Cornell University
Institute of Biotechnology
Biotechnology Resource Center

Workshop
March 18, 2013

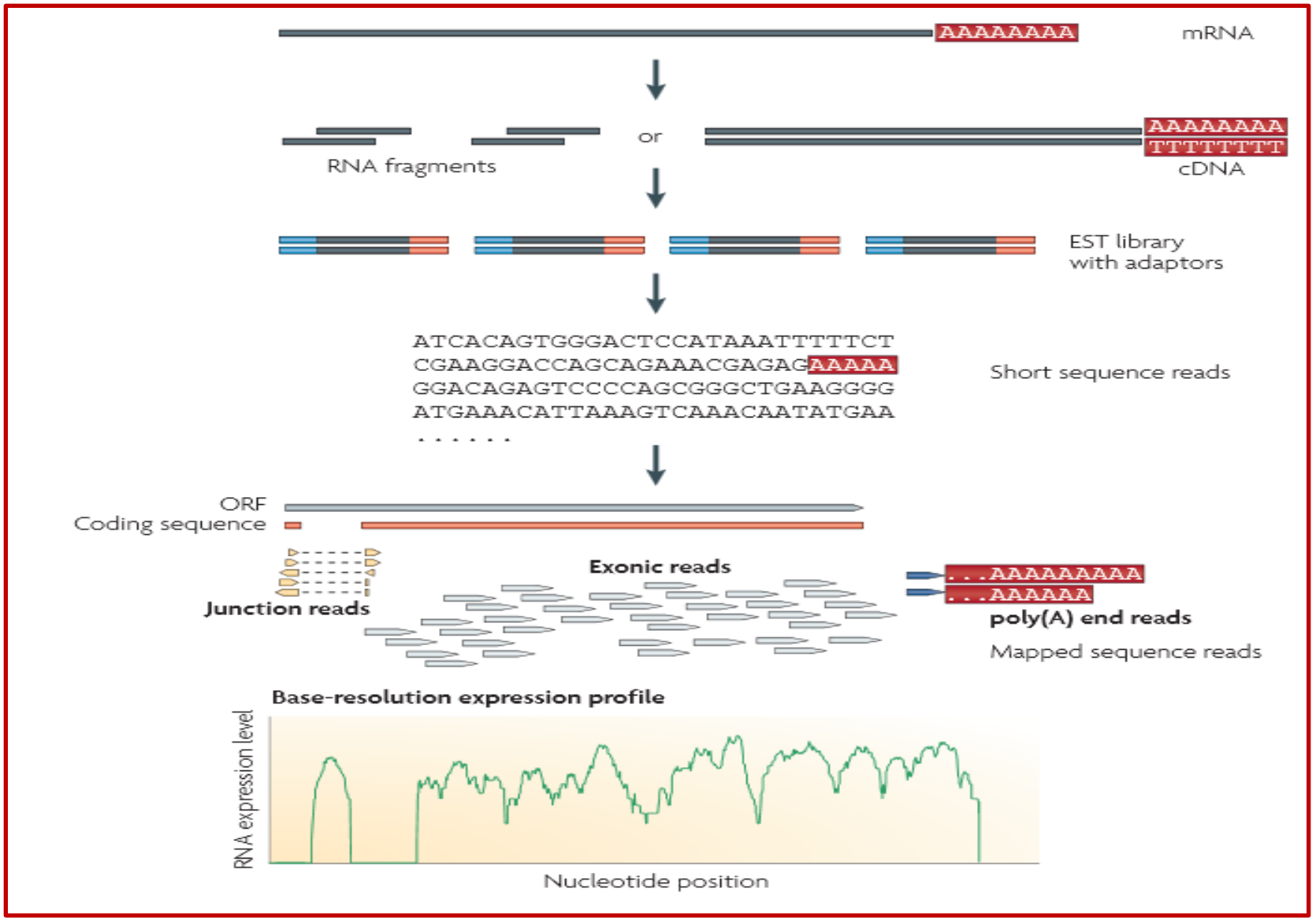
Reference Based RNA-Seq Data Analysis

Computational Biology Service Unit (CBSU)
Hsiao-Pei Yang
hy31@cornell.edu

Overview

- What is RNA-seq?
- Why RNA-seq?
- How to detect differential expression (DE) by RNA-seq?
  - Read Mapping
  - Summarization
  - Normalization
  - DE testing
- CBSU RNA-seq analysis pipeline

RNA-Seq: a revolutionary tool for transcriptomics



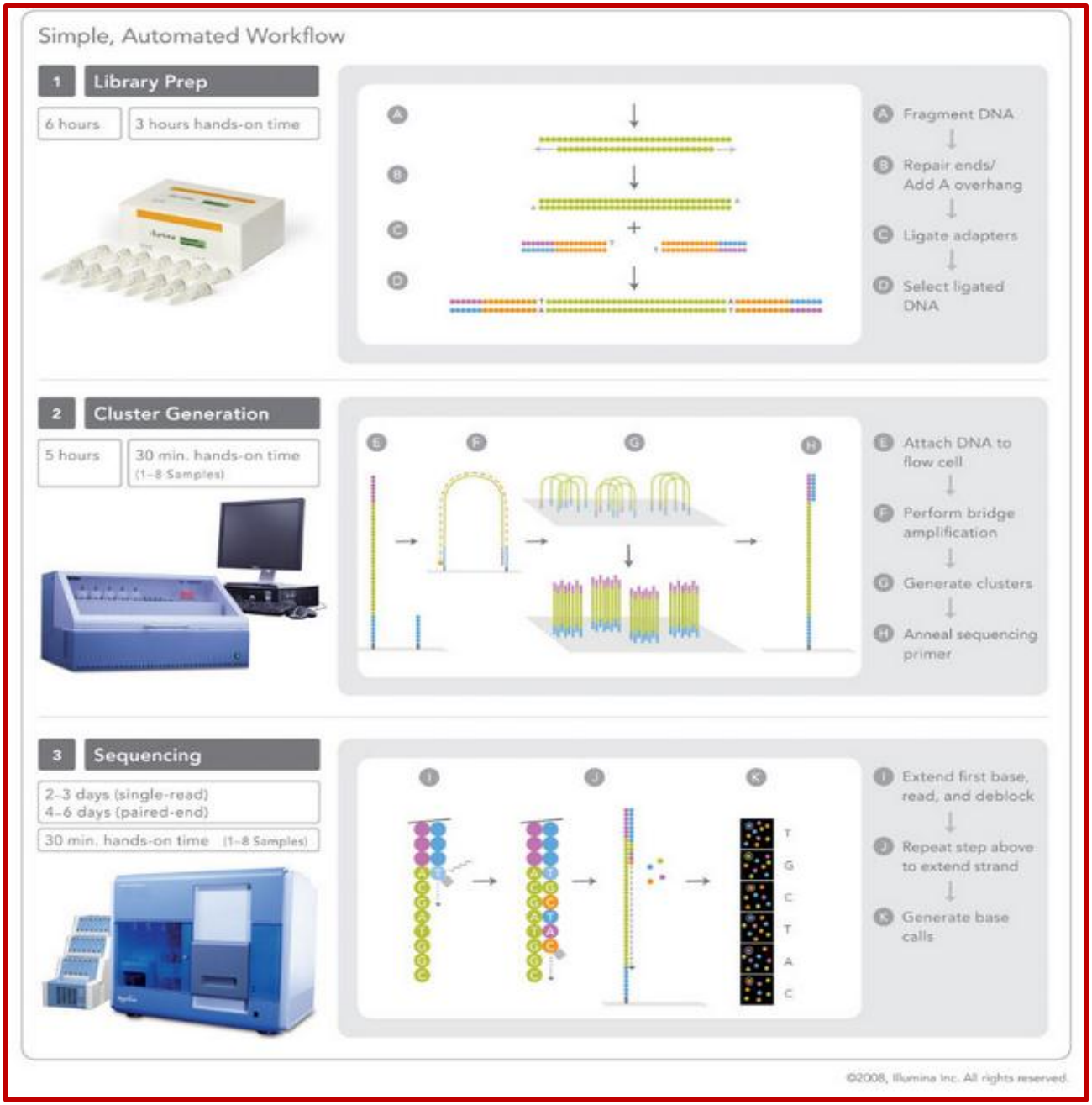
Wang *et al*., 2009 Nature Review Genetics 10:57

How RNA-seq was generated?

Examples of NGS Instrumentation

- Roche 454 sequencer
- Illumina Genome Analyzer (Solexa sequencing)
- Applied Biosystems SOLiD sequencer

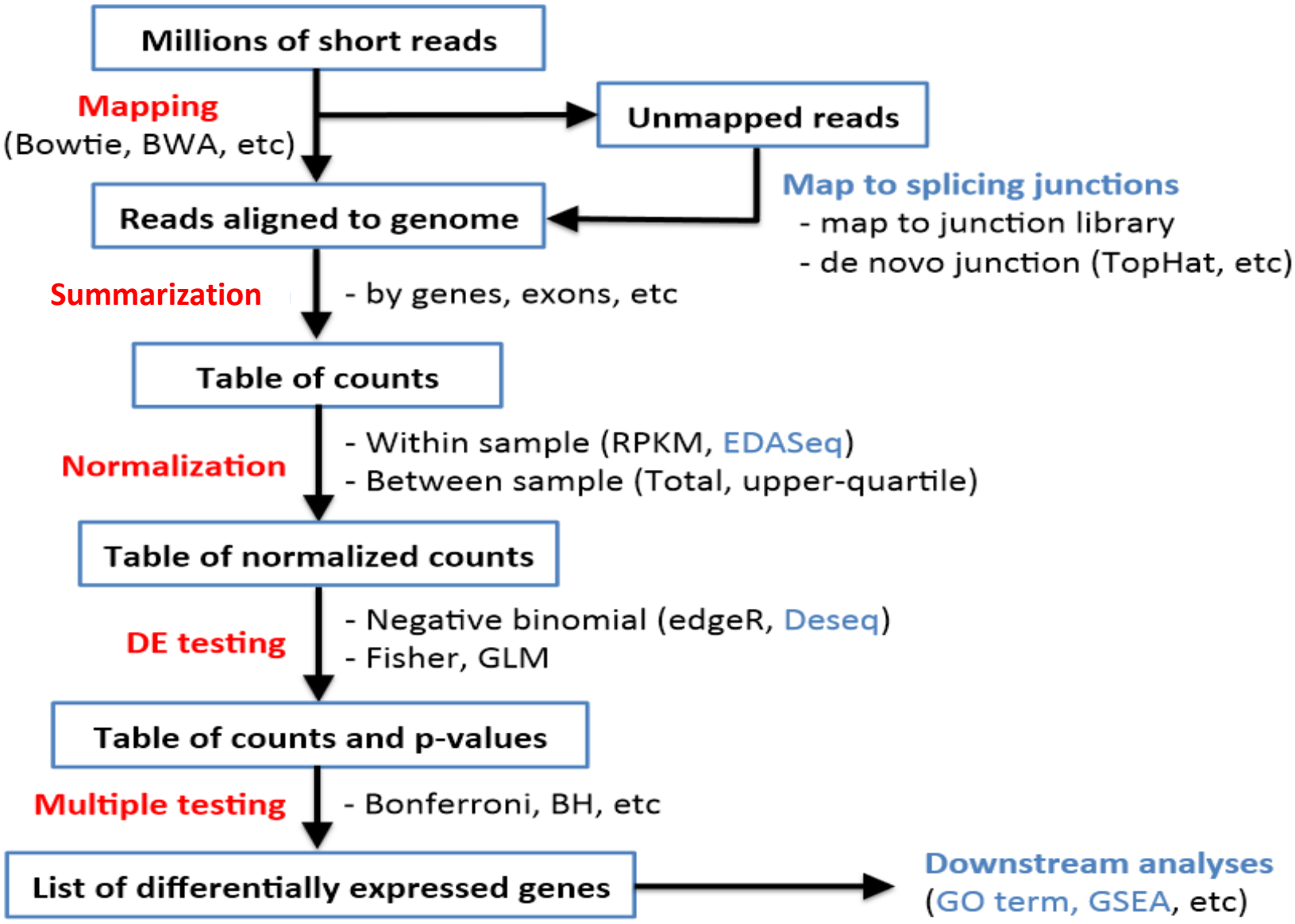
Illumina sequencing plateform



Applications for RNA-seq Analysis

- **Transcripts quantification**
- **Splicing sites discovery and quantification**
- **Gene discovery**
- **SNP/INDEL detection**
- **Allele specific expression**

Overview


Millions of short reads
Mapping
(Bowtie, BWA, etc)
Unmapped reads
Map to splicing junctions
- map to junction library
- de novo junction (TopHat, etc)
Reads aligned to genome
Summarization
- by genes, exons, etc
Table of counts
Normalization
- Within sample (RPKM, EDASeq)
- Between sample (Total, upper-quartile)
Table of normalized counts
DE testing
- Negative binomial (edgeR, Deseq)
- Fisher, GLM
Table of counts and p-values
Multiple testing
- Bonferroni, BH, etc
List of differentially expressed genes
Downstream analyses
(GO term, GSEA, etc)

Selected list of RNA-seq analysis programs

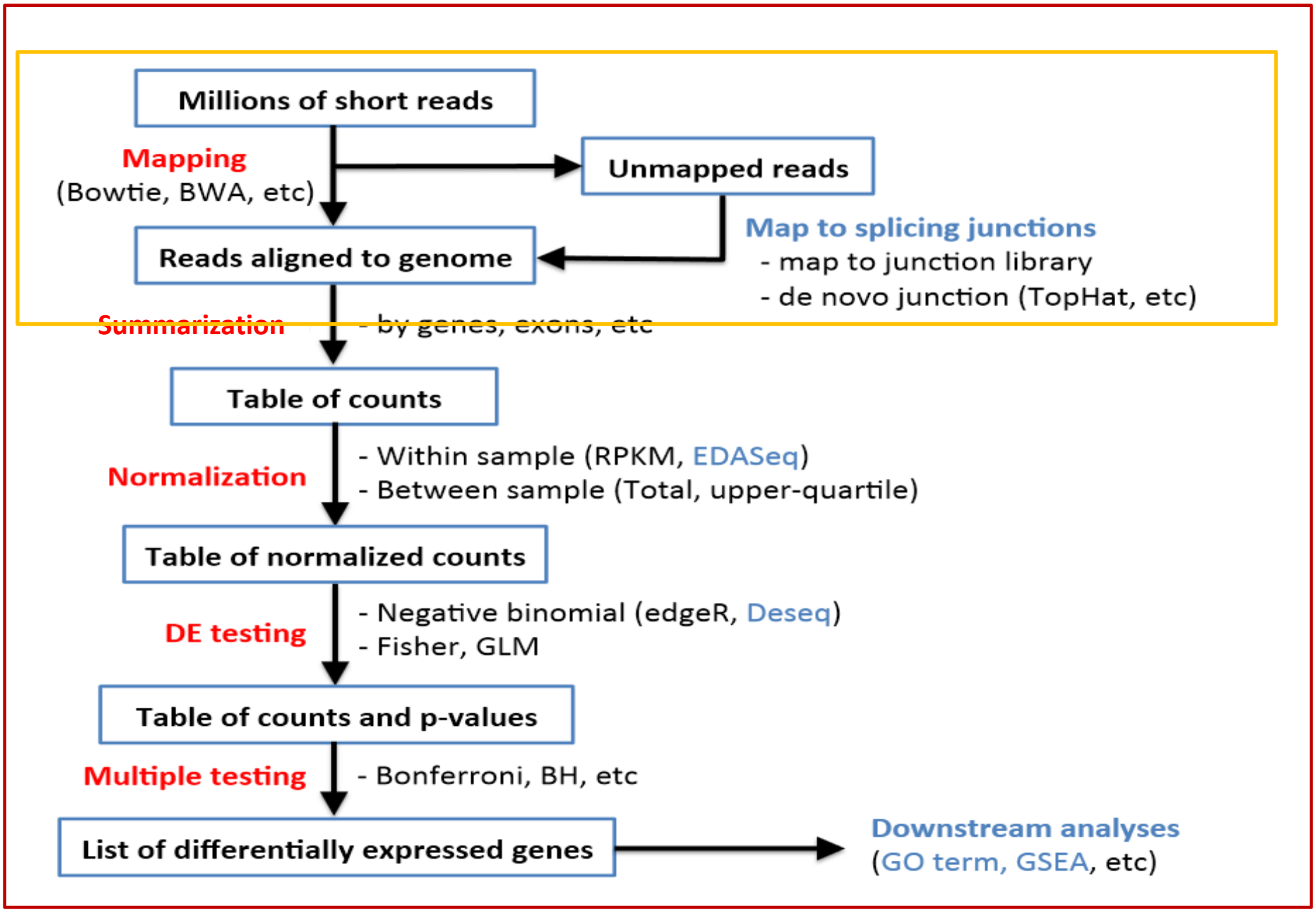
Table 1 | Selected list of RNA-seq analysis programs

Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy[39]	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie[43]			
		BWA[44]	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice[52]	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap[50]			
		TopHat[51]	Uses Bowtie alignments		
	Seed-extend methods	GSNAP[53]	Can use SNP databases		
		QPALMA[54]	Smith-Waterman for large gaps		
Transcriptome reconstruction					
Genome-guided reconstruction	Exon identification	G.Mor.Se	Assembles exons	Identifying novel transcripts using a known reference genome	Alignments to reference genome
	Genome-guided assembly	Scripture[28]	Reports all isoforms		
		Cufflinks[29]	Reports a minimal set of isoforms		
Genome-independent reconstruction	Genome-independent assembly	Velvet[61]	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
		TransABySS[56]			
Expression quantification					
Expression quantification	Gene quantification	Alexa-seq[47]	Quantifies using differentially included exons	Quantifying gene expression	Reads and transcript models
		Enhanced read analysis of gene expression (ERANGE)[20]	Quantifies using union of exons		
		Normalization by expected uniquely mappable area (NEUMA)[82]	Quantifies using unique reads		
	Isoform quantification	Cufflinks[29]	Maximum likelihood estimation of relative isoform expression	Quantifying transcript isoform expression levels	Read alignments to isoforms
		MISO[33]			
		RNA-seq by expectaion maximization (RSEM)[69]			
Differential expression		Cuffdiff[29]	Uses isoform levels in analysis	Identifying differentially expressed genes or transcript isoforms	Read alignments and transcript models
		DegSeq[79]	Uses a normal distribution		
		EdgeR[77]			
		Differential Expression analysis of count data (DESeq)[78]			
		Myrna[75]	Cloud-based permutation method		

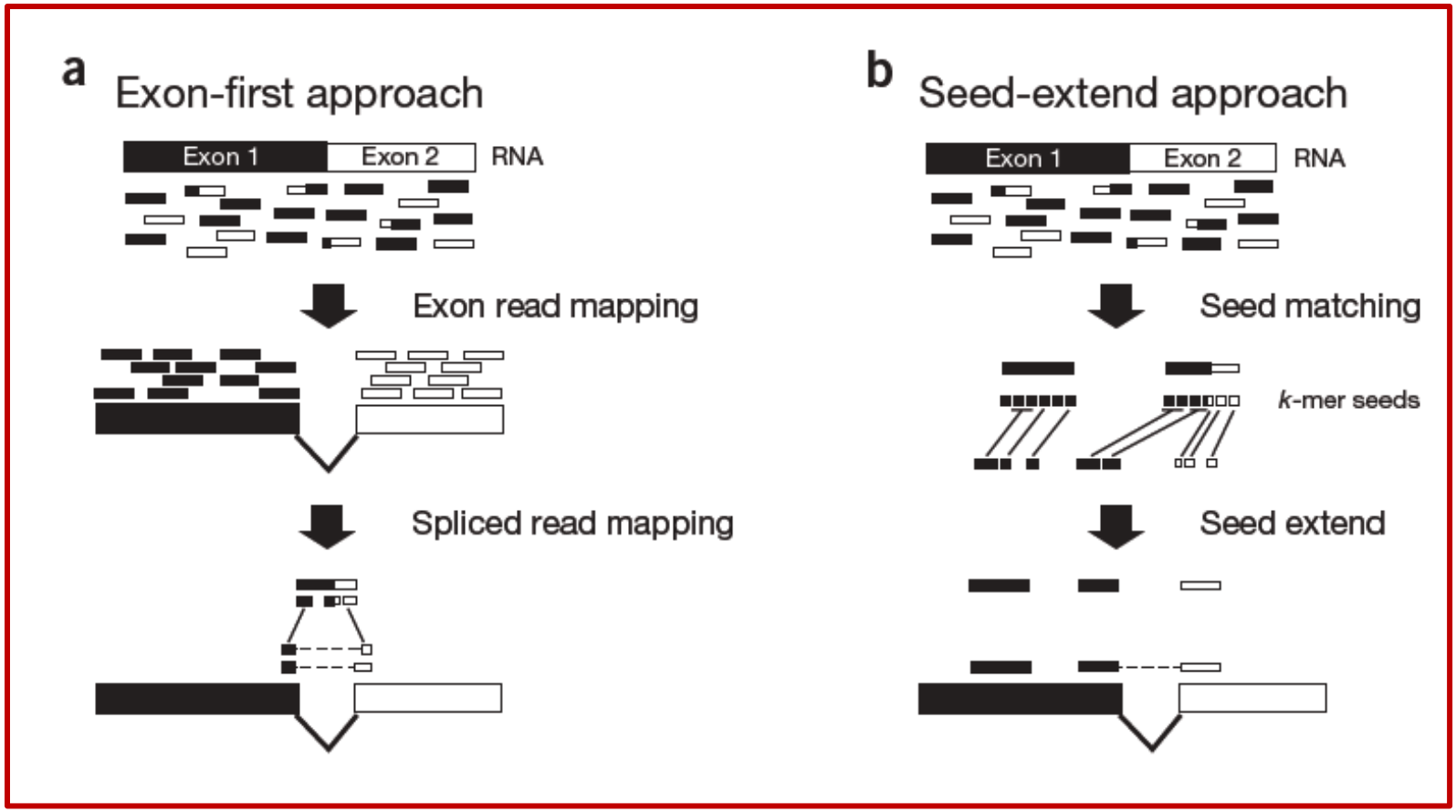
[a]This list is not meant to be exhaustive as many different programs are available for short-read alignment. Here we chose a representative set capturing the frequently used tools for RNA-seq or tools representing fundamentally different approaches.

Gaber *et al.*, 2011, Nature Methods 8:469

Overview

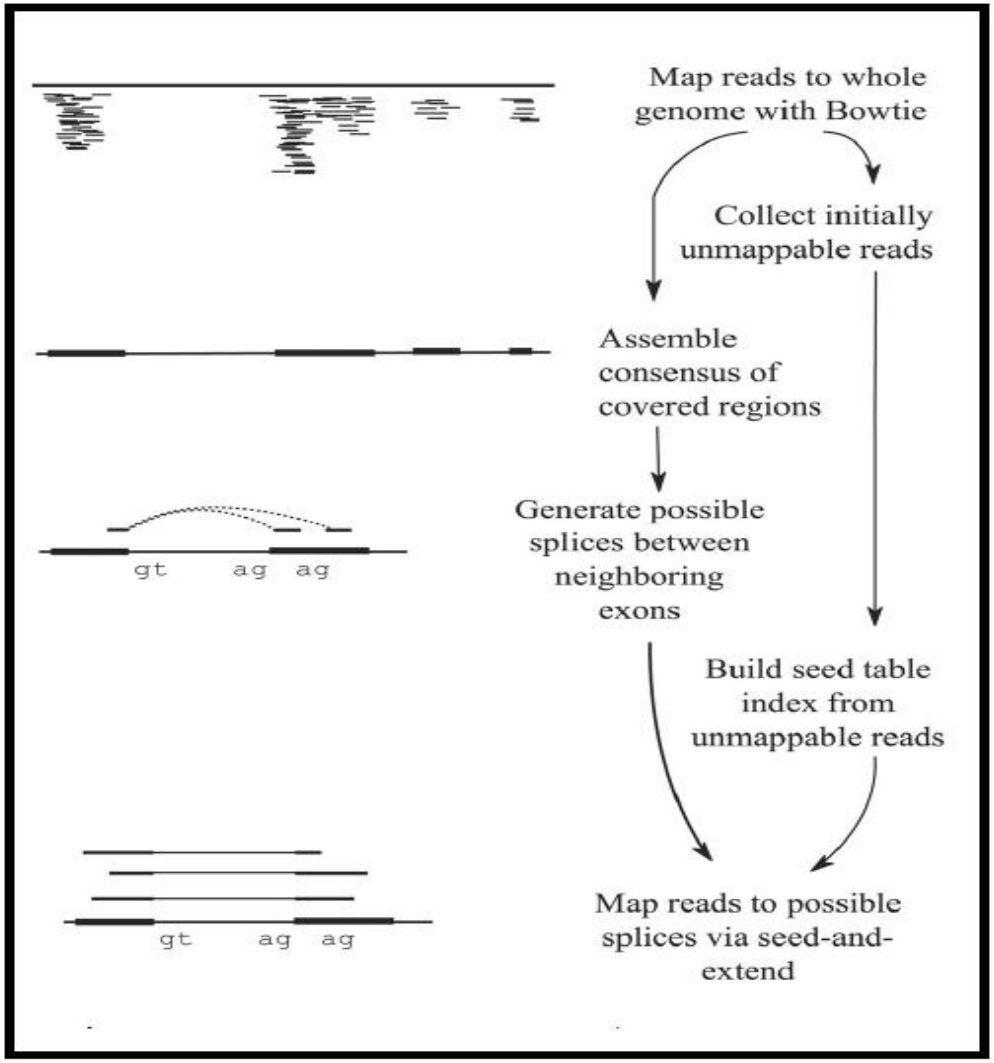

Millions of short reads
Mapping
(Bowtie, BWA, etc)
Unmapped reads
Map to splicing junctions
- map to junction library
- de novo junction (TopHat, etc)
Reads aligned to genome
Summarization
- by genes, exons, etc
Table of counts
Normalization
- Within sample (RPKM, EDASeq)
- Between sample (Total, upper-quartile)
Table of normalized counts
DE testing
- Negative binomial (edgeR, Deseq)
- Fisher, GLM
Table of counts and p-values
Multiple testing
- Bonferroni, BH, etc
List of differentially expressed genes
Downstream analyses
(GO term, GSEA, etc)

Strategies for gapped alignments of RNA-seq reads to the genome



Example: **TopHat** **QPALMA**

Map reads with Tophat



Trapnell C et al. Bioinformatics 2009;25:1105-1111

Limitation of Tophat

Two-step approach

- **If a read can be mapped to the genome without splicing, it would not be evaluated for spliced mapping.**
- Can be corrected with “--read-realign-edit-dist” option

Canonical junctions only

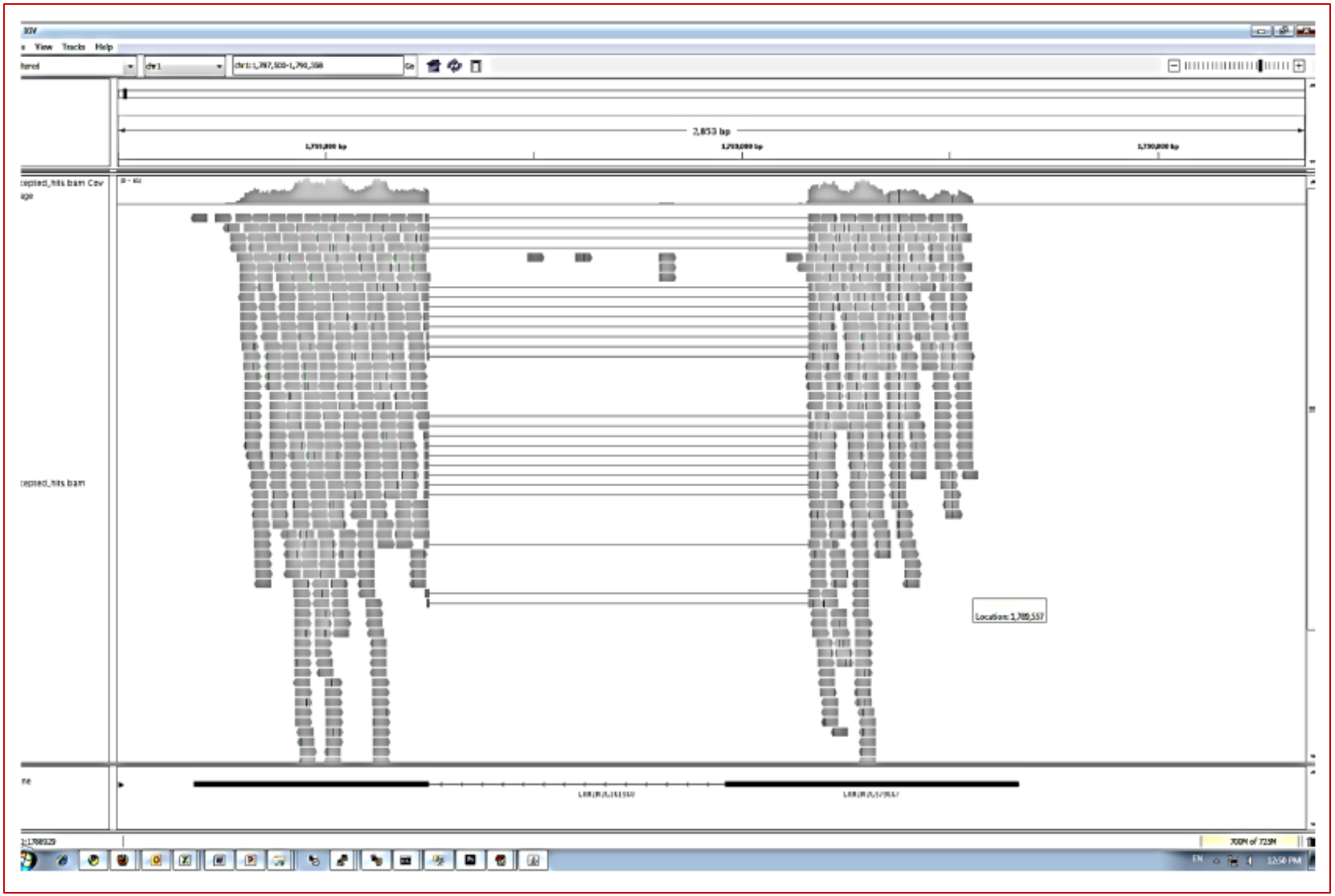
- **Reads < 75 bp, "GT-AG" introns**
- **Reads >=75bp, "GT-AG", "GC-AG" and "AT-AC“ introns**

Mapping with an aligner that allows for divergent reads

Stampy

- Maps single and paired Illumina reads to a reference genome/transcriptome
- High sensitivity for indels and divergent reads, up to 10-15%
- Input: Fastq and Fasta; gzipped or plain; SAM and BAM
- Output: SAM, Maq's map file

