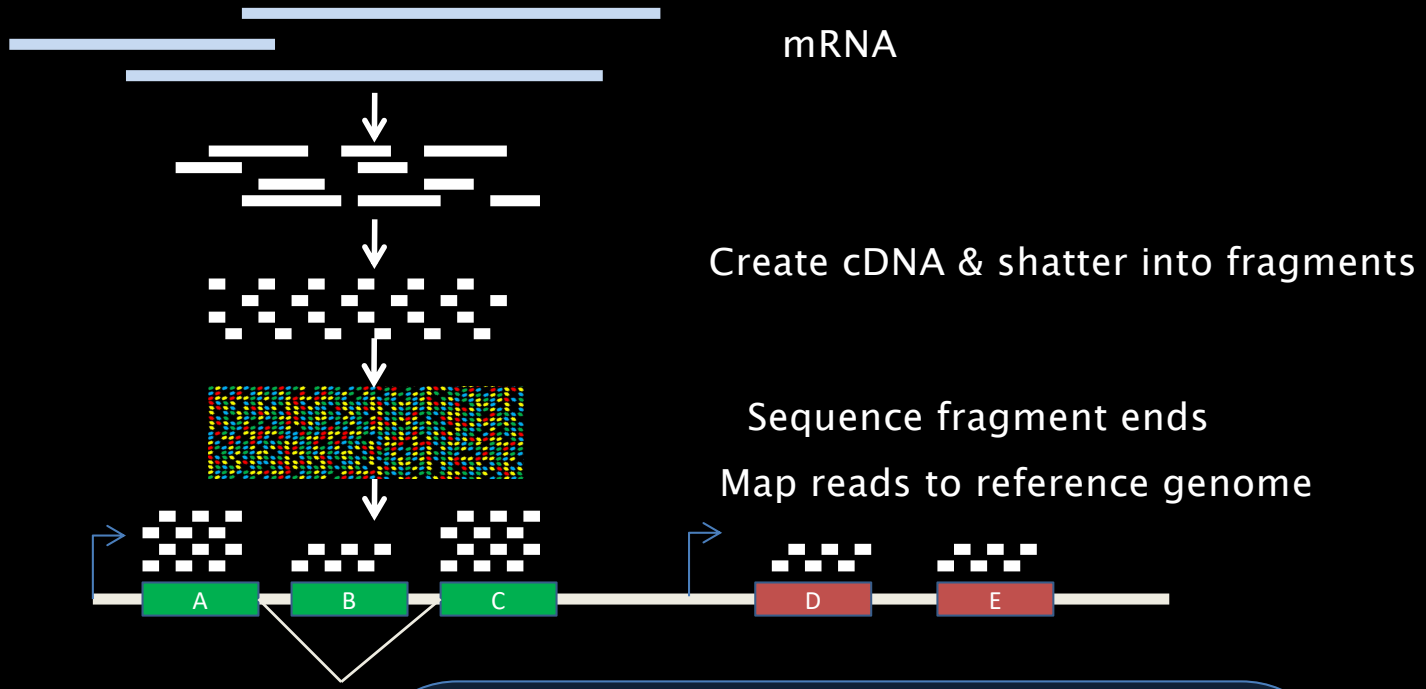
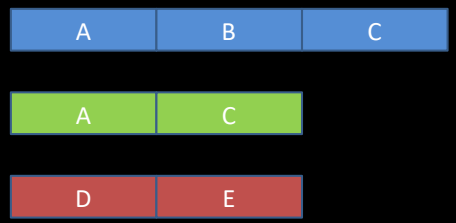
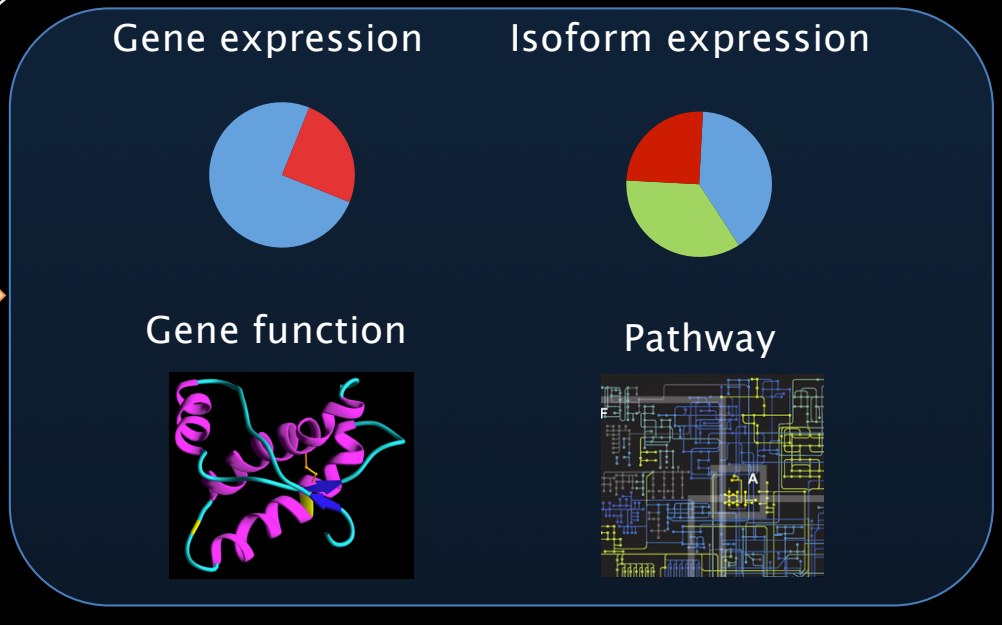
RNA-Seq Analysis

Transcriptome sequencing and data processing



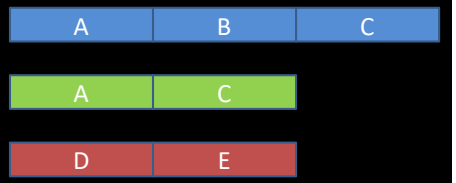
Transcript assembly



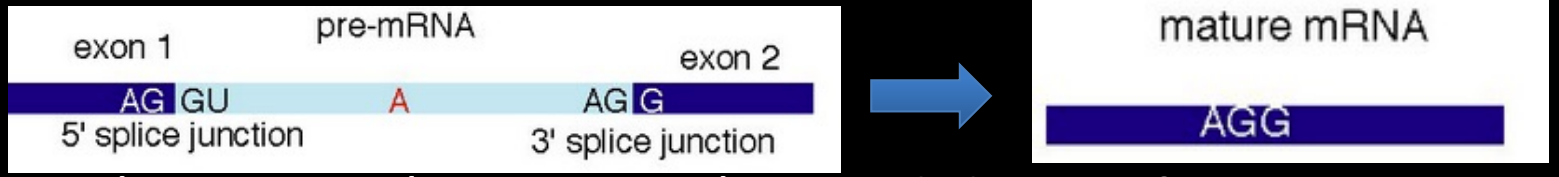


Common terms in transcriptome assembly

- **Isoforms**— transcripts from same gene but with different forms (alternative splicing)

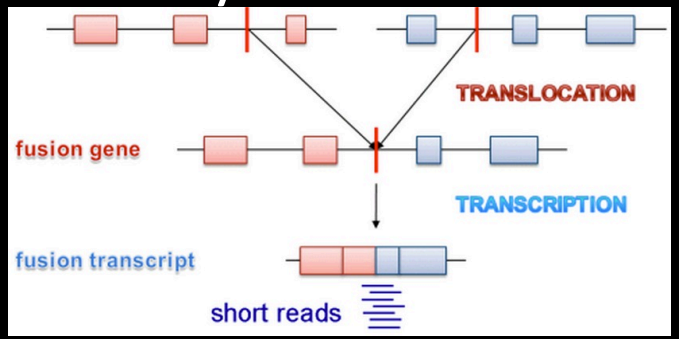


- Splice junction— boundary of alternative splicing



- RPKM (*FPKM*)— Reads (*Fragments*) Per Kilobase of transcript per Million mapped reads

- Fusion transcript— chimeric transcript encoded by a fusion gene or by two different genes



Difficulty and general strategy in transcriptome assembly

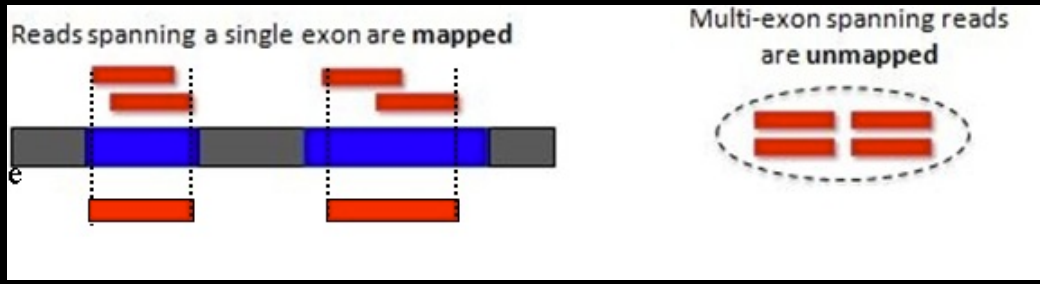
- Difficulties
  - Highly variable abundance
  - Multiple isoforms (different transcripts expressed from the same gene)
- Strategies
  - Reference based assembly
  - De novo assembly
  - Hybrid mode assembly

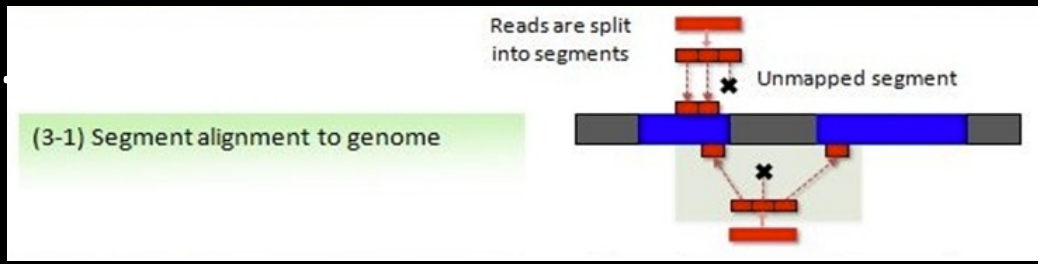
Softwares for transcriptome mapping assembly

