UNIVERSITY of CALIFORNIA, SAN DIEGO
SCHOOL OF MEDICINE

CENTER FOR COMPUTATIONAL BIOLOGY & BIOINFORMATICS

RNA-Seq

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 Case Studies**
  - Case Study 1 – Breast cancer precision medicine trial
  - Case Study 2 – Uveal melanoma neoantigen discovery and molecular subtype classification

- **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

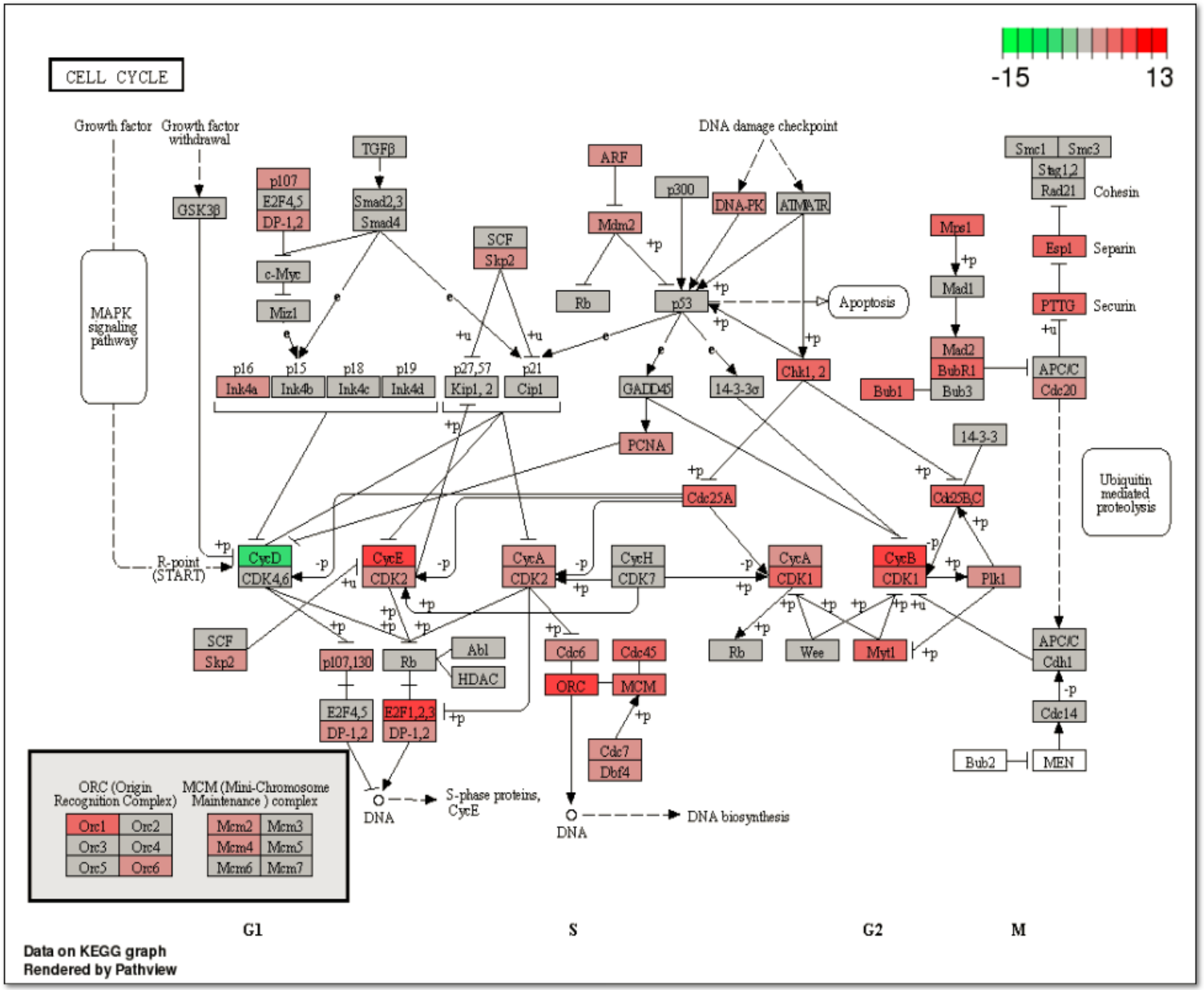
Case Study 1 – Breast Cancer Precision Medicine Trial

MARIO TAMA, GETTY IMAGES

- **Medical history**
  - 74 y/o female
  - Diagnosed April 2013
    - Stage IIIB inflammatory ER positive breast cancer
    - Bulky axillary disease
- **Previous treatments**
  - Dose-dense Neoadjuvant chemotherapy
    - AC → T
  - Modified radical mastectomy
- **Status prior to genomically guided treatment**
  - One month post surgery: dermal mets
  - Enrolled in clinical trial

2 Week Turnaround Time

Patient X – Gene Expression & Pathways



Patient X – Variants

- RNAseq
  - 5,389 mutations affecting genes
  - 26,364 SNPs/INDELS
- Targeted sequencing
  - NRAS – unknown function
  - TP53* – unknown function
  - FOXL2 – Function not altered

RNAseq: Known cancer mutation (COSMIC) Gene Exp: Upregulated	
CHR	12
STRAND	+
START	94772742
ALLELE	C/T
SYMBOL	**CCDC41**
PROTEIN_CHANGES	R209Q,R176Q
XREFS	dbSNP:rs2271979

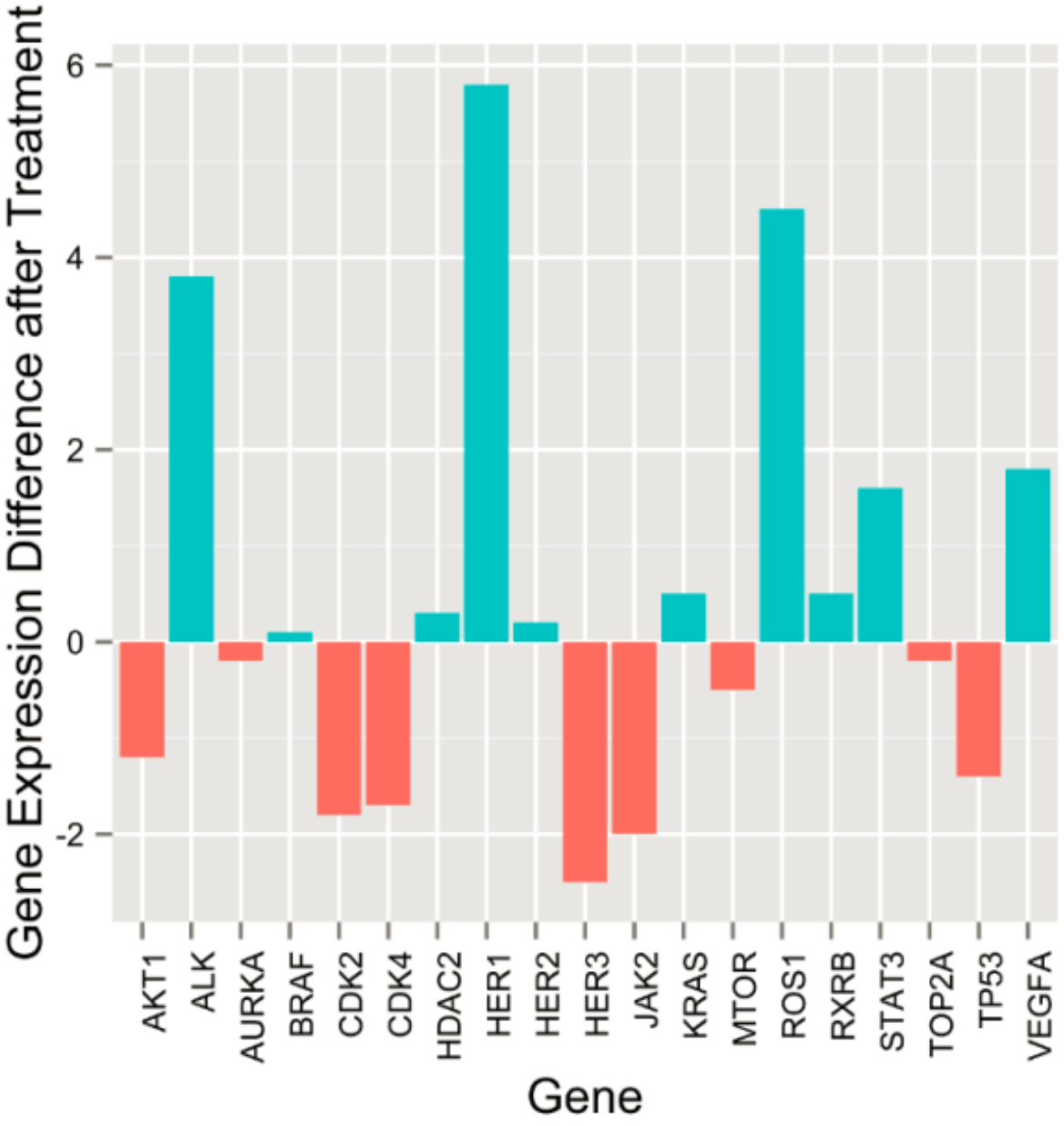
*Supported by RNAseq data

Patient X – RNAseq-based Drug Matching

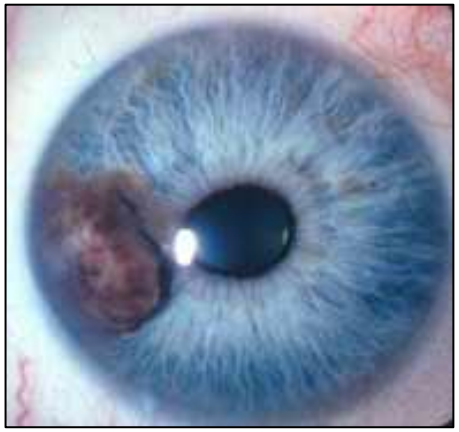
Drug	Gene Target	Gene log2FC	Gene Target w/ Variant	Variant Impact	Pathway	Evidence Score
Adriamycin	TOP2A TOP2B	5.02 0.92	TOP2A	high		2
Etoposide	TOP2A TOP2B	5.02 0.92	TOP2A	high		2
LEE011	CDK4	1.28	CDK4	high		2
PD-0332991	CDK4	1.28	CDK4	high		2
Gemcitabine	RRM1	1.02	RRM1	high		2
BGJ398	FGFR1	-2.85	FGFR2	high		2
XL184	MET	-1.79	KDR MET	high high		2
AMG337	MET	-1.79	MET	high		2
Crizotinib	MET	-1.79	MET	high		2
Paclitaxel	BCL2	-1.39			hsa04540	2
Pemetrexed	ATIC DHFR GART	-1.14 3.27 1.61	DHFR	high		2
Vorinostat			HDAC1 HDAC2	high high	hsa04110	2
MLN8237	AURKA	4.55				1
Dacarbazine	POLA2	1.15				1
Alitretinoin	RXRB RARG	-3.3				1
Interferon			IFNAR1	medium		1
Sorefenib			RAF1 KDR	high high		1

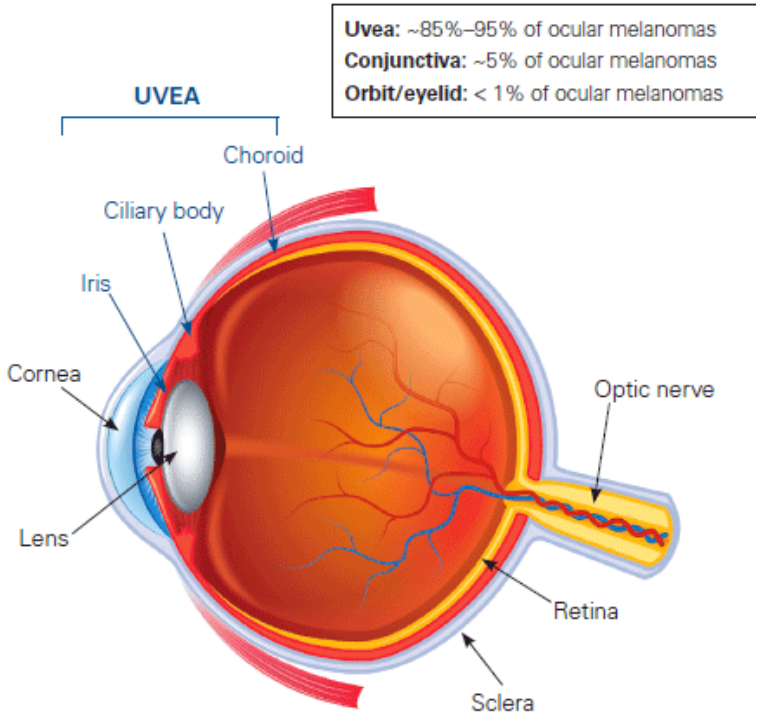
Patient X – Treatment & Rebiopsy

- Actionable Findings
  - ER-positive
  - PI3K/Akt/mTOR pathway activation
- Therapy
  - Combination therapy
    - Exemestane & Everolimus
- Rebiopsy – 8 days



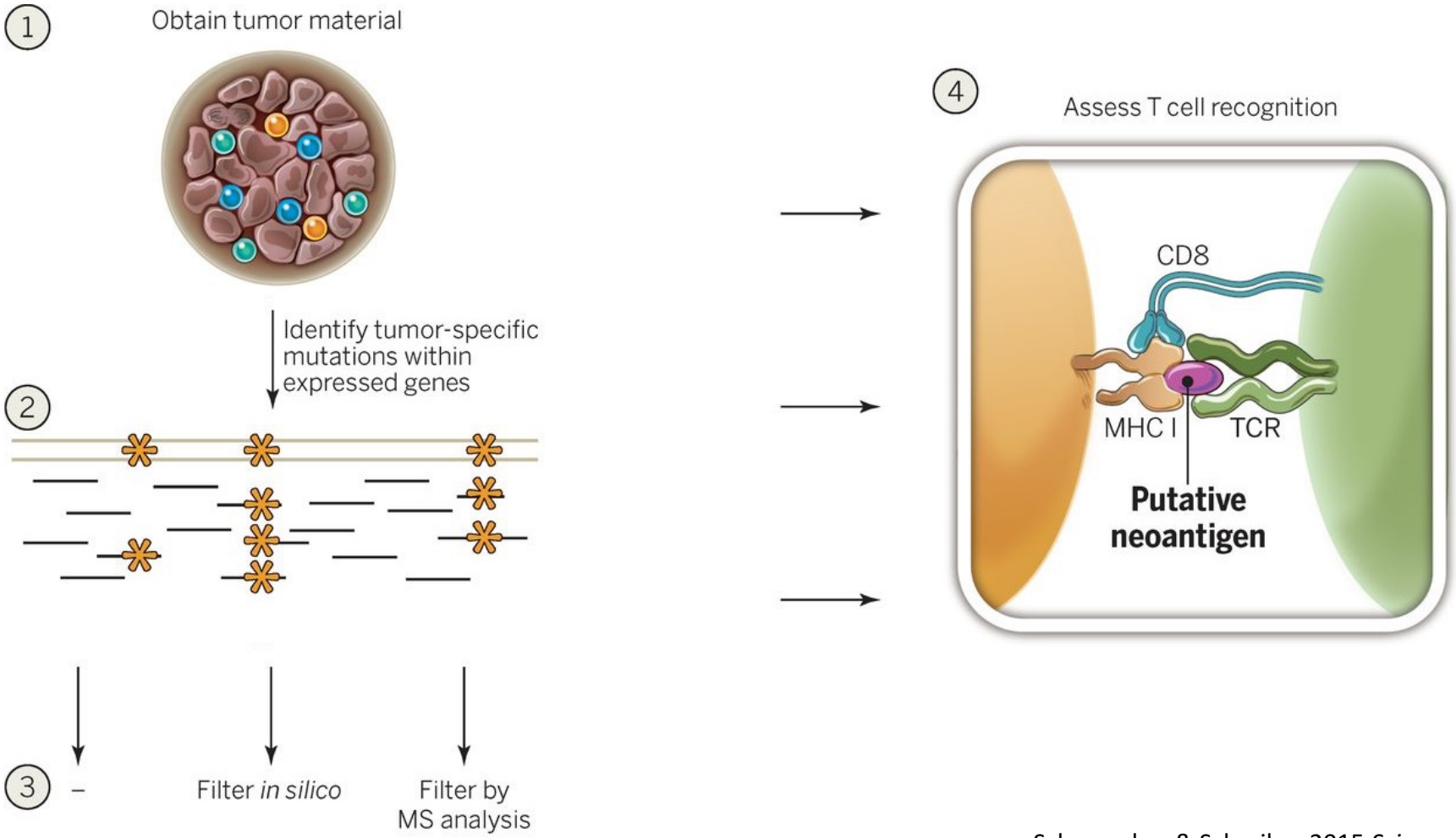
Case Study 2: Uveal Melanoma Neoantigen Discovery & Molecular Subtype Classification





Uveal Melanoma

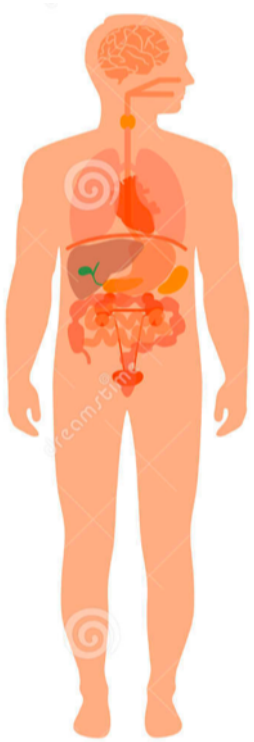
Translational Medicine – Neoantigen Discovery