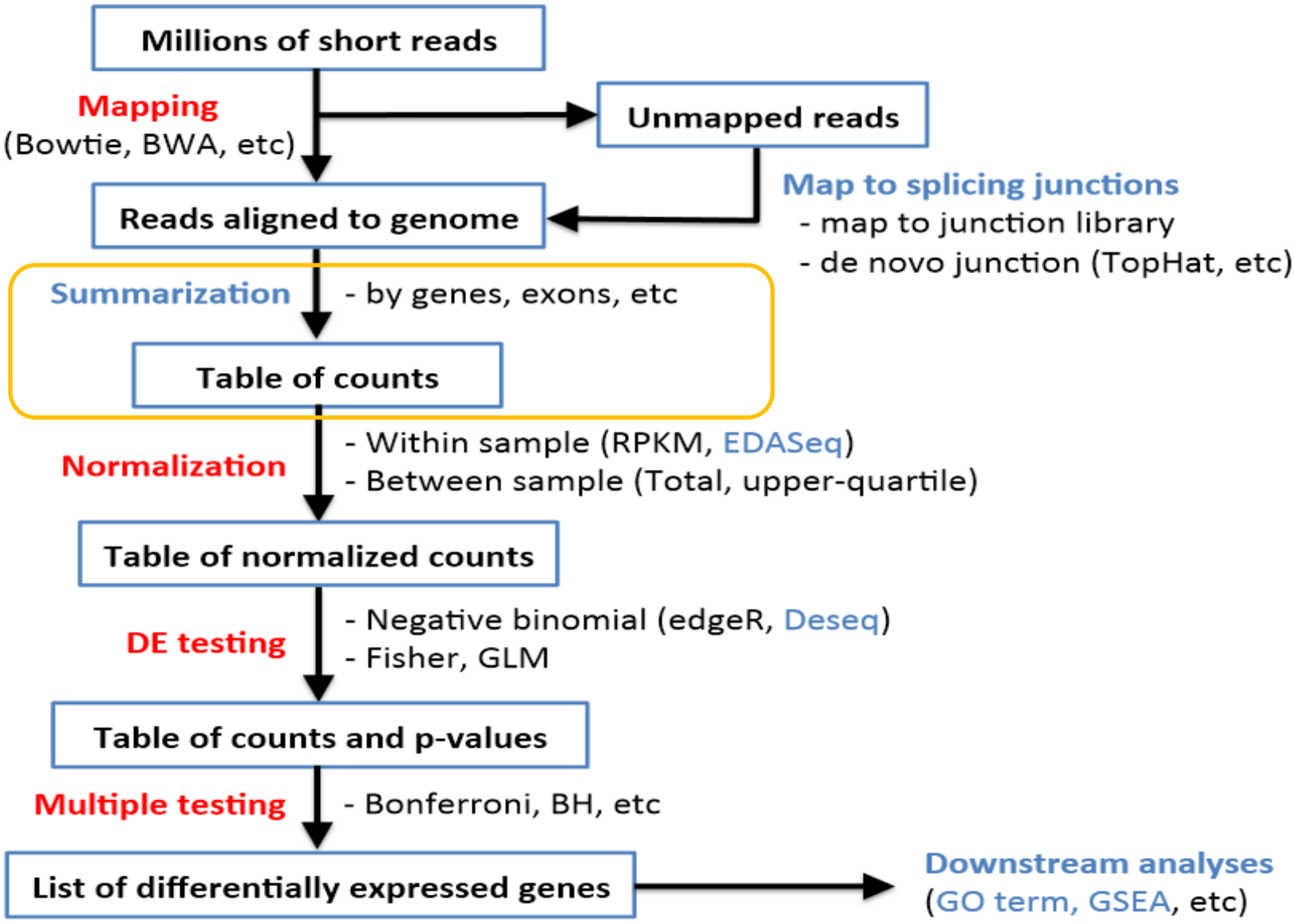
Visualization of read alignment with IGV



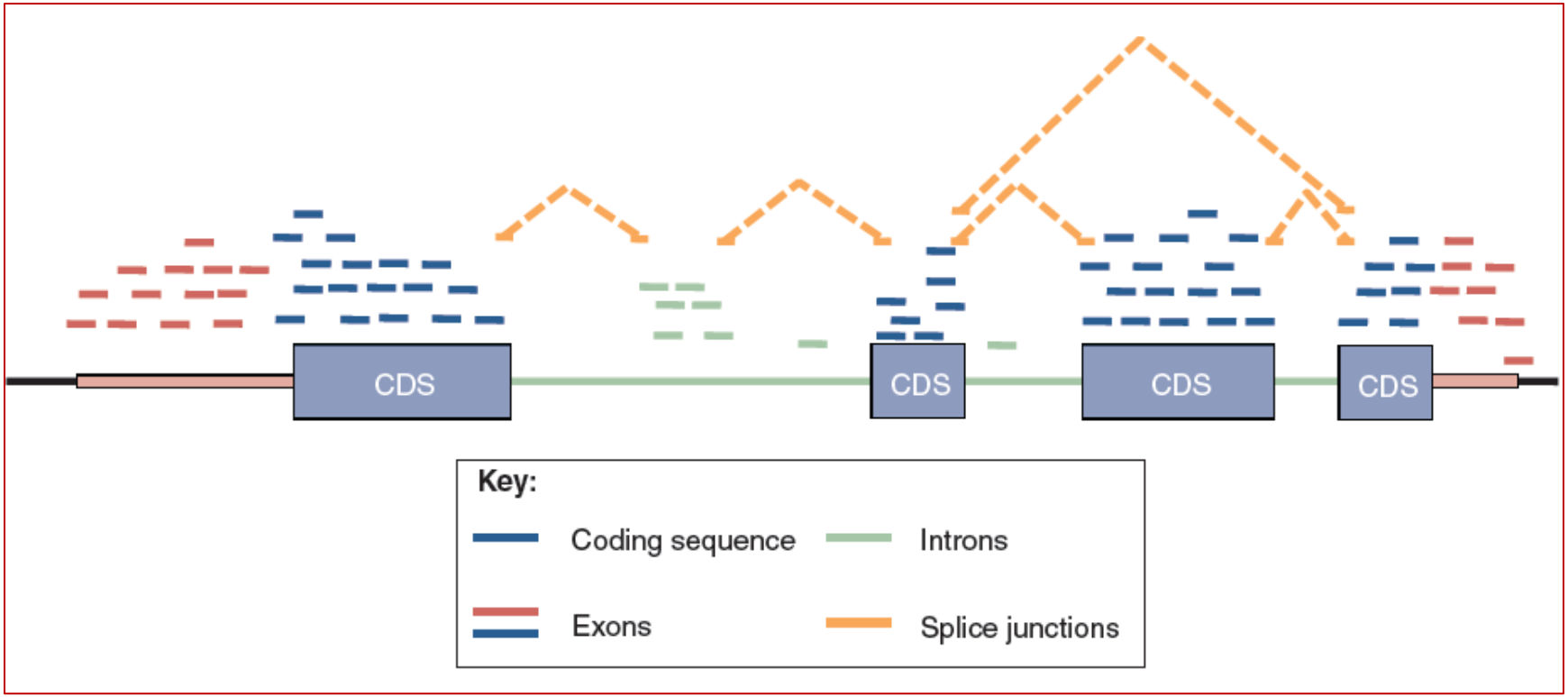
SAM & BAM files

- A SAM file (**.sam**) is a tab-delimited text file that contains sequence alignment data
- A BAM file (**.bam**) is the binary version of a SAM file
- **SAMtools** (http://en.wikipedia.org/wiki/SAMtools)
  - a set of utilities for interacting with and post-processing short DNA sequence read alignments in the SAM/BAM format
  - commands
    - ***view*** filters SAM or BAM formatted data
    - ***sort*** sorts a BAM file based on its position in the reference, as determined by its alignment
    - ***index*** creates a new index file that allows fast look-up of data in a (sorted) SAM or BAM
    - ***tview*** to visualize how reads are aligned to specified small regions of the reference genome (similar to IGV, but

Overview

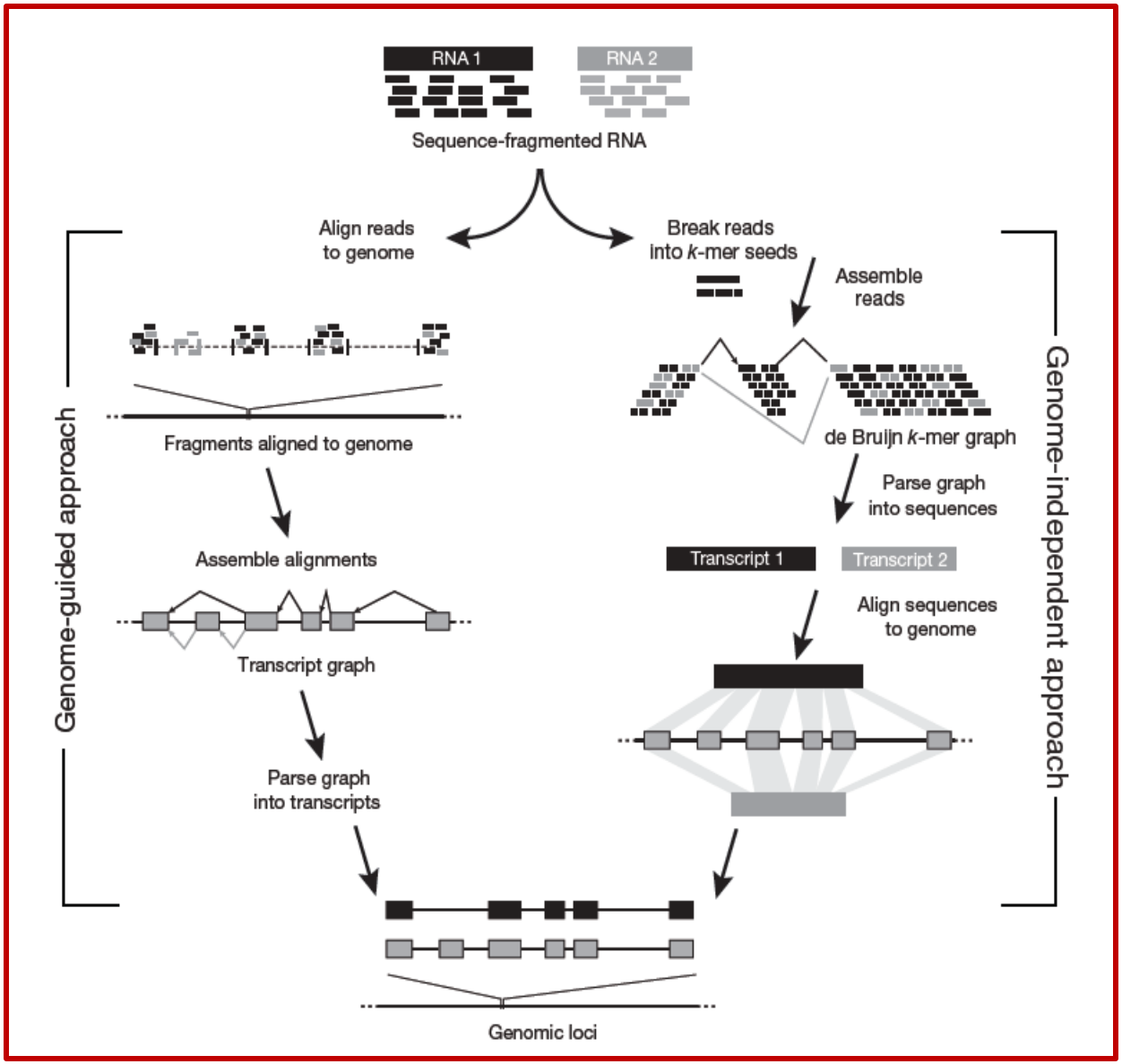

Millions of short reads
Mapping
(Bowtie, BWA, etc)
Unmapped reads
Map to splicing junctions
- map to junction library
- de novo junction (TopHat, etc)
Reads aligned to genome
Summarization
- by genes, exons, etc
Table of counts
Normalization
- Within sample (RPKM, EDASeq)
- Between sample (Total, upper-quartile)
Table of normalized counts
DE testing
- Negative binomial (edgeR, Deseq)
- Fisher, GLM
Table of counts and p-values
Multiple testing
- Bonferroni, BH, etc
List of differentially expressed genes
Downstream analyses
(GO term, GSEA, etc)

Summarizing mapped reads into a gene level count

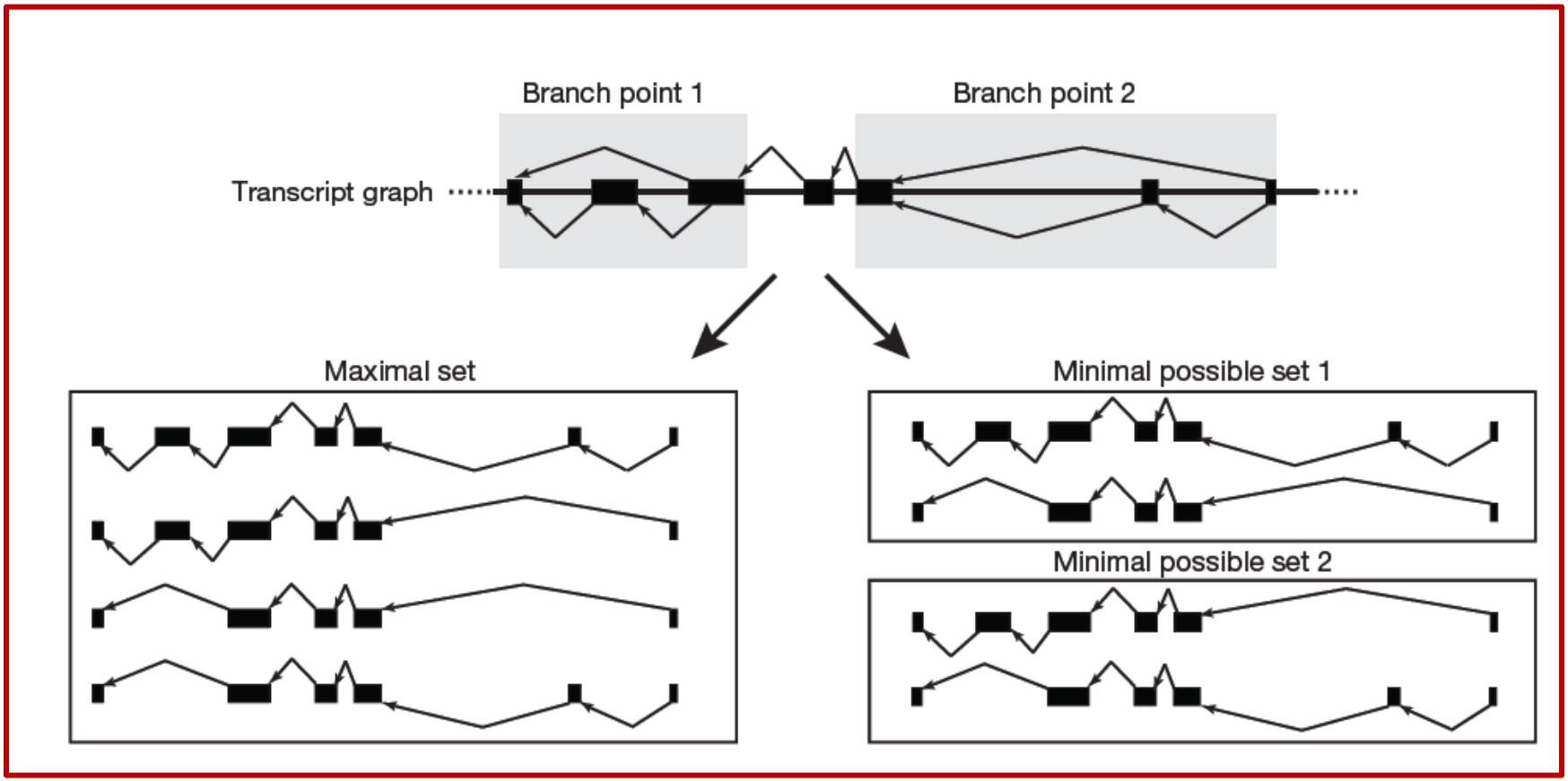


Different summarization strategies will result in the inclusion or exclusion of different sets of reads in the table of counts.