- **Tophat-cufflinks**
  - Produce excellent (accurate) assemblies with isoforms
  - Capable of detecting fusion transcript
  - Standard in mapping assembly
  - Full solution
- Scripture
  - Produce good assemblies
  - More robust statistical framework
  - Can handle chip-seq data
  - Only assembly
- Newbler
  - Produce excellent assemblies
  - Capable of detecting various aberrant transcripts, including fusion
  - Assembly only
  - No official support

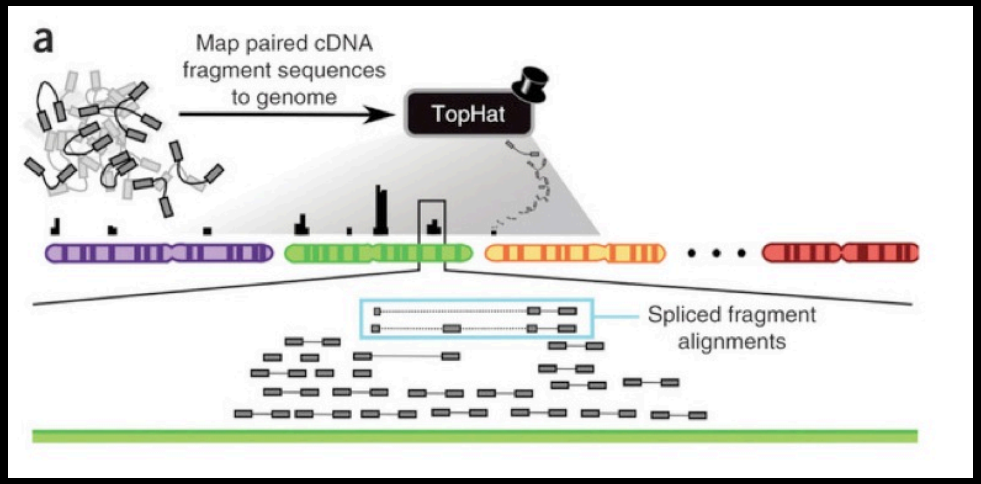
Critical step in reference-based transcriptome assembly-- mapping

- Two categories of softwares
  - Unspliced aligners
    - Map reads to reference
    - BWA, Bowtie …
  - Spliced aligner
    - Map reads to reference considering splicing junctions
    - TopHat, STAR, SpliceMap…



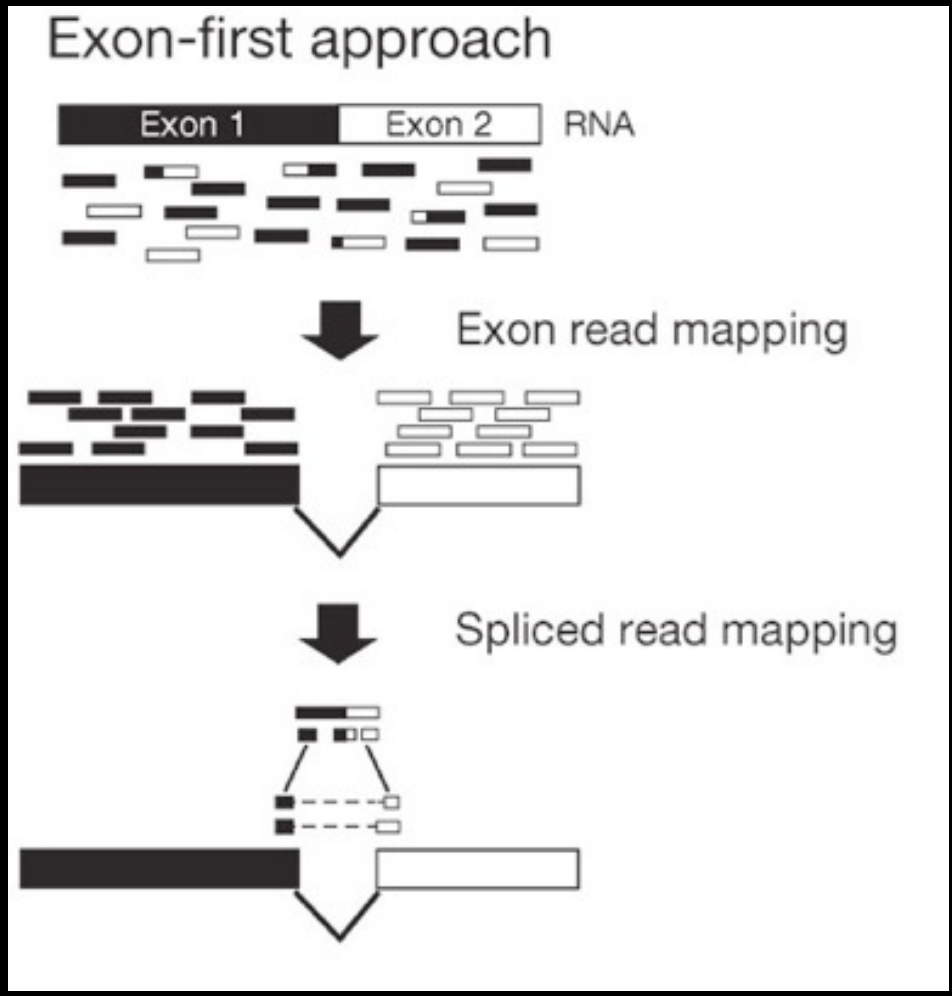


Mapping assembly based on reference genome, step 1 Tophat2 mapping



Tophat

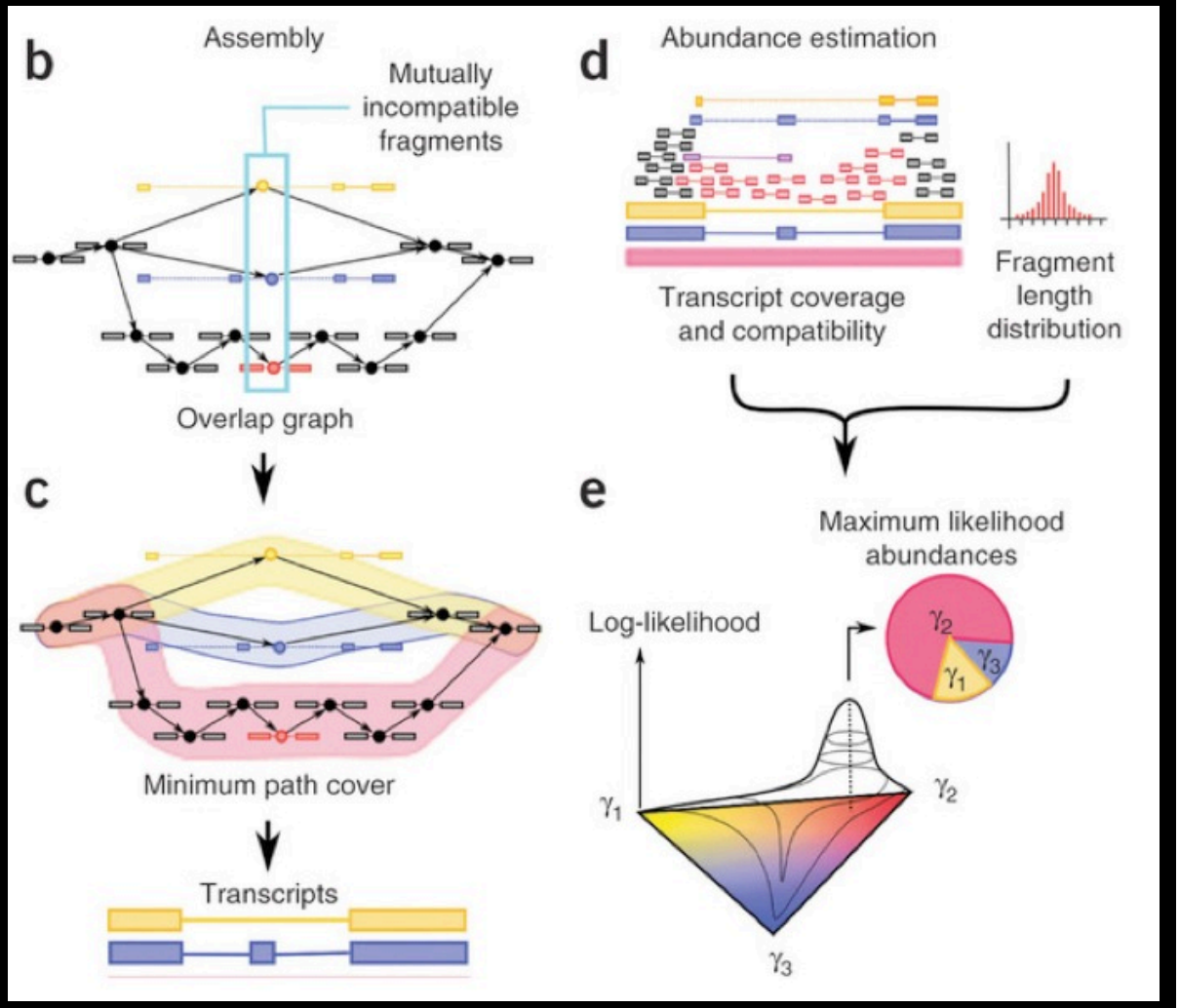
- Align NGS short reads using Bowtie (tophat) or Bowtie2 (tophat2) to reference genome
- Spliced read mapping
- Identify novel splicing sites



Mapping assembly based on reference genome, step 2 cufflinks assembly

cufflinks

- Detect different isofoms according to mapping result from spliced aligner, eg. TopHat
- Capable to detect fusion genes
- Capable to incorporate reference gene information
- Calculate coverage and FPKM