Schumacher & Schreiber 2015 *Science*

N of 1 Case Study: Patient Background

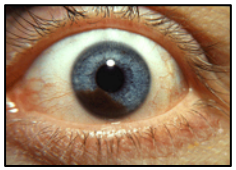
Female, Caucasian, mid-50s



Blood

Whole Genome Sequencing

Tumor



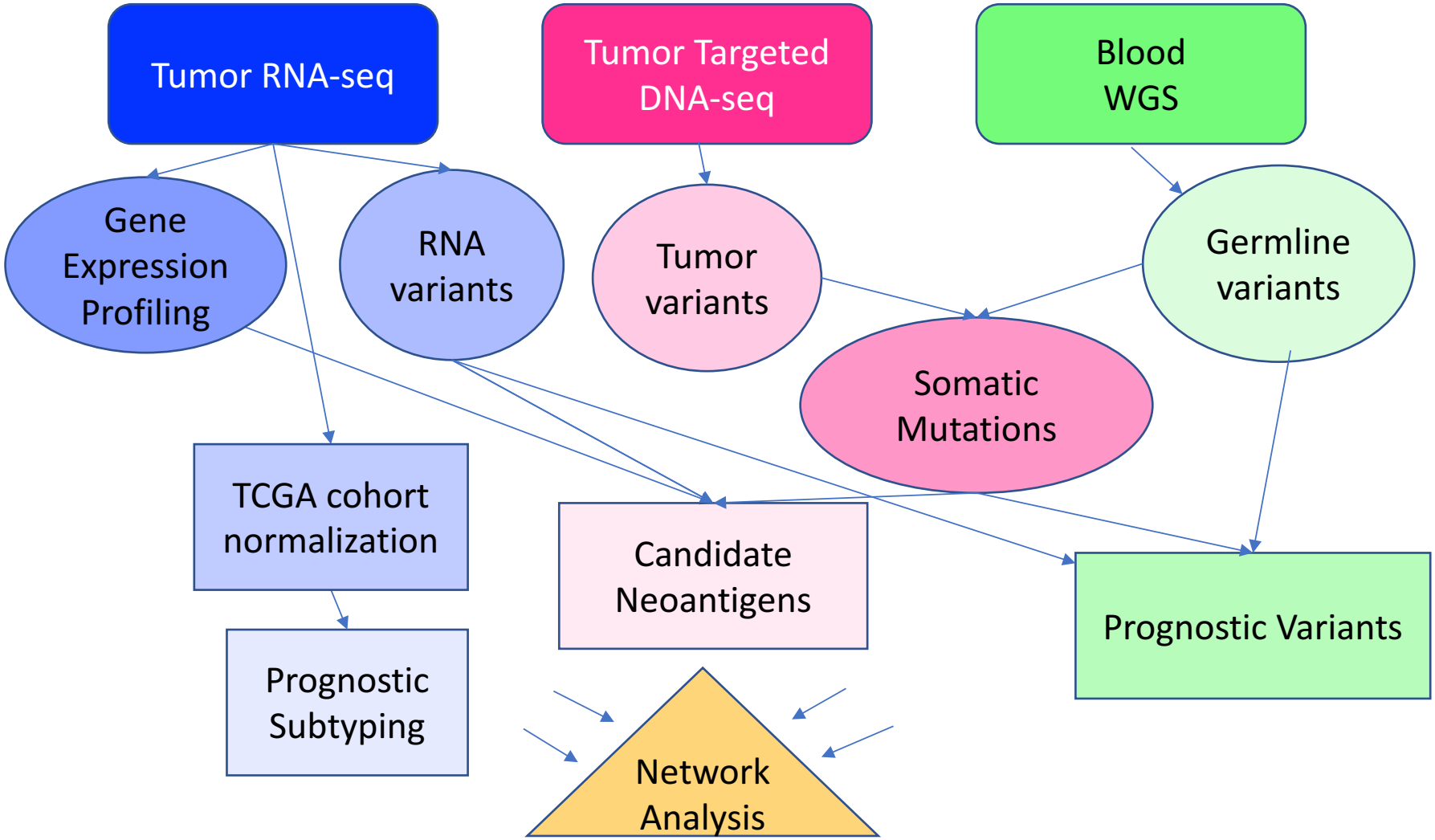
Targeted DNA sequencing

RNA sequencing

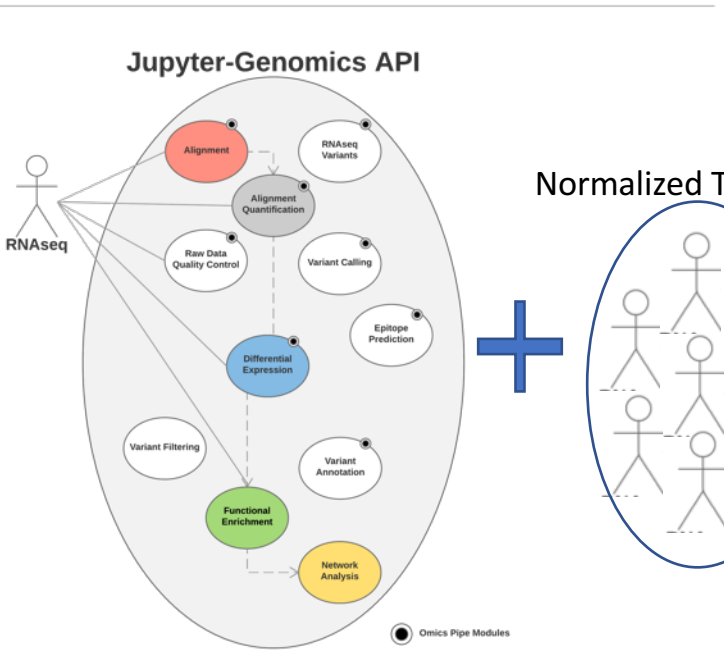
N of 1 Case Study: Motivation

1. Characterize molecular landscape of the patient's tumor
2. Identify prognostic markers
3. Identify tumor neoantigens for immunotherapy

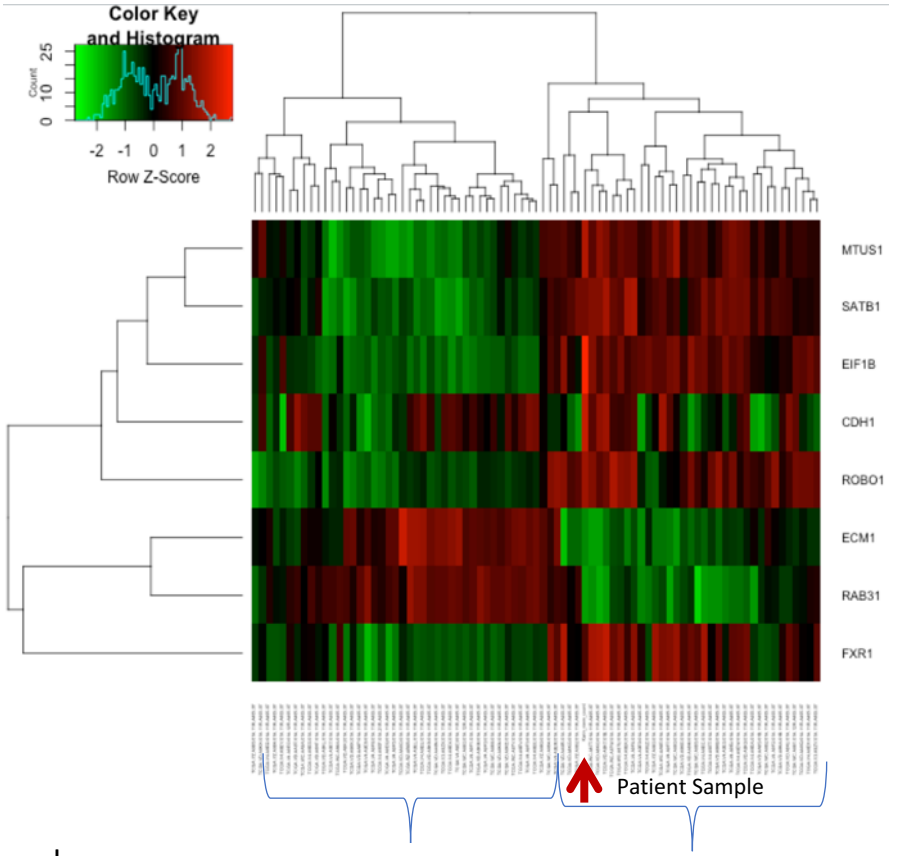
N of 1 Case Study: Analysis Workflow



N of 1 Case Study: Subtype Stratification

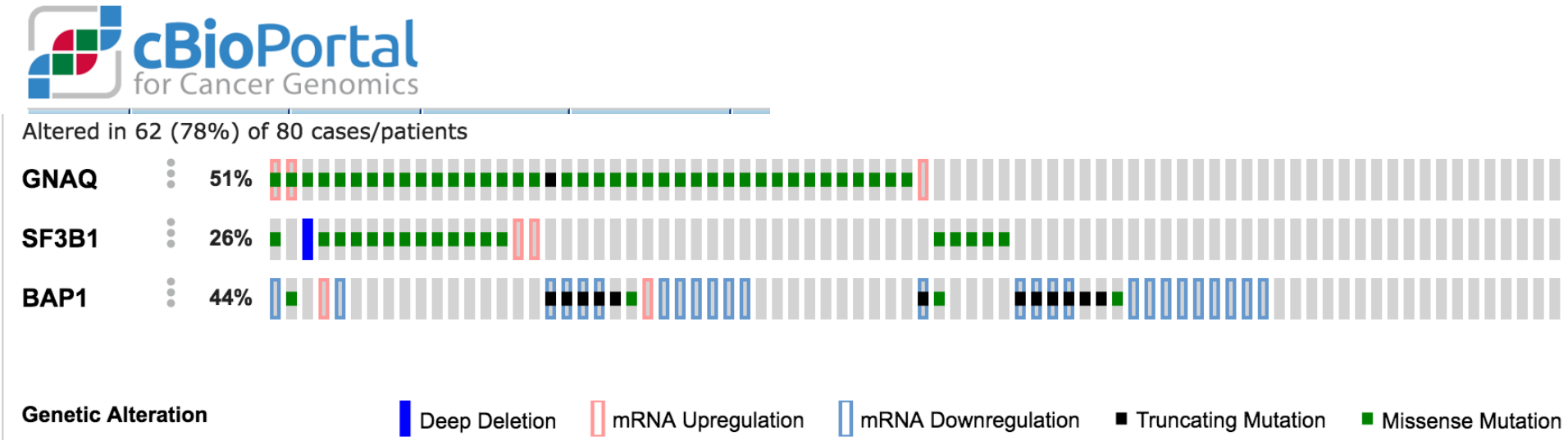


Normalized TCGA Cohort



N of 1 RNAseq subtype stratification with TCGA Uveal Melanoma Cohort

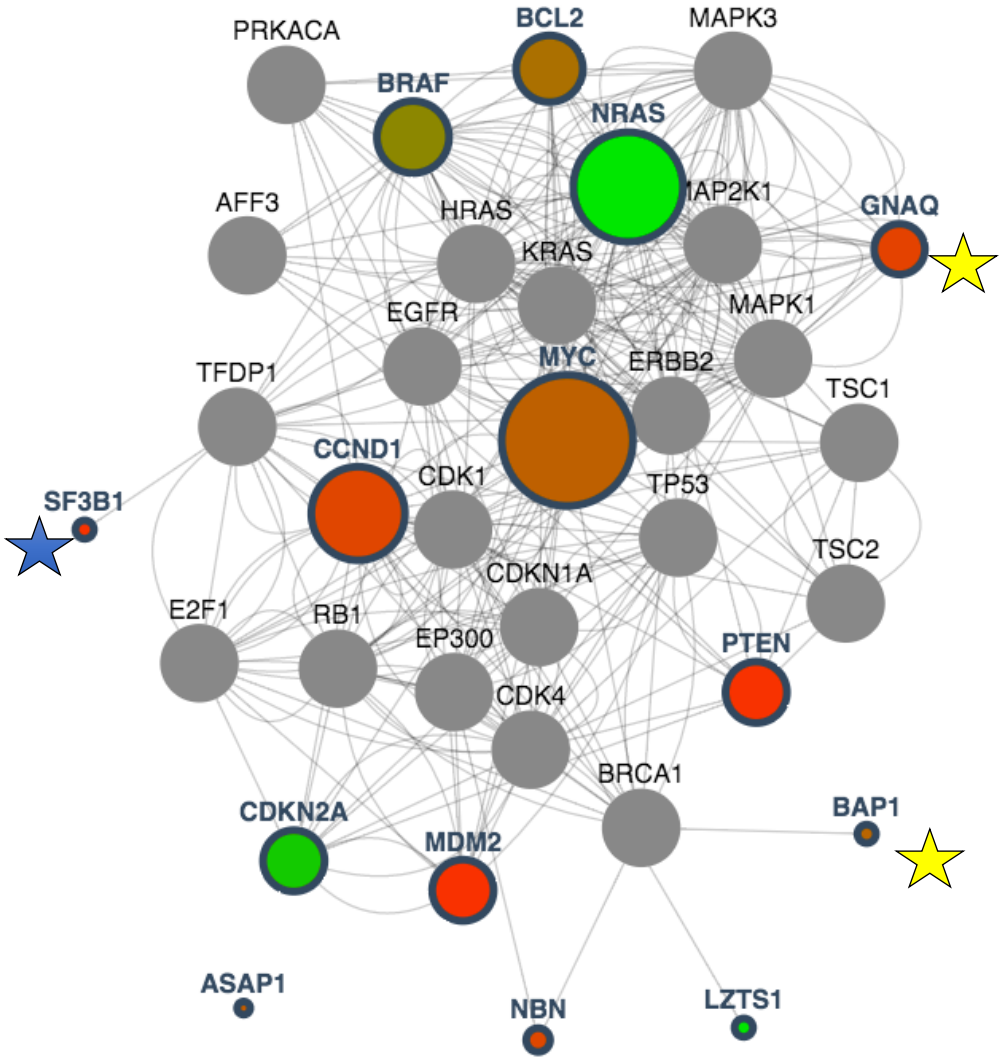
N of 1 Case Study: Pathogenic Variants



Patient Tumor Mutation Profile

GNAQ p.Q209L	**SF3B1 p.R625H**	**BAP1 p.G41S**
somatic	germline	somatic

N of 1 Case Study: Network Analysis

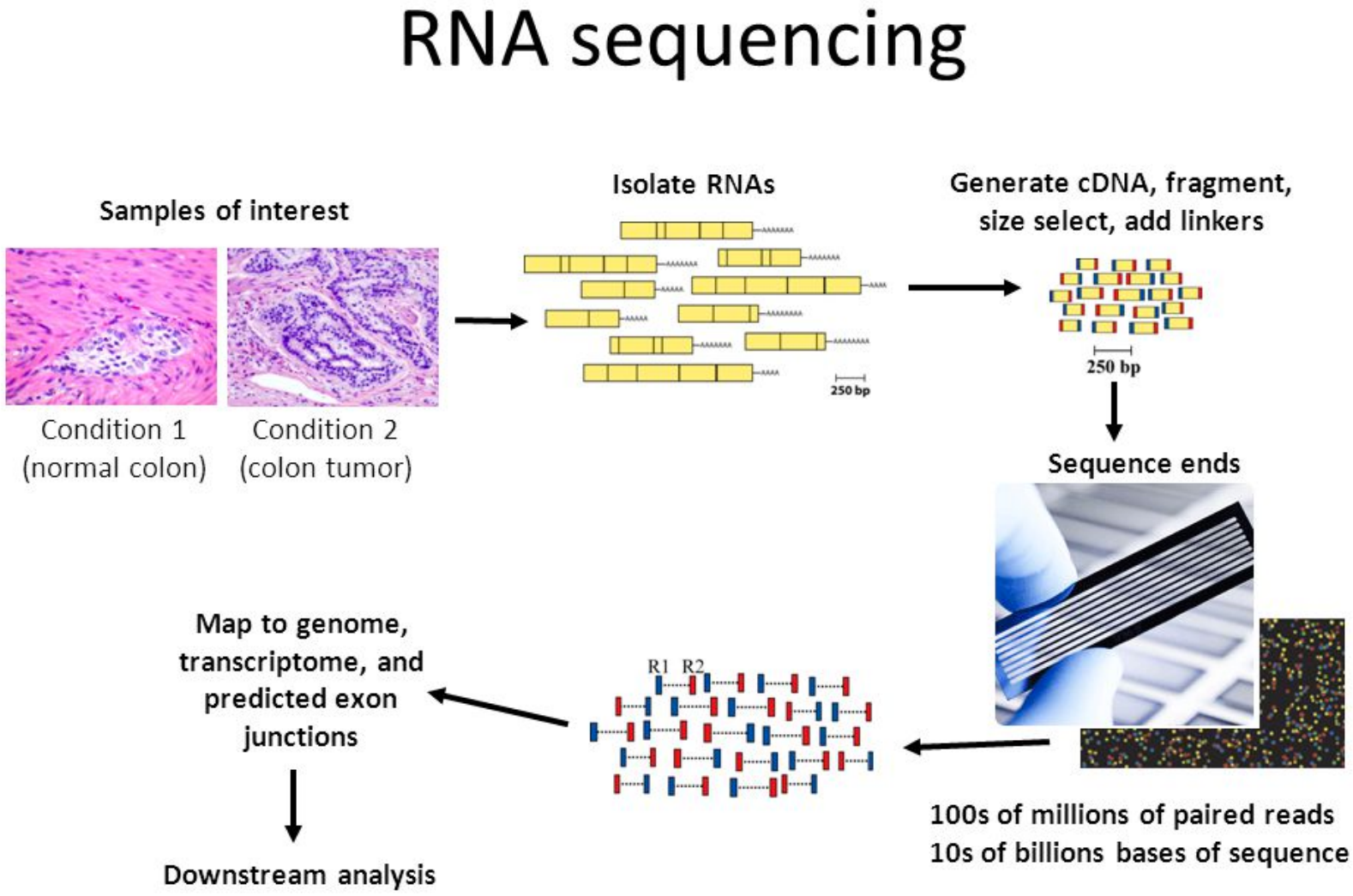


Legend

Red: High gene expression
Orange: Moderate gene expression
Green: Low gene expression
Blue Star: Germline variant
Yellow Star: Somatic variant

Network of UVM genes depicting relationships, relative gene expression and variants.

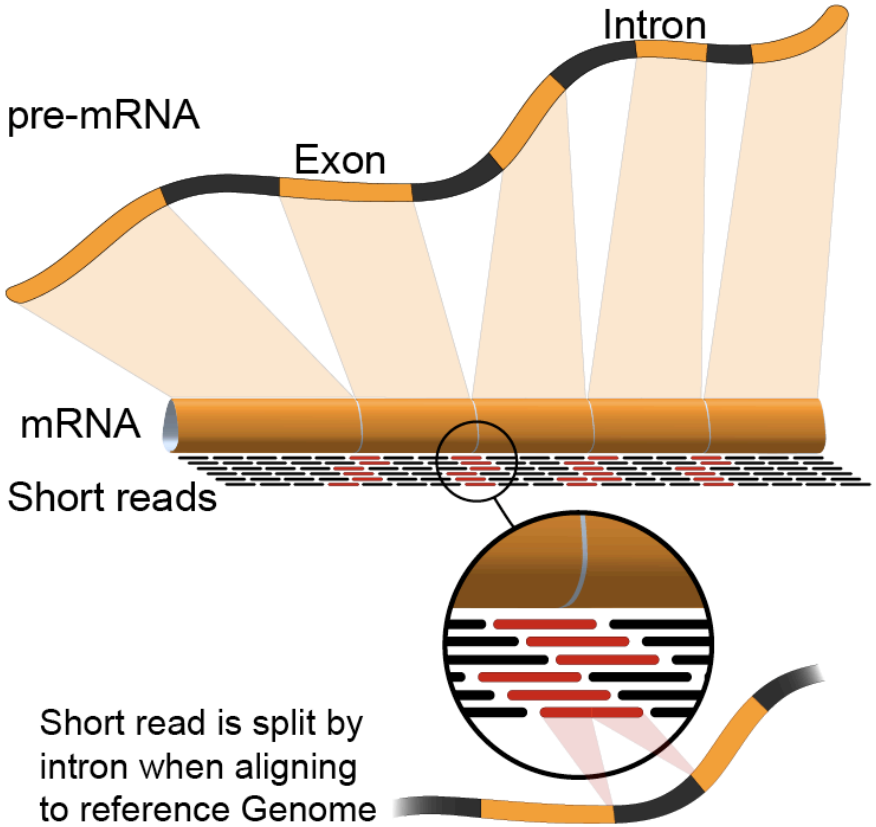
RNA-Seq Overview



Goldsby 2015 NESCent Academy

RNA sequencing Rationale

- Functional studies
  - An experimental condition may have a pronounced effect on gene expression
    - e.g. Drug treated vs. untreated cell line or KO vs WT
- Some molecular features can only be observed at the RNA level
  - Alternative isoforms, fusion transcripts, RNA editing
- Predicting transcript sequence from genome sequence is difficult
  - Alternative splicing, RNA editing, etc.

