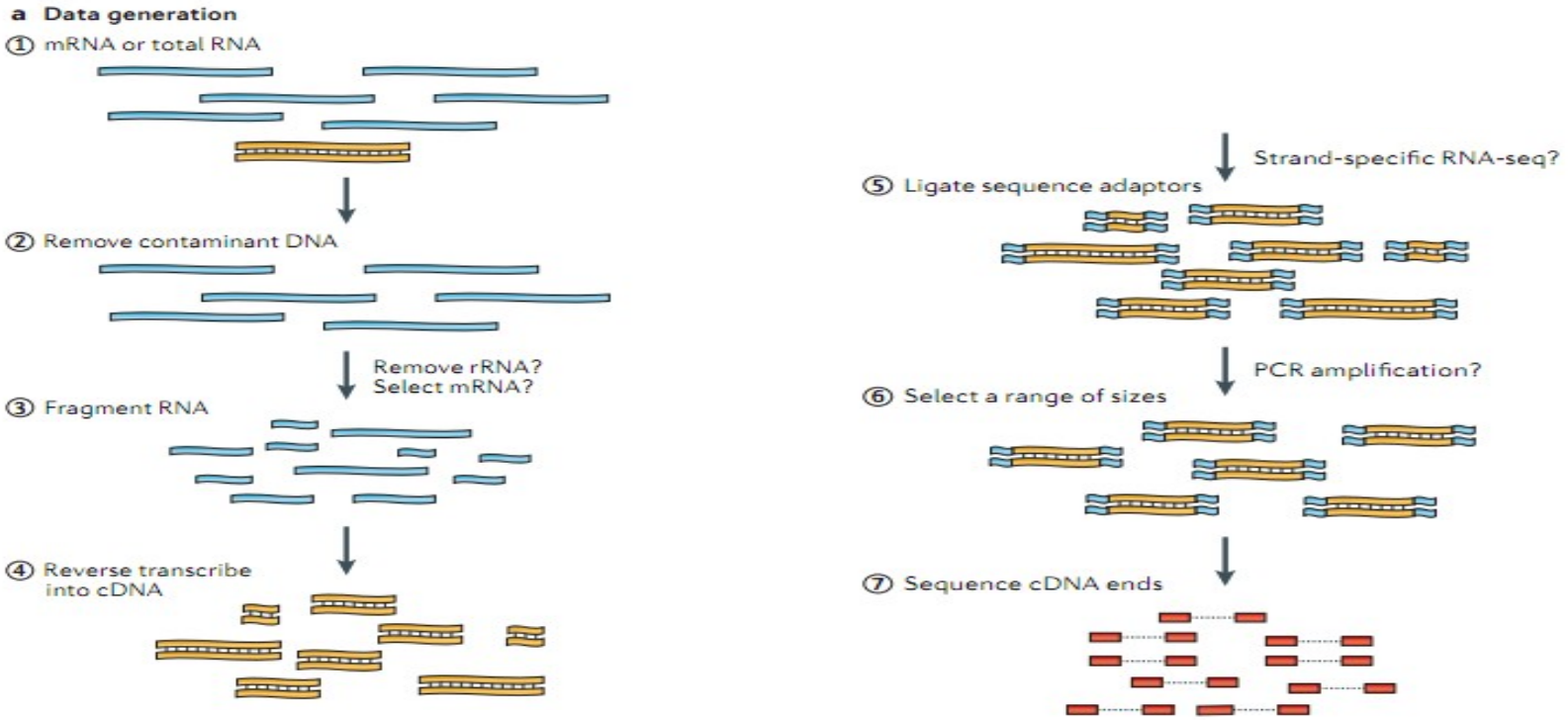
# Introduction to transcriptome analysis using High-Throughput Sequencing technologies (HTS)