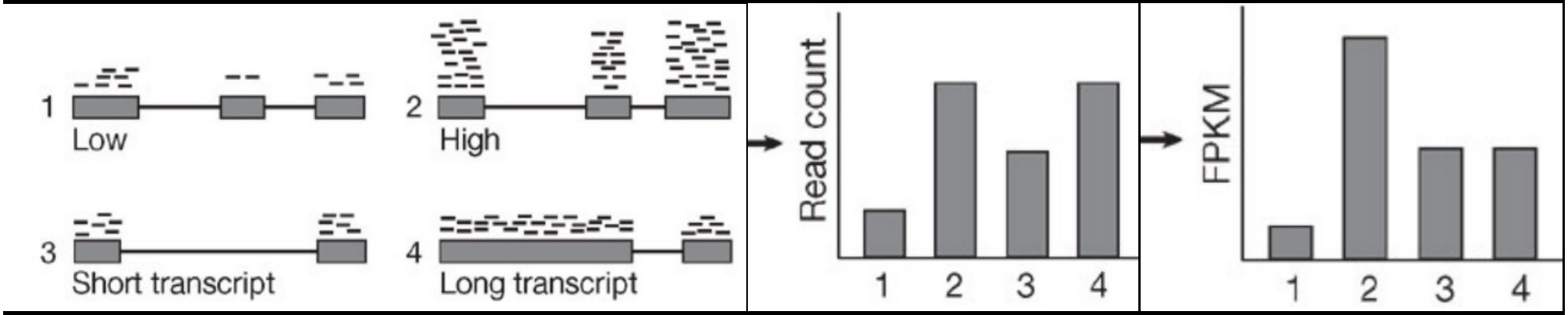
Merge assembly using cuffmerge

- Merge partial transcripts
  - If A is contained in B, A will be merged into B
  - If A and B are overlap and agree on splicing structure, A and B will be merged
  - Avoid merging fragments with disagree structure

Read count vs RPKM(FPKM)

- RPKM (FPKM): Reads (Fragments) per KB per million reads
  - C is the number of mapped reads
  - N is the total mappable reads(M)
  - L is the length of the transcript, bp

$$RPKM(X) = \frac{10^9 \cdot C}{N \cdot L}$$



Garber et al. (2011) Nature Methods 8:469–477.

Directly mapping of RNA-seq without assembly in metatranscriptome

- Why not assembling?
  - Higher complexity of gene pool
  - Extremely variable abundance
  - Existence of highly similar genes
  - Insufficient closely related reference genomes available
  - No suitable software or empirical parameters

- How to quantify the expression and annotate?
  - Select the representative reference genome (time-saving) for functional annotation
  - Map raw reads (rRNA depleted) to the reference genomes

Result of cufflinks

- genes.fpkm_tracking
  - Expression level (FPKM) of each gene
- isoforms.fpkm_tracking
  - Expression level (FPKM) of each isoform
- skipped.gtf
  - Skipped locus with too many segments mapped
- transcripts.gtf
  - Information of assembled isoforms

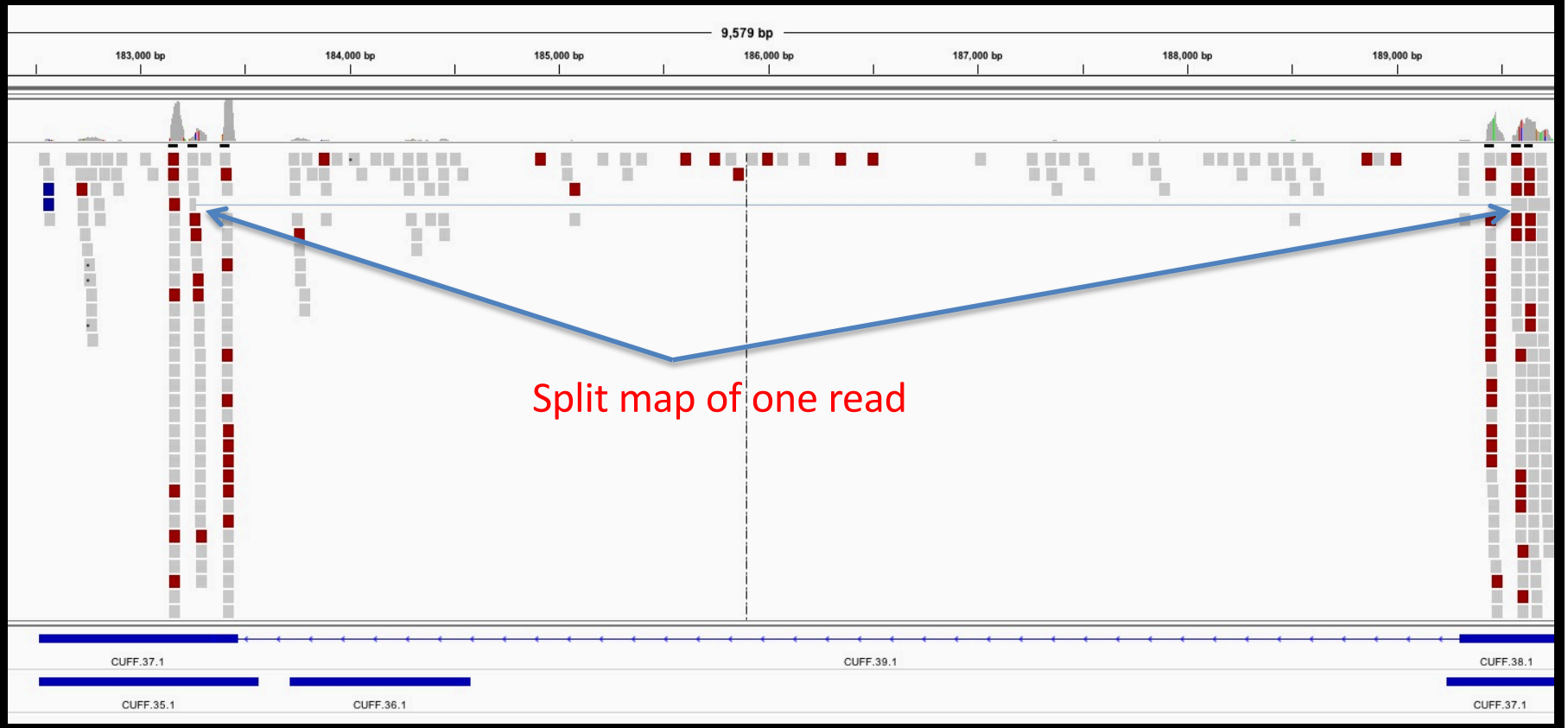
Explanations of the cufflinks assembly

```
gb|BK006935.2|  Cufflinks       transcript      33953   34910   1000    .       .       gene_id "CUFF.1"; transcript_id "CUFF.1.1";
FPKM "667.8460797974"; frac "1.000000"; conf_lo "280.802531"; conf_hi "1054.889629"; cov "1.558862";
gb|BK006935.2|  Cufflinks       exon    33953   34910   1000    .       .       gene_id "CUFF.1"; transcript_id "CUFF.1.1"; exon_num
ber "1"; FPKM "667.8460797974"; frac "1.000000"; conf_lo "280.802531"; conf_hi "1054.889629"; cov "1.558862";
gb|BK006935.2|  Cufflinks       transcript      35139   36320   1000    .       .       gene_id "CUFF.2"; transcript_id "CUFF.2.1";
FPKM "1373.1875806536"; frac "1.000000"; conf_lo "886.730732"; conf_hi "1859.644429"; cov "3.205245";
gb|BK006935.2|  Cufflinks       exon    35139   36320   1000    .       .       gene_id "CUFF.2"; transcript_id "CUFF.2.1"; exon_num
ber "1"; FPKM "1373.1875806536"; frac "1.000000"; conf_lo "886.730732"; conf_hi "1859.644429"; cov "3.205245";
gb|BK006935.2|  Cufflinks       transcript      45912   47983   1000    .       .       gene_id "CUFF.3"; transcript_id "CUFF.3.1";
FPKM "1280.8567352939"; frac "1.000000"; conf_lo "938.983840"; conf_hi "1622.729631"; cov "2.989730";
gb|BK006935.2|  Cufflinks       exon    45912   47983   1000    .       .       gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_num
ber "1"; FPKM "1280.8567352939"; frac "1.000000"; conf_lo "938.983840"; conf_hi "1622.729631"; cov "2.989730";
gb|BK006935.2|  Cufflinks       transcript      51829   52750   1000    .       .       gene_id "CUFF.4"; transcript_id "CUFF.4.1";
FPKM "935.0838994230"; frac "1.000000"; conf_lo "467.541950"; conf_hi "1402.625849"; cov "2.182639";
gb|BK006935.2|  Cufflinks       exon    51829   52750   1000    .       .       gene_id "CUFF.4"; transcript_id "CUFF.4.1"; exon_num
ber "1"; FPKM "935.0838994230"; frac "1.000000"; conf_lo "467.541950"; conf_hi "1402.625849"; cov "2.182639";
```

Attribute	Example	Description
gene_id	CUFF.1	Cufflinks gene id
transcript_id	CUFF.1.1	Cufflinks transcript id
FPKM	101.267	Isoform-level relative abundance in Fragments Per Kilobase of exon model per Million mapped fragments
frac	0.7647	Reserved. Please ignore, as this attribute may be deprecated in the future

conf_lo	0.07	Lower bound of the 95% confidence interval of the abundance of this isoform, as a fraction of the isoform abundance. That is, lower bound = FPKM * (1.0 - conf_lo)
conf_hi	0.1102	Upper bound of the 95% confidence interval of the abundance of this isoform, as a fraction of the isoform abundance. That is, upper bound = FPKM * (1.0 + conf_lo)
cov	100.765	Estimate for the absolute depth of read coverage across the whole transcript
full_read_support	yes	When RABT assembly is used, this attribute reports whether or not all introns and internal exons were fully covered by reads from the data.