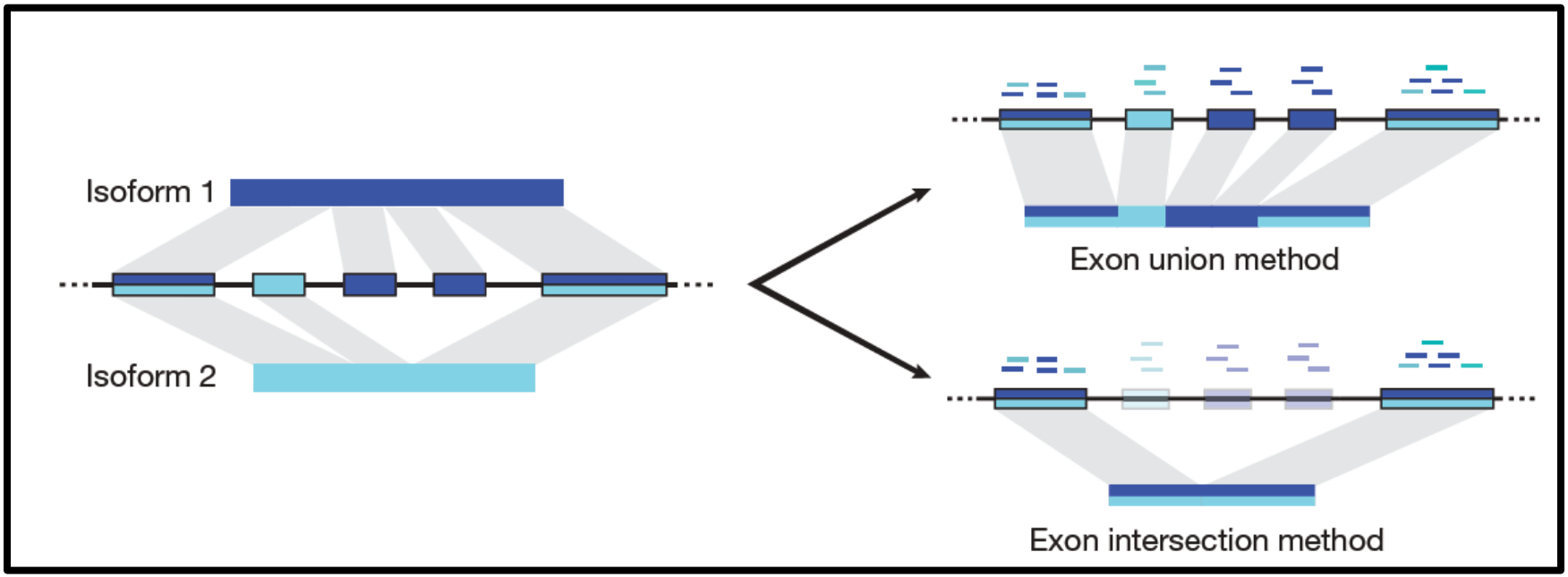
Transcriptome reconstruction methods



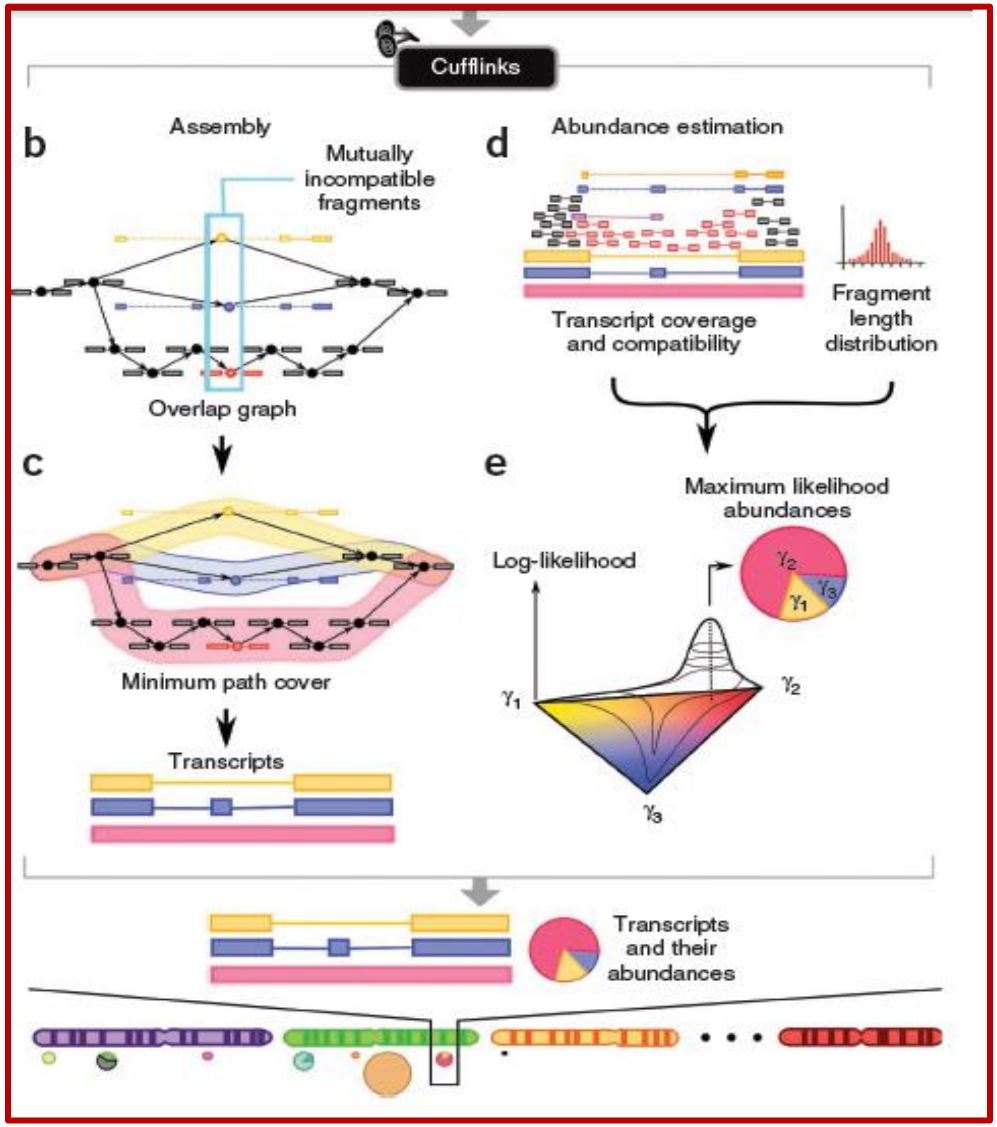
Methods summarizing transcript set



Two simplified gene models used for gene expression quantification



Transcript abundance estimate using Cufflinks "Isoform-expression methods"



- uses a statistical model in which the probability of observing each fragment is a linear function of the abundances of the transcripts from which it could have originated.
- incorporates distribution of fragment lengths to help assign fragments to isoforms.
- maximizes a function that assigns a likelihood to all possible sets of relative abundances
- reports abundances that best explain the observed fragments

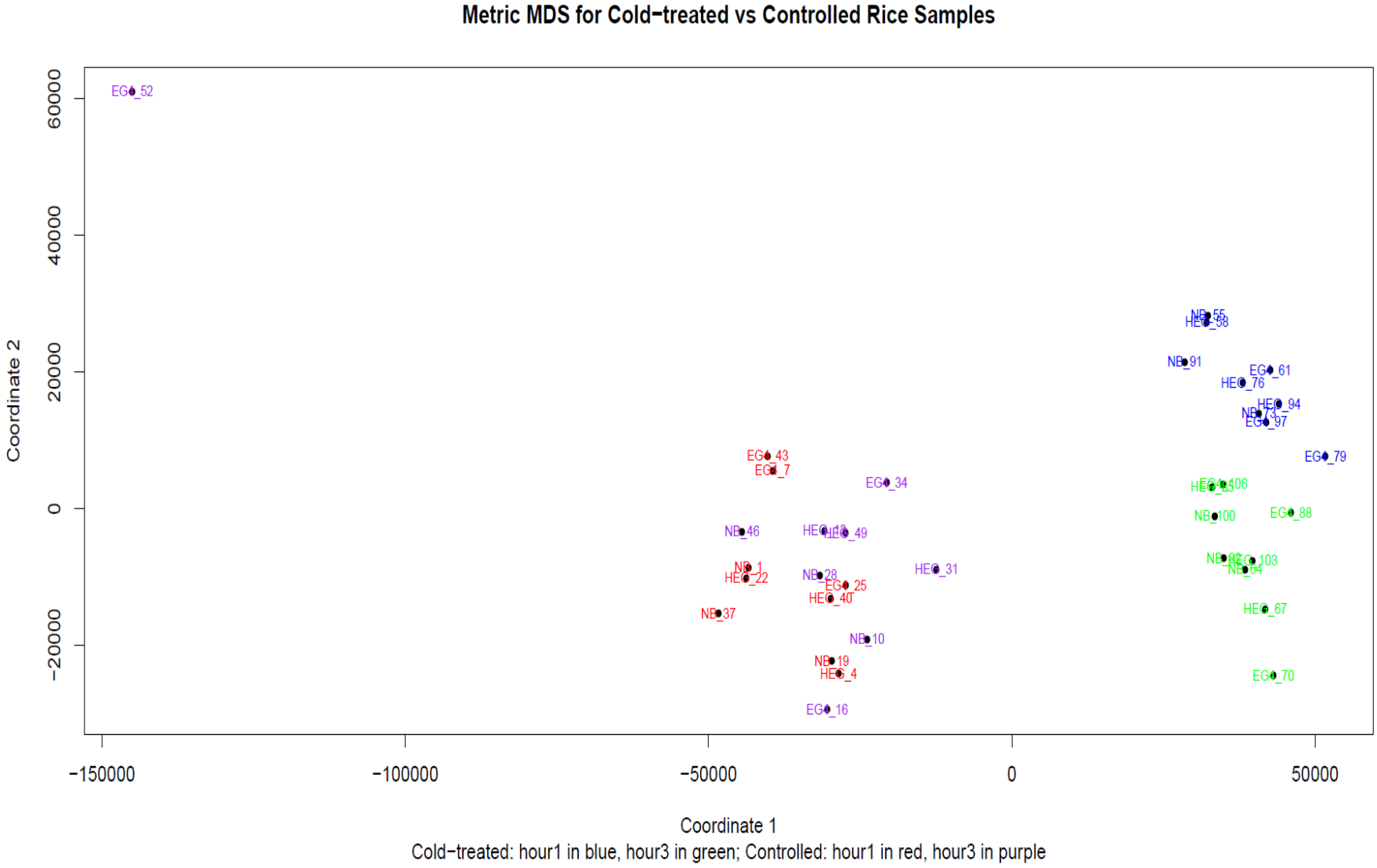
Trapnell *et al.*, 2010 Nat. Tech. 28:511.

Data QC

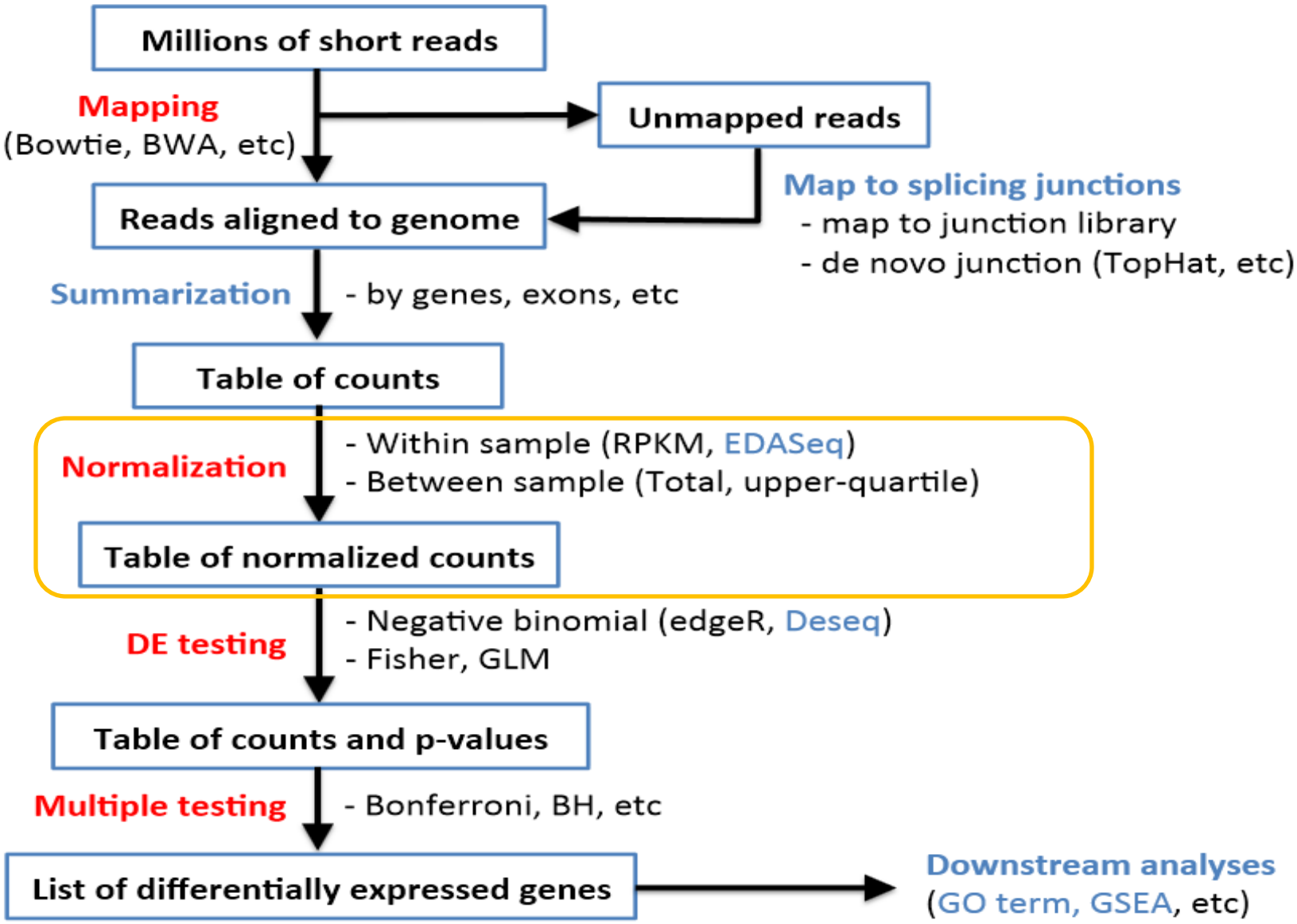
1. Check basic statistics of alignment results
   - Total reads
   - % reads mapped/unmapped
   - % reads mapped to unique site
   - % reads mapped to multiple sites

2. If the basic statistics looks good, check overall gene expression pattern among samples by clustering methods, such as MDS or PC.
   - to identify potential "outliers" due to contamination or other tech problem.
   - to check potential sample mixed-up (for example, samples from biological replicates are expected to be clustered with one another).
   - The clustering among samples may provide underlie biological explanations.