# A typical RNA-Seq experiment

- Library construction



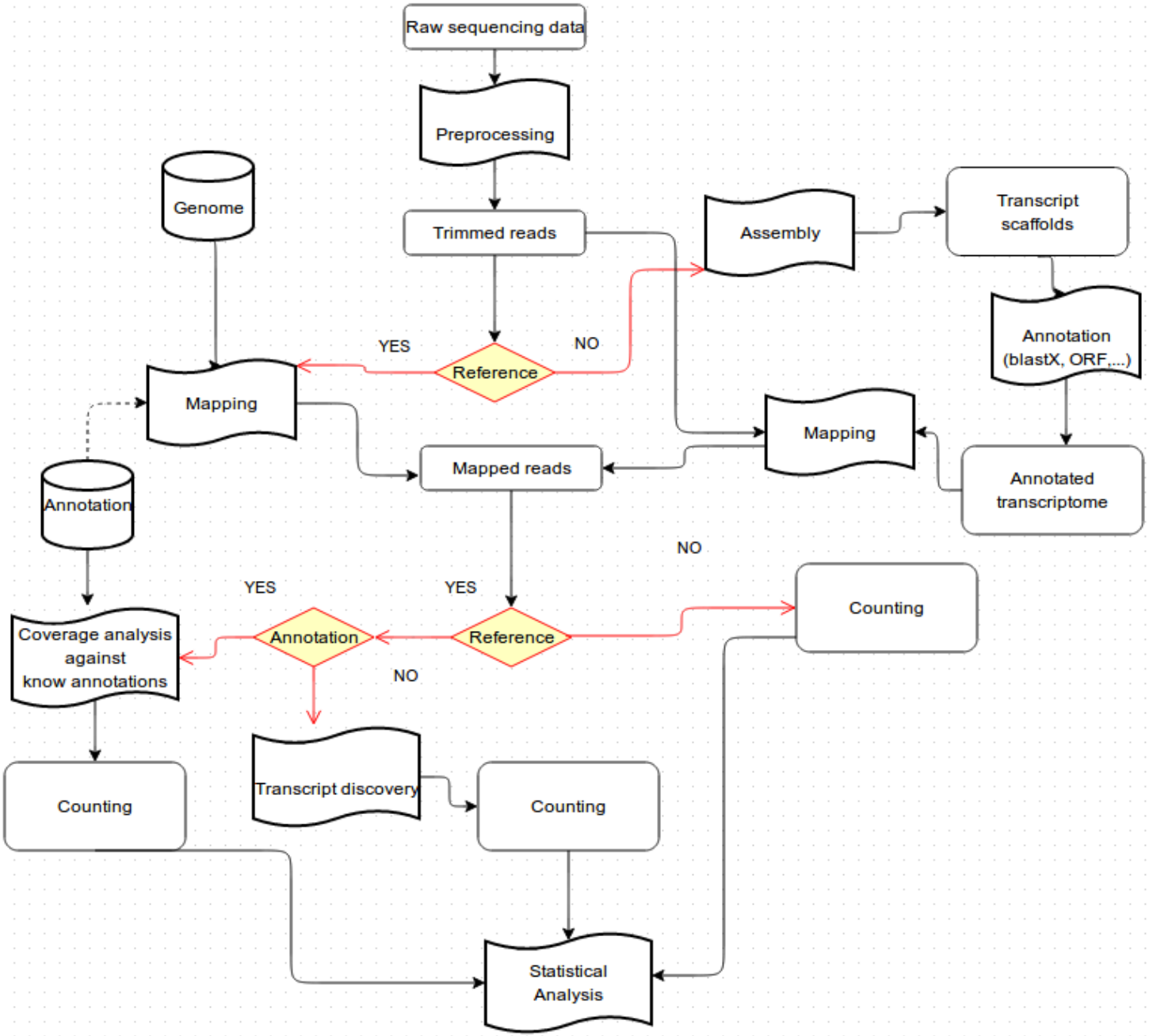

*Nature Reviews Genetics* **12**, 671-682 (October 2011) | doi:10.1038/nrg3068

ARTICLE SERIES: Study designs

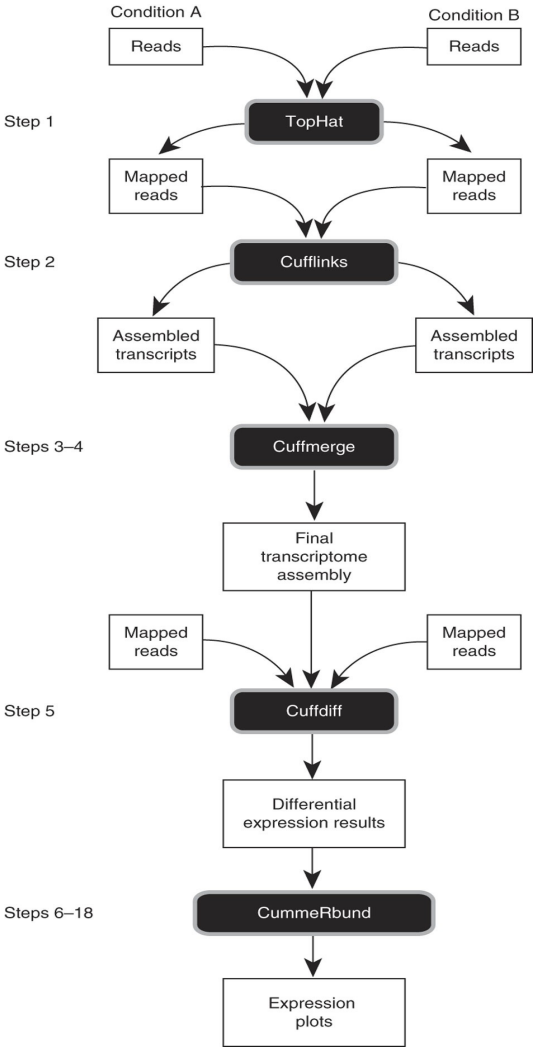
**Next-generation transcriptome assembly**

Jeffrey A. Martin[1] & Zhong Wang[1] About the authors

# Reference mapping and de novo assembly



# An overview of the Tuxedo protocol




Nat Protoc. 2012 Mar 1;7(3):562-78. doi: 10.1038/nprot.2012.016.

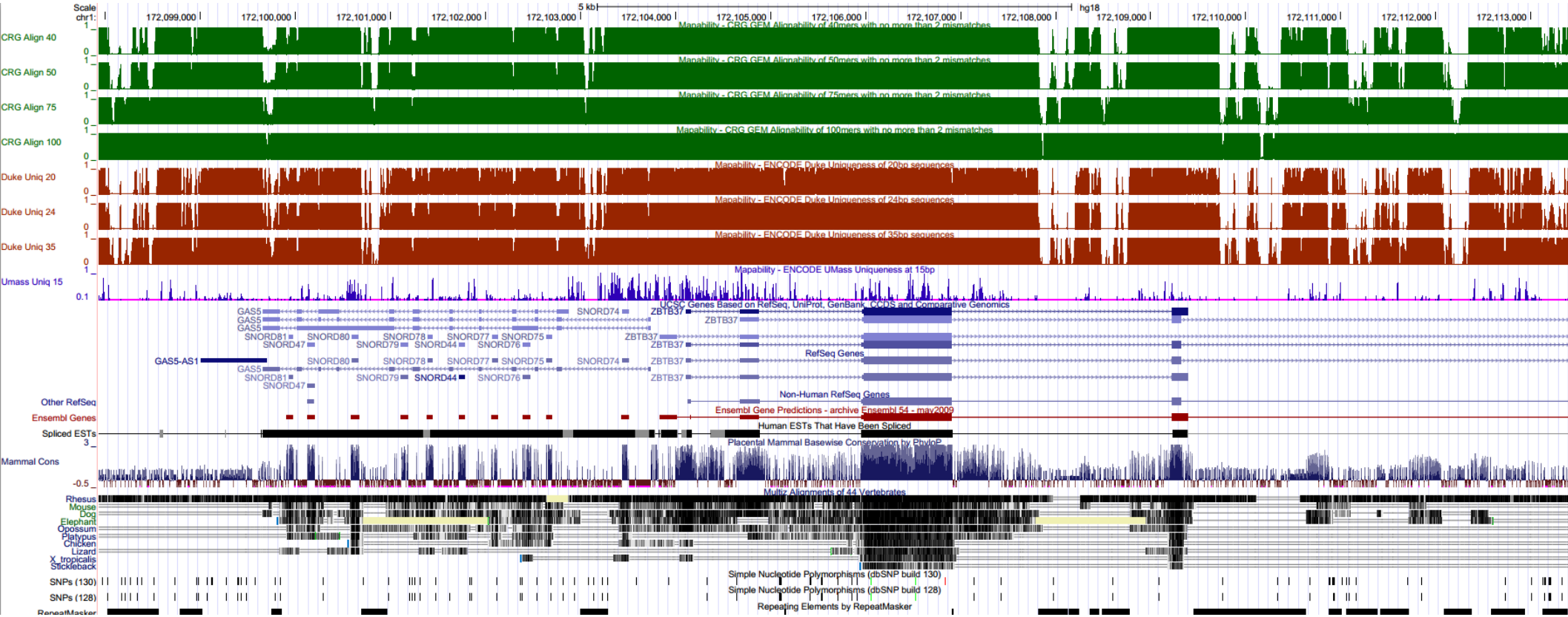
**Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.**

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L.

# Mapping reads

- Main Issues:
  - multi-hits (i.e multireads vs unireads)
    - Mapability issues
  - Number of allowed mismatches
  - PCR duplicates
  - Mates expected distance (mate/paired-sequencing)
- RNA-Seq specific
  - Considering exon junctions (RNA-Seq)

# Mappability




- These tracks display the level of sequence uniqueness of the reference NCBI36/hg18 genome assembly. They were generated using different window sizes, and high signal will be found in areas where the sequence is unique.

# Mapping reads to genome: general softwares

| Program | Algorithm | SOLiD | Long[a] | Gapped | PE[b] | Q[c] |
|---|---|---|---|---|---|---|
| Bfast | hashing ref. | Yes | No | Yes | Yes | No |
| Bowtie | FM-index | Yes | No | No | Yes | Yes |
| BWA | FM-index | Yes[d] | Yes[e] | Yes | Yes | No |
| MAQ | hashing reads | Yes | No | Yes[f] | Yes | Yes |
| Mosaik | hashing ref. | Yes | Yes | Yes | Yes | No |
| Novoalign[g] | hashing ref. | No | No | Yes | Yes | Yes |

[a]Work well for Sanger and 454 reads, allowing gaps and clipping.