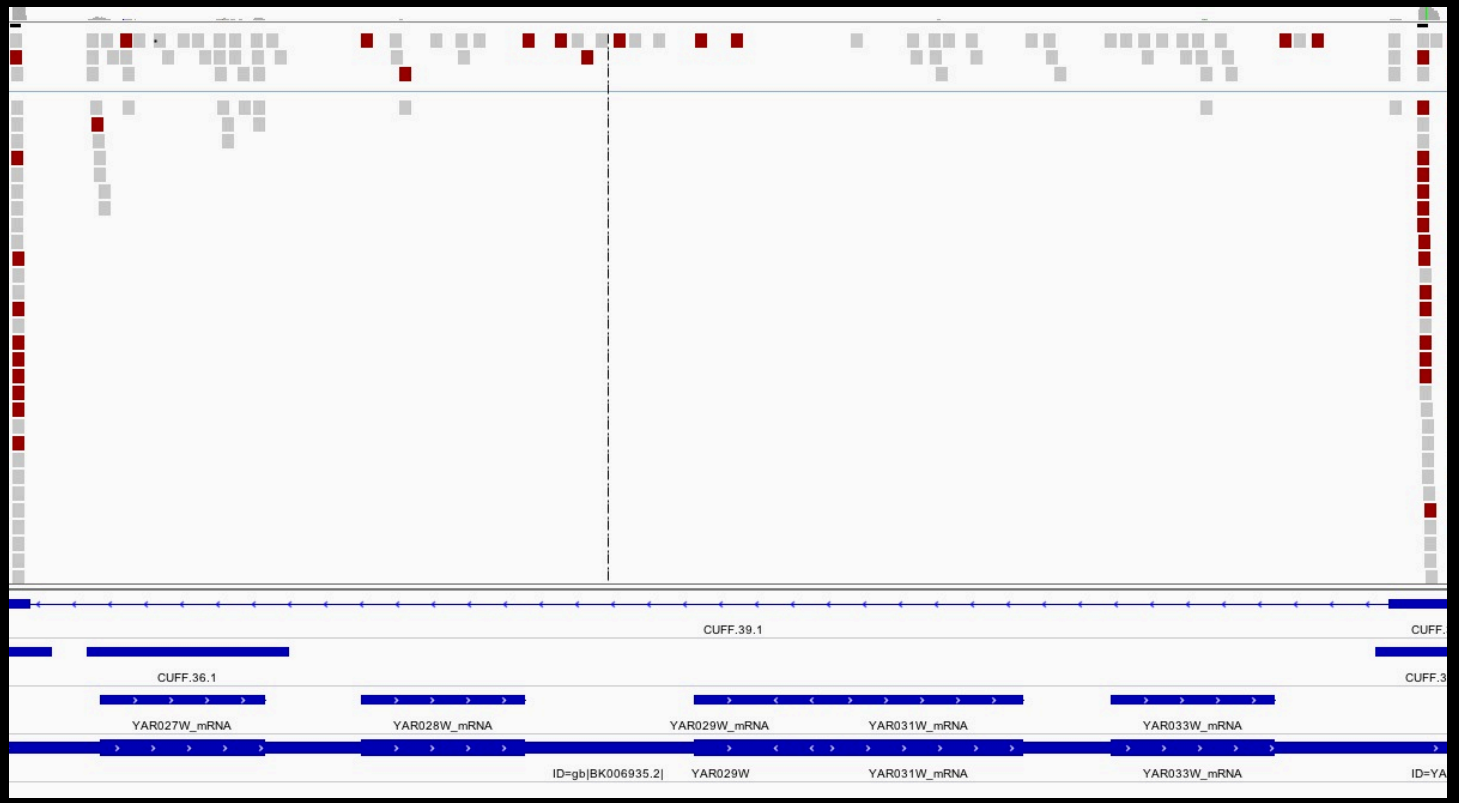
What's the difference between transcripts detected between unspliced aligner and spliced aligner?

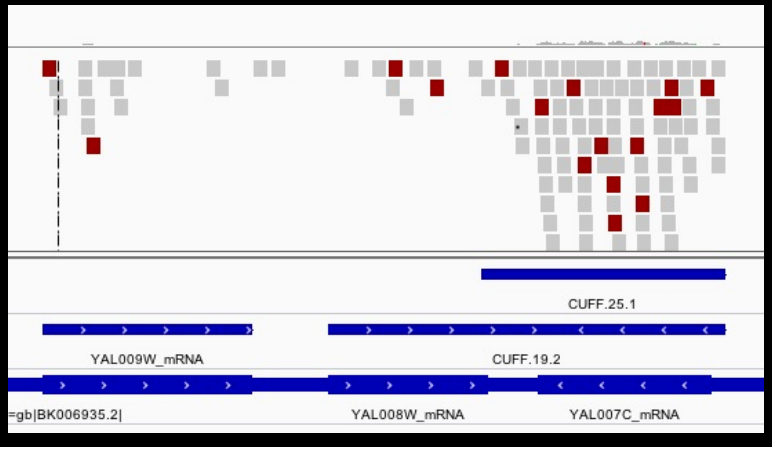


TopHat
Bwa

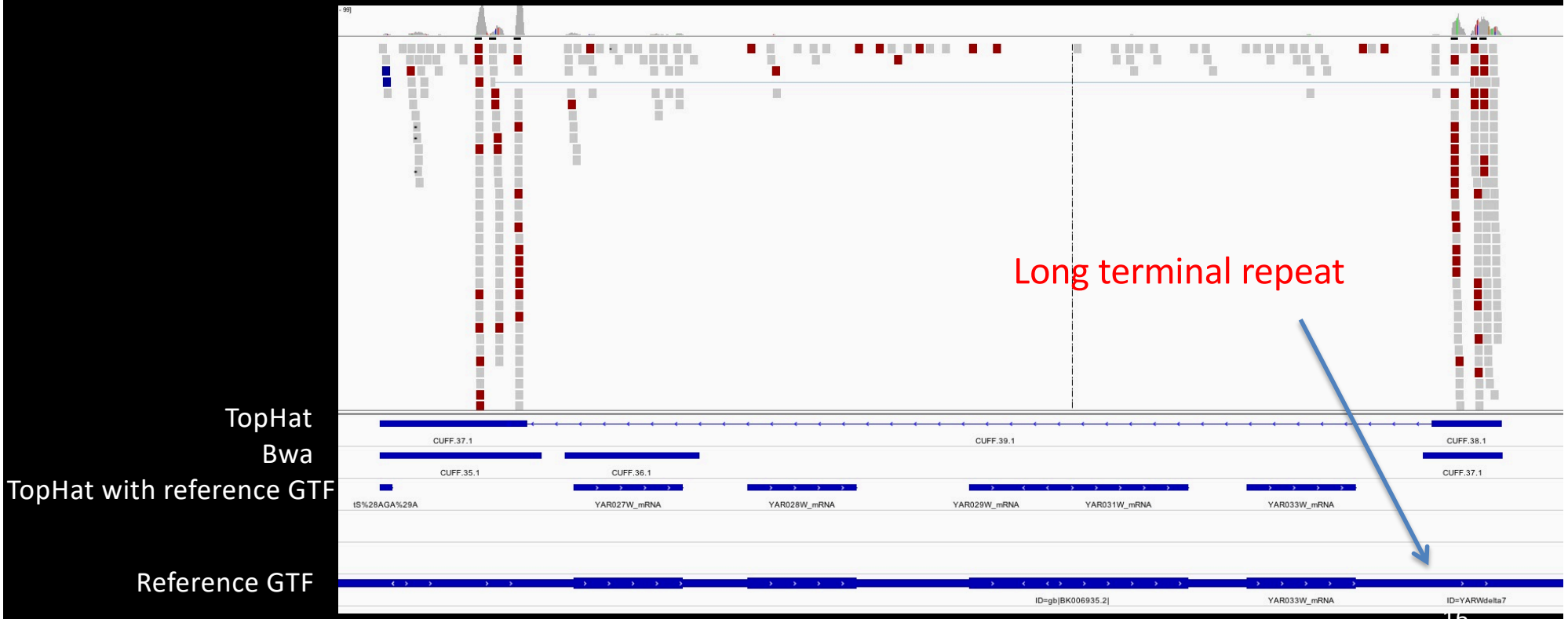
exon 183232 183268
exon 189552 189753

Incorporating reference GTF would increase completeness or novelty of the assembled transcripts

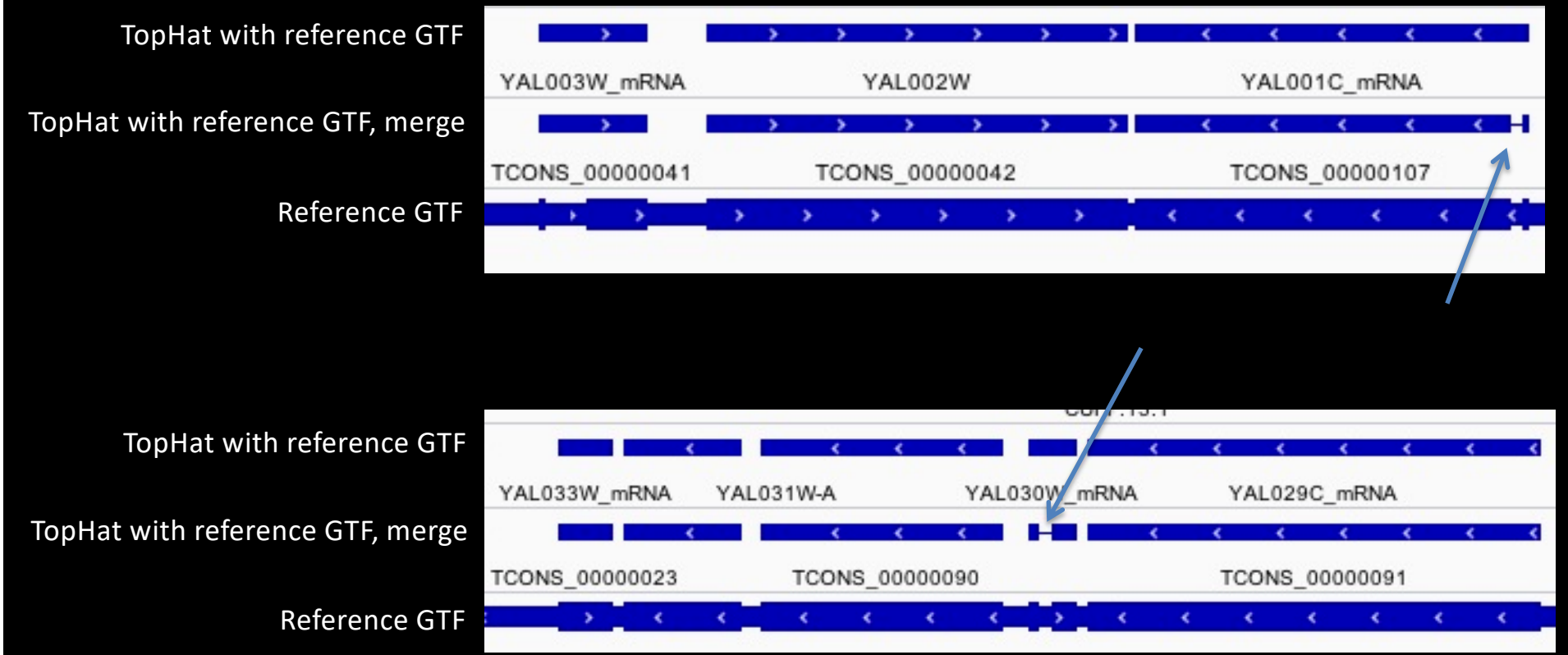




Incorporating reference GTF would reduce false positive detection of the transcripts