❖ Software for RNA-seq QC
   - FastQC
   - RNA-SeQC
   - ShortRead

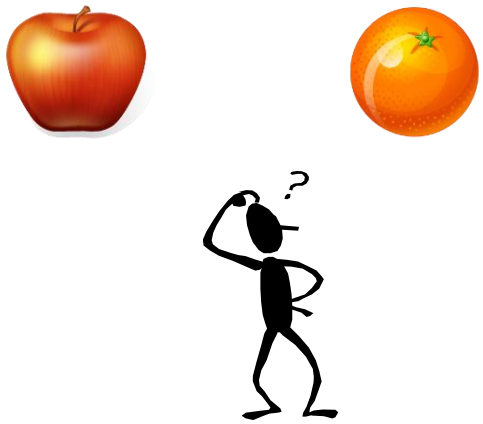

Metric MDS for Cold−treated vs Controlled Rice Samples
Coordinate 2
60000
40000
20000
0
−20000
−150000
−100000
−50000
0
50000
Coordinate 1
Cold−treated: hour1 in blue, hour3 in green; Controlled: hour1 in red, hour3 in purple

Overview


Millions of short reads
Mapping
(Bowtie, BWA, etc)
Unmapped reads
Map to splicing junctions
- map to junction library
- de novo junction (TopHat, etc)
Reads aligned to genome
Summarization
- by genes, exons, etc
Table of counts
Normalization
- Within sample (RPKM, EDASeq)
- Between sample (Total, upper-quartile)
Table of normalized counts
DE testing
- Negative binomial (edgeR, Deseq)
- Fisher, GLM
Table of counts and p-values
Multiple testing
- Bonferroni, BH, etc
List of differentially expressed genes
Downstream analyses
(GO term, GSEA, etc)

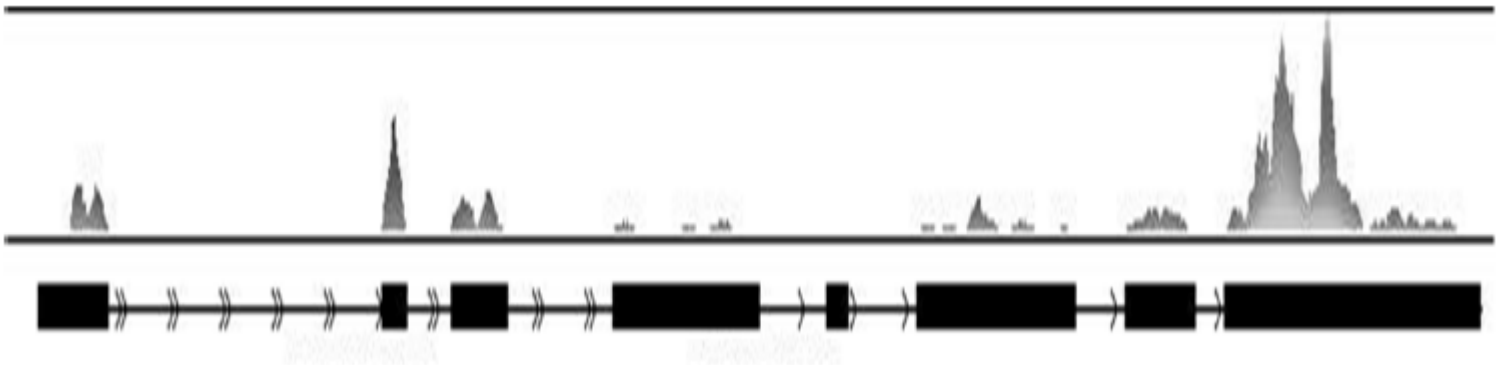
You have a list of counts, what next?

Gene	Condition A	Condition B
1	200	300
2	15	30
3	4000	4500
:	:	:



Factors affect RNA-seq read counts

1. Molar concentration of RNA molecules
2. Length of RNA molecules
3. Sequence-specific bias



Normalization for RNA-seq Data

The Aim:

To remove systematic technical effects in the data to ensure that technical bias has minimal impact on the results.

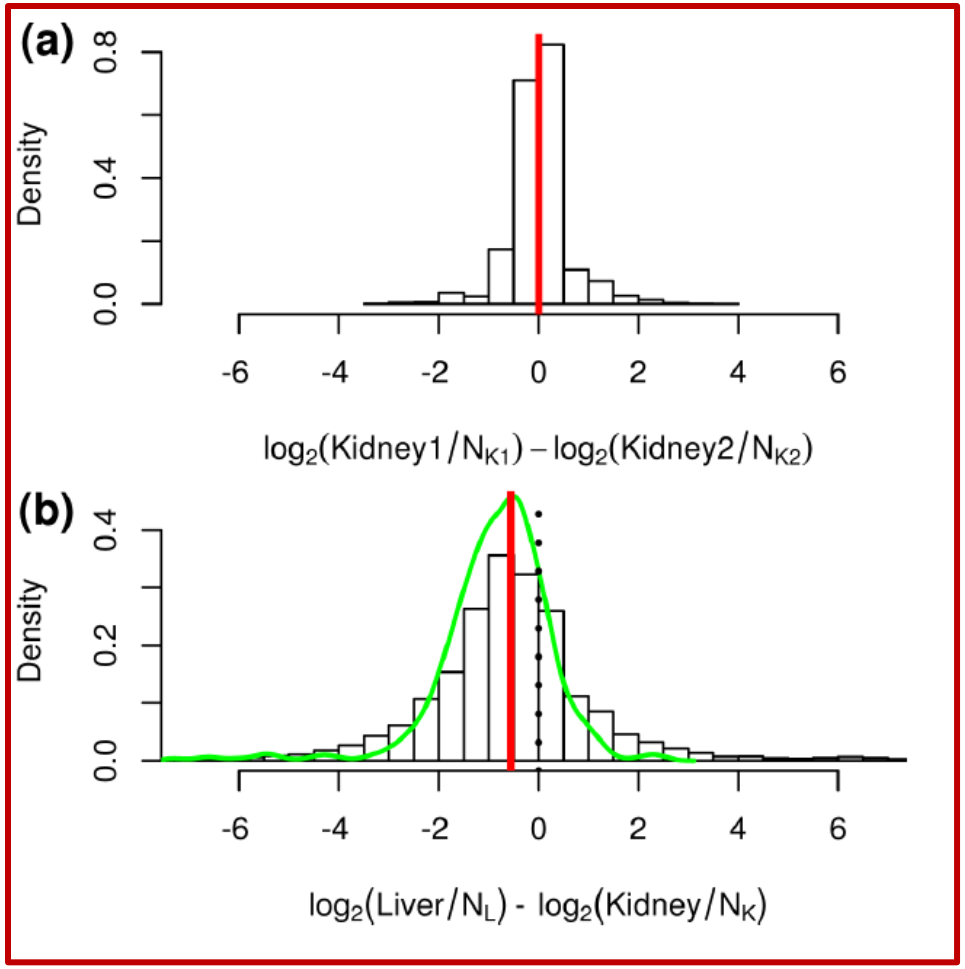
Normalization methods

- **Total-count normalization**
  - Low sensitivity in detecting DE, especially for low expressed genes
- **Upper-quantile (75%) normalization**
  - a small number of abundant, differentially expressed genes can create incorrect impression that less abundant genes are also differentially expressed
  - This issue can be mitigated by excluding these genes when normalizing expression values for the number of mapped reads in each sample.
  - use the number of reads mapping to the upper-quartile loci as normalization factor
- **Normalization by counts of stably expressed genes, such as housekeeping genes**
- **Trimmed mean (TMM) normalization**

For more discussion on normalization, see:

Bullard *et al*., 2010 Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 2010, 11:94.

Normalization for RNA-seq data



Technical replicates

Liver vs kidney

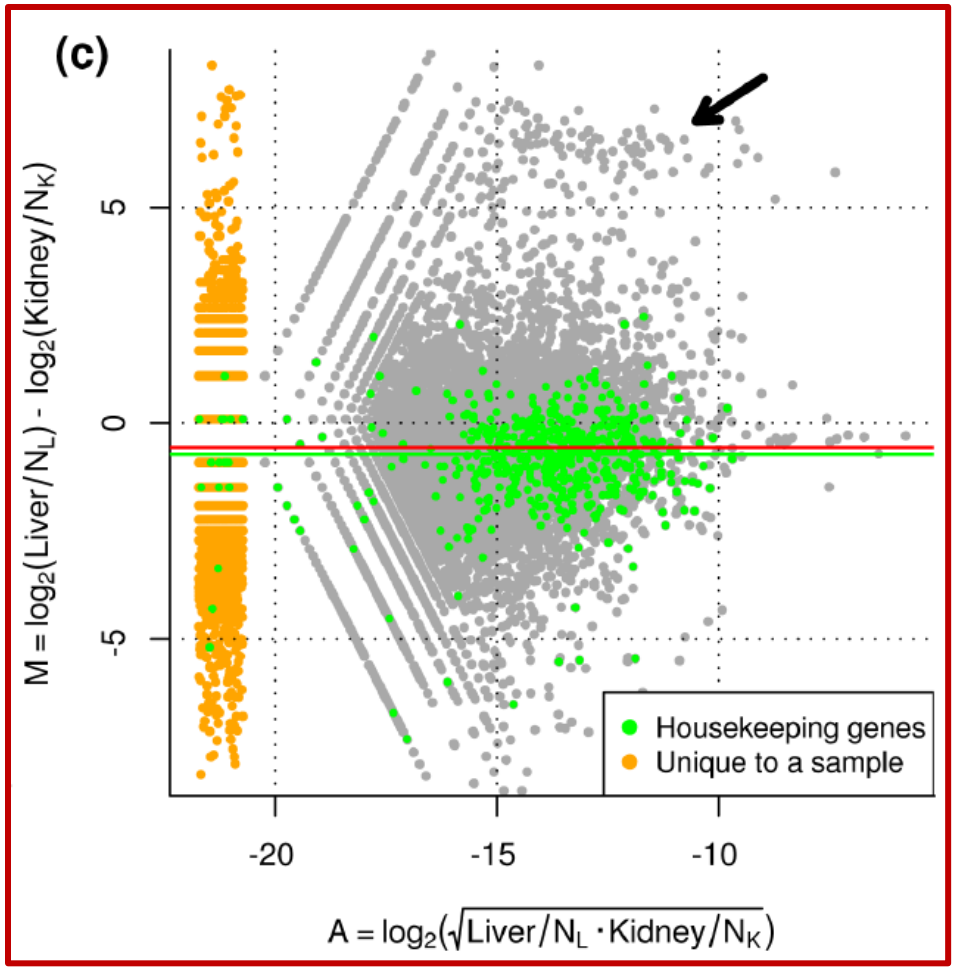
Normalized by total number of reads in each sample

smoothed distribution for logfold-changes of housekeeping genes

Robinson & Oshlack 2010 Genome Biology 2010, 11:R25.

Normalization for RNA-seq data

MA-plot



Median log-ratio of the housekeeping genes

Estimated TMM normalization factor

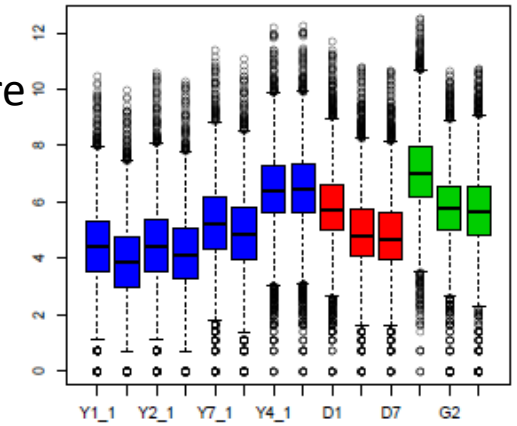
Robinson & Oshlack 2010 Genome Biology 2010, 11:R25.