[b]Paired end mapping.

[c]Make use of base quality in alignment.dBWA trims the primer base and the first color for a color read.

[e]Long-read alignment implemented in the BWA-SW module. fMAQ only does gapped alignment for Illumina paired-end reads.

[g]Free executable for non-profit projects only.

Brief Bioinform. 2010 Sep;11(5):473-83. Epub 2010 May 11.

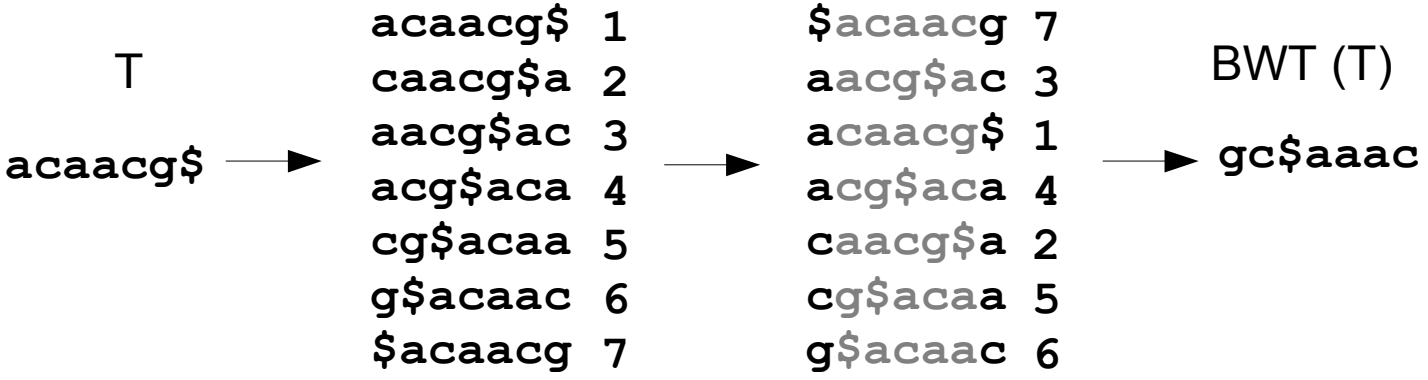
**A survey of sequence alignment algorithms for next-generation sequencing.**

Li H, Homer N.

Broad Institute, Cambridge, MA 02142, USA. hengli@broadinstitute.org

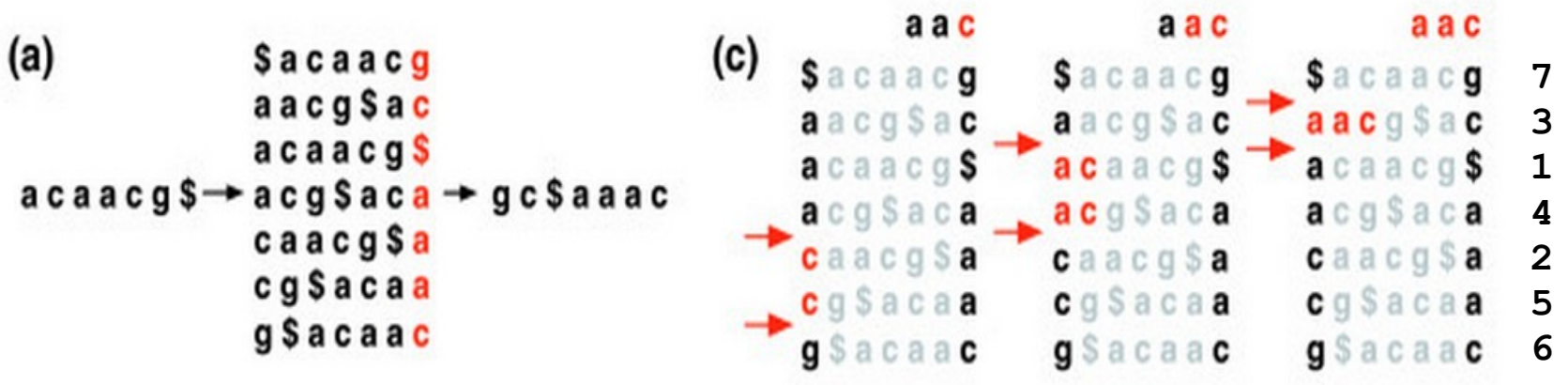
# Bowtie principle

- Use highly efficient compressing and mapping algorithms based on Burrows Wheeler Transform (BWT)
- The Burrows-Wheeler Transform of a text T, BWT(T), can be constructed as follows.
  - The character $ is appended to T, where $ is a character not in T that is lexicographically less than all characters in T.
  - The Burrows-Wheeler Matrix of T, BWM(T), is obtained by computing the matrix whose rows comprise all cyclic rotations of T sorted lexicographically.



# Bowtie principle

- Burrows-Wheeler Matrices have a property called the Last First (LF) Mapping.
  - The ith occurrence of character c in the last column corresponds to the same text character as the ith occurrence of c in the first column.
  - Example: searching ”AAC” in ACAACG




Genome Biol. 2009;10(3):R25. Epub 2009 Mar 4.

**Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.**

Langmead B, Trapnell C, Pop M, Salzberg SL.

# Mapping read spanning exons

- Limit of bowtie for RNA-Seq
  - mapping reads spanning exons
- Solution: splice-aware short-read aligners
  - RUM
  - MapSplice
  - Tophat (v1, v2)
  - GSTRUCT
  - STAR (Encode)
  - ...

Nat Methods. 2013 Nov 3. doi: 10.1038/nmeth.2722. [Epub ahead of print]

**Systematic evaluation of spliced alignment programs for RNA-seq data.**

Engström PG, Steijger T, Sipos B, Grant GR, Kahles A; The RGASP Consortium, Alioto T, Behr J, Bertone P, Bohnert R, Campagna D, Davis CA, Dobin A, Engström PG, Gingeras TR, Goldman N, Grant GR, Guigó R, Harrow J, Hubbard TJ, Jean G, Kahles A, Kosarev P, Li S, Liu J, Mason CE, Molodtsov V, Ning Z, Ponstingl H, Prins JF, Rätsch G, Ribeca P, Seledtsov I, Sipos B, Solovyev V, Steijger T, Valle G, Vitulo N, Wang K, Wu TD, Zeller G, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, Bertone P.

# TopHat pipeline

- RNA-Seq reads are mapped against the whole reference genome (bowtie).
- TopHat to allows up to ”*max-multihits*” alignments to the reference for a given read (choose the alignments based on their alignment scores if there are more than this number)
- Reads that do not map are set aside (initially unmapped reads, or IUM reads)
- TopHat then assembles the mapped reads using the assembly module in Maq. An initial consensus of mapped regions is computed.
- The ends of exons in the pseudoconsensus will initially be covered by few reads (most reads covering the ends of exons will also span splice junctions)
  - Tophat a small amount of flanking sequence of each island (default=45 bp).