Merged assembly reveal more subtle transcript structures



Practice

- Use Tophat2 and Cufflinks to do the mapping assembly
- Use unspliced aligner BWA and cufflinks to do the mapping assembly
- Visualize and compare two assemblies
- Assemble based on genome annotation
- Calculate gene expression levels in a metatranscriptomic community

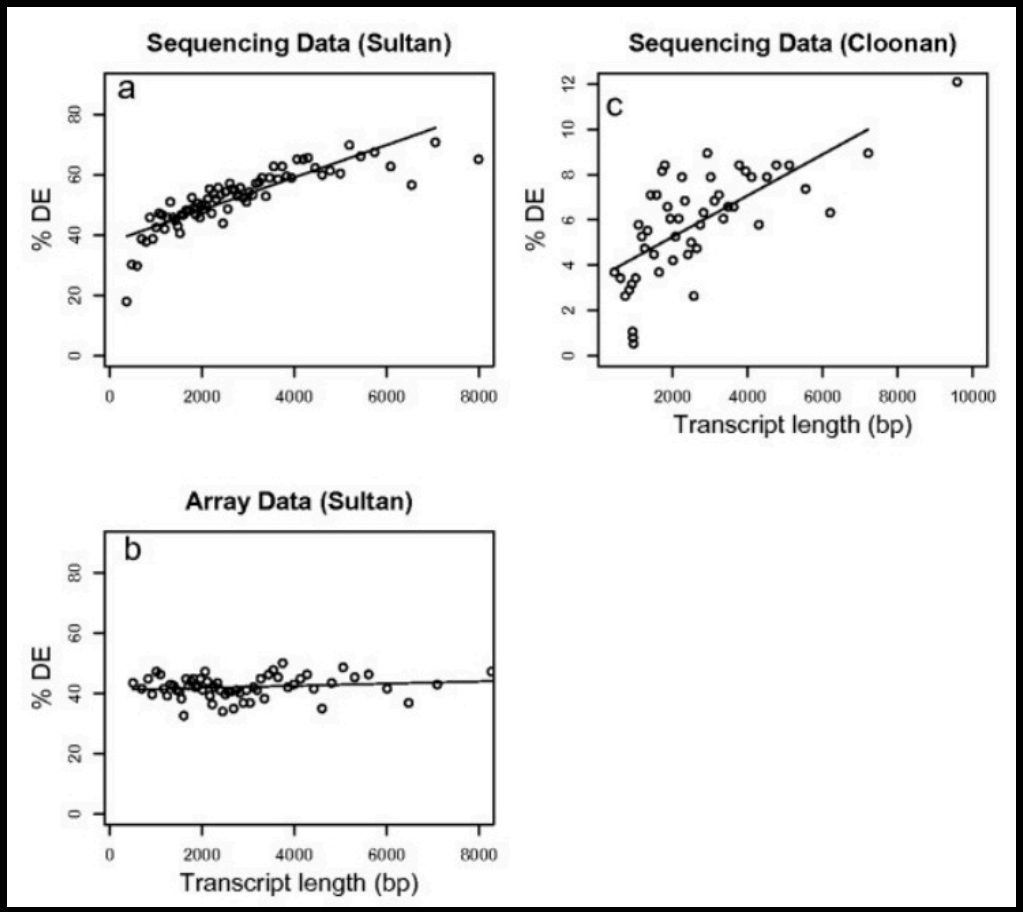
Identification of differentially expressed genes (DEGs)

- ~~Direct~~ goal of some experiment

- Up- or Down-regulated genes are usually related to the phenotype/condition difference

- Some regulatory elements could be detected within co-expression DE network

Is the length correction in RPKM solving the bias in DE analysis?



Not completely

Oshlack 2009

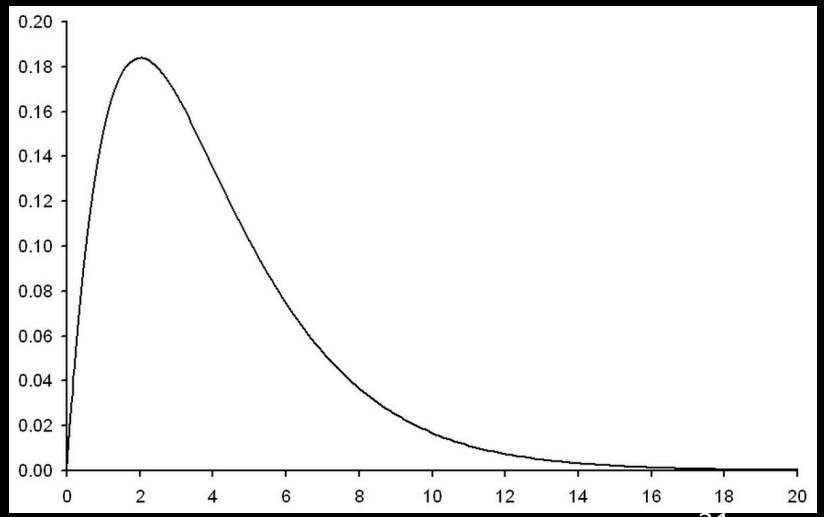
Differential expression, 2 x 2 contingency table for read count

	condition 1	condition 2	Total
Gene x	n_{11}	n_{12}	$n_{11} + n_{12}$
Remaining genes	n_{21}	n_{22}	$n_{21} + n_{22}$
Total	$n_{11} + n_{21}$	$n_{12} + n_{22}$	N

Fisher's exact test or chi-square test– check whether the proportion of gene x was significant varied between two conditions

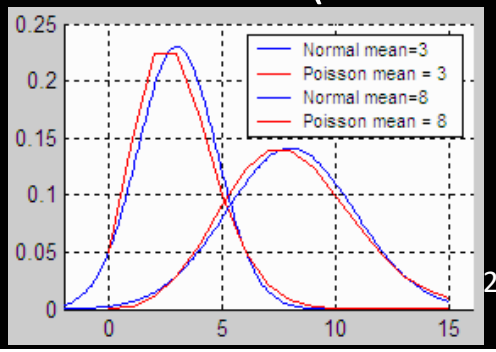
- The null hypothesis: the proportions of counts for certain gene x amongst two samples are the same

- Fundamental problem: a complete lack of knowledge about biological variation

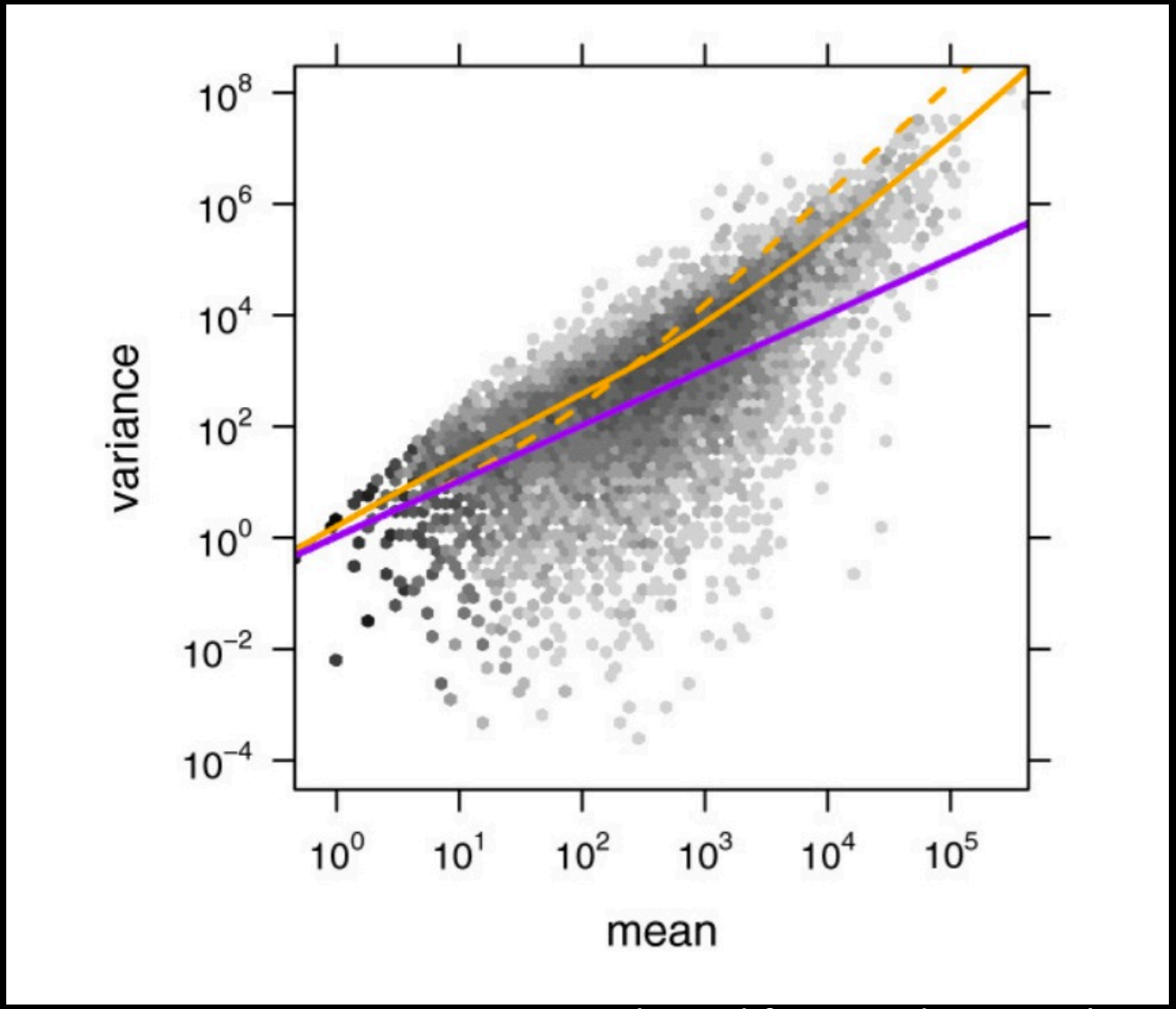


Detecting differentially expression: Poisson process

- Why Poisson?
  - Each read from certain gene has extremely low probability to be observed.
  - The sampling size is enormous in NGS.
  - The mean and variation can be estimated from the biological replicates.
  - T-test based on normal distribution is special case when expected read count is large enough (symmetry) and replicates are sufficient (>20 to estimate stable mean and variance)

