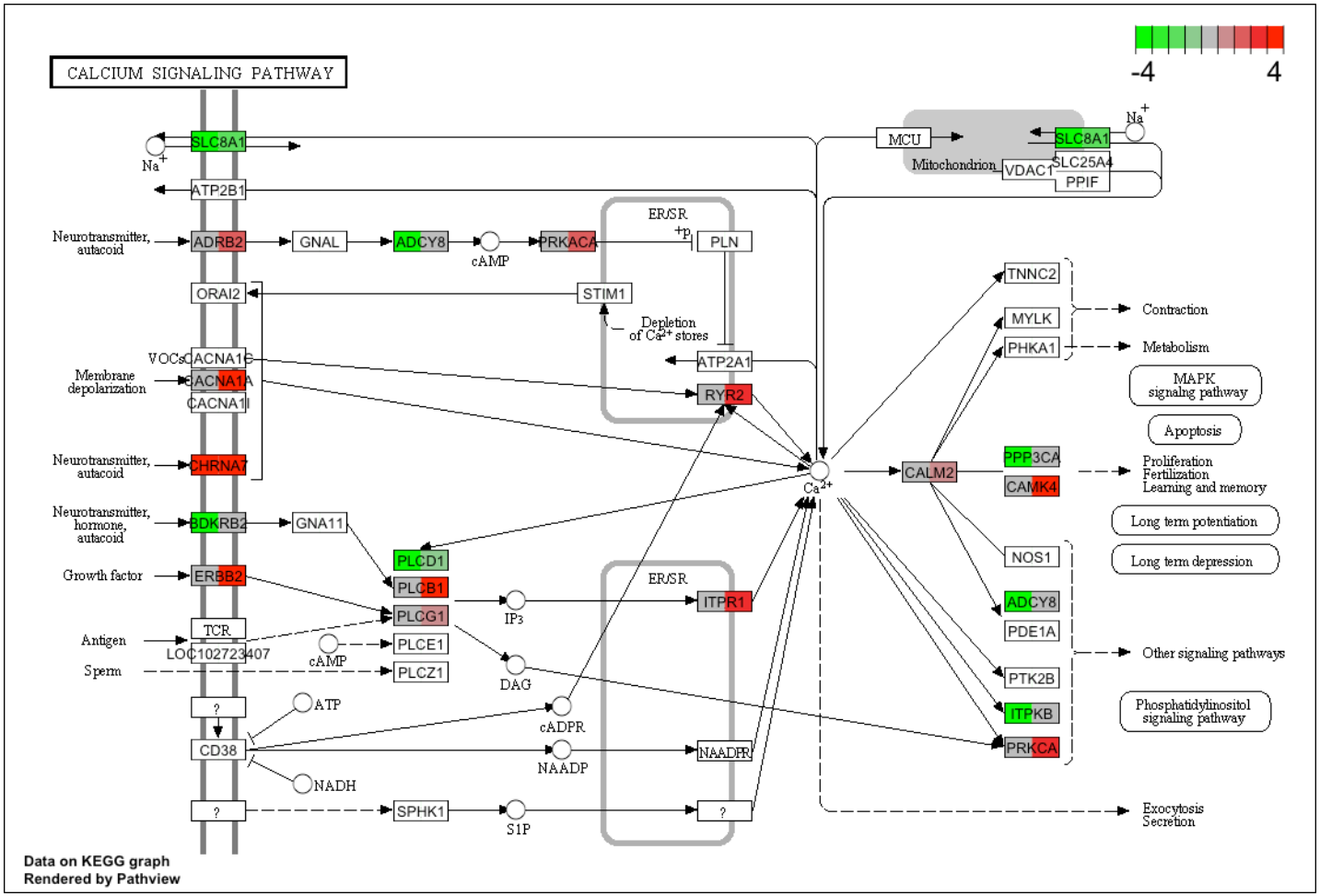
Downstream Analysis & Interpretation: Pathway Analysis



- Databases
  - Ex. KEGG, WikiPathways, Reactome, PathwayCommons, BioCarta
- Tools
  - Ex. Webgestalt, Signaling Pathway Impact Analysis, ToppGene, WikiPathways