Bioinformatics. 2009 May 1;25(9):1105-11. Epub 2009 Mar 16.

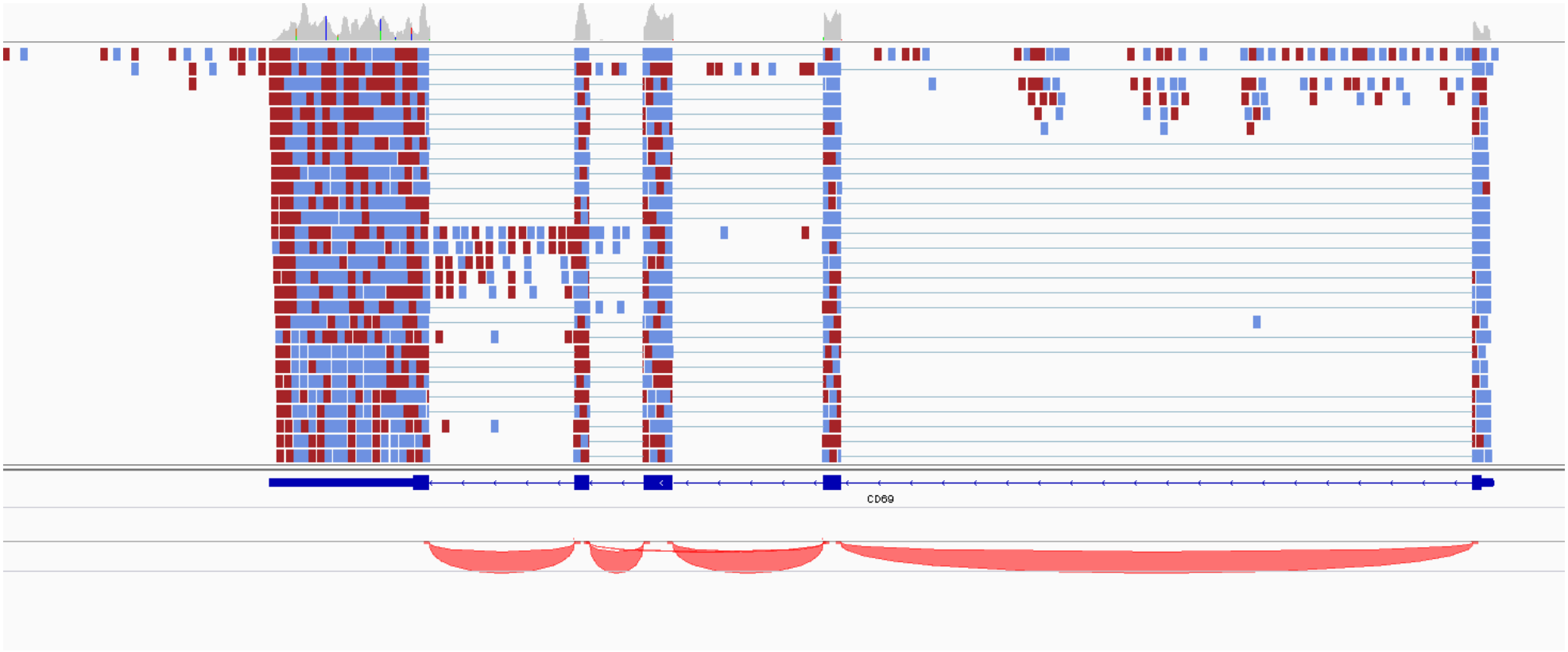
**TopHat: discovering splice junctions with RNA-Seq.**

Trapnell C, Pachter L, Salzberg SL.