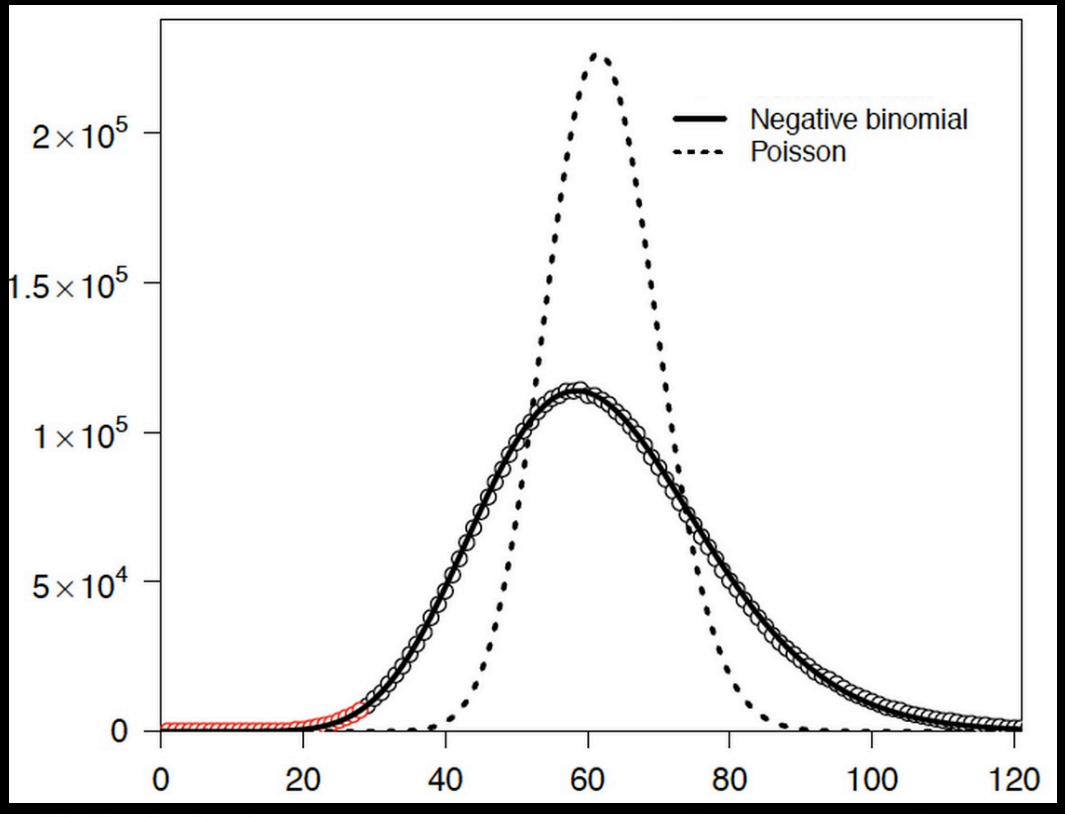


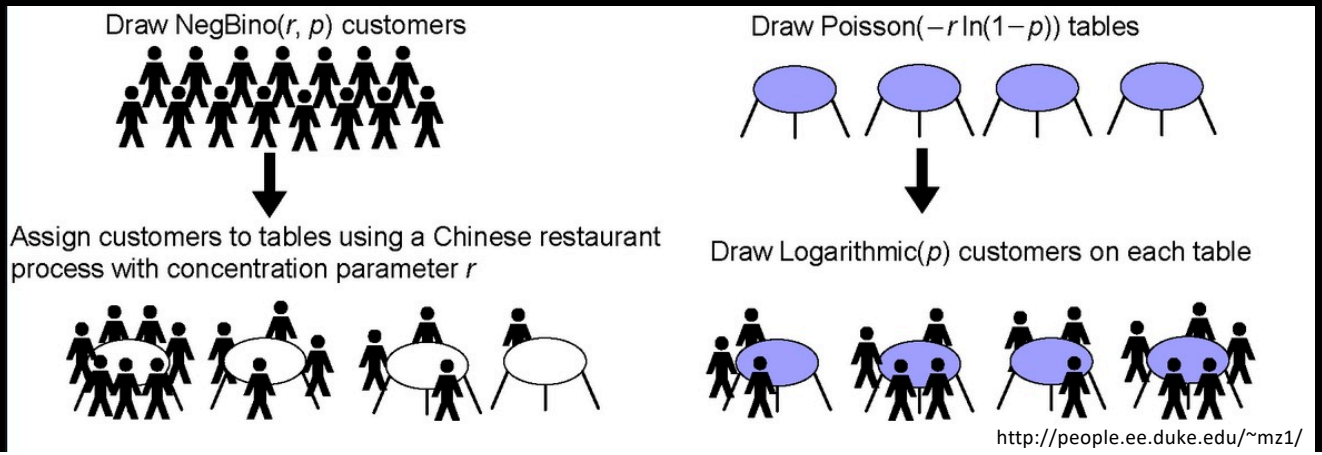
Problem of Poisson distribution



Adapted from Anders & Huber, 2010

Negative binomial distribution to model over-dispersion in sequence count

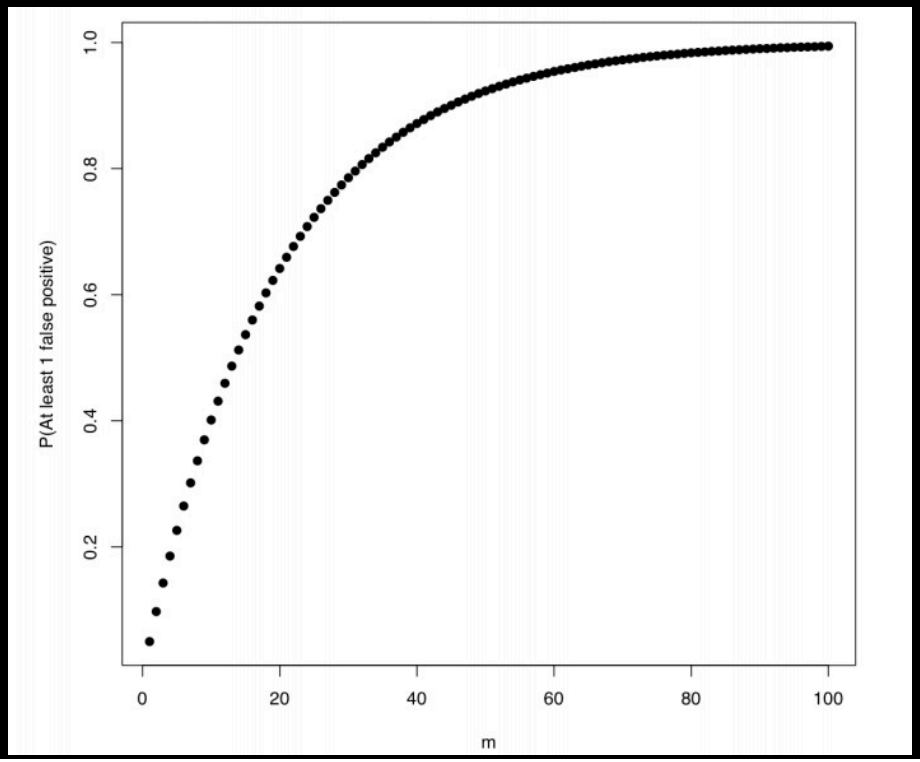




Problem of multiple testing

- Thousands of genes to test their expression change
- Find lots of false positive DE genes (Type-I error)
- Correction is needed to control Type-I error

$$P(\text{at least one Type I error among the 10 tests}) = 1-(1-\alpha)(1-\alpha)\cdots(1-\alpha) = 1-(1-\alpha)^{10}$$

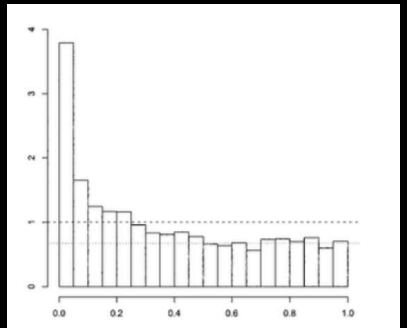


Method to correct multiple testing

- **Bonferroni correction**
  - P<=α/N
  - N is the number of tests, α is the false discovery rate (FDR)

ID	P-value	Rank	iα/N (α=0.05)
Gene_1	**0.001**	1	**0.0025**
Gene_2	**0.0012**	2	**0.0050**
Gene_3	**0.0034**	3	**0.0075**
Gene_4	**0.0039**	4	**0.0100**
Gene_5	**0.0067**	5	**0.0125**
Gene_6	**0.012**	6	**0.0150**
Gene_7	**0.023**	7	0.0175
Gene_8	**0.036**	8	0.0200
Gene_9	**0.043**	9	0.0225
Gene_10	**0.05**	10	0.0250
Gene_11	0.062	11	0.0275
Gene_12	0.069	12	0.0300
Gene_13	0.07	13	0.0325
Gene_14	0.076	14	0.0350
Gene_15	0.1	15	0.0375
Gene_16	0.11	16	0.0400
Gene_17	0.12	17	0.0425
Gene_18	0.14	18	0.0450
Gene_19	0.15	19	0.0475

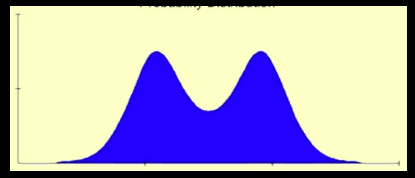
- **Benjamini-Hochberg**
  - P <= iα/N
  - N is the number of tests, α is the type-I error
  - i is the rank (from smallest to largest) of raw p-value
  - Extended as q-value, FDR adjusted p-value