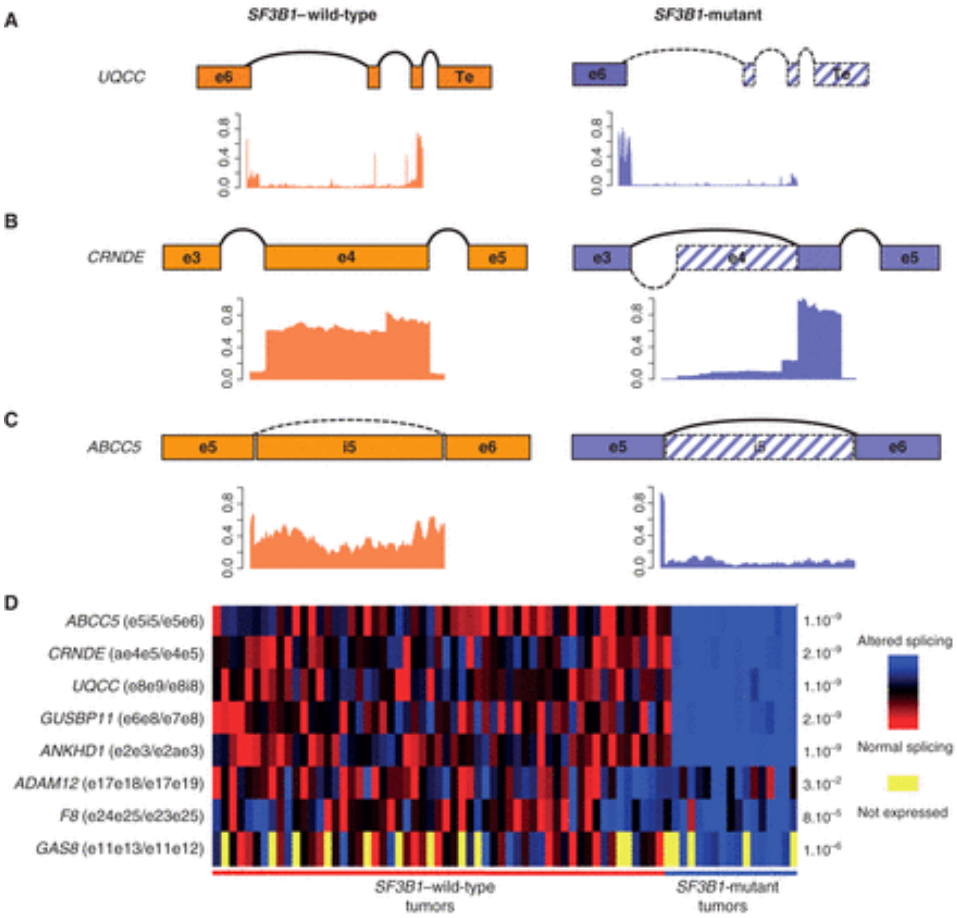


Adapted from Griffith & Griffith 2013

RNA sequencing Rationale

- Interpreting mutations that do not have an obvious effect on protein sequence
  - 'Regulatory' mutations that affect what mRNA isoform is expressed and how much
    - e.g. splice sites, promoters, exonic/intronic splicing motifs, etc.
- Prioritizing protein coding somatic mutations (often heterozygous)
  - If the gene is not expressed, a mutation in that gene would be less interesting
  - If the gene is expressed but only from the wild type allele, this might suggest loss-of-function (haploinsufficiency)
  - If the mutant allele itself is expressed, this might suggest a candidate drug target

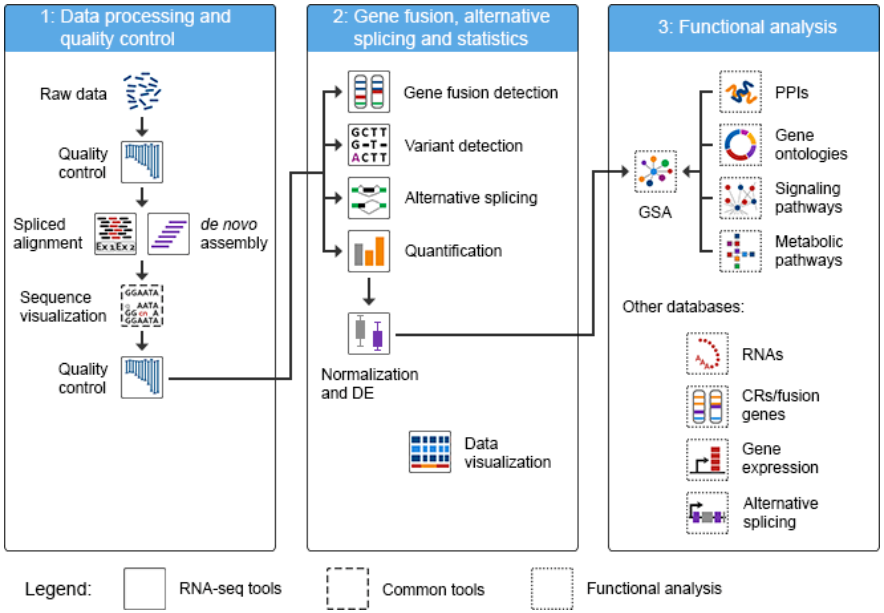
Adapted from Griffith & Griffith 2013



Furney et al. 2013 Cancer Discovery

RNA-Seq Analysis Goals

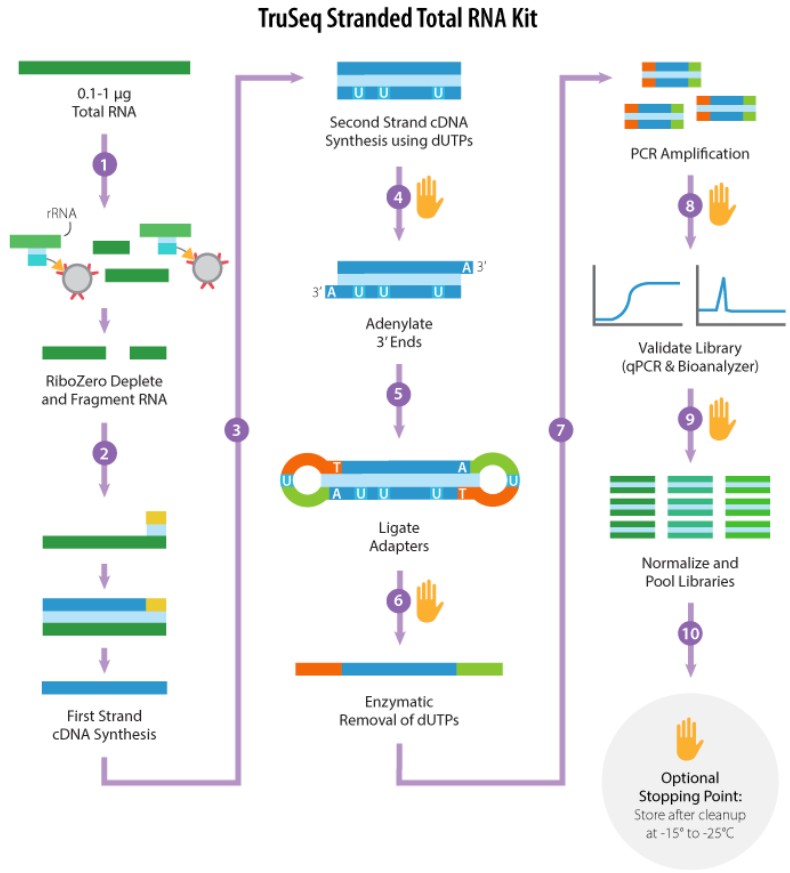
- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- lncRNA
- Allele specific expression
- Mutation discovery
- Fusion detection
- RNA editing
- miRNAseq
- Single cell



Omictools.com

RNA-seq library preparation and sequencing considerations

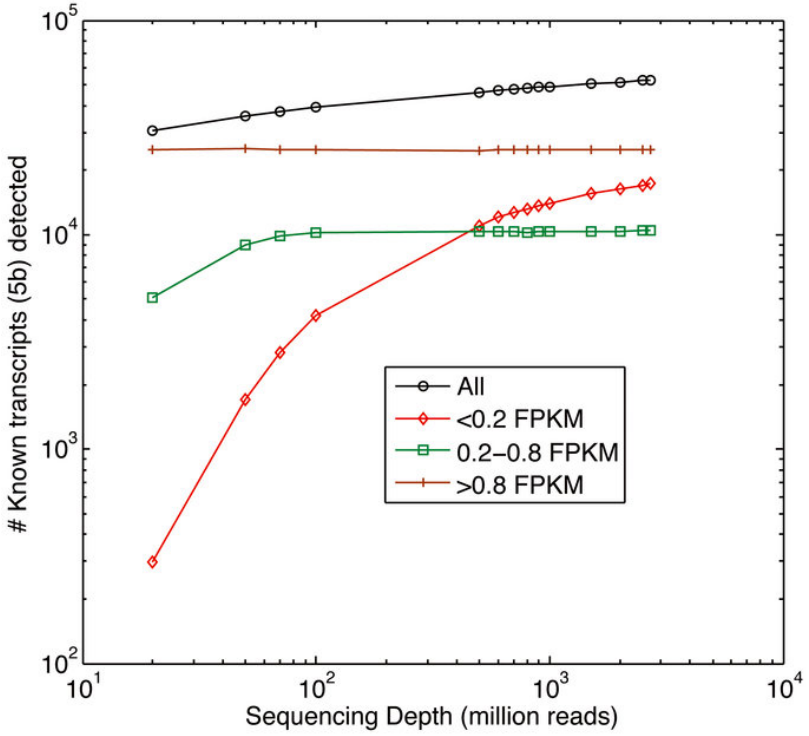
- Quality control and library-size normalization
  - Adding exogenous reference transcripts ('spike-ins')
- Minimize bias
  - Use of adapters with random nucleotides at the extremities
  - Use of chemical-based fragmentation instead of RNase III-based fragmentation
- Minimize batch effects
  - Randomize samples across library preparation batches and lanes
  - Or, individually barcode samples and include all samples in each lane



Illumina

RNA-Seq Experimental Design

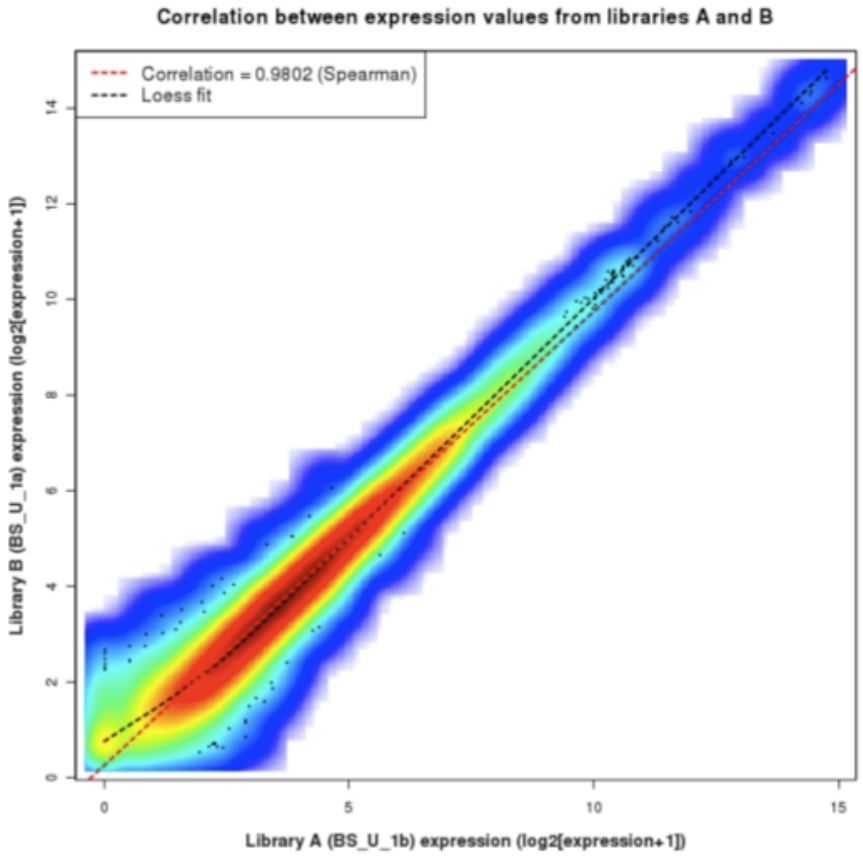
1. RNA extraction protocol
2. Stranded protocols
3. Single-end (SE) vs paired-end (PE) reads
4. Sequencing Depth aka library size
5. **Number of replicates**



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
  - Correlation Coefficient 0.92-0.98



Adapted from Griffith & Griffith 2013

RNA-Seq Experimental Design: Number of Replicates

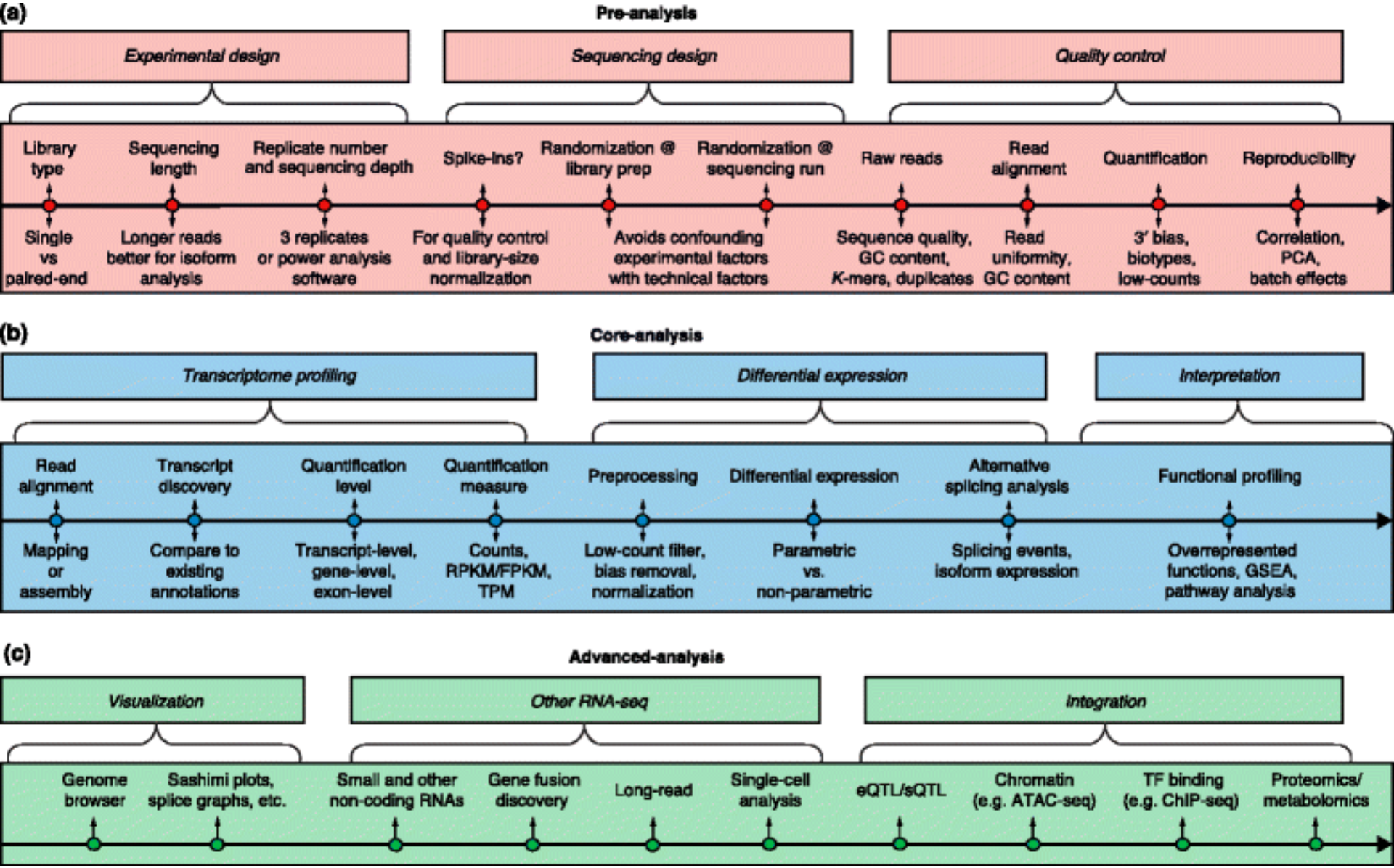
Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	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 %

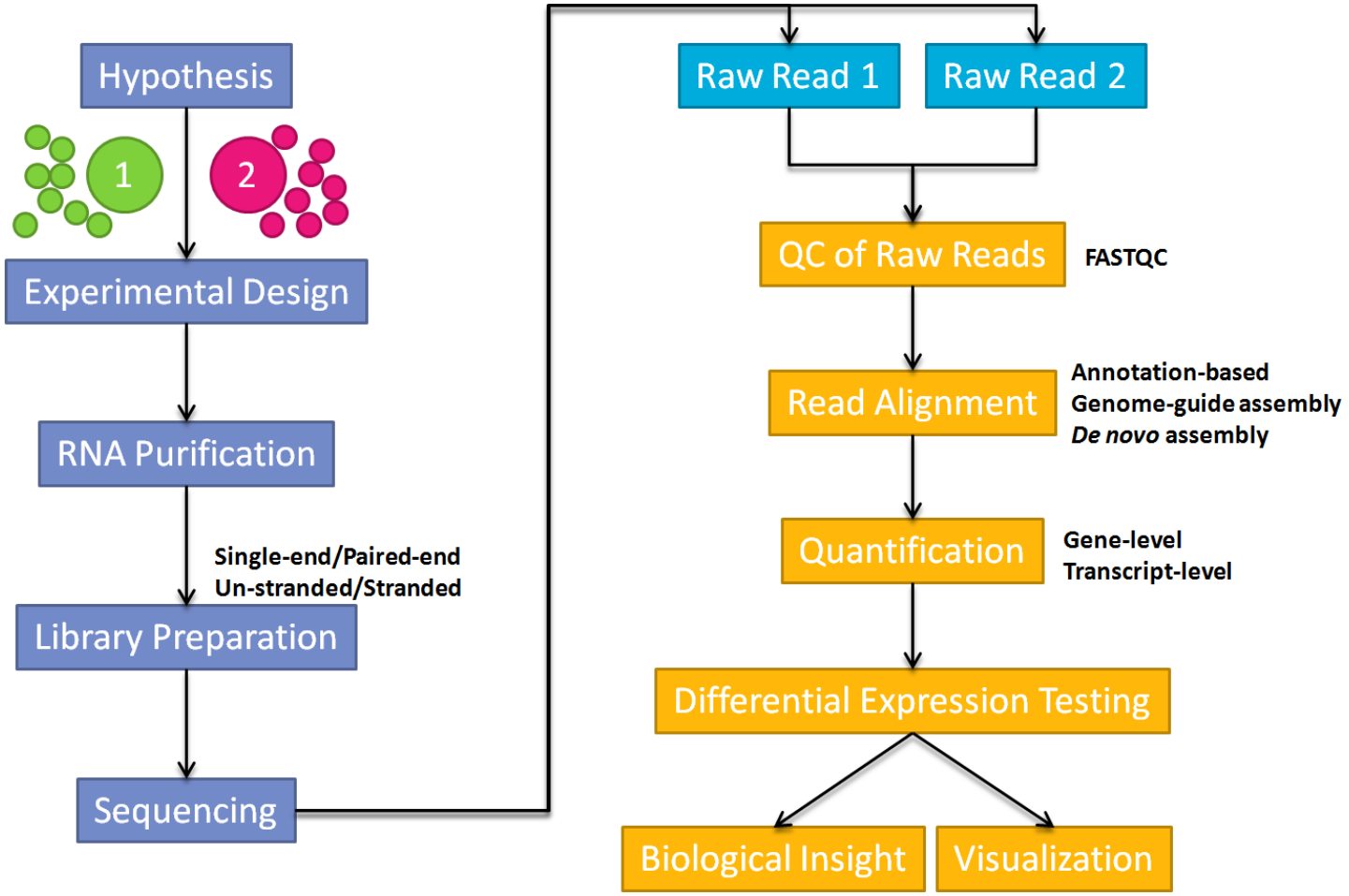
Example of calculations for the probability of detecting differential expression in a single test at a significance level of 5 %, for a two-group comparison using a Negative Binomial model, as computed by the RNASeqPower package of Hart et al. [190]. For a fixed within-group variance (package default 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