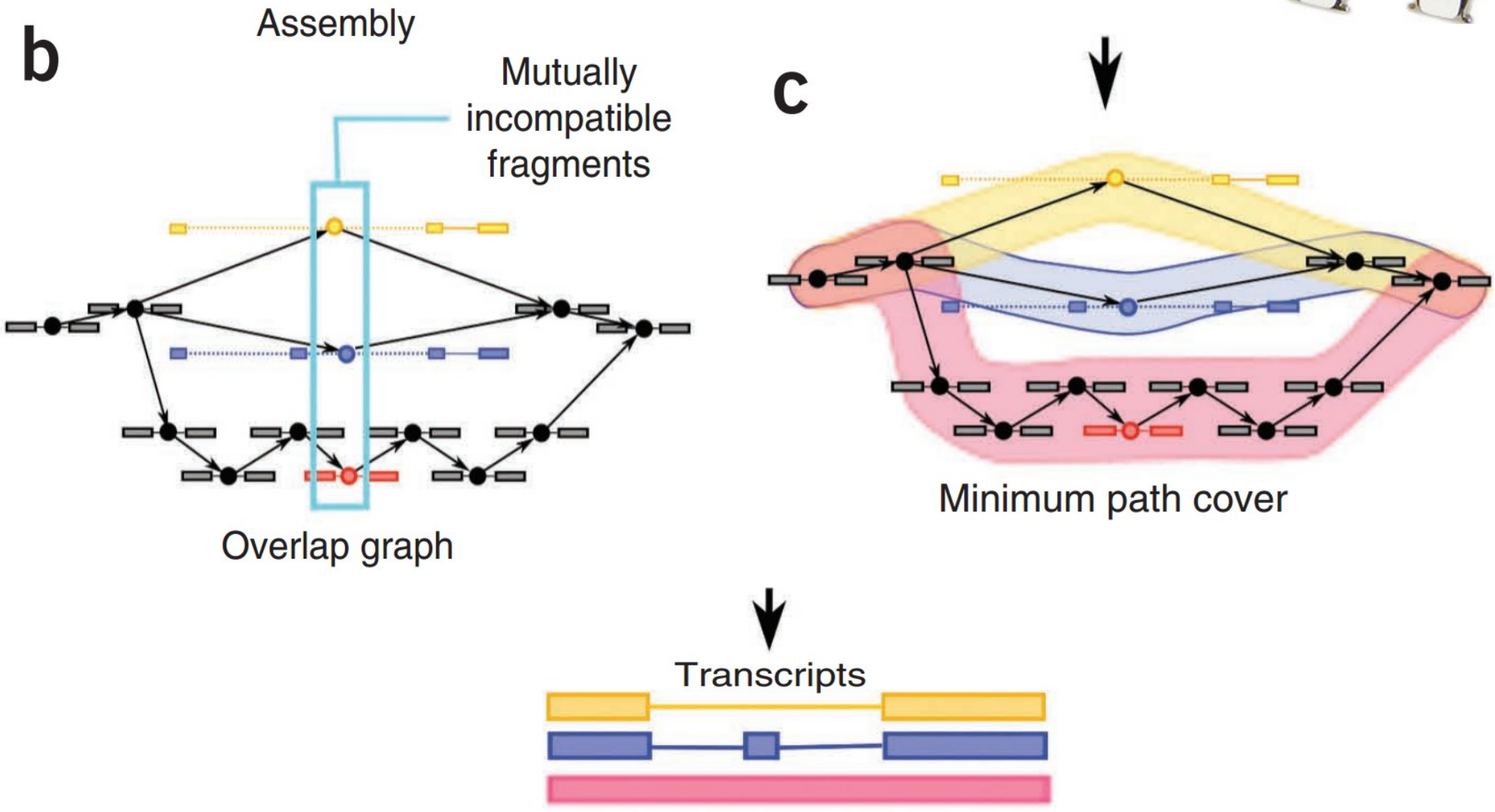
# TopHat pipeline

- Weakly expressed genes should be poorly covered
  - Exons may have gaps
- A parameter controls when two distinct but nearby exons should be merged into a single exon.
  - Introns shorter than 70 bp are rare in mammalian genome
  - To be conservative, the TopHat default is 6 bp
- To map reads to splice junctions, TopHat first enumerates all canonical donor and acceptor sites within the island sequences (as well as their reverse complements)
- Next, it considers all pairings of these sites that could form canonical (GT–AG) introns between neighboring (but not necessarily adjacent) islands.
  - By default, TopHat examines potential introns longer than 70 bp and shorter than 20 000 bp (more than 93% of mouse introns in the UCSC known gene set fall within this range)
- Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form potential splice junctions.

# TopHat example results



# Cufflinks: transcript assembly




Nat Biotechnol. 2010 May;28(5):511-5. doi: 10.1038/nbt.1621. Epub 2010 May 2.

**Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.**

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L.

# Cufflinks: transcript assembly

- -g : Tells Cufflinks to use the supplied reference annotation (GFF) to guide the assembly
- -G : Tells Cufflinks to use the supplied reference annotation (a GFF file) to estimate isoform expression.
- ! -g AND ! -G : Cufflinks will perform assembly without any reference annotation (warning with short reads).

# Cuffcompare / Cuffmerge

- Cuffcompare
  - Compare discovered transcript to reference transcript
  - Output classcode
    - e.g
      - *i* A transfrag falling entirely within a reference intron
      - *u* Unknown, intergenic transcript
      - *j* potentially novel isoform
- Cuffmerge
  - Used to merge several transcript models obtained from several samples.
  - Delete some transcripts model based on classcodes

# Quantification

- Cufflinks
  - FPMK
- Cuffdiff
  - Estimation of counts
  - FPKM
  - Differential expression
- HTSeq-count
  - A python package
  - Htseq-count (python script)

# Normalization ?

- **Several methods proposed**

- **Reads Per Kilobase per Million mapped reads** (RPKM): This approach was initially introduced to facilitate comparisons between genes within a sample.
  - Not sufficient (need to be combined with inter-sample normalization method)
- Quantiles (Q): First proposed in the context of microarray data, this normalization method consists in **matching distributions** of gene counts across lanes.
  - Use with caution when comparing distantly related tissues.
- Upper Quartile (UQ): the total counts are replaced by the **upper quartile** of counts different from 0 in the computation of the normalization factors.
  - Very similar in principle to TC (but really more powerful)
- Trimmed Mean of M-values (TMM): This normalization method is implemented in the **edgeR Bioconductor** package (version 2.4.0). Scaling is based on a subset of M values
  - TMM seems to provide a robust scaling factor.
- RLE: This normalization method is included in the **DESeq Bioconductor package** (version 1.6.0). Close to TMM.

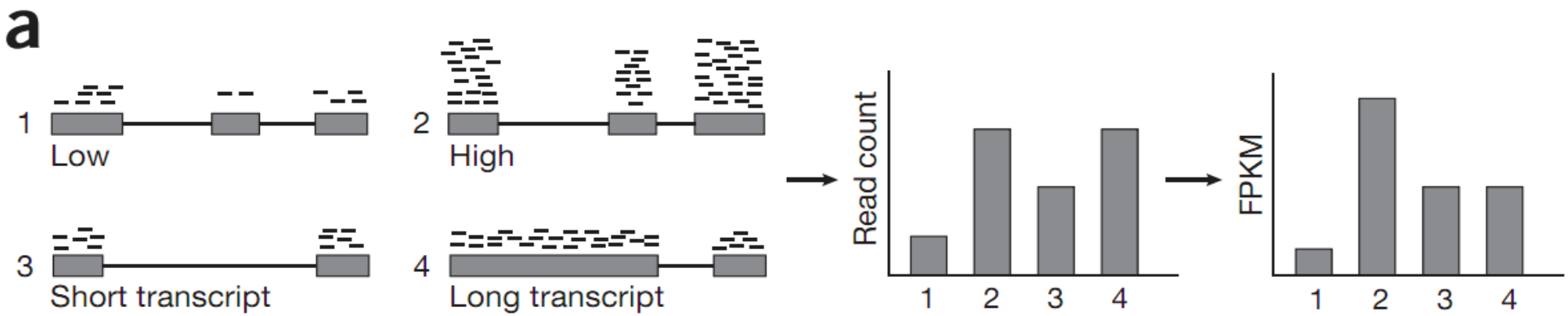
Brief Bioinform. 2012 Sep 17. [Epub ahead of print]

**A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.**

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F; on behalf of The French StatOmique Consortium.

# RPKM / FPKM

- Transcrits of different length have different read count




- Tag count is normalized for **transcrit length and total read number** in the measurement (RPKM, Reads Per Kilobase of exon model per Million mapped reads)
- 1 RPKM corresponds to approximately one transcript per cell
- FPKM, Fragments Per Kilobase of exon model per Million mapped reads (paired-end sequencing)