Werner 2008 Current Opinion Biotechnology

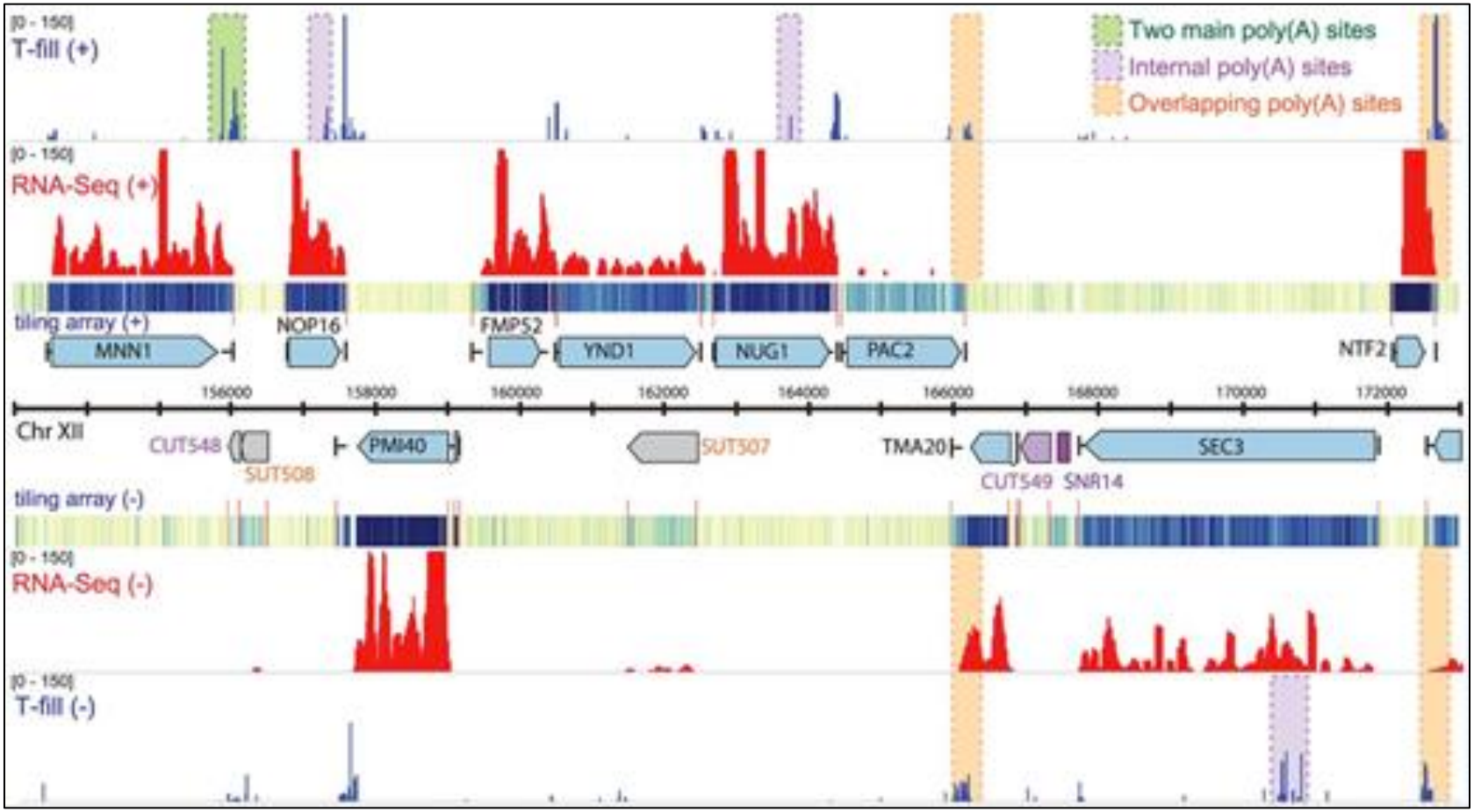
Downstream Analysis & Interpretation: Pathway Analysis