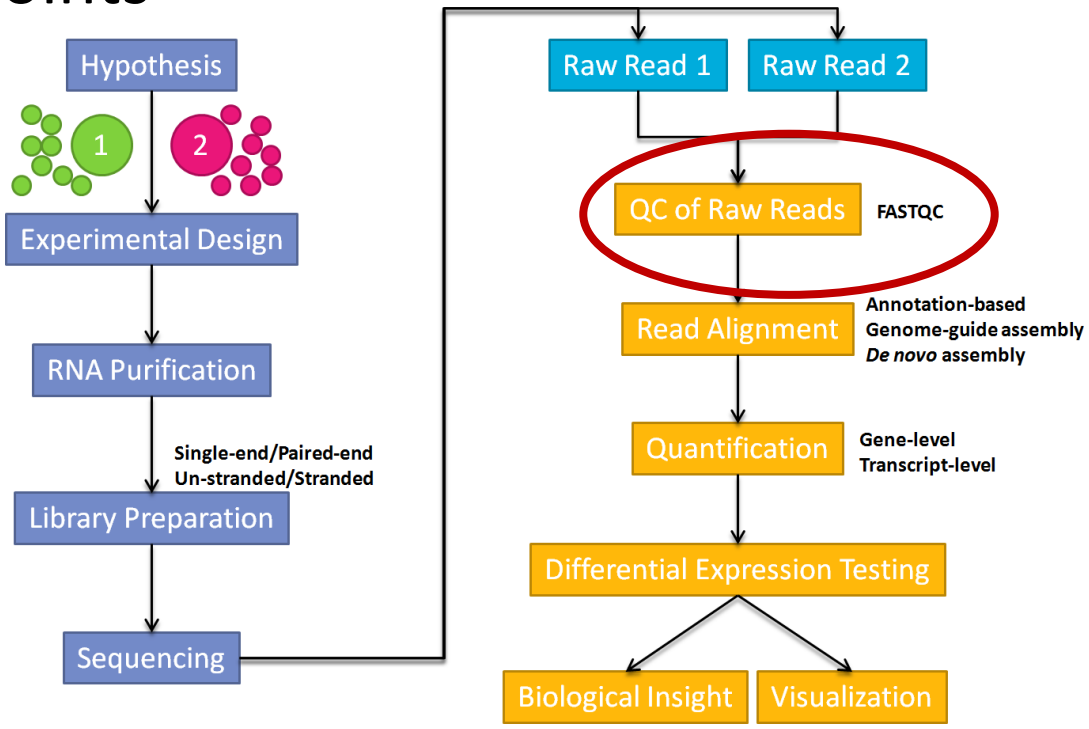
RNA-Seq Analysis Overview



Lin 2015 NGS Workshop

RNA-Seq Analysis QC

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

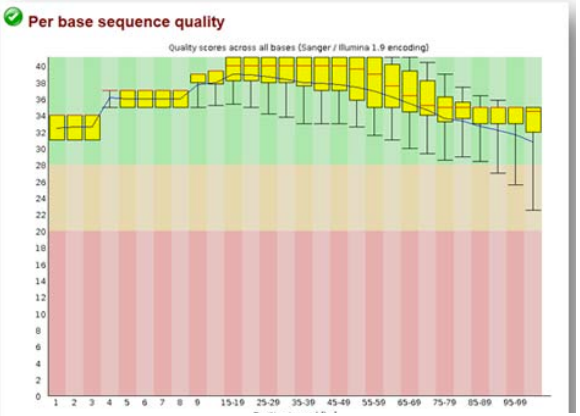


RNA-Seq Analysis QC

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

FASTQC Report

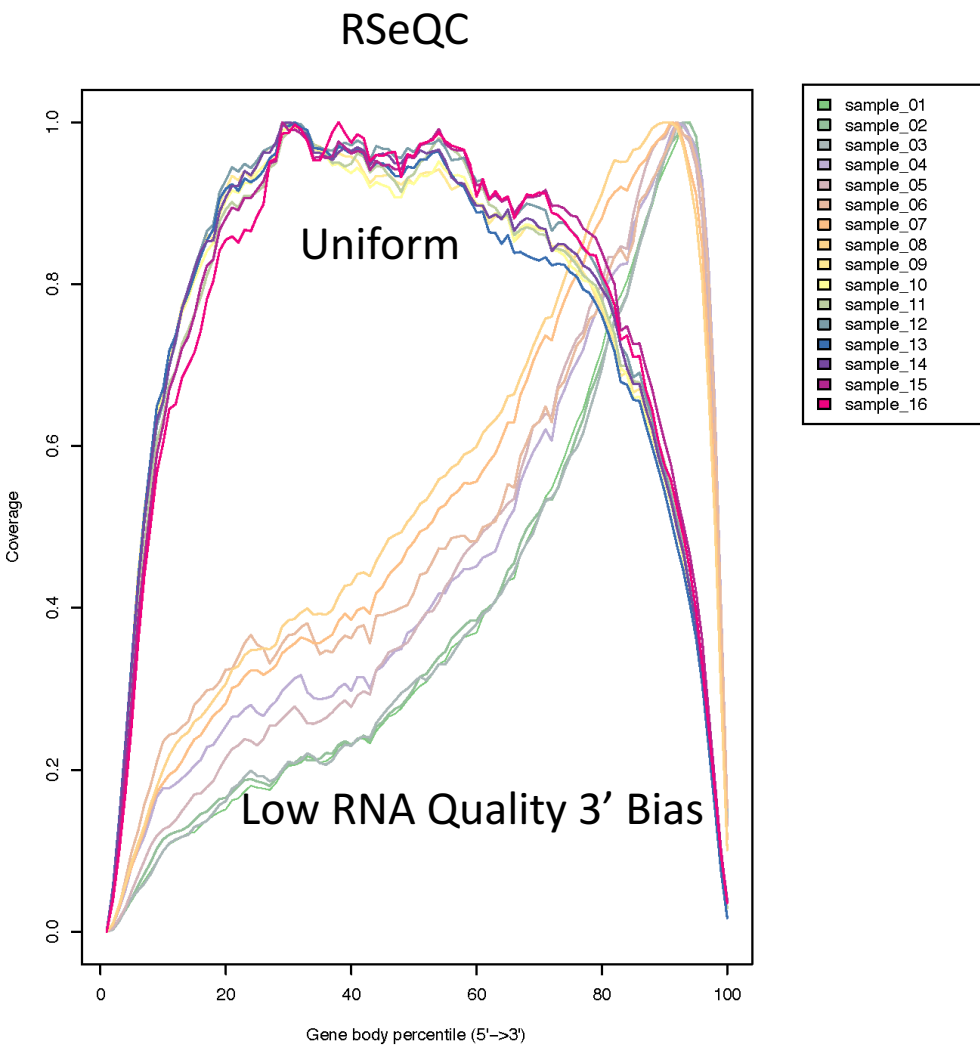
Good Sequence Quality



Poor Sequence Quality at 3' Ends

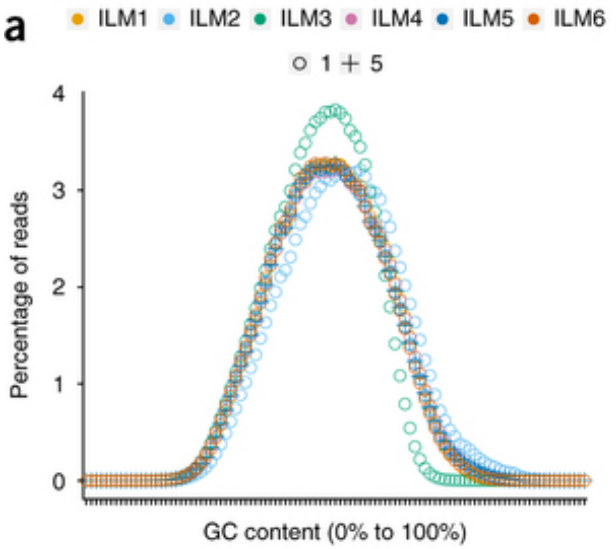
RNA-Seq Analysis QC

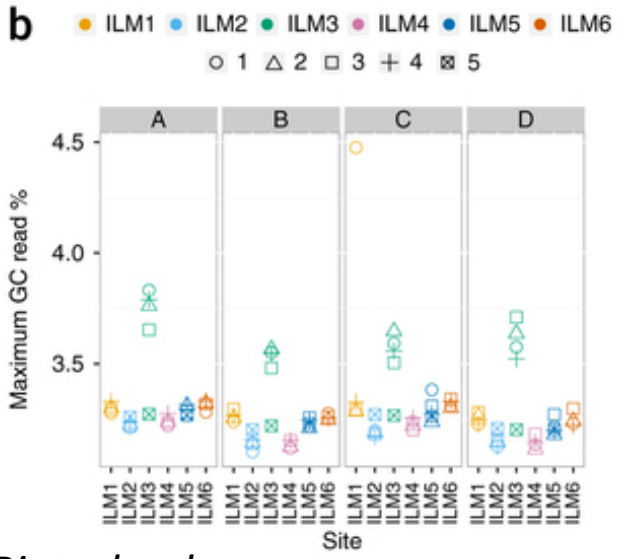
- Quality-control checkpoints
  - Raw reads
  - Read alignment
    - % Mapped reads
    - Uniformity of read coverage on exons and mapped strand
    - Ideal: 70-90% mapped to human genome
    - Tools: Picard, RSeQC, Qualimap
    - Goals:
      - Global indicator of overall sequencing accuracy and presence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



RNA-Seq Analysis QC

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

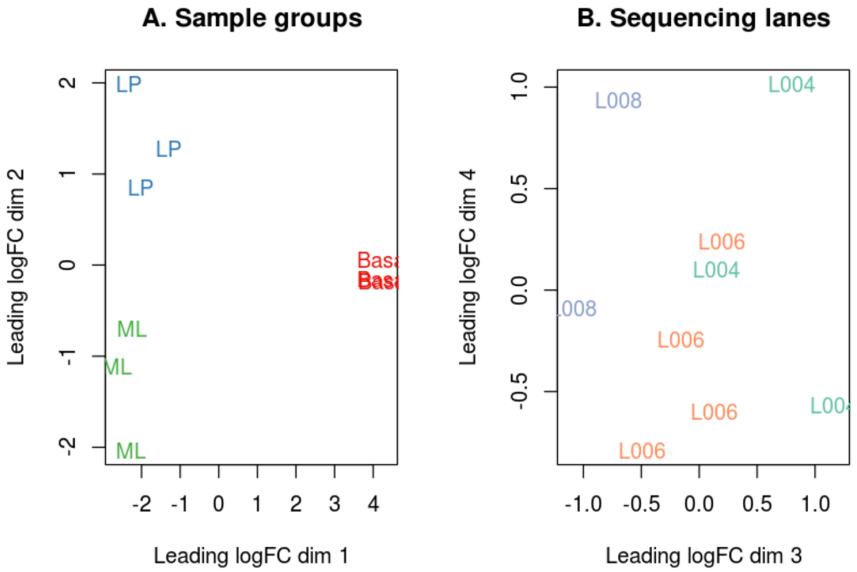




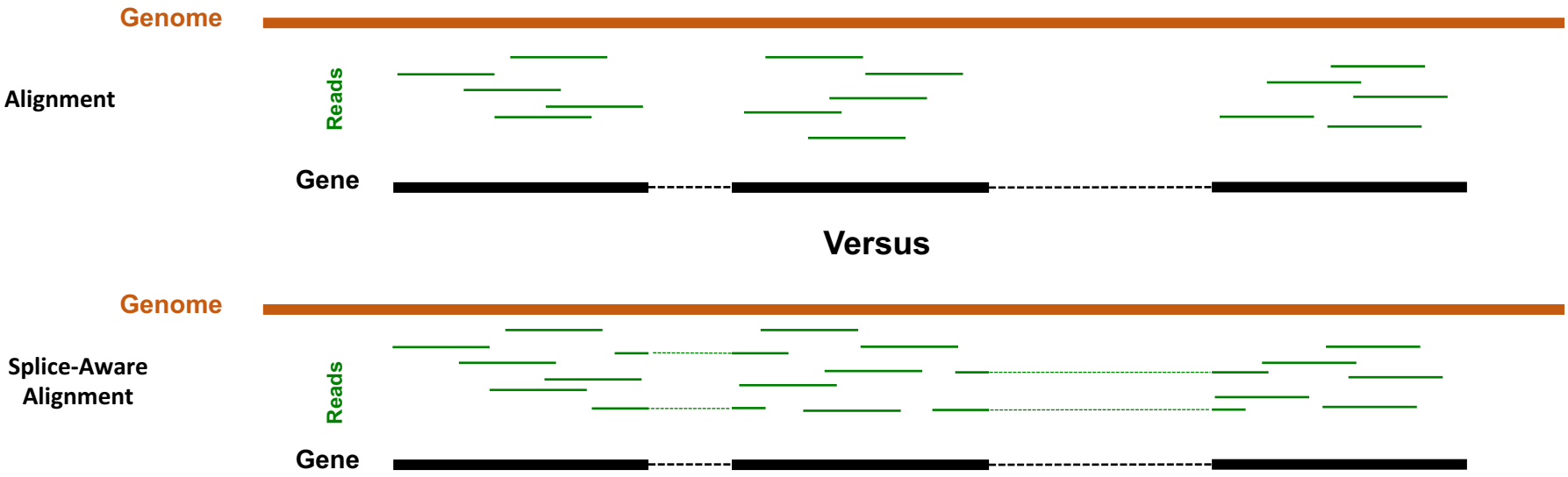
Li *et al*. 2014. *Nature Biotechnology*

RNA-Seq Analysis QC

- Quality-control checkpoints
  - Raw reads
  - Read alignment
  - Quantification
  - Reproducibility
    - Checking on reproducibility among replicates and for possible batch effects
    - Tool: Principal component analysis (PCA)
    - Goal: Assess global quality of RNA-seq dataset

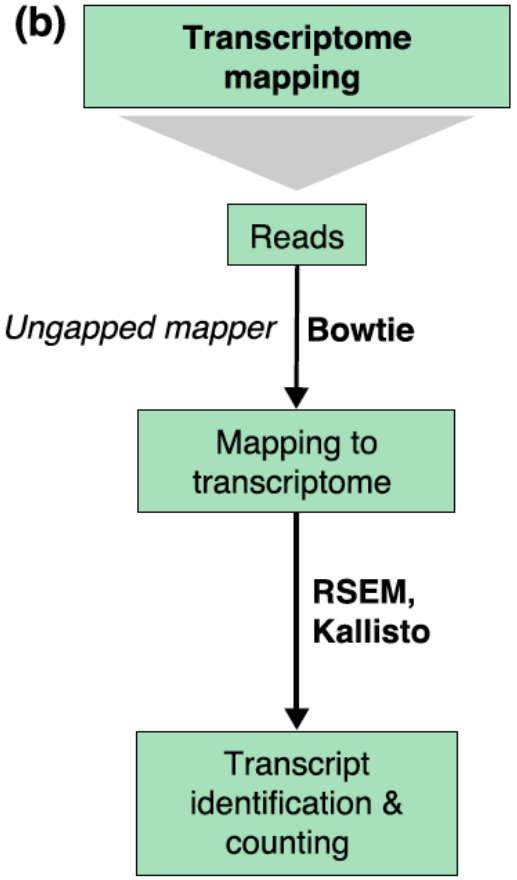


RNA-Seq Analysis -- Alignment



RNA-Seq Analysis
Alignment – Transcriptome Mapping

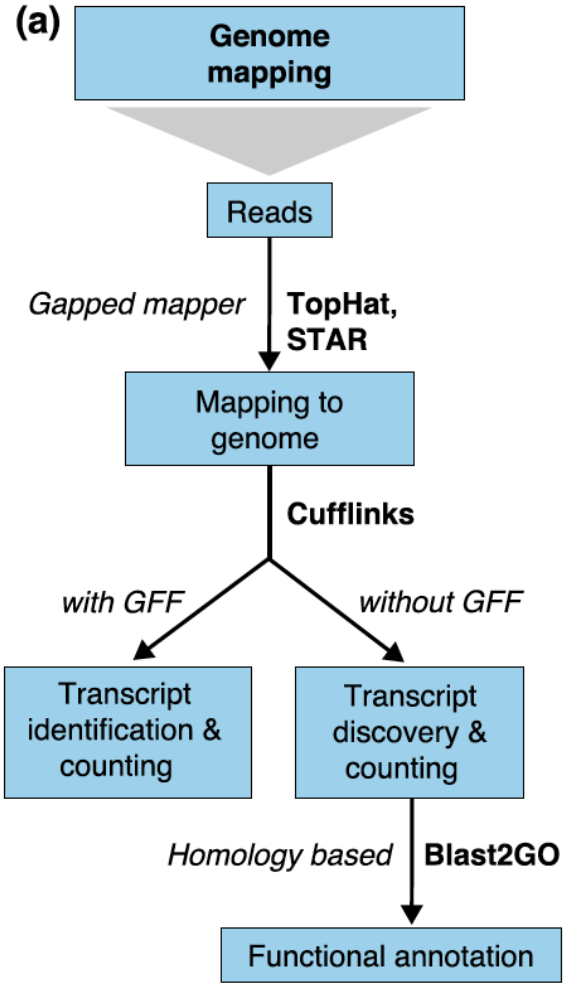
- Align reads to annotated human transcriptome
- Faster than alignment to genome
- Does not allow de novo transcript discovery
- Uniquely mapped reads
- Multi-mapped reads
  - Arises more often than with genome mapping as a read may map equally well to all gene isoforms in the transcriptome that share the exon.