**Computational methods for transcriptome annotation and quantification using RNA-seq**

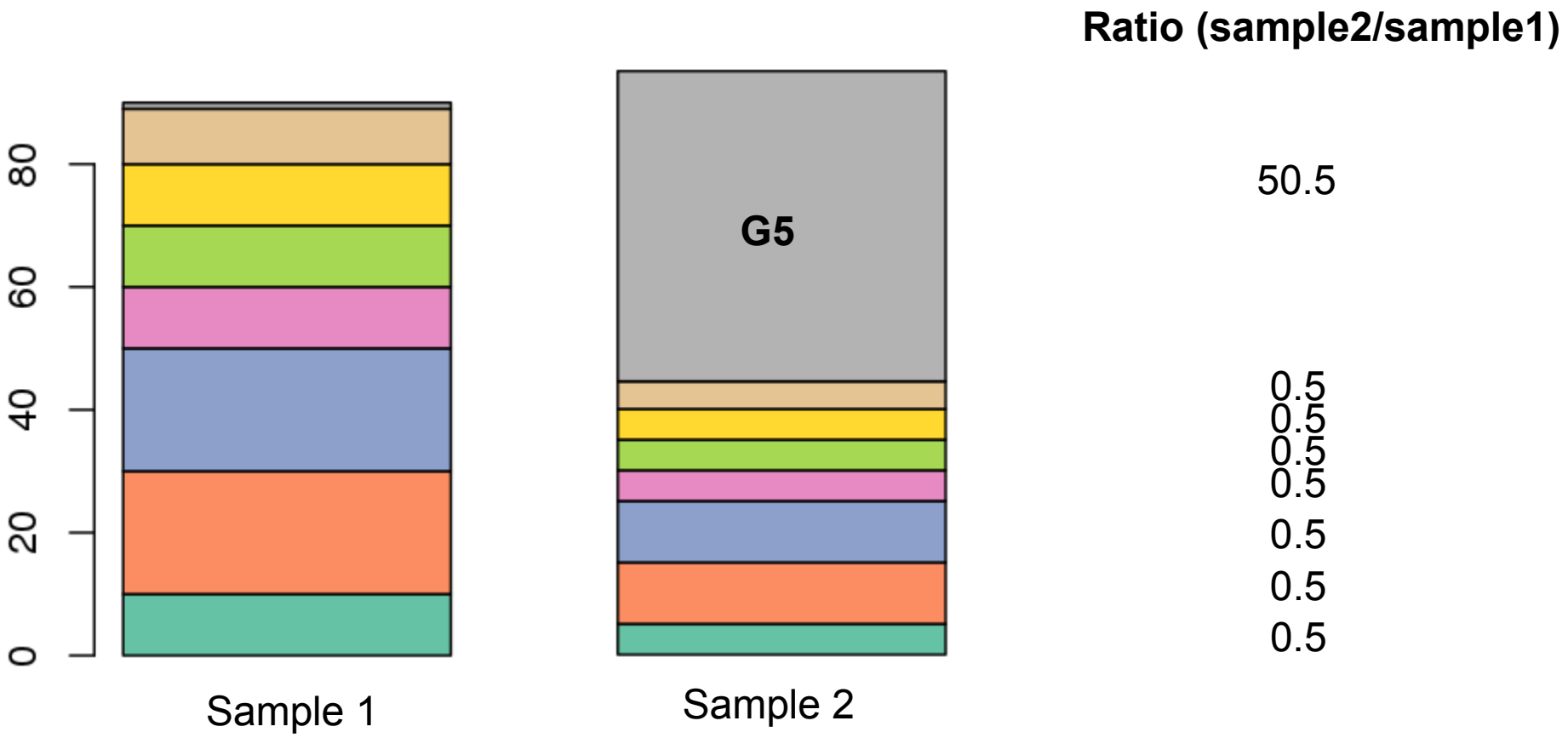
Manuel Garber[1], Manfred G Grabherr[1], Mitchell Guttman[1,2] & Cole Trapnell[1,3]

**Accurate quantification of transcriptome from RNA-Seq data by effective length normalization**

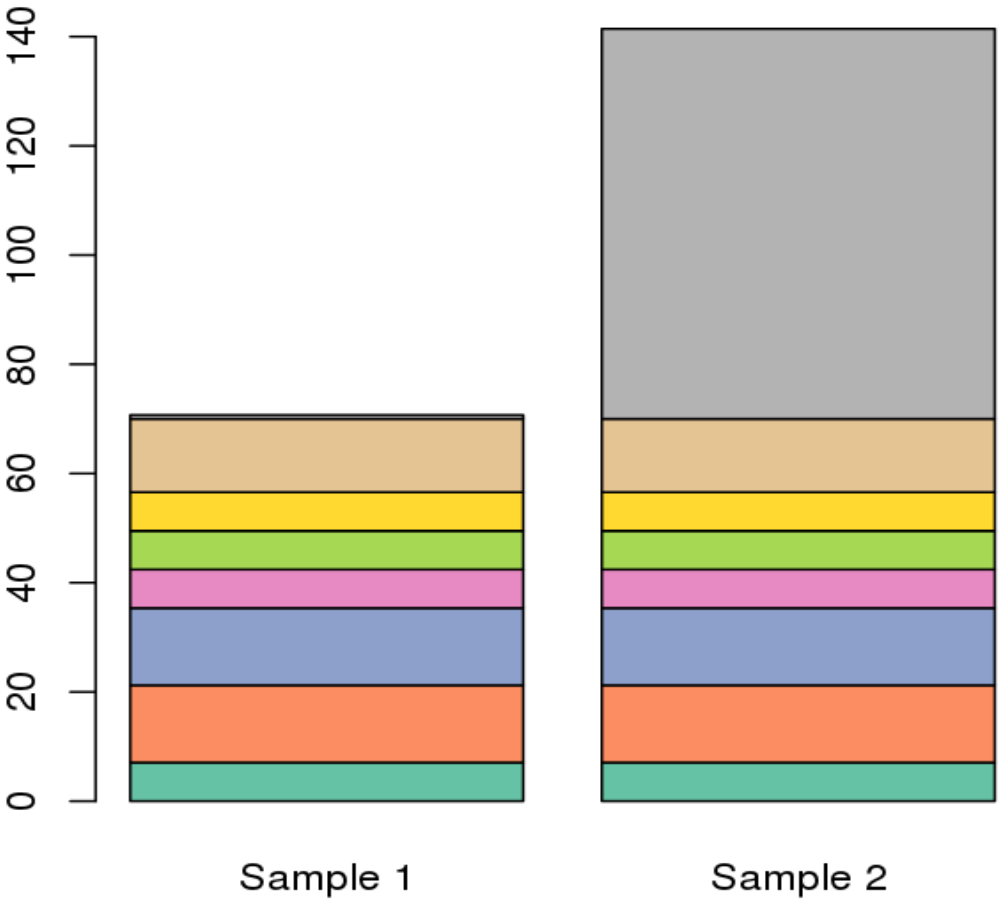
Soohyun Lee[1], Chae Hwa Seo[1], Byungho Lim[2], Jin Ok Yang[1], Jeongsu Oh[1], Minjin Kim[2], Sooncheol Lee[2], Byungwook Lee[1], Changwon Kang[2] and Sanghyuk Lee[1,3,*]

# Main issues in RNA-Seq normalization

- Highly abundant genes:
  - E.g; All genes unchanged but G5
    - Total count → repression of all other genes by a factor 2 !



# TMM Normalization (Robinson and Oshlack, 2010)




- Outline
  - Compute the M values (log ratio).
    - Take the trimmed mean of the M value as scaling factor.
    - Multiply read counts by scaling factor (they multiply to one)
    - If more than two columns
      - The library whose 3rd quartile is closest to the mean of 3rd quartile is used.
    - **Very similar to RLE**

Genome Biol. 2010;11(3):R25. Epub 2010 Mar 2.

**A scaling normalization method for differential expression analysis of RNA-seq data.**

Robinson MD, Oshlack A.

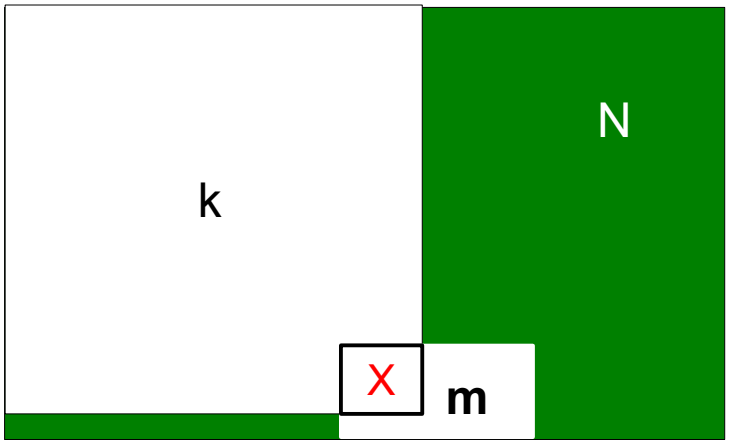
# Differential Expression

- Several methods proposed
  - Fisher, EdgeR, DESeq, NOISeq, Cuffdiff...