Detect differentially expressed genes using DESeq or Wilcoxon test

- Why DESeq? (normal RNA-seq from single species)
  - One of the best and commonly used
  - Robust statistical framework; read count based (NBD)
  - Non-replicated samples accepted (not recommended)

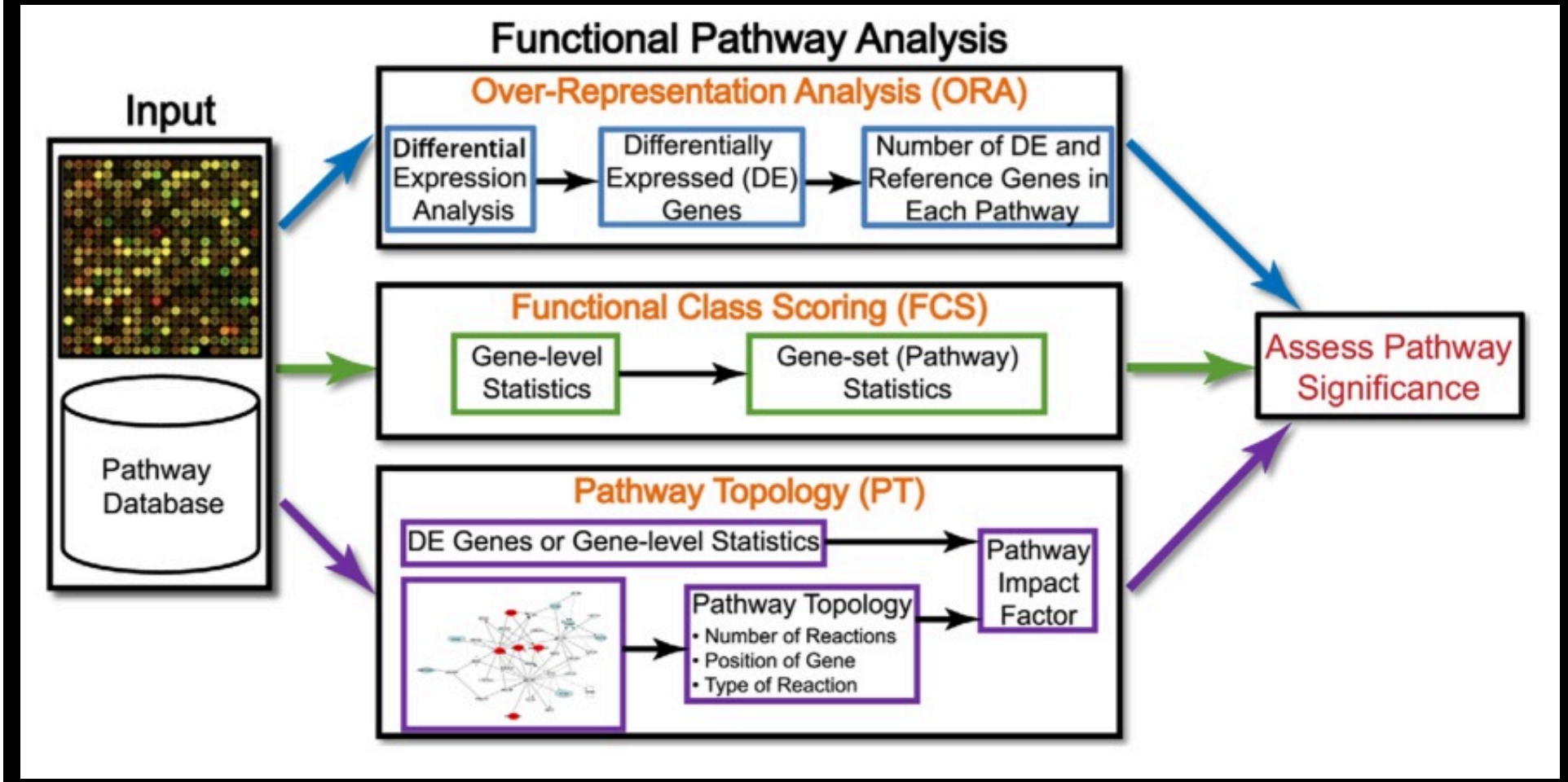
- When Wilcoxon rank-sum test (or other non-parametric test)? (metatransciptome)
  - No replicates to estimate the parameters in the assumed distribution
  - The variability of gene expression is too huge or the distribution may be multimodal



Practice

- Detect differentially expressed genes between two yeast transcriptome using DESeq
- Detect differentially expressed genes between gut metatranscriptome data using Wilcoxon rank sum test

Enrichment of gene sets (eg. pathway)

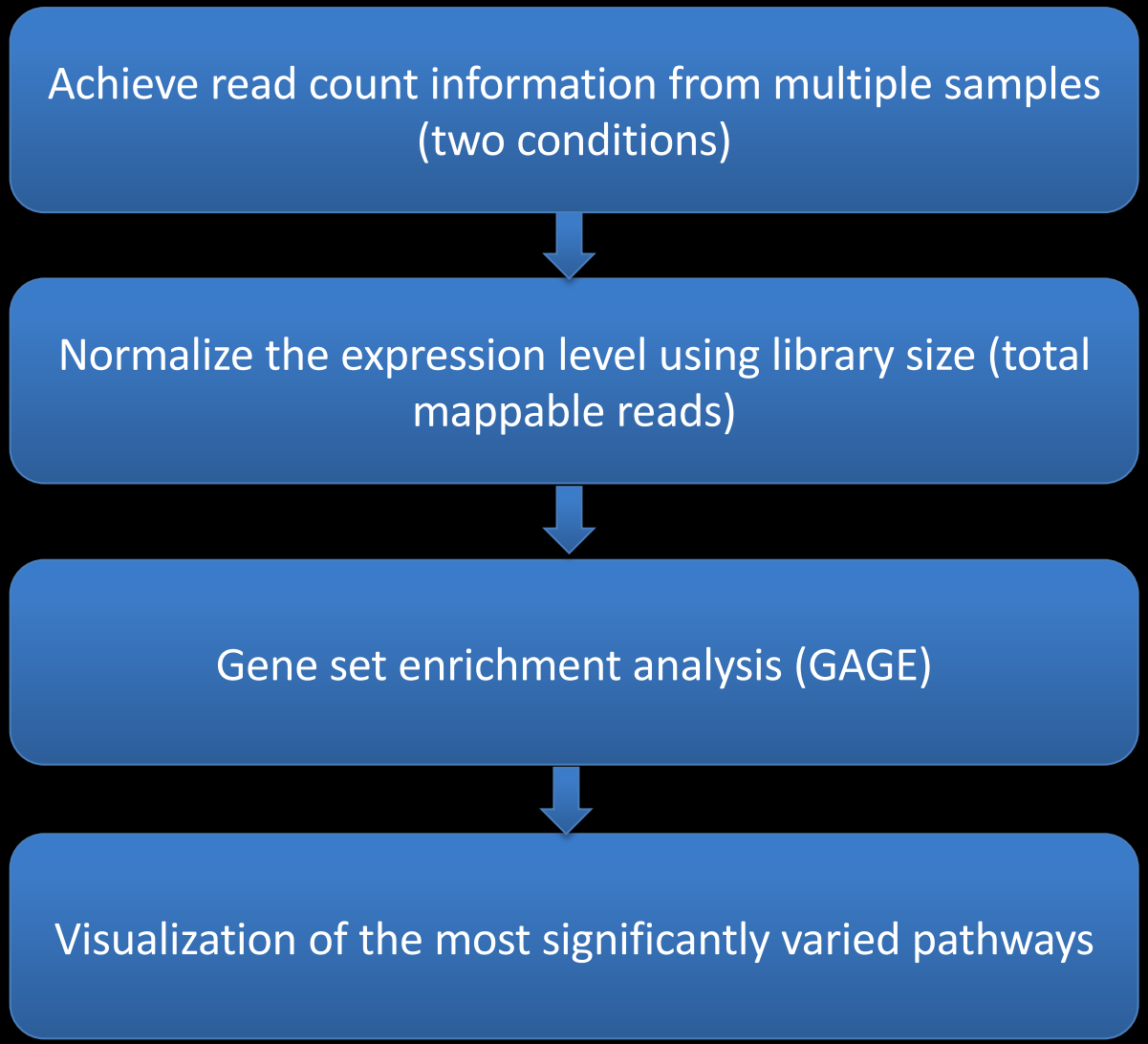


P Khatri - 2012

Tools, their advantages and limitations in pathway (gene sets) enrichment analysis

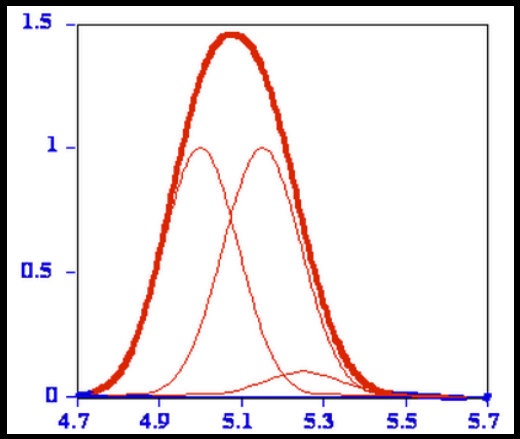
- Over-Representation Analysis (ORA)
  - GOstat, WEGO, GOFFA, etc
  - Hypergeometric, chi-square or binomial distribution based
  - Advantage: simple, quick and easy to interpret
  - Limitation: treat gene equally; use DE genes only; genes are independent; gene sets are independent
- Functional Class Scoring (FCS)
  - GESA, GSVA, GAGE, Globaltest, PADOG, CAMERA, etc
  - Gene metric (z-score, fold change) to pathway metric (KS, wilcoxon, etc), interaction of gene could be incorporated
  - Advantage: quick, robust and provide high confident results; applicable to any gene sets
  - Limitation: Information may not be complete; gene sets are independent
- Pathway Topology (PT)
  - ScorePAGE, SPIA , Pathway-Express, etc
  - Incorporate the information topology and gene-gene interactions from knowledge based database (KEGG, MetaCyc,)
  - Advantage: prior biological knowledge utilized
  - Limitation: pathway topology is flexible; pathway interaction is weakly considered