Normalization using EDASeq package

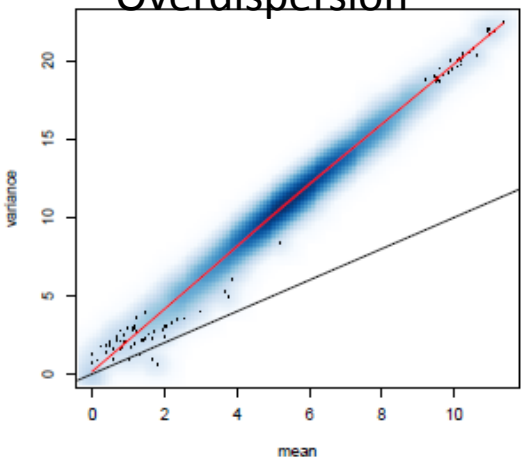
Before

After

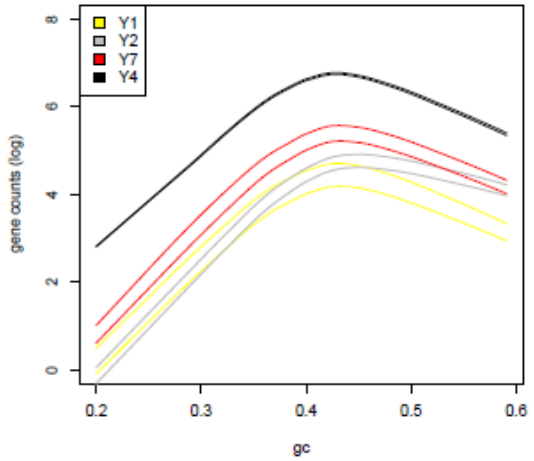
Gene-level count

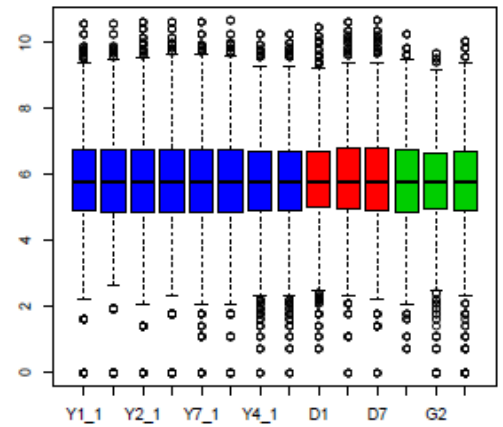


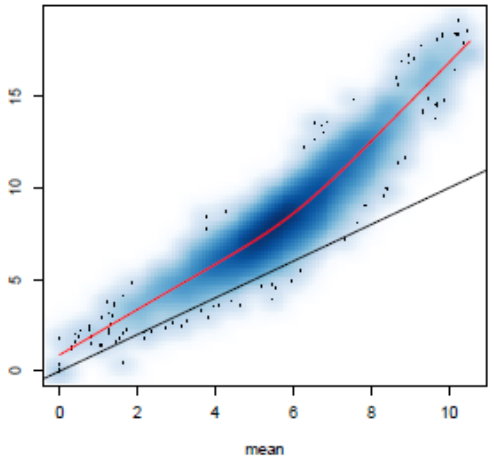
Overdispersion

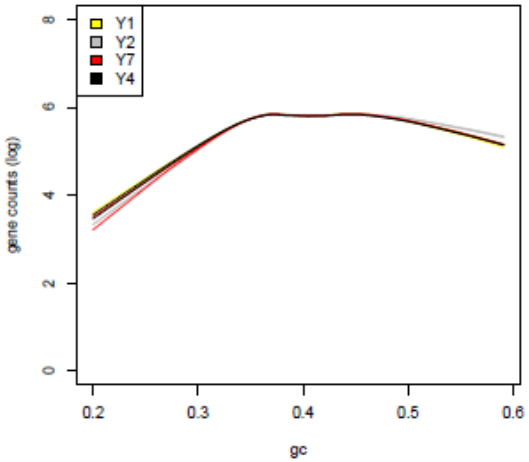


GC content



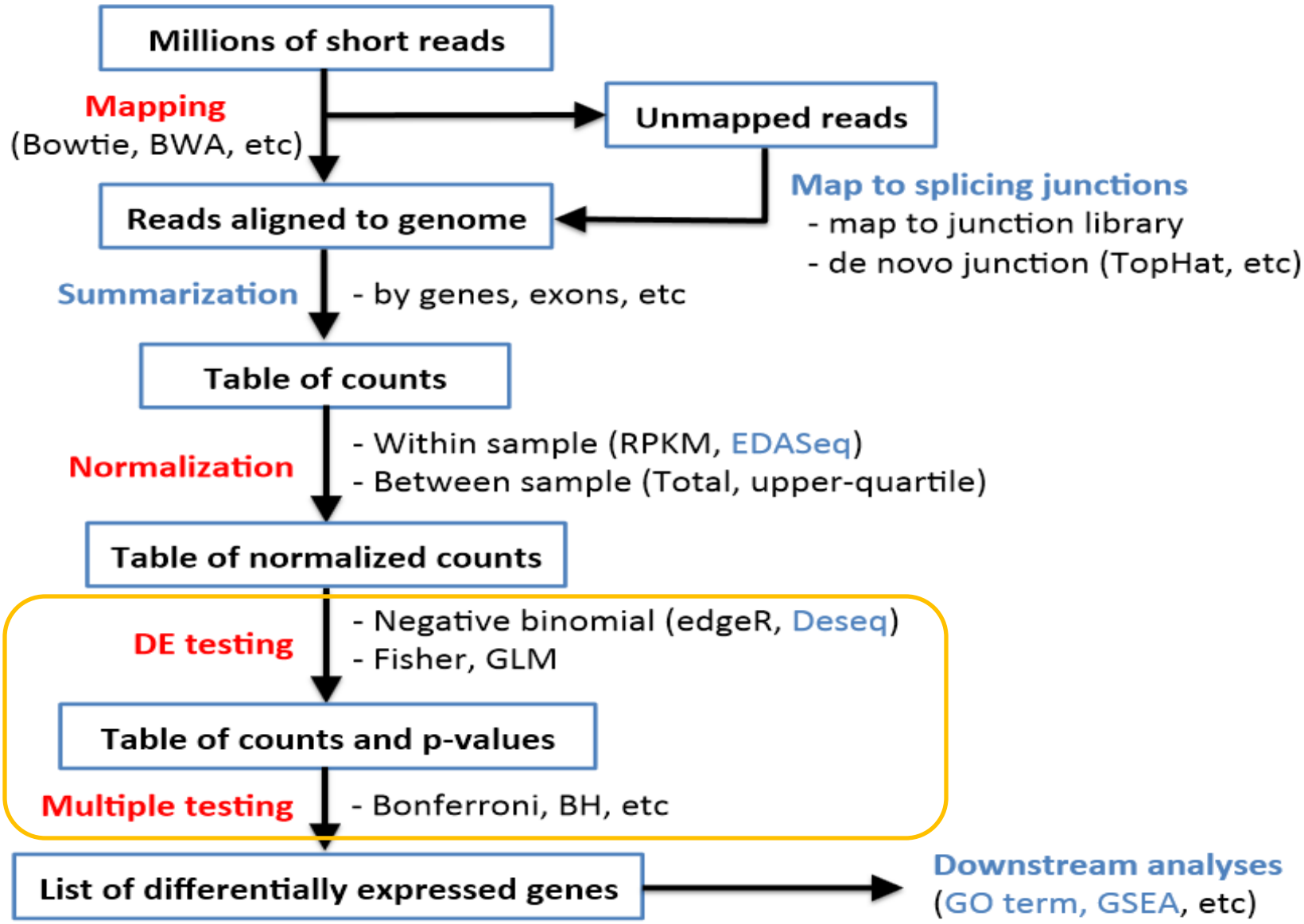






Risso, D. and Dudoit, S. (2011). EDASeq: Exploratory Data Analysis and Normalization for RNA-Seq. R package v 1.2.0

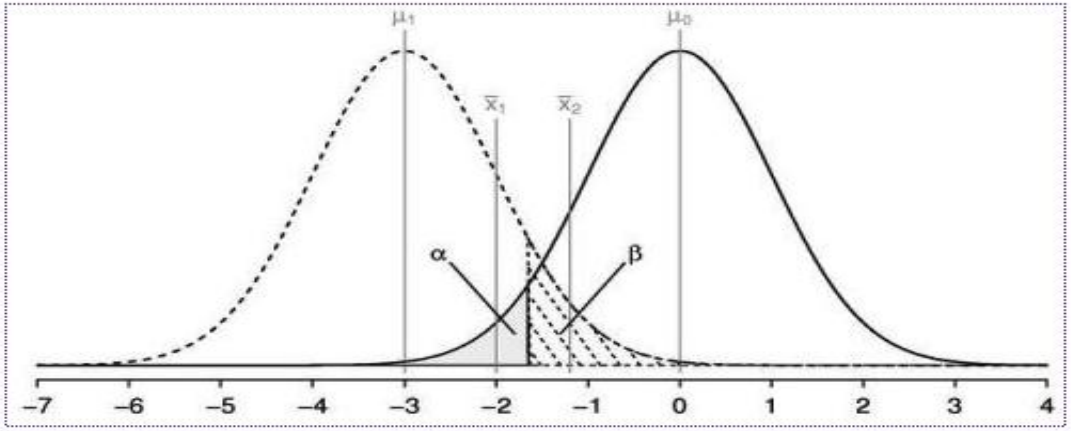
Overview


Millions of short reads
Mapping
(Bowtie, BWA, etc)
Unmapped reads
Map to splicing junctions
- map to junction library
- de novo junction (TopHat, etc)
Reads aligned to genome
Summarization
- by genes, exons, etc
Table of counts
Normalization
- Within sample (RPKM, EDASeq)
- Between sample (Total, upper-quartile)
Table of normalized counts
DE testing
- Negative binomial (edgeR, Deseq)
- Fisher, GLM
Table of counts and p-values
Multiple testing
- Bonferroni, BH, etc
List of differentially expressed genes
Downstream analyses
(GO term, GSEA, etc)

Statistic framework to detect DE genes

- Which genes are being expressed at different levels in different conditions?
- In statistical terms:
  - Do our measurements for the expression of a gene in different RNAseq experiments come from two different distributions or the same distribution?

Hypothesis Testing



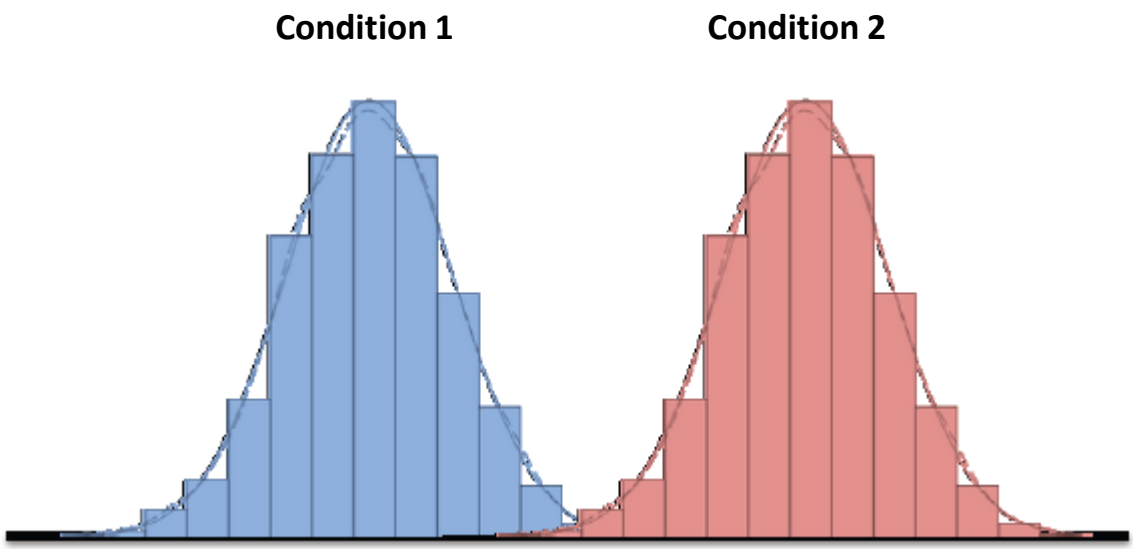
		DECISION	
		Reject H_o	*Fail to Reject H_o*
ACTUAL	*H_o True*	Type I Error *Producer Risk* α-Risk False Positive	**Correct Decision** Confidence Interval = 1- α
	H_a True	**Correct Decision** **Power = 1- β**	Type II Error *Consumer Risk* β-Risk False Negative

H_o: Null Hypothesis H_a: Alternative Hypothesis

H_0: The measurements come from the same distribution (i.e. the gene is being expressed at the same level across conditions.)

A p-value that represents the probability of the null hypothesis is calculated.

How to estimate variance (dispersion)

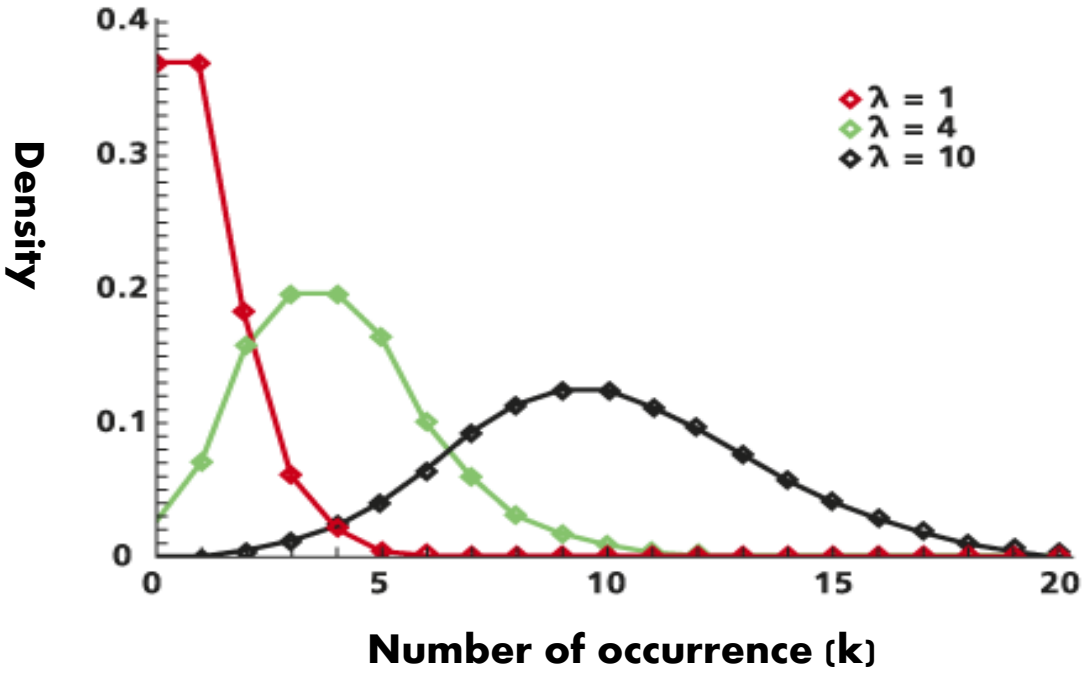


It is unrealistic to have more than a few RNA-seq replicates.

We need to make some assumptions about dispersion.

Model RNA-seq data under Poisson distribution

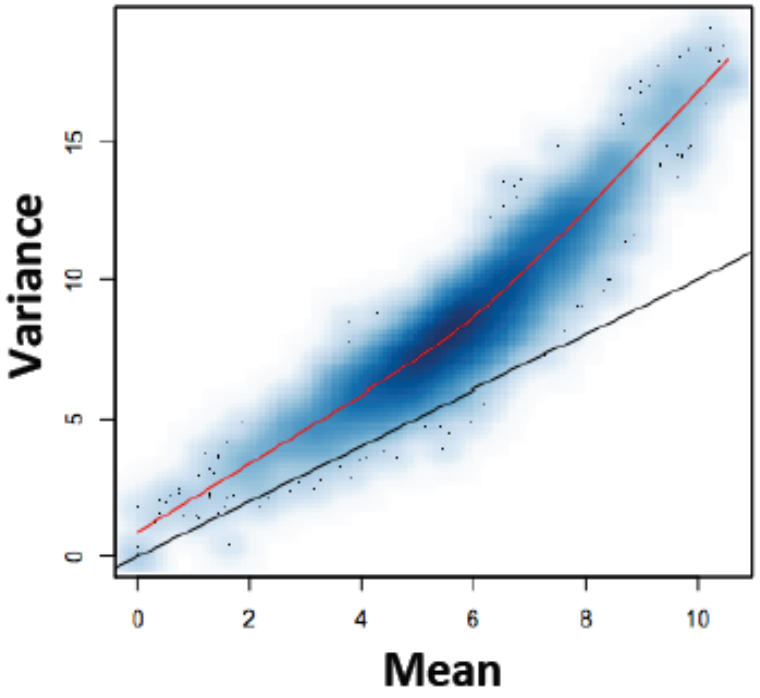
RNA-seq are counts --> counts follows Poisson distribution



$$f(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

$$\lambda = \mathrm{E}(X) = \mathrm{Var}(X).$$

Problem of overdispersion



Source of variation:
- Biological variation
- Technical variation from library prep
- GC bias, transcript length bias
- Flowcell effect

etc

Generalized linear model (GLM)

allows incorporation of known additional variations

Negative binomial

models unexplained variance as

$$\text{Variance} = \text{Mean} + \phi\,\text{Mean}^2$$

Generalized Linear Model (GLM)

- Linear regression that allows distributions such as Poisson
- Can incorporate replicates and other variables

Gene	Untreated				Treated			
	Lib Prep 1		Lib Prep 2		Lib Prep 1		Lib Prep 2	
	FC1	FC2	FC1	FC2	FC1	FC2	FC1	FC2
Gene 1	95	105	110	83	313	301	325	295
Gene 2	10	7	12	5	19	18	24	20
Gene 3	4930	4990	5050	4850	4549	4529	4869	4497
:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:
Total	10M	11M	11M	8M	10M	9M	12M	10M

log(Counts) ~ Treatment + Lib_Prep + Flowcell

Generalized Linear Model (GLM)