# Fisher's exact test (or hypergeometric test)

- Simple two-library comparison
- If read counts for a gene g are balanced we should expect ~ same number of read in both conditions.

| | Cont | Treated | |
|---|---|---|---|
| Reads from gene G | x | m-x | m (white) |
| Remaining reads | k-x | n-(k-x) | n (black) |
| | k | N-k | N |

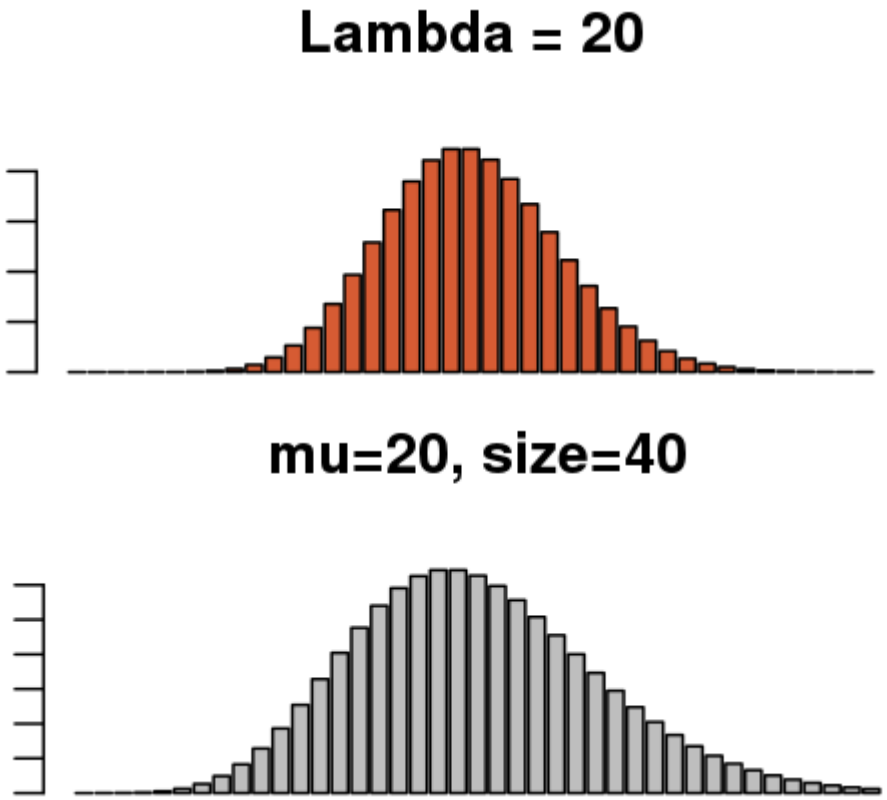



- x follows a hypergeometric distribution with parameter N, K, n

# Two-class Differential expression analysis

- Problematic
  - What is the underlying distribution of read counts
    - If reads for gene *g* were obtained from a population of samples with equal expression level one could model read counts of *g* as a **poisson distribution**
    - However, depending on samples, expression level may vary in each class according to :
      - Genes type (*e.g, stress-responsive genes*)
      - Biological samples (*e.g*, purity)
      - → Overdispersion
    - Poisson distribution predict smaller dispersion than observed in the data
      - incorrectly optimistic p values

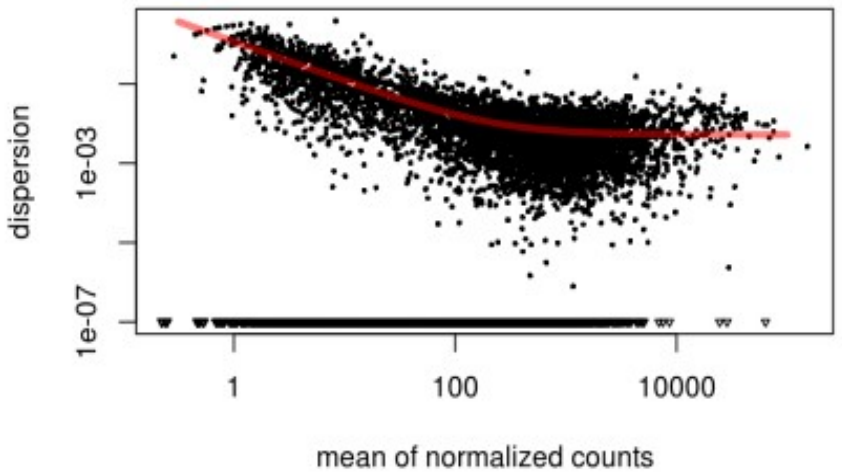
# Negative binomial




- Poisson
  - One parameter, $\lambda$
  - Variance is equal to $\lambda$
- Negative binomial
  - Has two parameters mean ($\mu$) and variance ($\sigma^2$).
  - Can be used as an alternative model to the Poisson distribution when sample variance exceeds the sample mean.

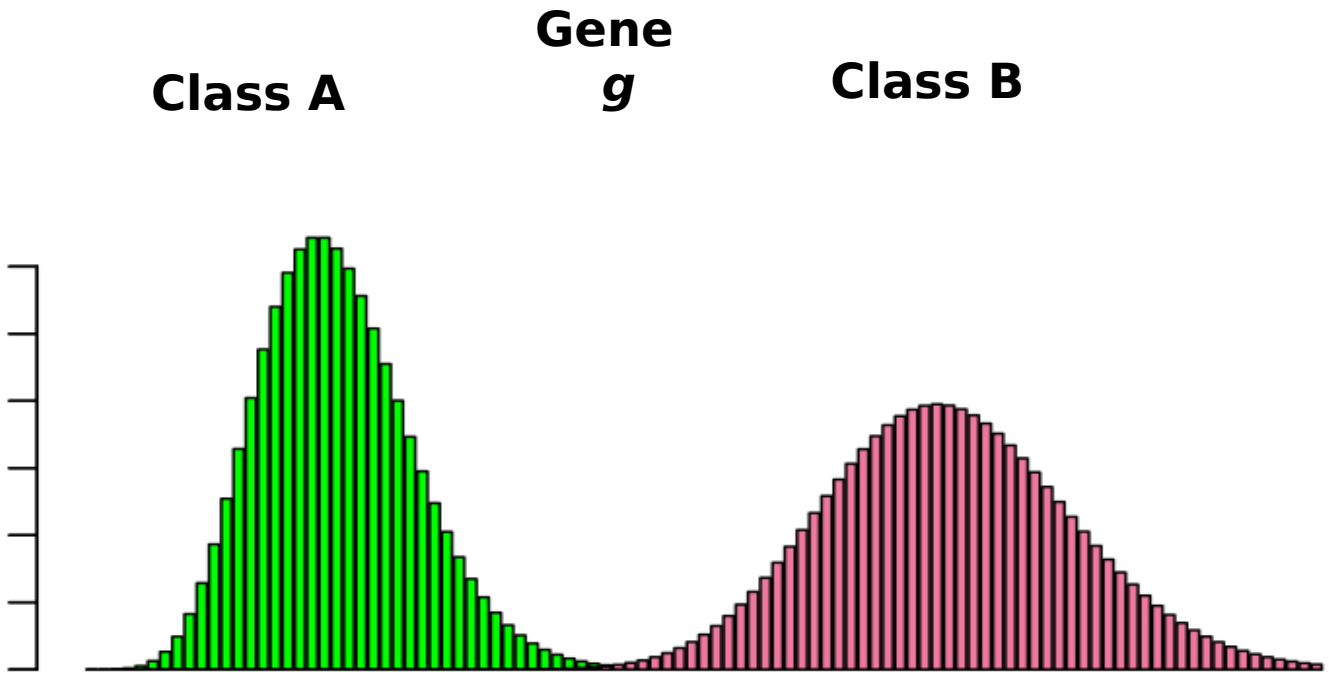
# Estimating dispersion (DESeq)