Workflow of enrichment analysis for pathways



Computational deconvolution

- Deconvolution
  - Dissecting original signal from a mixture of signals



Practice

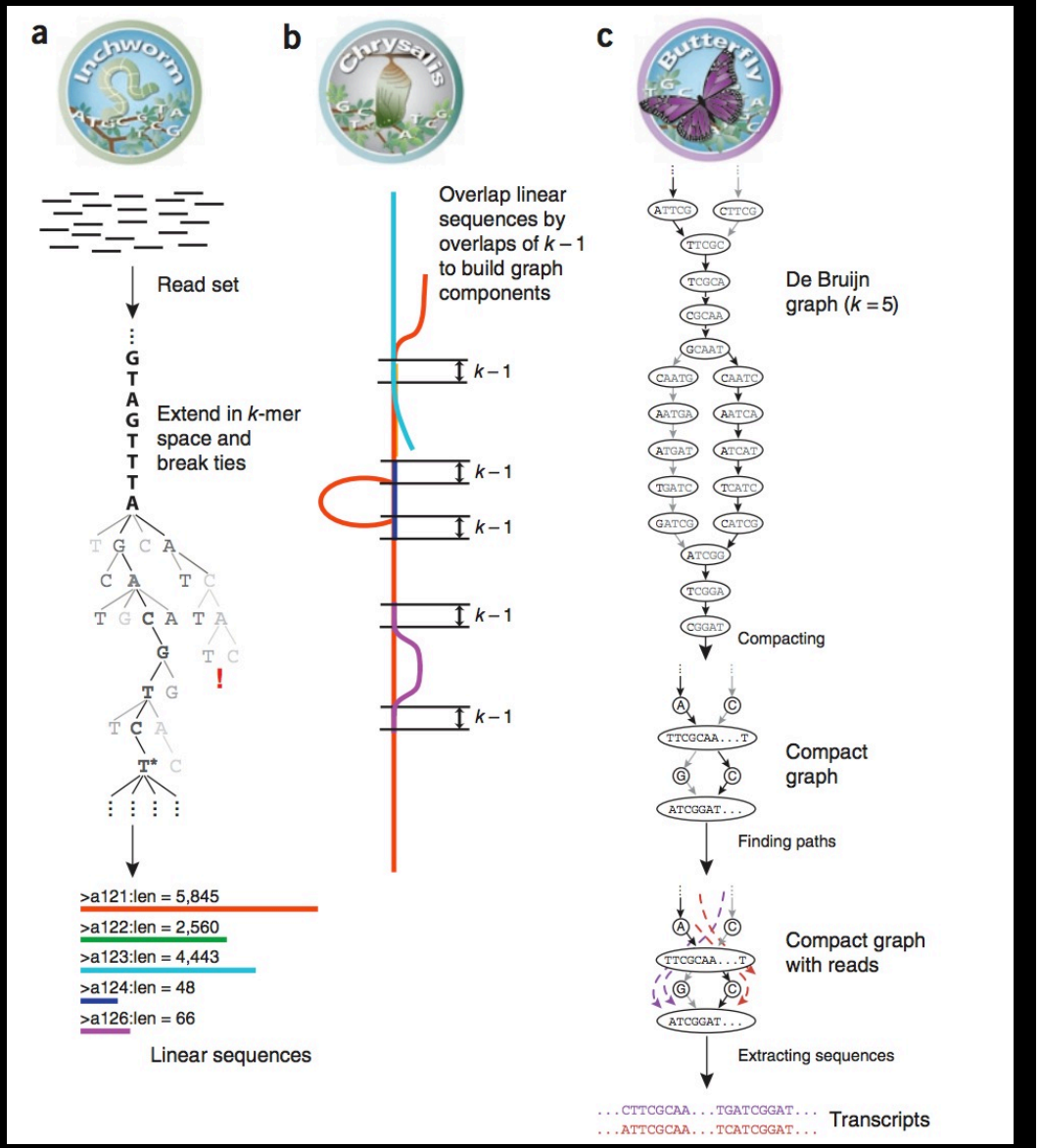
- Pathway enrichment analysis for two groups of human samples
- Deconvolution of gene expression profile in 10 samples, which were composed of 5 tissues.
  - We got the expression profile (>25000 genes) in each sample
  - We have the tissue-specific gene expression profile
  - **Question is: What's the proportion of each tissue in each sample?**

Software for transcriptome de novo assembly

- **Trinity**
  - Produce overall best assembly (except that Newbler works on long reads)
  - Time-consuming
- Oases
  - High accuracy
  - Merge assemblies from multiple k-mer
  - Huge memory requirement
- Soapdenovo
  - Acceptable memory requirement and speed
  - Unclear principles (and inaccurate) to extend the contigs
- Newbler
  - Produce the best assembly for long reads
  - Time-consuming and huge memory requirement
  - No official support
- …

How Trinity works

- Inchworm
  - Fast assembly based greedy k-mer extension
  - Only one representative isoform will be reported at full length per gene
  - The reported contigs (isoforms) provide information for further transcript construction
- Chrysalis
  - Cluster overlapped contigs into connected components
  - Construct the complete de Bruijn graph for each component
- Butterfly
  - Reconstruct the full-length isoforms
  - Reconcile the de Brujin graphs



Quality evaluation of transcriptome assembly

- Criteria
  - N50, assembly length Not good
  - Compactness (too many of bases or contigs are not good)
  - Support from reads

- Commonly used software
  - ALE
  - Genovo
  - RSEM-EVAL in **DETONATE**
    - deal with varied abundance on similar sequences
    - assess how well the assembly explains the RNA-Seq reads
    - consider the fragment size distribution

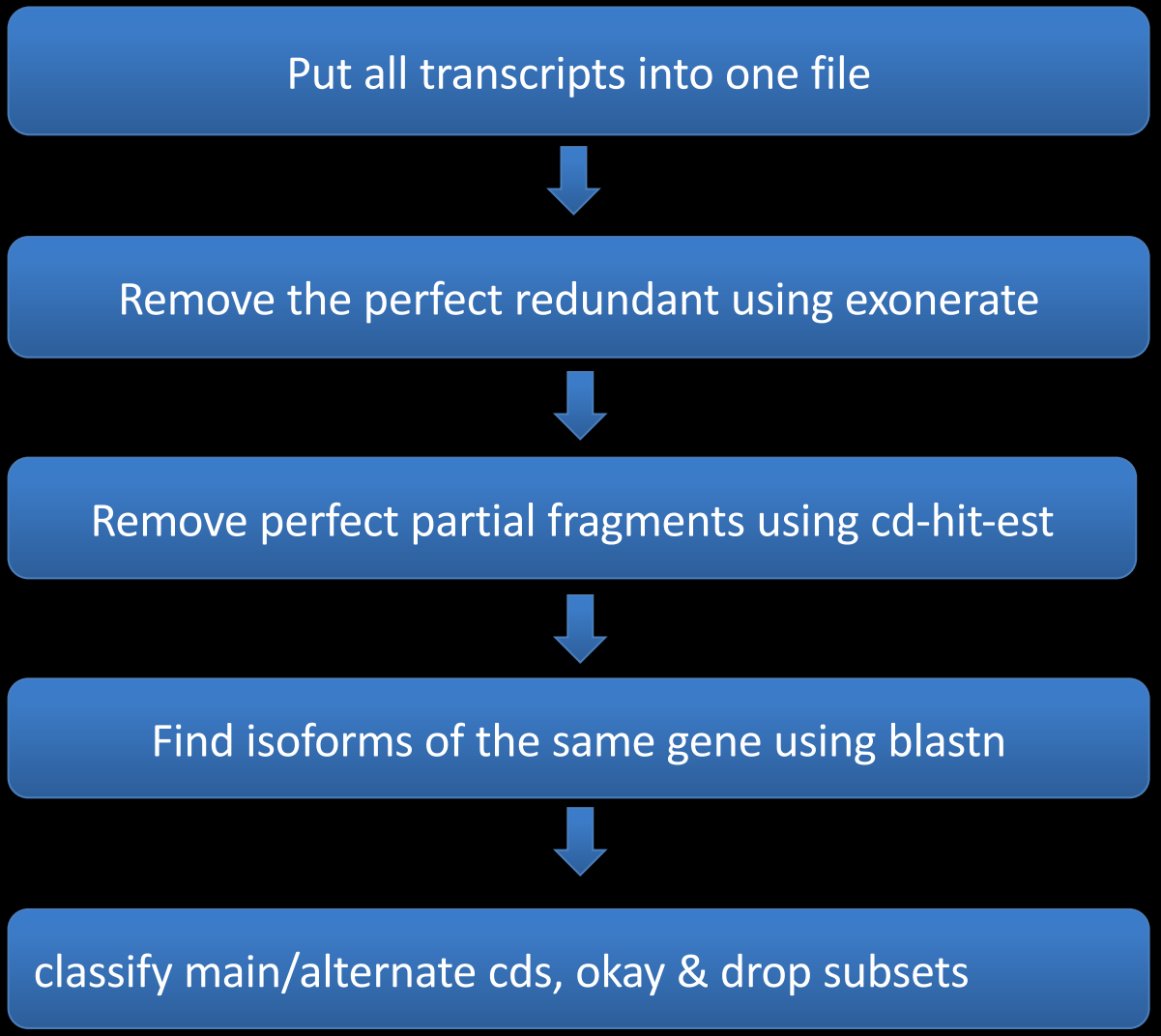
De novo assembly of Camelina sativa

- No reference genome sequence available
- Hexaploid genome
- Genetically close to Arabidopsis thaliana

Merge assembles

- Why merging
  - Complexity of the Transcriptome make different software/options produce different best transcripts set
  - Transcripts from multiple samples provide more information
- Two strategies
  - Transcripts reconstruction
    - Cap3 (merge assembly) or Oases (from multiple k-mer)
    - Cufflinks (cuffmerge)
    - Combine multiple samples
  - Transcripts cluster
    - **Cd-hit-est**
    - **EvidentialGene**

Merge assembly using EvidentialGene



Practice

- Assemble Camelina transctiptome using Trinity and Newbler
- Get the summary statistics from the assemblies
- Quality evaluation of the de novo transcriptome assemblies
- Merge multiple denovo assemblies for Camelina using EvidentialGene

Assembled transcripts from Trinity