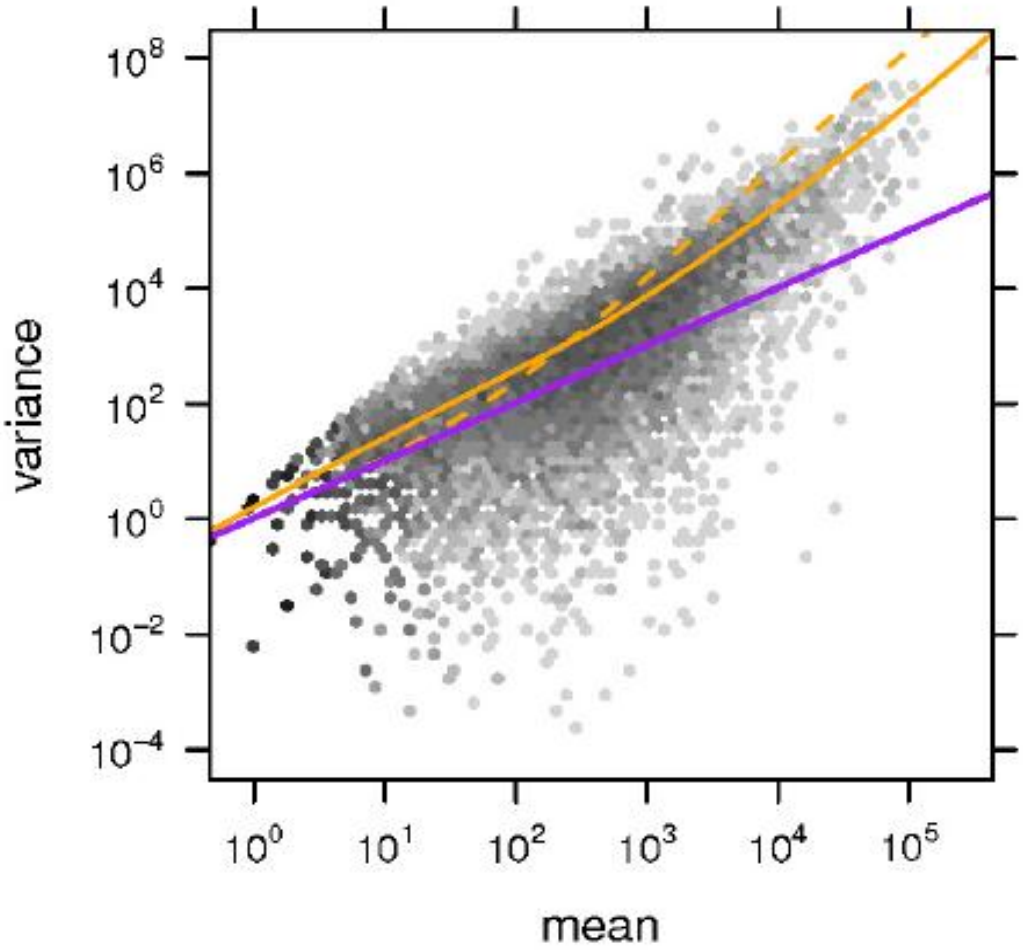
log(Counts) ~ log(Total) + Treatment + Lib_Prep + Flowcell

Design matrix

Treatment	Lib_Prep	Flowcell	Count	Total reads
1	1	1	95	10
1	1	2	105	11
1	2	1	110	11
1	2	2	83	8
2	1	1	313	10
2	1	2	301	9
2	2	1	325	12
2	2	2	295	10

```
> counts <- c(95,105,110,83,313,301,325,295)
> treatment <- c(1,1,1,1,2,2,2,2)
> lib_prep <- c(1,1,2,2,1,1,2,2)
> flowcell <- c(1,2,1,2,1,2,1,2)
> norm.factor <- c(10,11,11,8,10,9,12,10)
> glm.gene1 <- glm(counts ~ treatment + lib_prep + flowcell,
                   family=poisson(),offset=log(norm.factor))
> summary(glm.gene1)
```

Overdisperssion problem



Poisson

Negative binomial (DESeq)

Negative binomial (edgeR)

Anders & Huber, 2010, *Genome Biology* **11**: R106

edgeR

Robinson *et al*., 2009

- Estimates the gene-wise dispersions by maximum likelihood, conditioning on the total count for that gene.
- An empirical Bayes procedure is used to shrink the dispersions towards a consensus value, effectively borrowing information between genes.
- Differential expression is assessed for each gene using Fisher's exact test.

Fisher's exact test

- Very easy to use
- Used with 2x2 contingency table
- Based on hypergeometric distribution

	Untreated	Treated	Total
Gene 1	100	250	350
Other genes	9,999,900	12,999,750	24,999,650
Total	10M	13M	25M

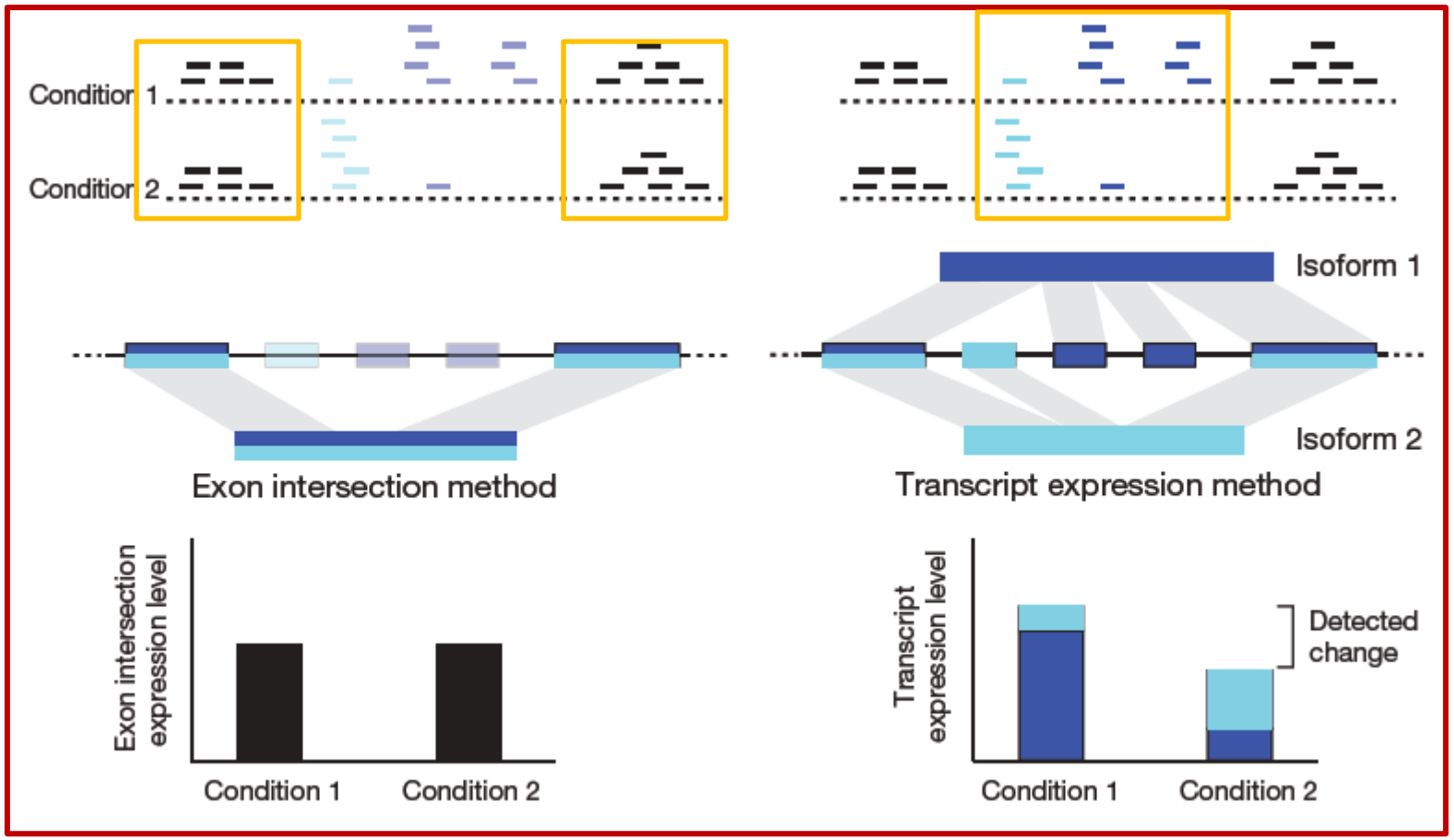
Multiple test correction

- The problem of multiplicity:
  - arises from the fact that as we increase the number of hypotheses in a test, we also increase the likelihood of witnessing a rare event, and therefore, the chance to reject the null hypotheses when it's true (type I error or False-positive).

- Solution: **Bonferroni correction**
  - The most naive way to correct multiplicity
  - If the significance level for the whole family of tests is α, then the Bonferroni correction would be to test each of the individual tests at a significance level of α/n, where n is the number of tests.

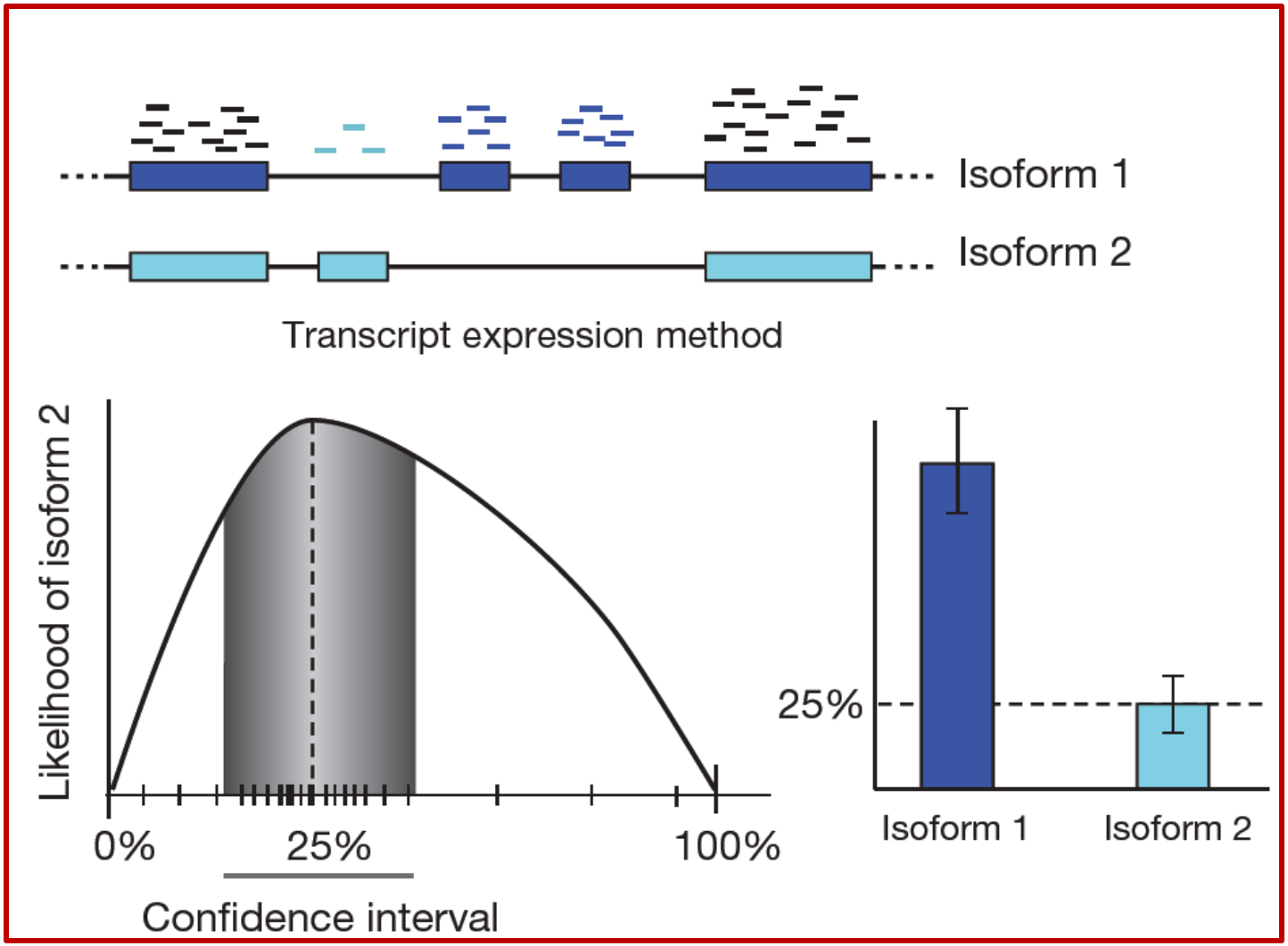
Problem with isoforms

"Read assignment uncertainty" affects expression quantification accuracy



Cufflinks

Isoform-expression methods



DE testing with Cuffdiff

- Based on FPKM (Fragments per kb per million reads)
- Cuffdiff compares the log-ratio of gene's expression in two conditions (a & b) against 0
  - Suppose we write the ratio of expression of a transcript "t" in condition a versus condition b as

$$Y = \frac{FPKM_a}{FPKM_b}$$

-The test statistic T :

$$T = \frac{E[\log(Y)]}{Var[\log(Y))]}$$

  - T is approximately normally distributed and can be calculated as:

$$T = \frac{E[\log(Y)]}{\mathrm{Var}[\log(Y)]} \approx \frac{\log\left(\frac{FPKM_a}{FPKM_b}\right)}{\sqrt{\frac{\mathrm{Var}[FPKM_a]}{FPKM_a^2} + \frac{\mathrm{Var}[FPKM_b]}{FPKM_b^2}}}$$

Cuffdiff vs **count-based packages**

Cuffdiff uses **beta negative binomial** to model
overdispersion and **fragment assignment uncertainty** simultaneously

- Cuffdiff deals with problem of **overdispersion** across replicates
  - Uses LOCFIT to fit a model for fragment count variances in each condition, similar methods as Deseq.
  - If only one replicate is available in each condition, Cuffdiff pools the conditions together to derive a dispersion model
  - Use the variances of fragment *counts* to calculate the variances on a gene's *relative expression level* across replicates
  - Use *relative expression level variances for DE testing.*