- Variance observed between counts is the sum of two components
  - Sample-to-sample variation, biological variation (**dispersion**, dominate in highly expressed genes)
  - Uncertainty in measure (**shot noise**, dominate in weakly expressed genes)
- Variance is estimated in each class by using a shrinkage method



# Test for differential expression (DESeq)

- Intuition

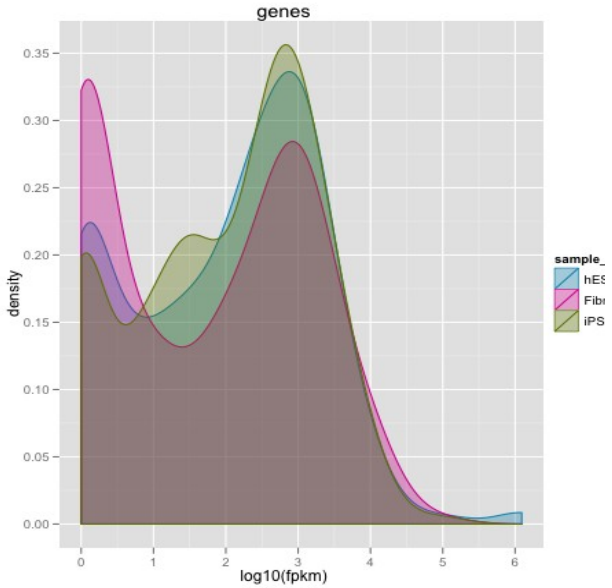


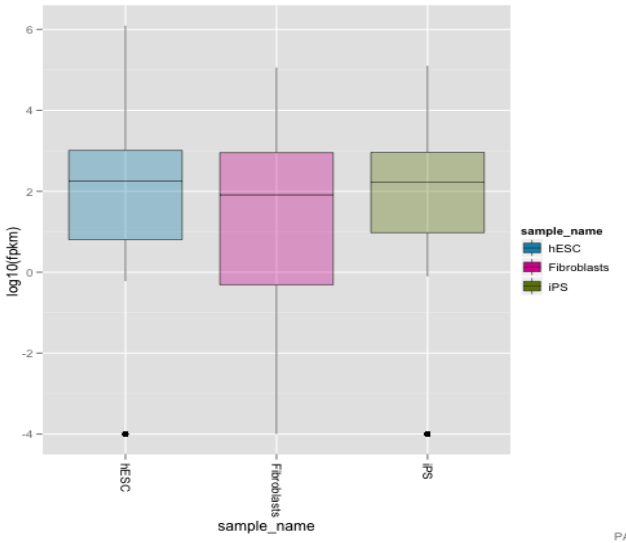

- The test implemented in DESeq is based on the sums of counts in class A and B (that are NB distributed variables)

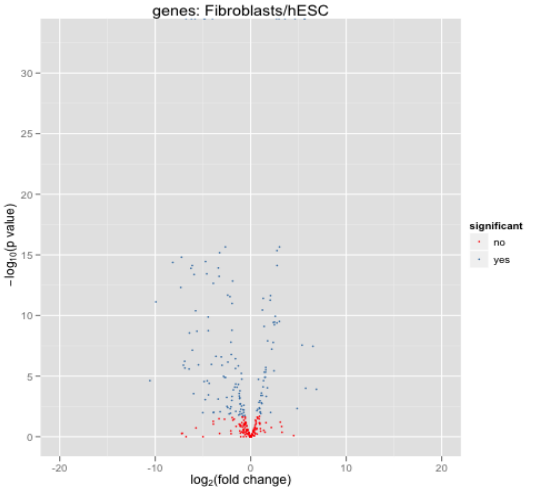
# Cuffdiff

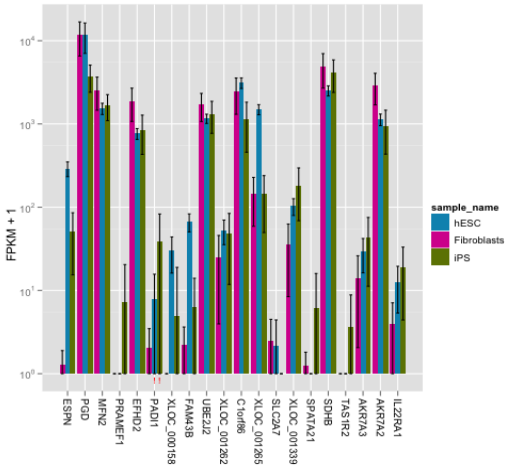
- Differential expression
  - Gene
  - Alternative transcripts
  - Alternative 5' UTRs
  - ...

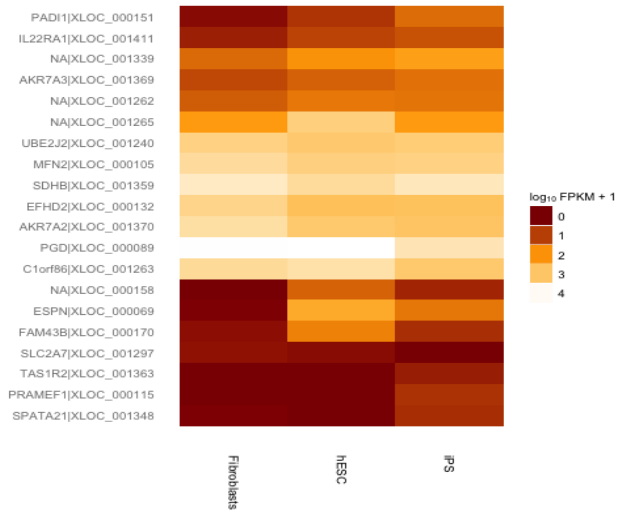
# CummeRbund




- cummeRbund is a visualization package for Cufflinks high-throughput sequencing data.