Conesa *et al.* 2016 *Genome Biology*

RNA-Seq Analysis
Alignment – Genome Mapping

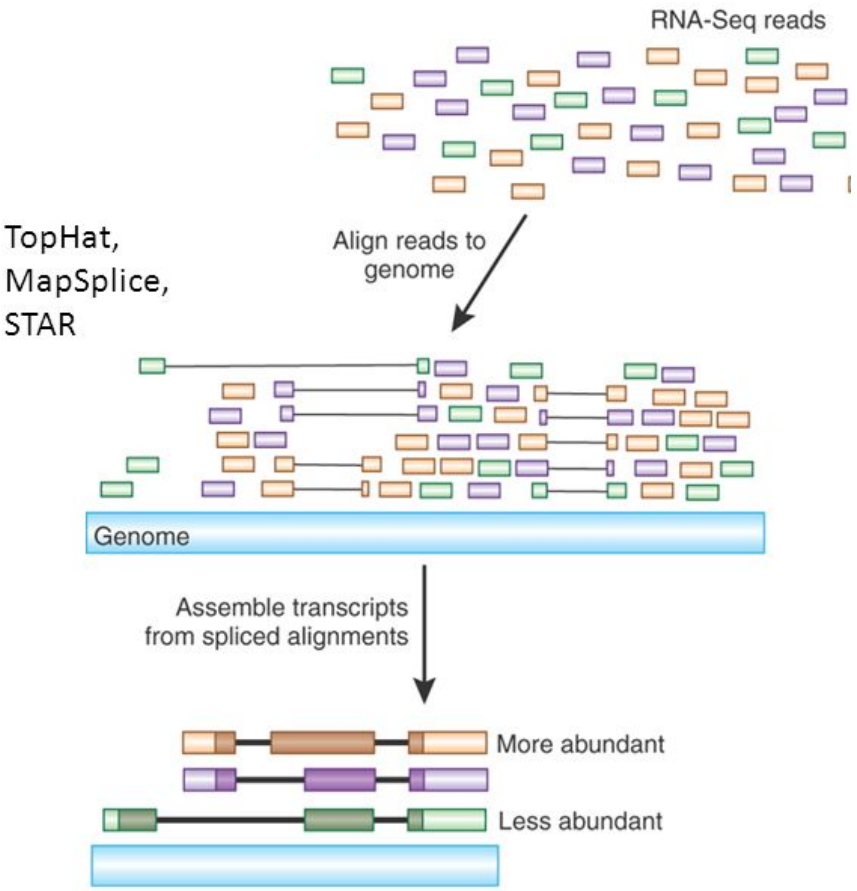
- Align reads to annotated human genome
- Requires use of a gapped or spliced mapper as reads may span splice junctions
- Uniquely mapped reads
- Multi-mapped reads
  - Primarily due to repetitive sequences or shared domains of paralogous genes
  - Account for a significant fraction of mapped output and should not be discarded



Conesa *et al.* 2016 *Genome Biology*

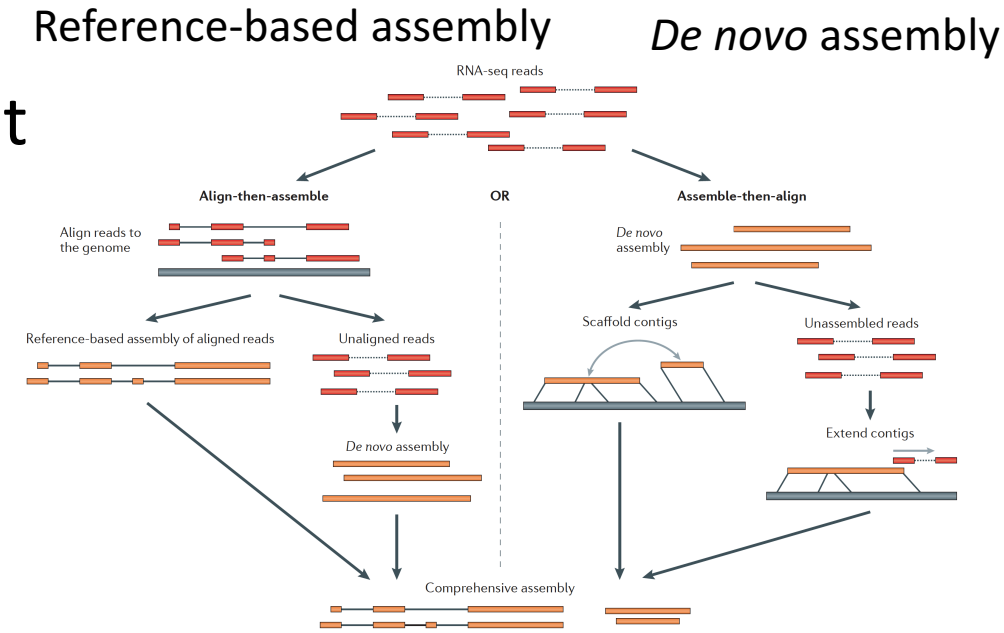
RNA-Seq Analysis
Alignment -- Important Parameters

- 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

- Paired End (PE) strand-specific sequencing and long reads more informative
- Reference-based transcript reconstruction
- De novo transcript reconstruction
- Long-read technologies (PacBio) can sequence complete transcripts



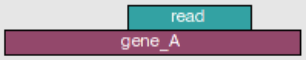
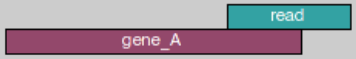
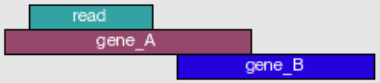
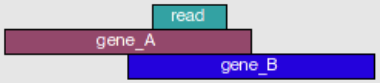
Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

RNA-Seq Analysis
Read Quantification

- Most common application of RNA-seq
  - Estimate gene and transcript expression
- Based on the number of reads that map to each transcript/gene
- Aggregation of raw counts of mapped reads
- Gene-level vs Transcript-level expression algorithms

RNA-Seq Analysis
Gene-level Quantification

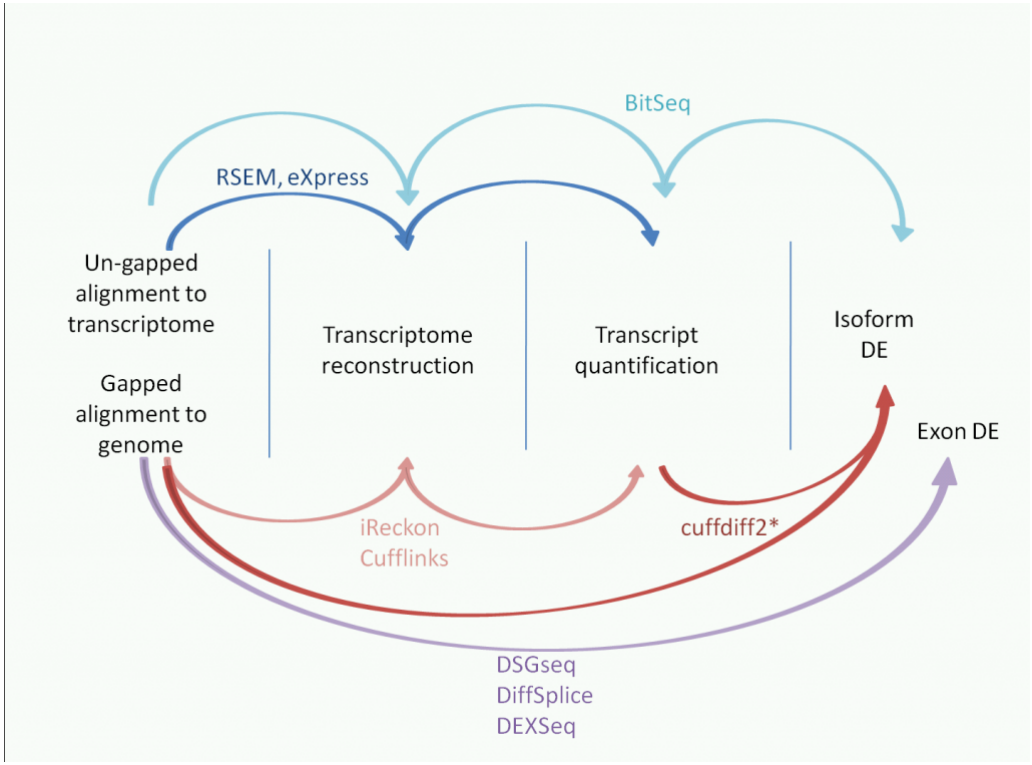
- Aggregation of raw counts of mapped reads
  - HTSeq-count or featureCounts
  - Gene-level approach based on GTF gene coordinates
  - Discard multimappers

	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

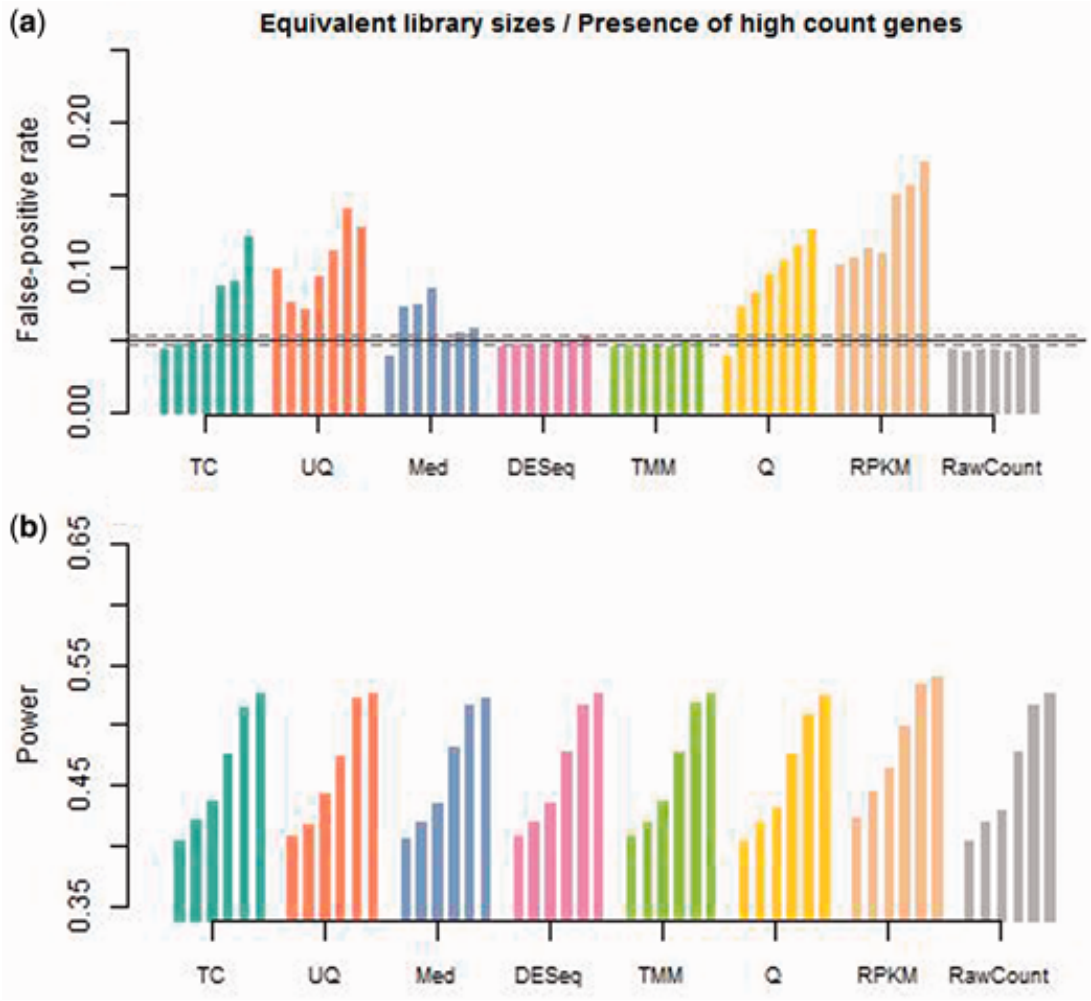
HTSeq-count

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



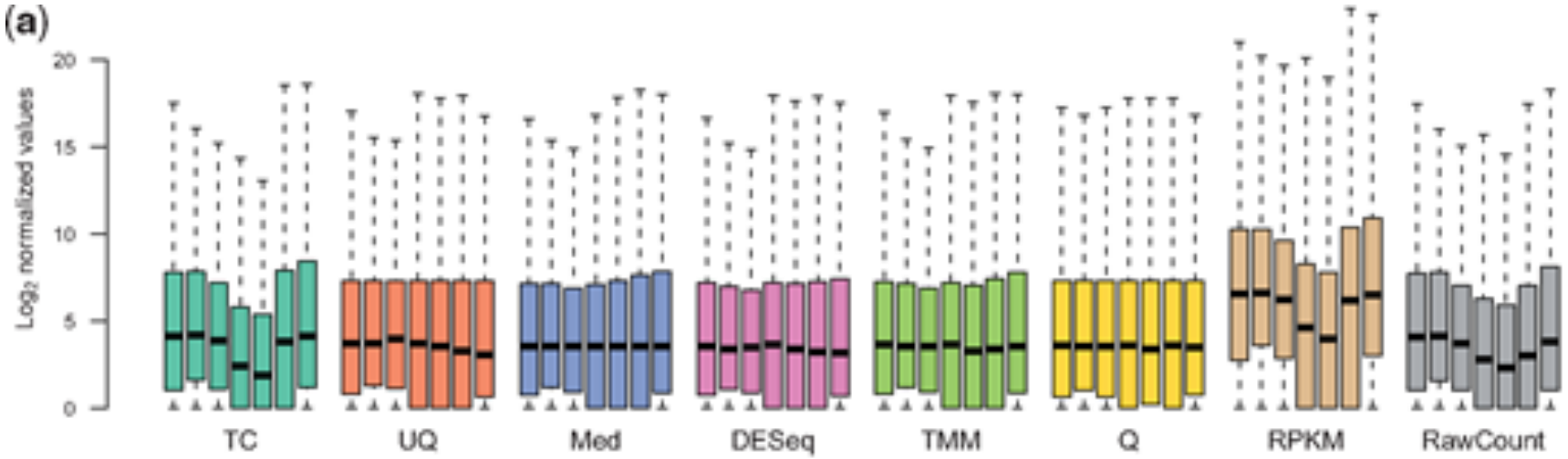
Dillies et al. 2013 *Briefings in Bioinformatics*

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

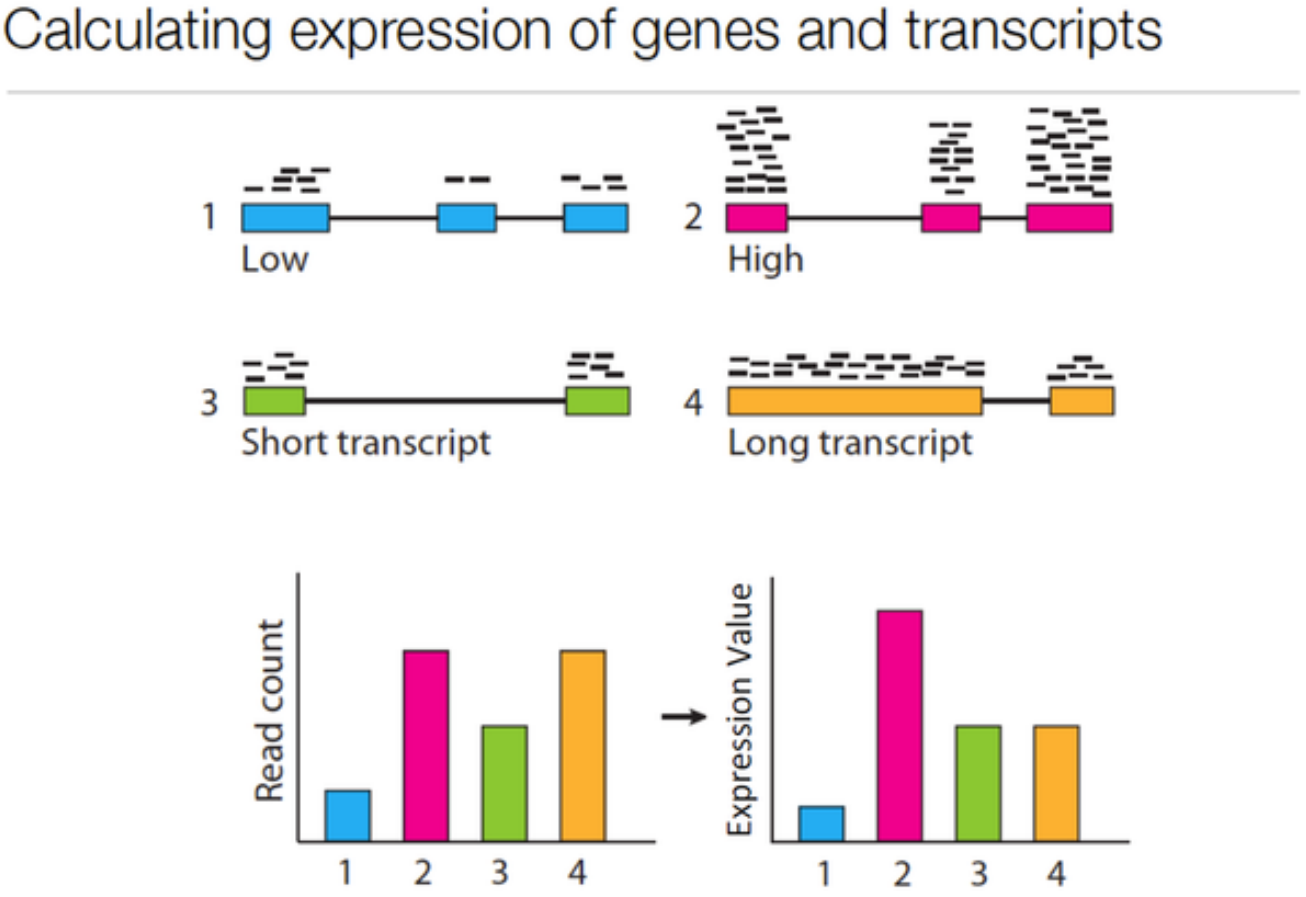
RNA population 1

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

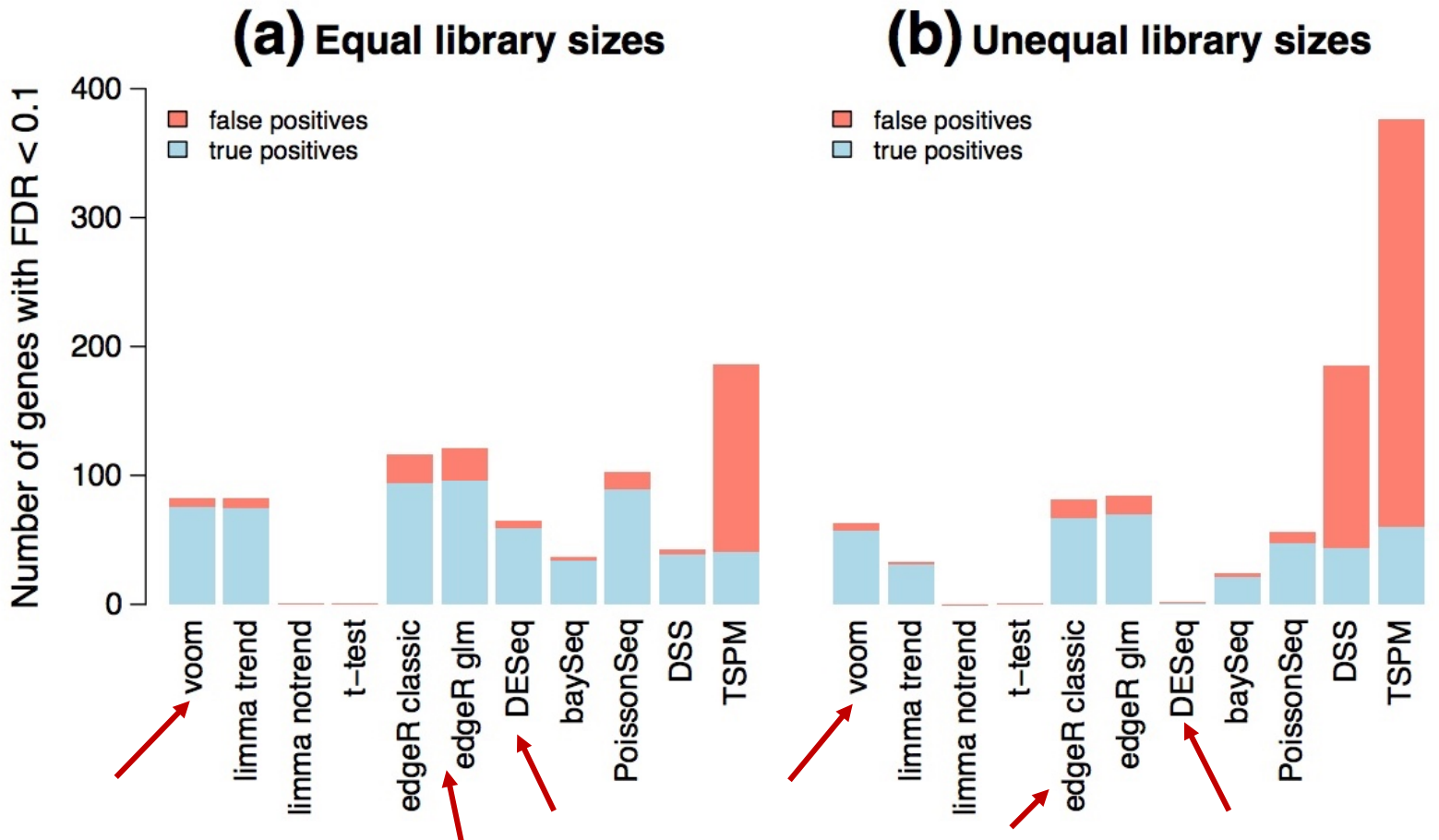


Figure 4 Power to detect true differential expression. Bars show the total number of genes that are detected as statistically significant (FDR < 0.1) **(a)** with equal library sizes and **(b)** with unequal library sizes. The blue segments show the number of true positives while the red segments show 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