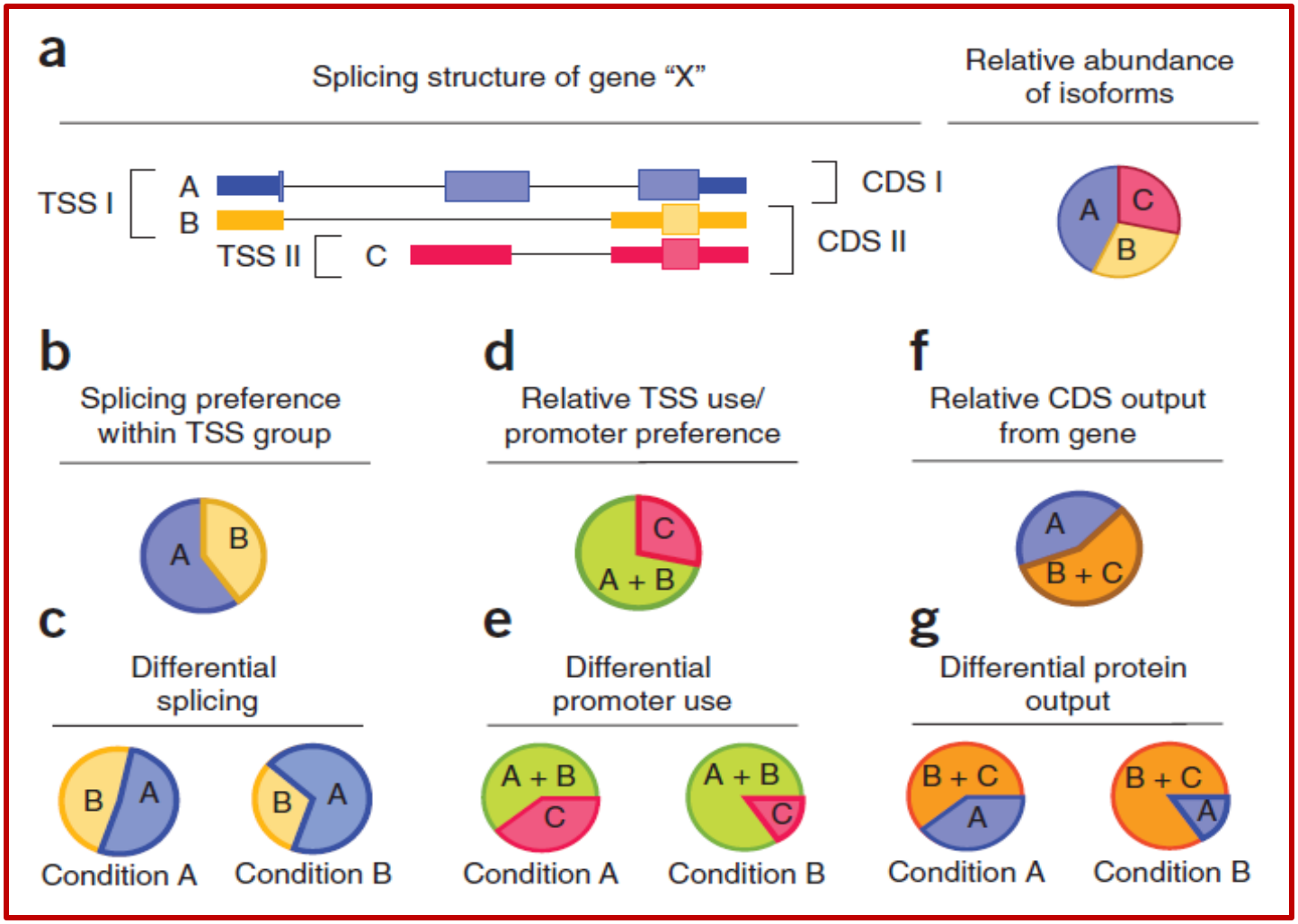
Cuffdiff vs **count-based packages**

Cuffdiff uses **beta negative binomial** to model
overdispersion and **fragment assignment uncertainty** simultaneously

- Cuffdiff uses replicates to capture **fragment assignment uncertainty** between alternative isoforms across replicates
  - pools fragments from replicates and then examines the likelihood surface of the replicate pool.
  - estimated from the bootstrapping procedure to set the parameters of a beta negative binomial distribution as the variance model

Differential analysis with Cuffdiff

Analyzing different groups of transcripts to identify differentially regulated genes



Trapnell et al., 2012 Nat. Protocol 7:562

Other important features in Cufflinks

- How does Cufflinks handle multi-mapped reads?
  - uniformly divide each multi-mapped read to all of the positions it maps to.
  - If multi-mapped read correction is enabled (-u/--multi-read-correct), Cufflinks will improve its estimation by dividing each multi-mapped read probabalistically based on the initial abundance estimation of the genes it maps to, the inferred fragment length, and fragment bias (if bias correction is enabled).

- How does Cufflinks identify and correct for sequence bias?
  - Sequence bias is usually caused by primers used either in PCR or reverse transcription, it appears near the ends of the sequenced fragments.
  - Cufflinks correct this bias by "learning" what sequences are being selected for (or ignored) in a given experiment, and including these measurements in the abundance estimation.
  - Cufflinks will **not** bias correct reads mapping to transcripts with unknown strandedness.
  - For more details, see http://cufflinks.cbcb.umd.edu/howitworks.html#hmul

Downstream data analysis

Functional analysis of DE genes

1. Function annotation: Gene Ontology (GO)

2. Function enrichment test for differential expressed gene set

3. Pathway mapping

4. Profiling clustering

…

Gene Ontology (GO)

- Describes properties of gene products in a structured, standardized way
  - Biological process
  - Molecular function
  - Cellular component
- Hierarchical: broader terms lead to more specific terms
- Can be applied to any species
- www.geneontology.org

Fisher's exact test

for functional enrichment of DE genes

	Genes in category	Genes not in category	Sums
Differentially expressed genes	**k**	m-k	**m**
Not differentially expressed genes	n-k	N-m-n+k	N-m
Sums	**n**	N-n	**N**

k: # of DE genes are in category

m: # of total DE genes

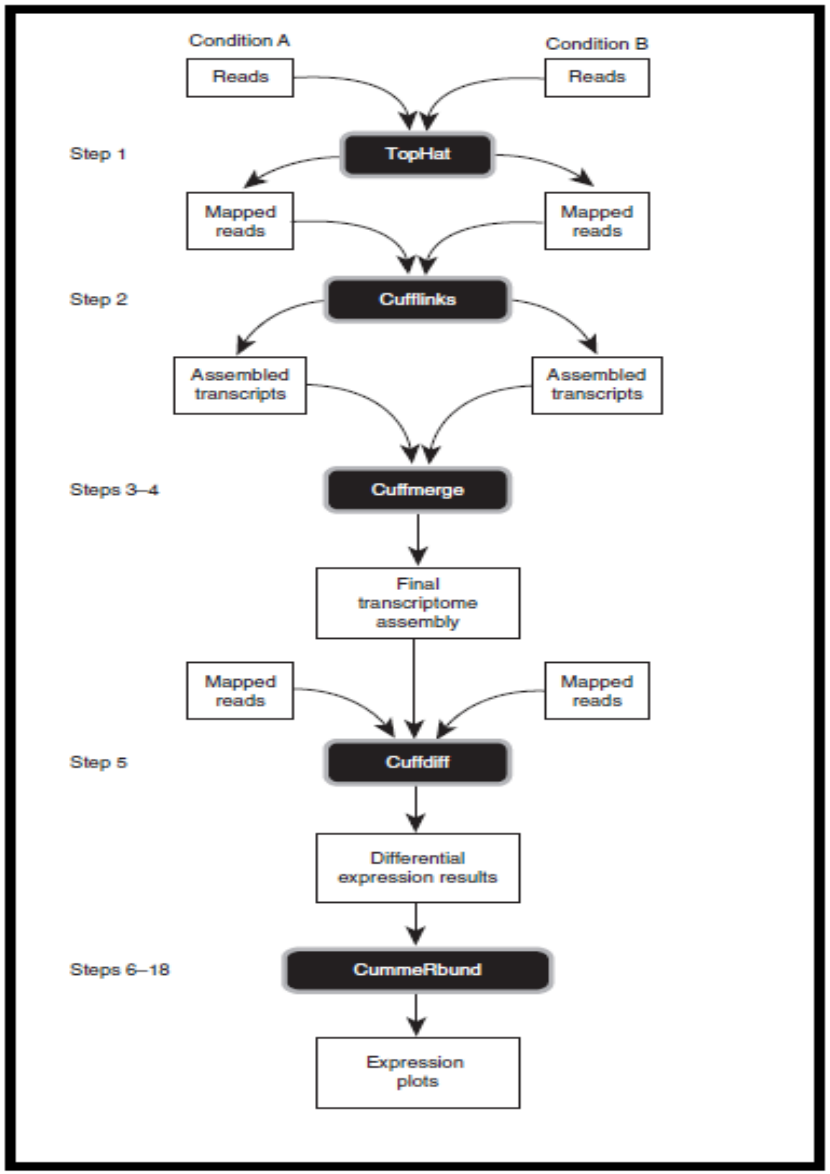
n: # of total genes in category

N: # of genes with valid data in your study

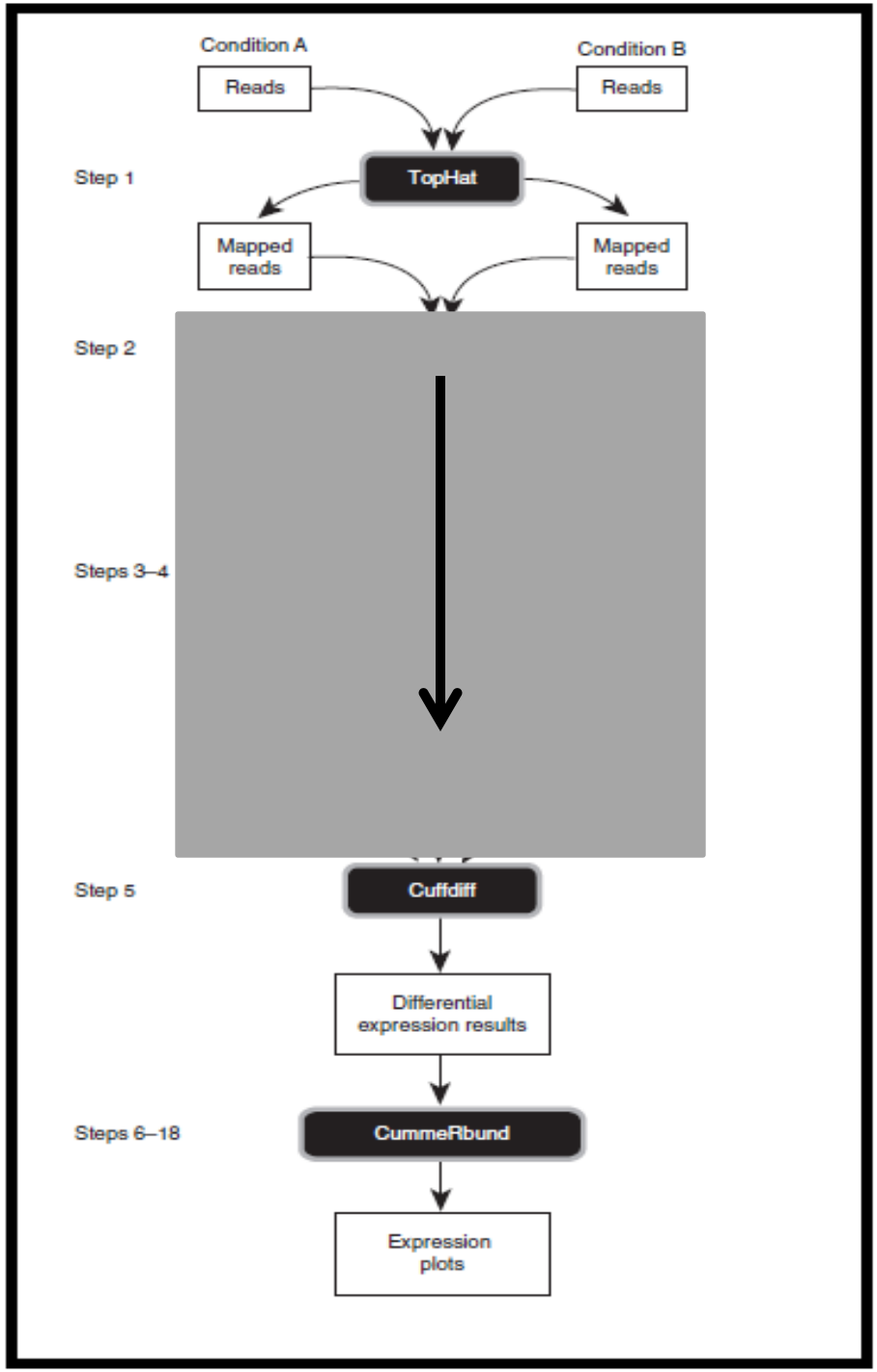
CBSU pipeline for RNA-seq data analysis

- **The Tuxedo protocol**
  - **TopHat**
  - **Cufflinks**
  - **Cuffmerge**
  - **Cuffdiff**
    - To compute FPKM and counts
    - Use FPKM data for DE testing
  - **CummeRbund**
- **edgeR**
  - use count data for DE testing

The Tuxedo protocol



Trapnell *et al.*, 2012 Nat. Protocols 7:562.



Lab exercise:

Differential analysis without gene and transcript discovery

Running Tophat

1. **Reference Genome**
   - **FASTA file**
2. **indexed by bowtie-build**
   - **Genome Annotation**
   - **GFF or GTF files**
   - **optional**
3. **Sequence data file**
   - **FASTQ or FASTA**

Using Tophat through Command line

1. Reformat and index the genome fasta file

```
bowtie-build maize.fa maize &
```

2. Do alignment (with or without annotation)

```
tophat –p 3 -o s1_guided -G ZmB73_5a_WGS.gtf --no-novel-juncs
maize s_1_sequence.txt &

tophat –p 3 -o s1_unguided maize s_1_sequence.txt &
```

Manual: http://tophat.cbcb.umd.edu/manual.html

Tophat parameters

- **Library type**
  - fr-unstranded : standard illumina
  - fr-firststrand : strand specifid dUTP method
  - fr-secondstrand : SOLiD
- **Novel junctions**
  - Default: novel junctions.
  - Use --no-novel-juncs to turn it off

Tophat parameters

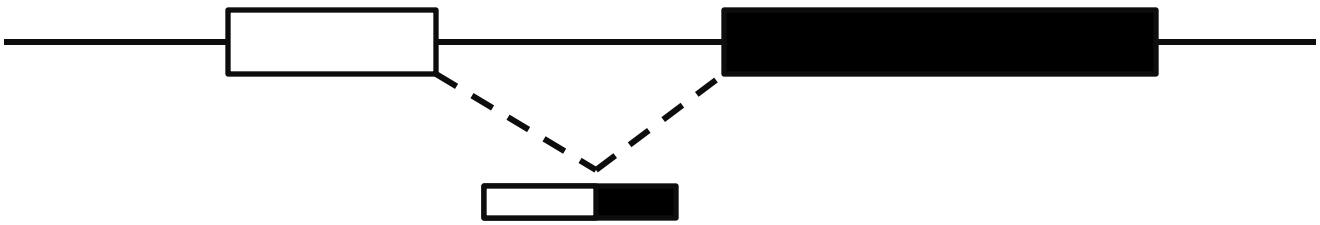
- **For novel junctions**

  -i/--min-intron-length 70 bp

  -I/--max-intron-length 500 kb

  -a/--min-anchor-length 8 bp

  -m/--splice-mismatches 0



Tophat parameters

- **Other parameters**

  -p : number of threads

  -g : maximum number of hits

  --report-secondary-alignments

Running Cuffdiff

Input files

- Tophat output (.bam) from multiple samples.

(biological duplicates should be defined as a single **comma-separated** list)

- GTF/GFF3: gene annotation file

Cuffdiff Parameters

- **Quantification or Assembly**
  - -G: quantification only
  - -g: annotation guided assembly
  - -M: novel transcripts
- **Library type**
  - – fr-unstranded : standard illumina
  - – fr-firststrand : strand specifid dUTP method
  - – fr-secondstrand : SOLiD

Running Cuffdiff

Output files

- Run info
- Read group info
- Read group tracking
  - FPKM tracking files
  - Count tracking files
- Differential expression files

Four attributes: genes, isoforms, tss_groups, and cds.

Computational Resource at Cornell

CBSU / 3CPG BioHPC Laboratory (625 Rhodes Hall)

Office Hour: 1:00 to 3:00 PM every Monday.

Email cbsu@cornell.edu to get an BioHPC lab account

References

- Oshlack *et al.* 2010 From RNA-seq reads to differential expression results. *Genome Biology* 11:220.
- Garber *et al*., 2011 Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods 8:469
- Trapnell *et al*., 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protocols 7:562.
- Robinson & Oshlack 2010 A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 2010, 11:R25.
- Bullard *et al*., 2010 Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 2010, 11:94.
- Robinson *et al*., 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139.
- Anders & Huber 2010 Differential expression analysis for sequence count data. Genome Biol. 11:R106.