- De Novo Mutations (Left Half of Rectangles)
  - Red (4) = hot DNM
  - Green (-4) = not hot DNM
  - Gray (0) = no DNM
- Hot Genes (Right half of rectangles)
  - Red (4) = hot in autism & epilepsy
  - Red (3)= hot in epilepsy
  - Red (2)= hot in autism
  - Pink (1)= hot novel gene
  - Gray (0) = not hot
  - Light green (-1) = not hot but in autism
  - Green (-2) = not hot but in epilepsy

Tool: Bioconductor Pathview

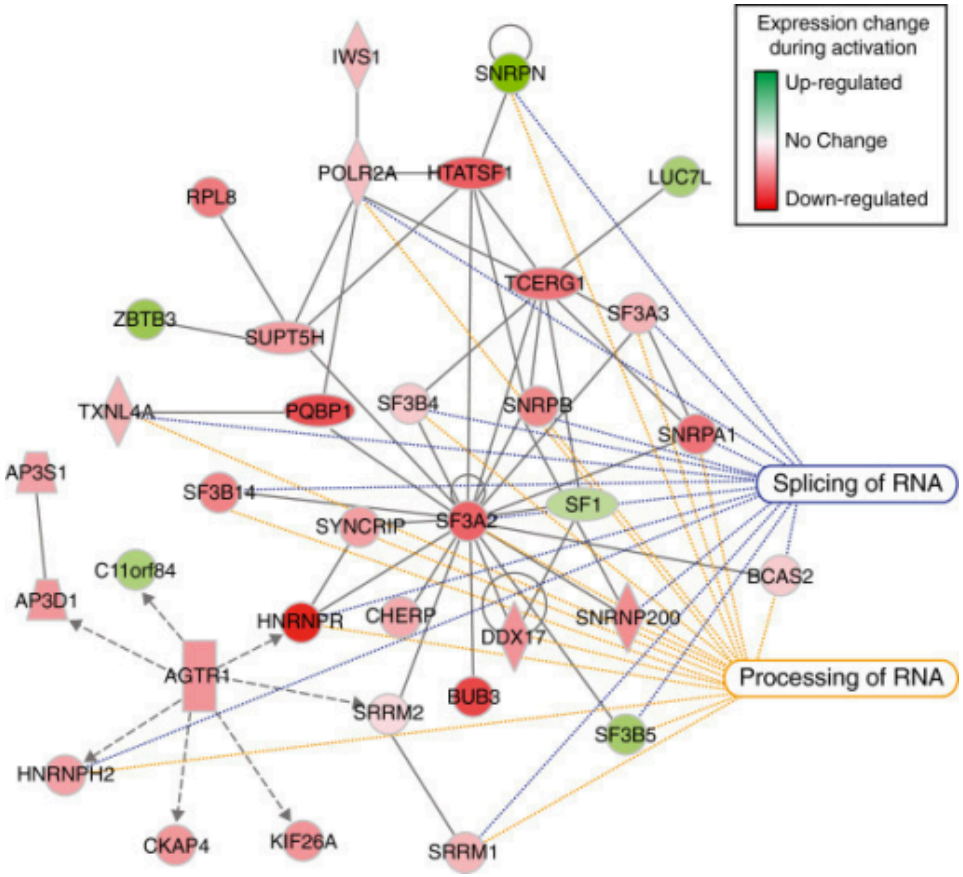
Downstream Analysis & Interpretation: Visualization with IGV and/or GenePattern



Wilkening et al. 2013 *Nucleic Acids Research*

Network Analysis

- Databases
  - PPI
  - Physical interactions
  - Indirect associations
  - Coexpression
  - Literature
  - Experimental
- Tools
  - Cytoscape, StringDB, GeneMania, NetworkX



Farina et al. 2012 Skeletal Muscle

Recommended Reading

A survey of best practices for RNA-seq data analysis

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang and Ali Mortazavi

Genome Biology 2016 17:13 | DOI: 10.1186/s13059-016-0881-8 | © Conesa et al. 2016

Review Article | Published: 24 July 2020

mRNAs, proteins and the emerging principles of gene expression control

Christopher Buccitelli & Matthias Selbach

Nature Reviews Genetics **21**, 630–644(2020) | Cite this article

https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html