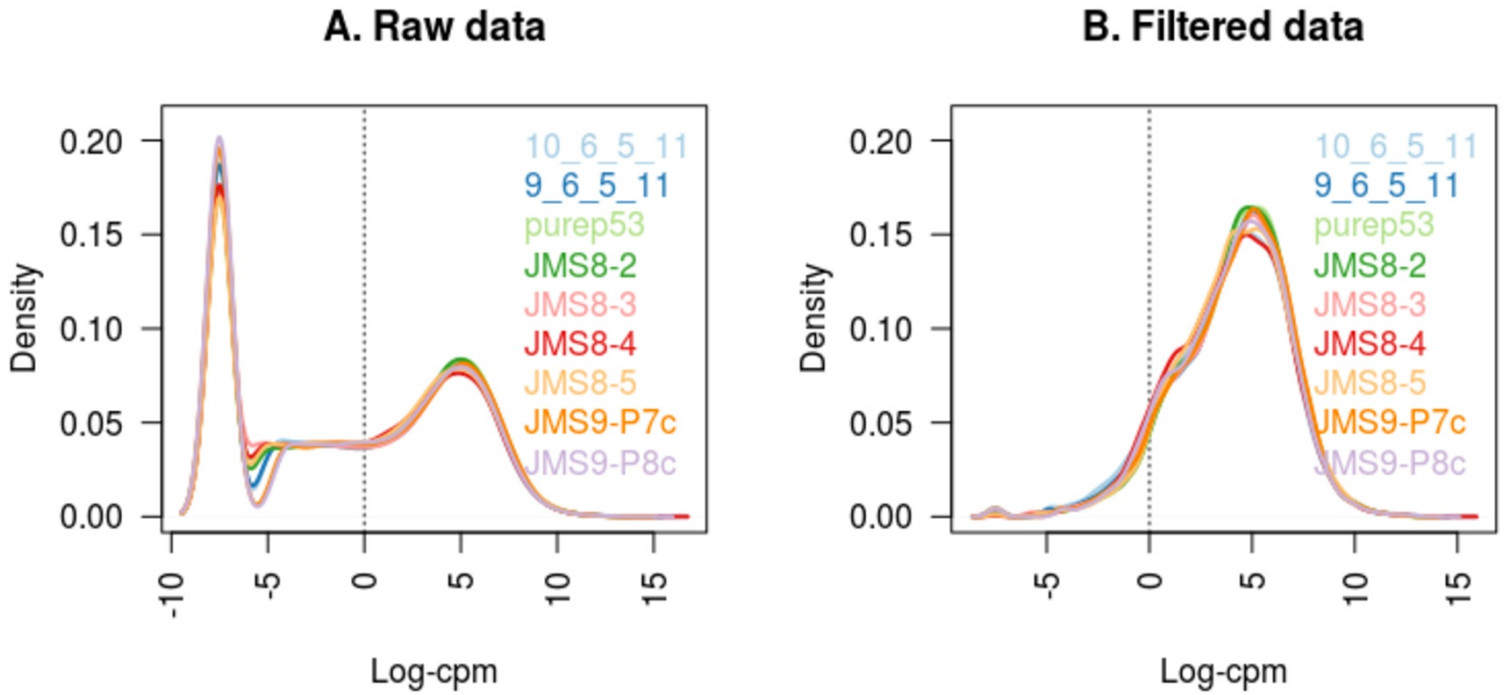
Data pre-processing
- Transformations from the raw-scale
- Removing genes that are lowly expressed
- Normalising gene expression distributions
- Unsupervised clustering of samples

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

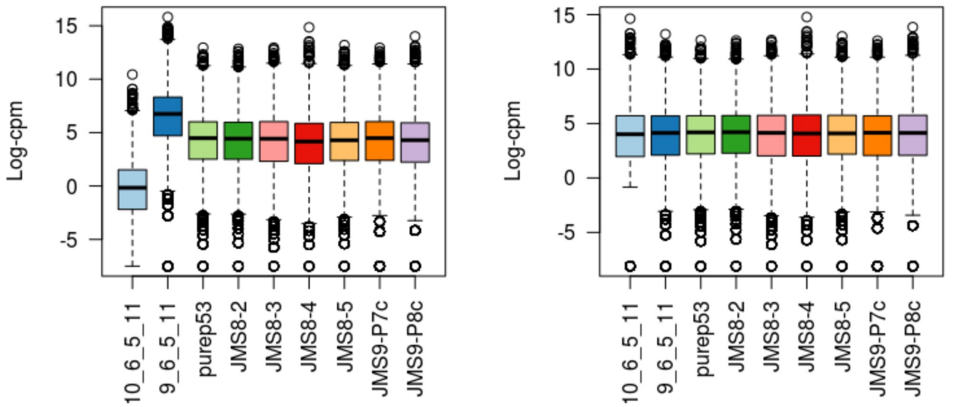
RNA-Seq – Differential Expression Analysis: Data Pre-processing

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)

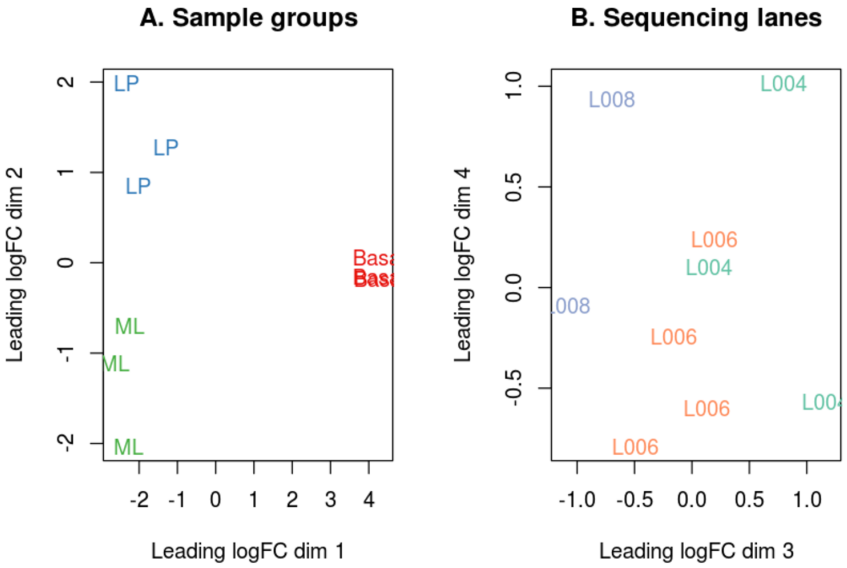


RNA-Seq – Differential Expression Analysis: Data Pre-processing

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

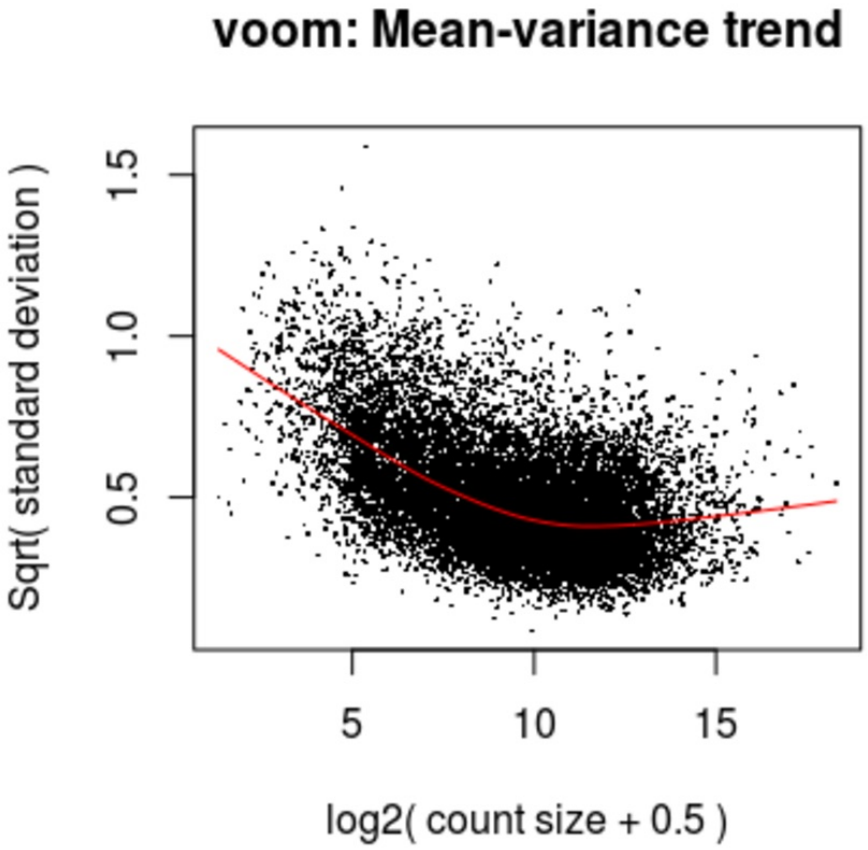
1. Create design matrix and contrasts

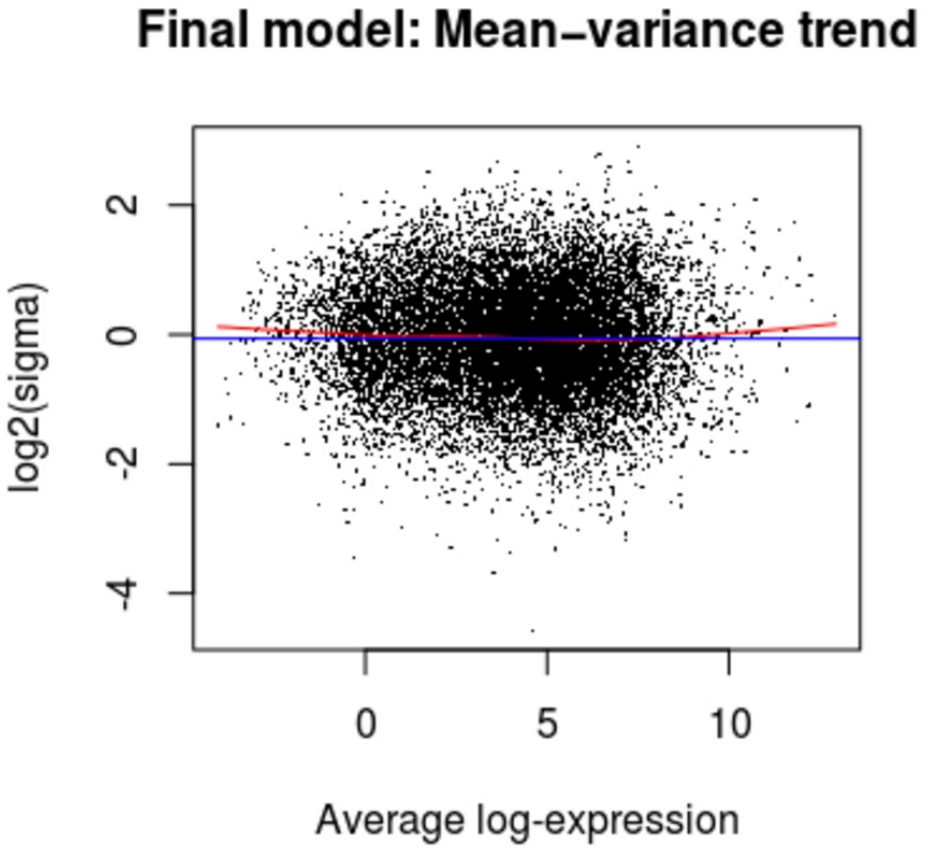
```
design <- model.matrix(~0+group+lane)
colnames(design) <- gsub("group", "", colnames(design))
design

##   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

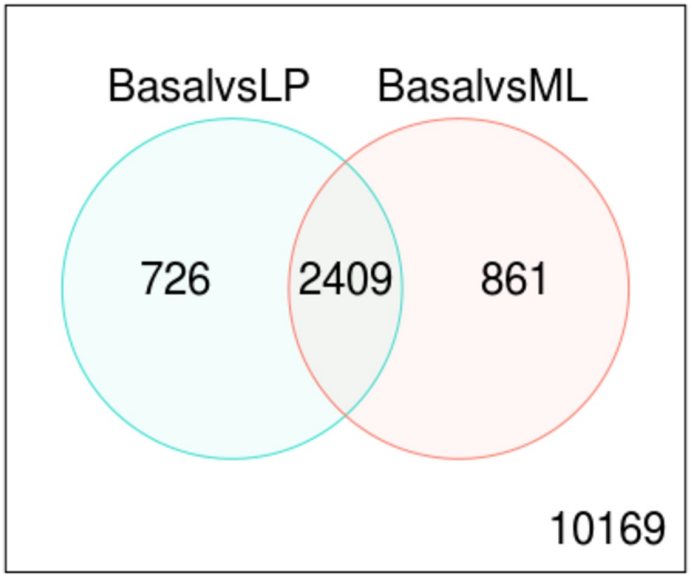




RNA-Seq – Differential Expression Analysis

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
```



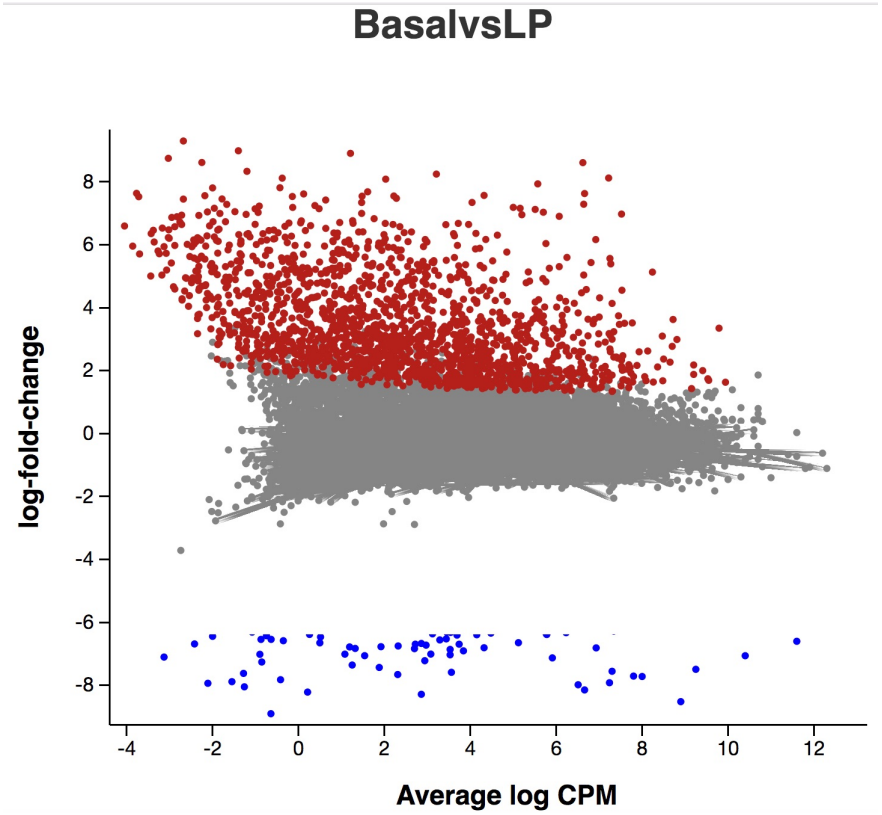
RNA-Seq – Differential Expression Analysis

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
```

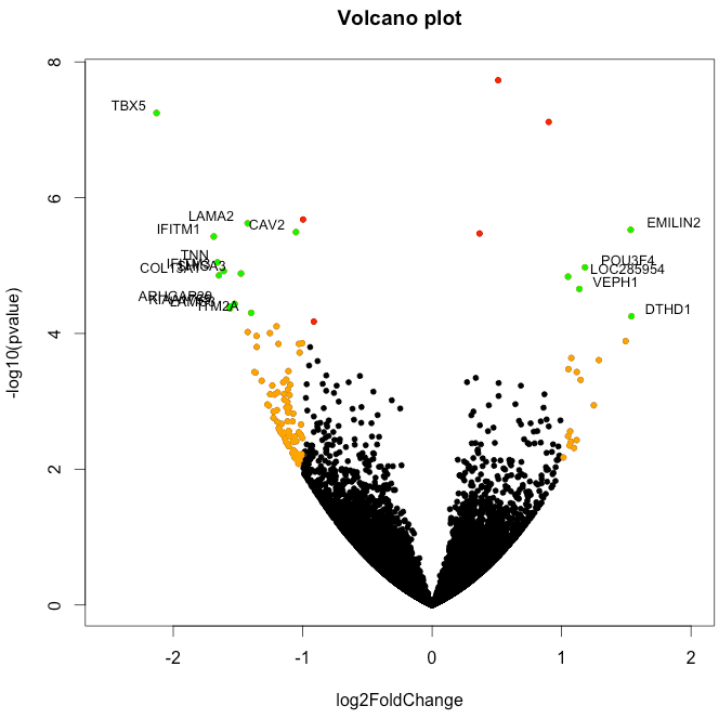
RNA-Seq – Differential Expression Analysis

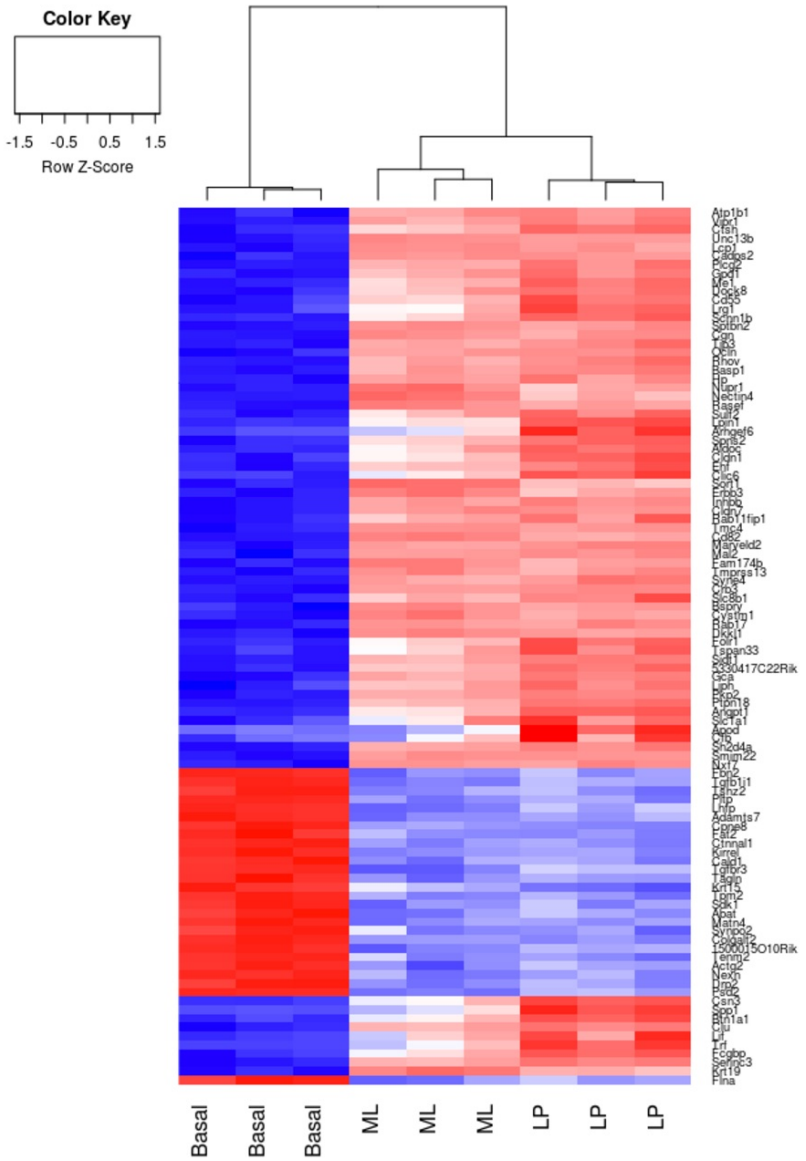
6. Useful graphical representations of differential expression



RNA-Seq – Differential Expression Analysis

6. Useful graphical representations of differential expression





Downstream Analysis & Interpretation

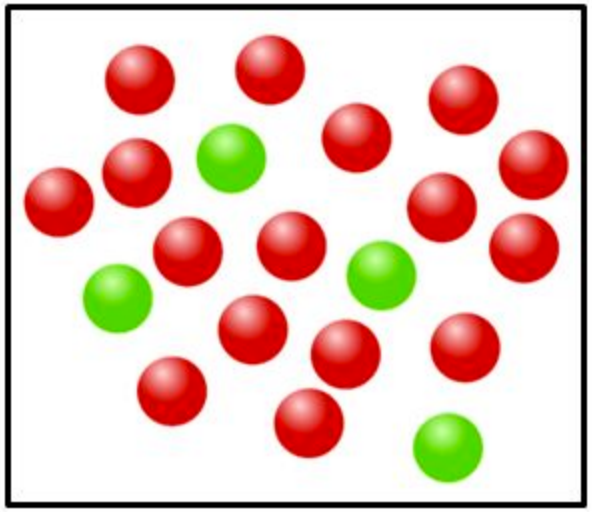
- Hypergeometric test and overrepresentation analysis
- Functional Gene Set Enrichment Analysis
- Pathway Analysis
- Visualize Alignments with IGV
- Network Analysis

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:



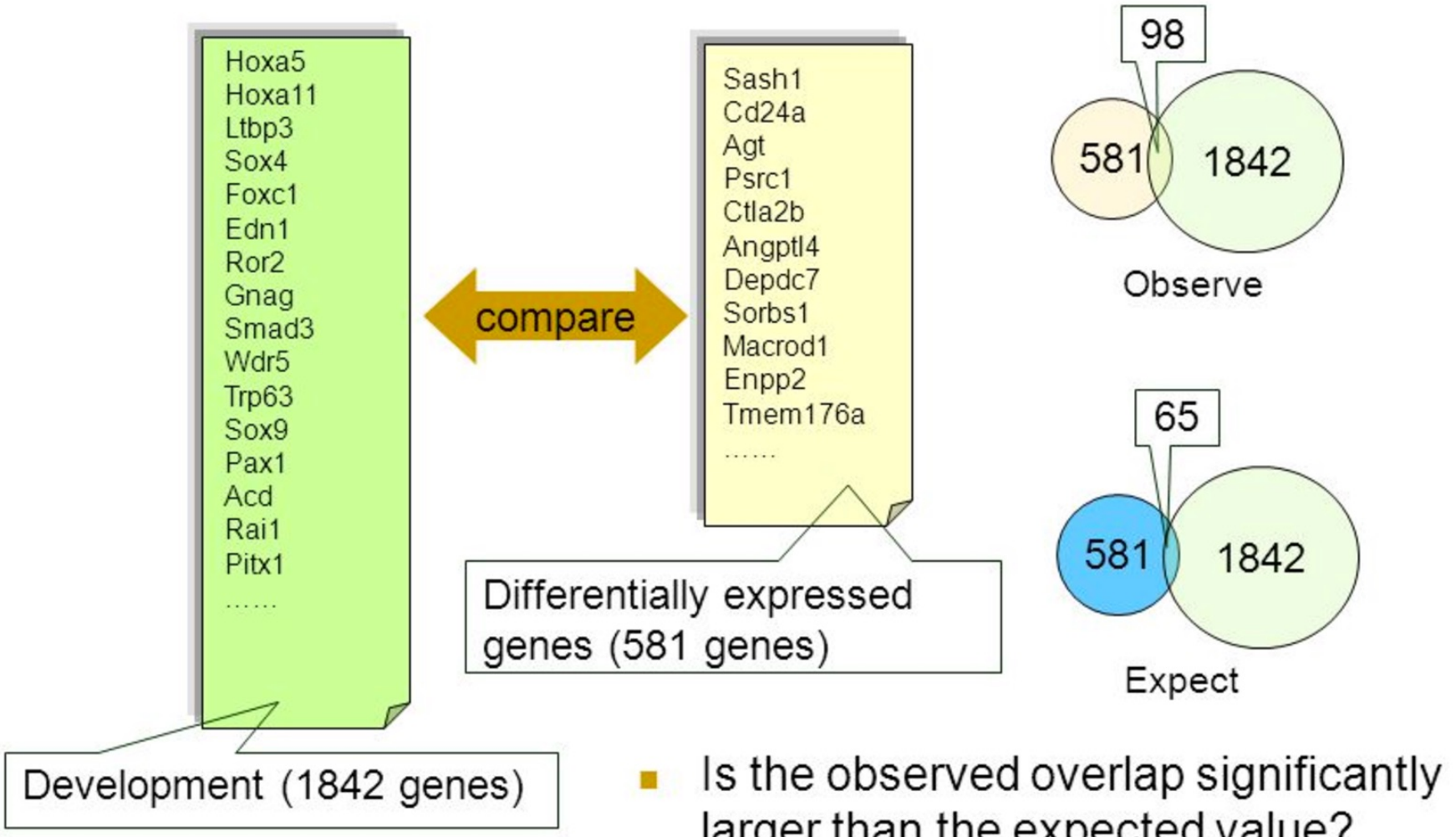
Imagine that there are 4 green and 16 red marbles in a box.

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

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

Babraham Bioinformatics

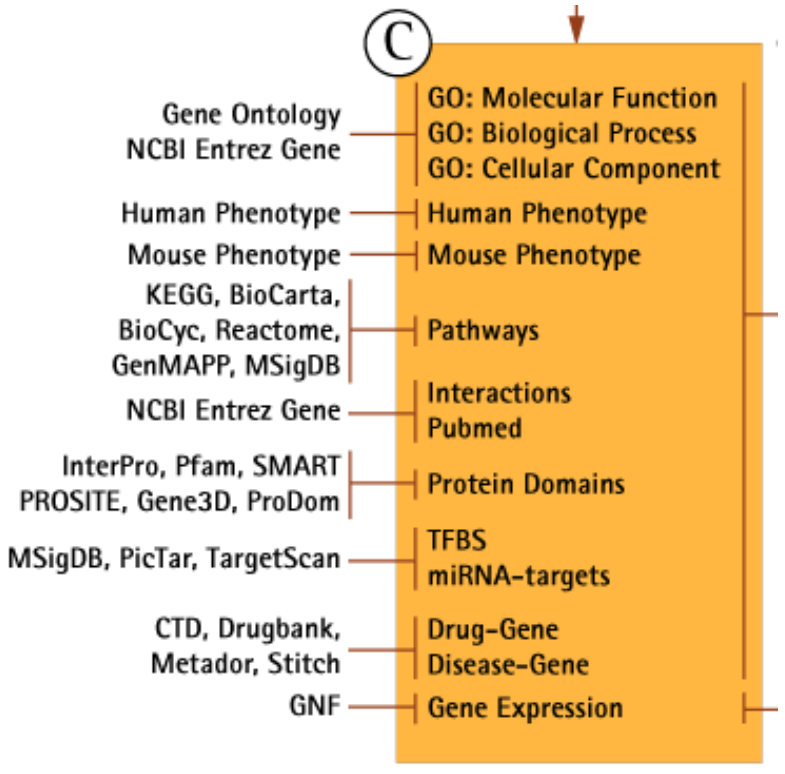
Hypergeometric Test for Overrepresentation Analysis



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

Downstream Analysis & Interpretation: Functional Enrichment

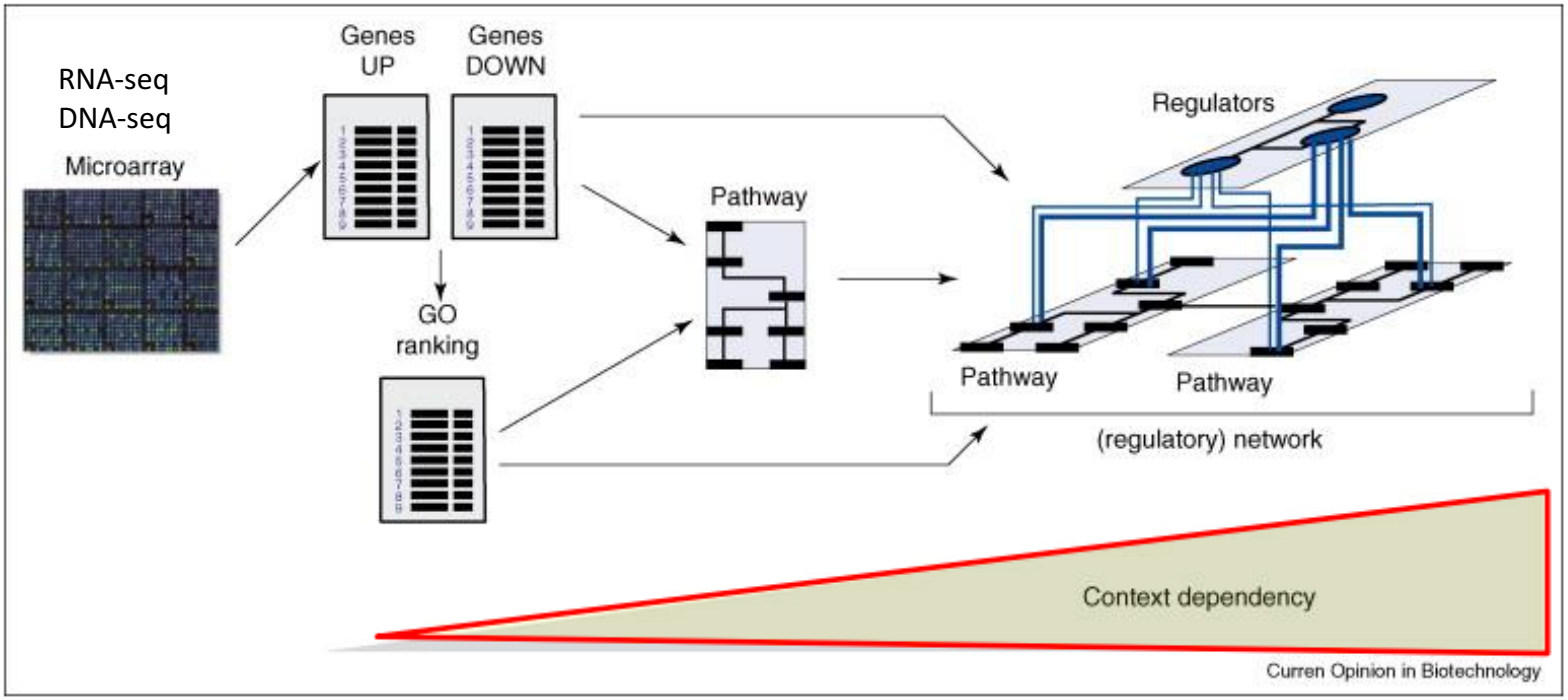
- Gene list enrichment analysis (Hypergeometric test) based on functional annotations
- Tools
  - ToppGene, GSEA, Webgestalt, DAVID



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

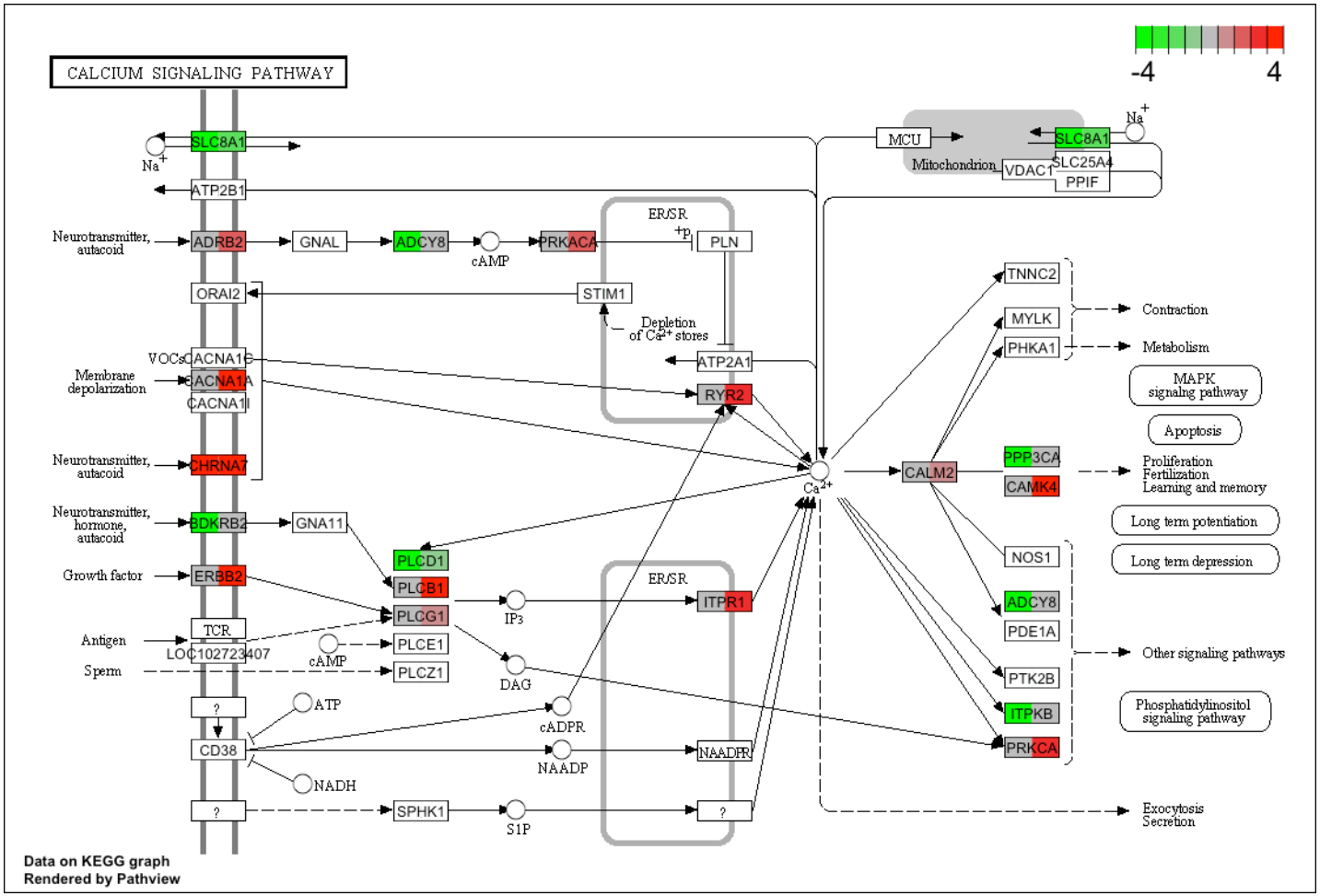
Downstream Analysis & Interpretation: Pathway Analysis



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

Werner 2008 Current Opinion Biotechnology

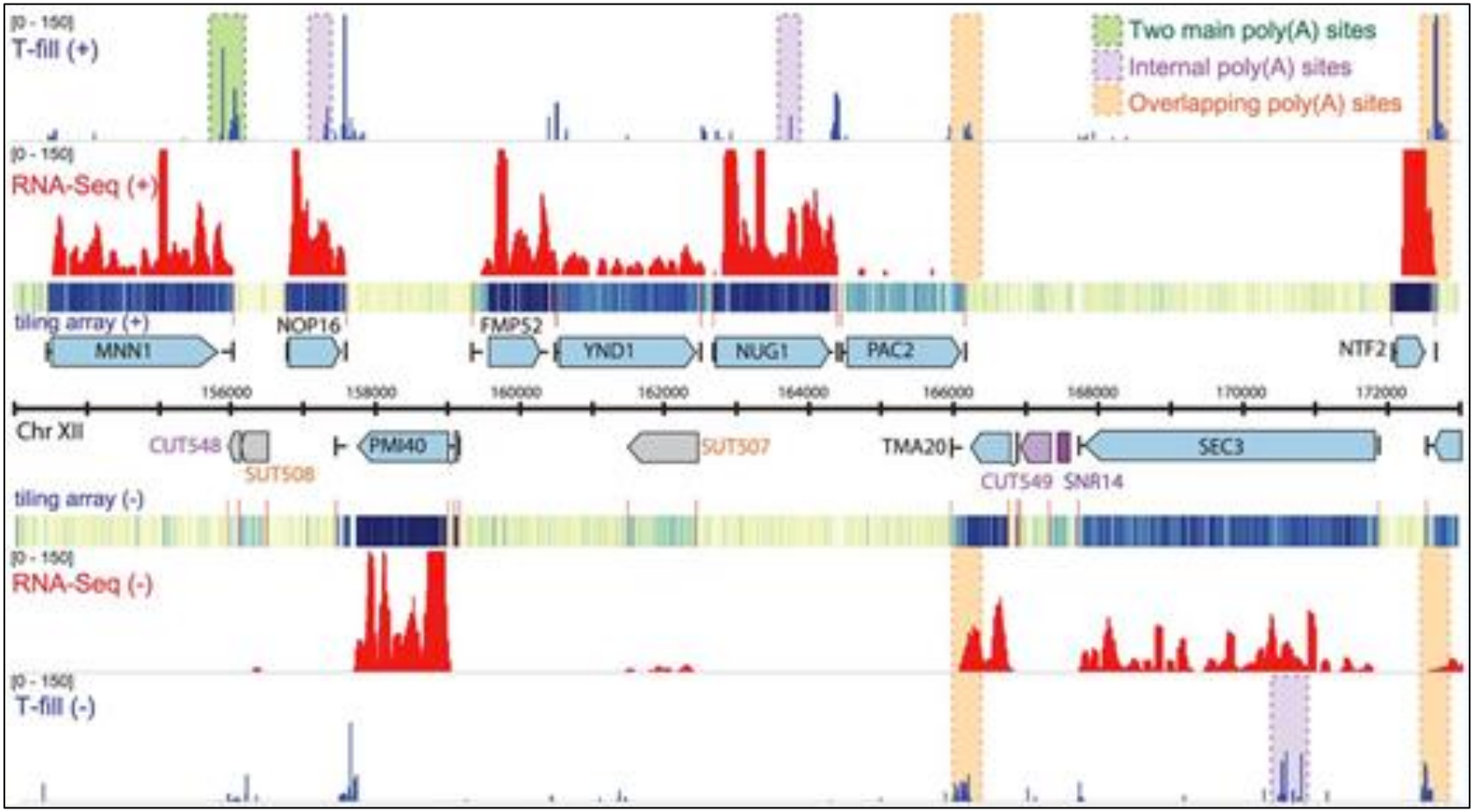
Downstream Analysis & Interpretation: Pathway Analysis



- 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

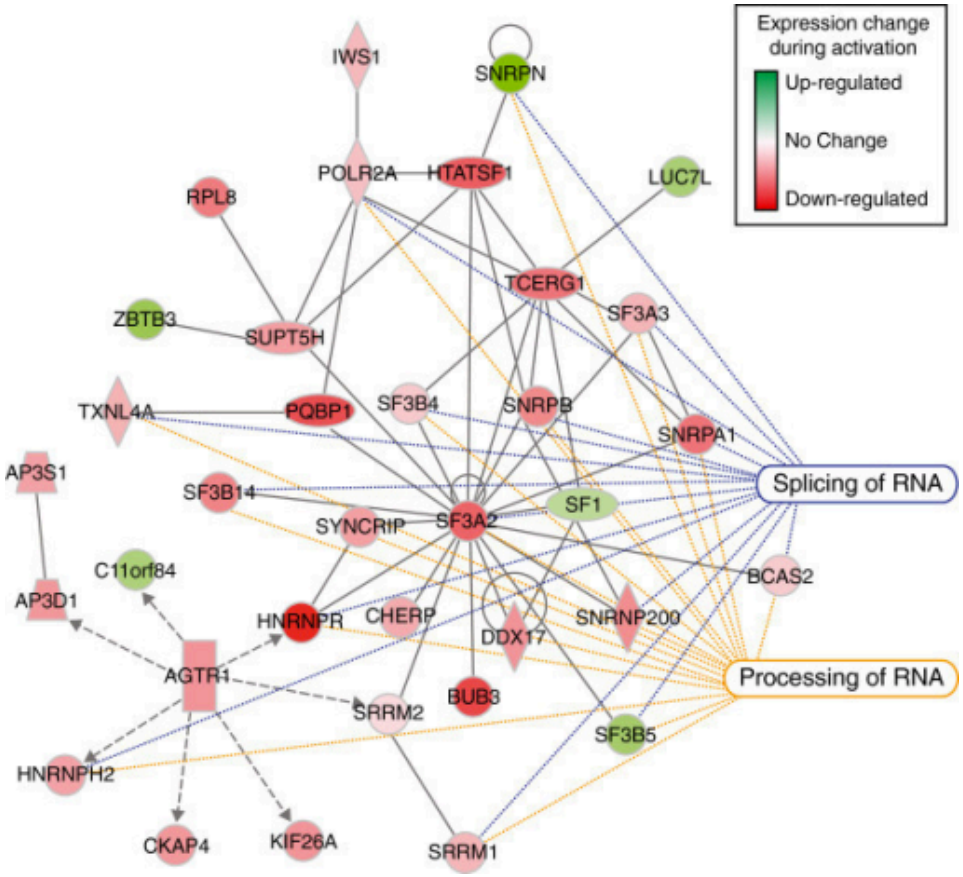
Downstream Analysis & Interpretation: Visualization with IGV and/or GenePattern



Wilkening et al. 2013 *Nucleic Acids Research*

Network Analysis

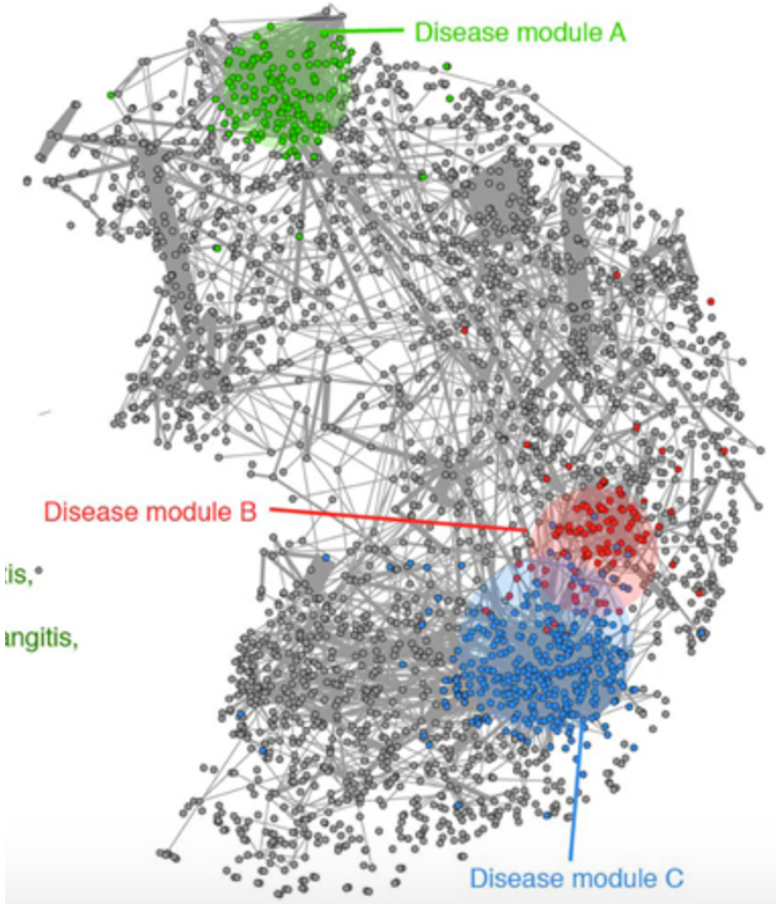
- Databases
  - PPI
  - Physical interactions
  - Indirect associations
  - Coexpression
  - Literature
  - Experimental
- Tools
  - Cytoscape, StringDB, GeneMania, NetworkX



Farina et al. 2012 Skeletal Muscle

Network Analysis – Choice of Interactome

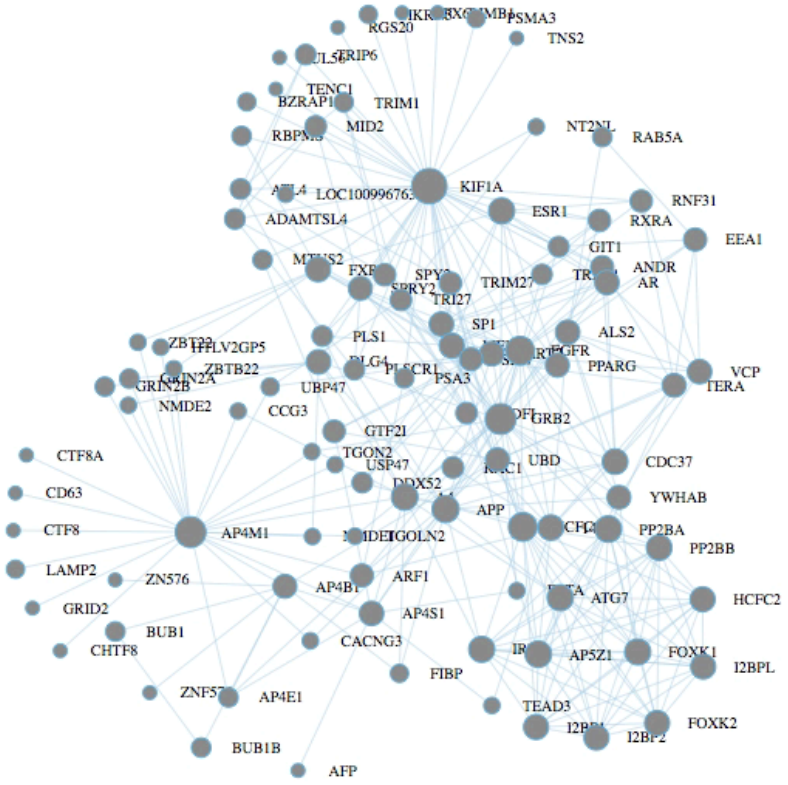
- **Menche et al[1]**: (binary addition of 7 literature and PPI based networks)
  - Regulatory interactions (TRANSFAC)
  - Yeast-two-hybrid (IntAct, MINT)
  - Literature curated interactions (IntAct, MINT, BIOGRID, HPRD)
  - Metabolic enzyme-coupled interactions (KEGG, BIGG)
  - Protein complexes (CORUM)
  - Kinase network (PhosphositePlus)
  - Signaling interactions (Vinayagam et al 2011)
  - Note: paper published in 2015, so network constructed from recently updated databases.
- **Multinet[2]** (Gerstein lab- binary addition):
  - PPI (BIOGRID)
  - Metabolic (KEGG)
  - Genetic (BIOGRID)
  - Phosphorylation
  - Regulatory (ENCODE)
  - Signaling (SignaLink)
  - Note: “we find that out of ~110,000 interactions in our data set, only 881 interactions occur in more than one network”
  - Potential problem: This paper was published in 2013, so the network is constructed from databases > 3 years old.



[1] Menche, Jörg, et al. "Uncovering disease-disease relationships through the incomplete interactome." *Science* 347.6224 (2015): 1257601.

[2] Khurana, Ekta, et al. "Interpretation of genomic variants using a unified biological network approach." *PLoS Comput Biol* 9.3 (2013): e1002886.

Network Propagation



$$F_t = \alpha\, W' * F_{t-1} + (1-\alpha)\, Y$$

Spread 'heat' to nearby nodes

Diffusion coefficient

Degree-normalized adjacency matrix

Seed nodes

Vanunu et al:
Heat not conserved

$$W' = \frac{W_{ij}}{\sqrt{d_i d_j}}$$

Leiberson et al:
Heat is conserved

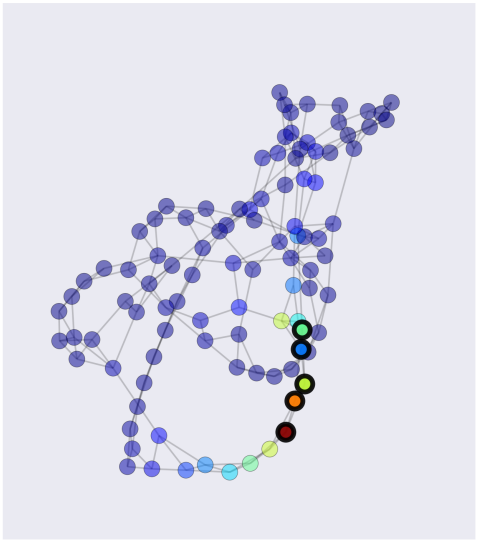
$$W' = \frac{W_{ij}}{d_i}$$

1) Vanunu, Oron, et al. "Associating genes and protein complexes with disease via network propagation." *PLoS Comput Biol* 6.1 (2010): e1000641.
2) Leiserson, Mark DM, et al. "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes." *Nature genetics* 47.2 (2015): 106-114.

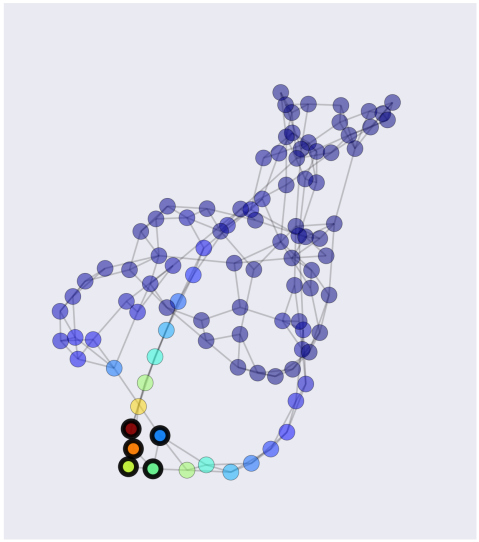
Disease Co-Localization Method

- Propagate heat from nodes in two sets (A and B: Bold nodes)
- Resulting heat vectors (H_A, H_B)- nodes below are color-coded by heat
- Sum the product of these heat vectors (e.g. take the dot product)
- If nodes in A and B are co-localized, this dot product will be larger
- Compare to baseline from heat vectors on degree preserving edge-shuffled network

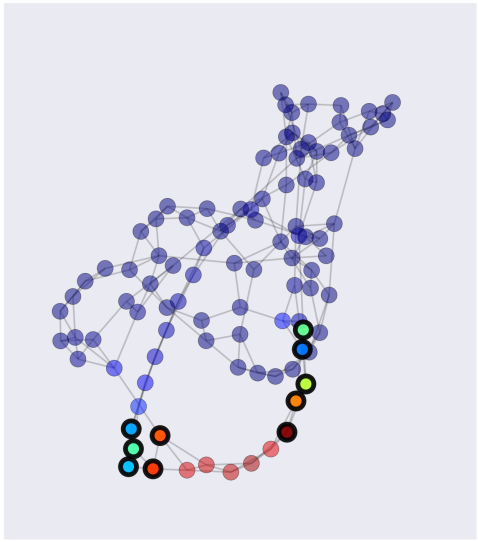
Heat: nodes A (H_A)



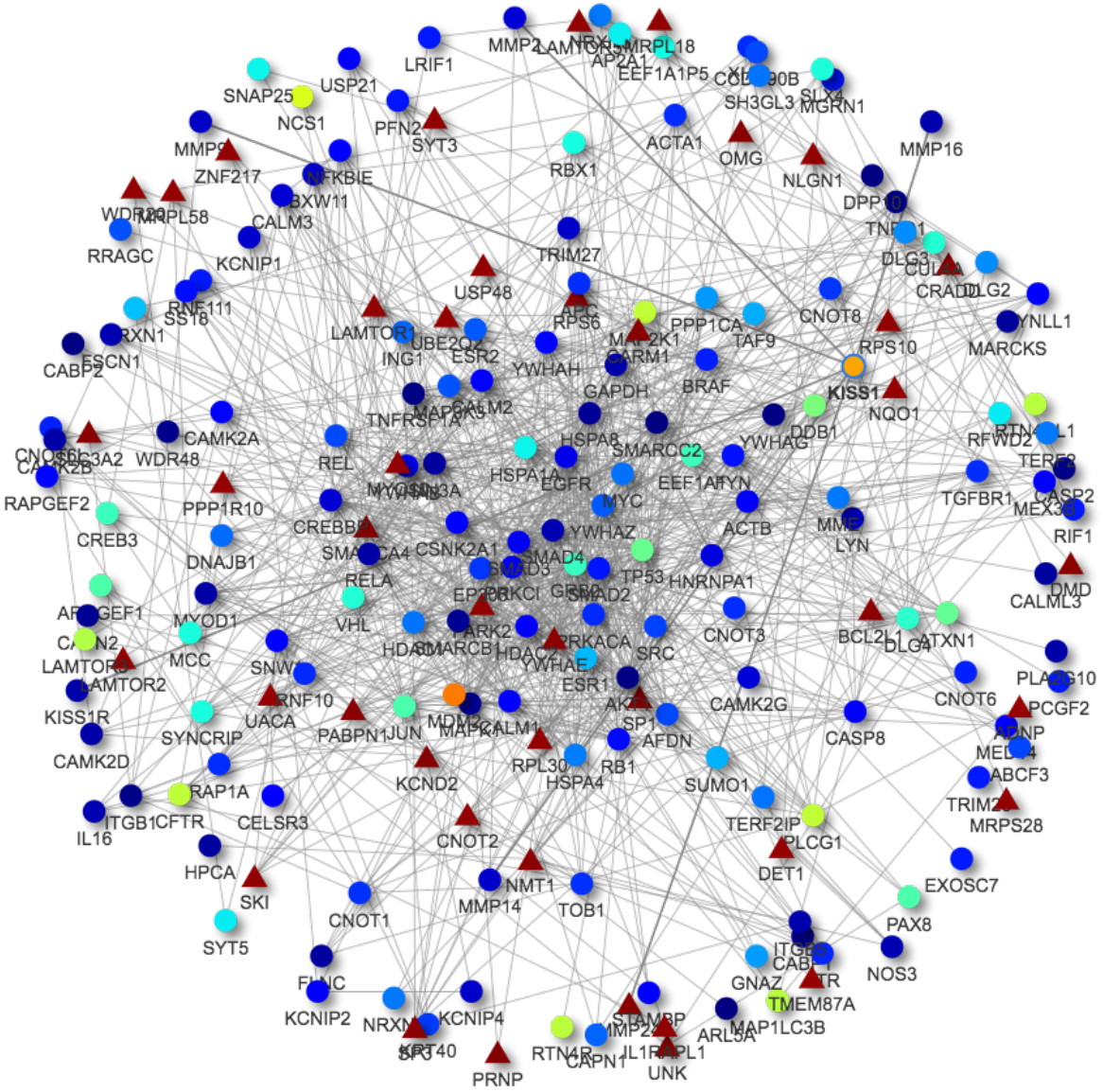
Heat: nodes B (H_B)



$H_A \cdot H_B$



Downstream Analysis & Interpretation: Network Analysis



Summary

- **RNA-Seq Case Studies**
  - Case Study 1 – Breast cancer precision medicine trial
  - Case Study 2 – Uveal melanoma neoantigen discovery and molecular subtype classification
- **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

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

Bioconductor RNAseq123: https://www.bioconductor.org/help/workflows/RNAseq123/#data-pre-processing