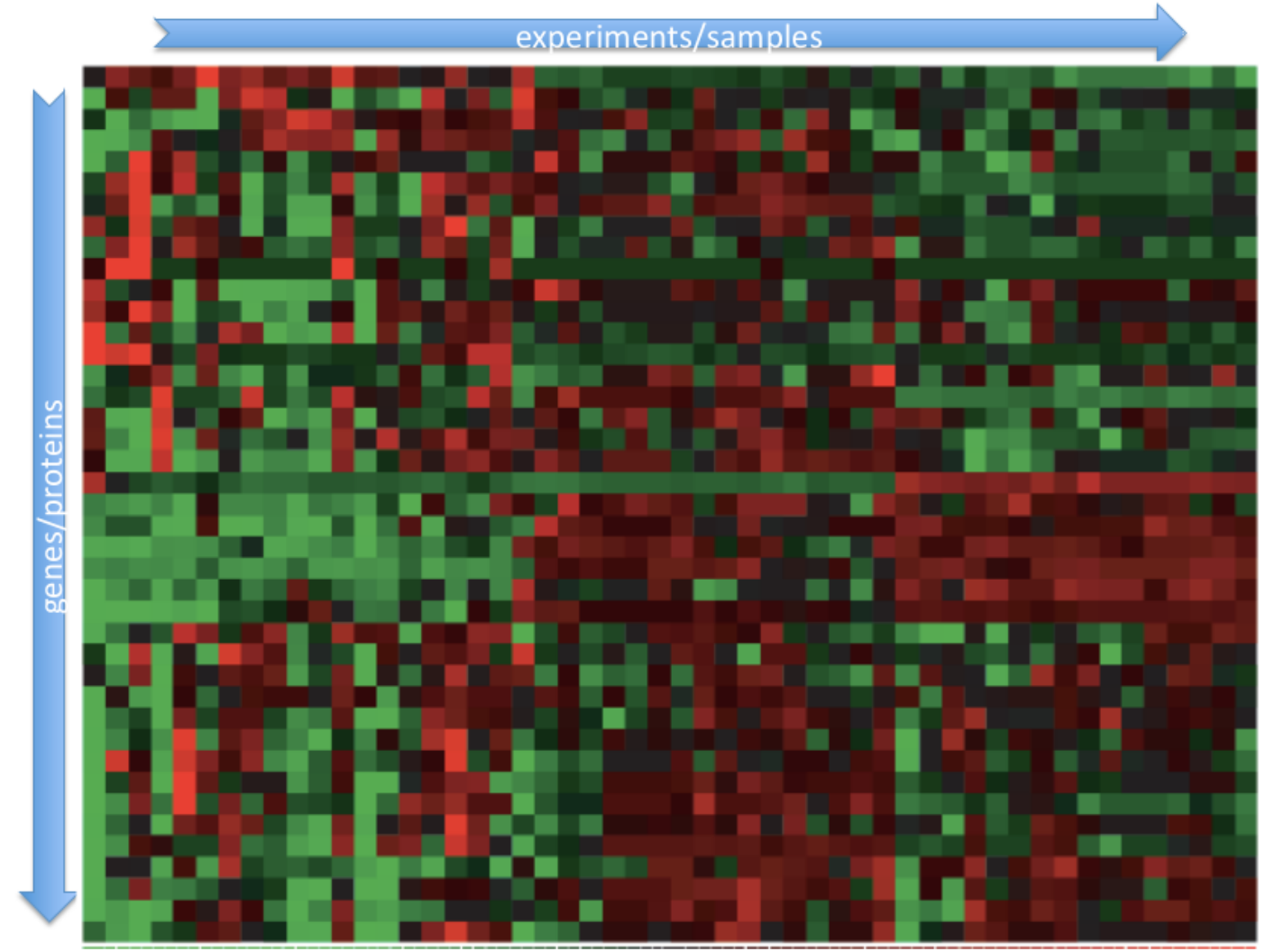
Clustering

- In differential gene expression, you are looking for genes that behave differently between one sample and another, either up- or down-
- Once you get your DE gene set, you group the genes according to similar expression, and the outliers become more obvious

Clustering

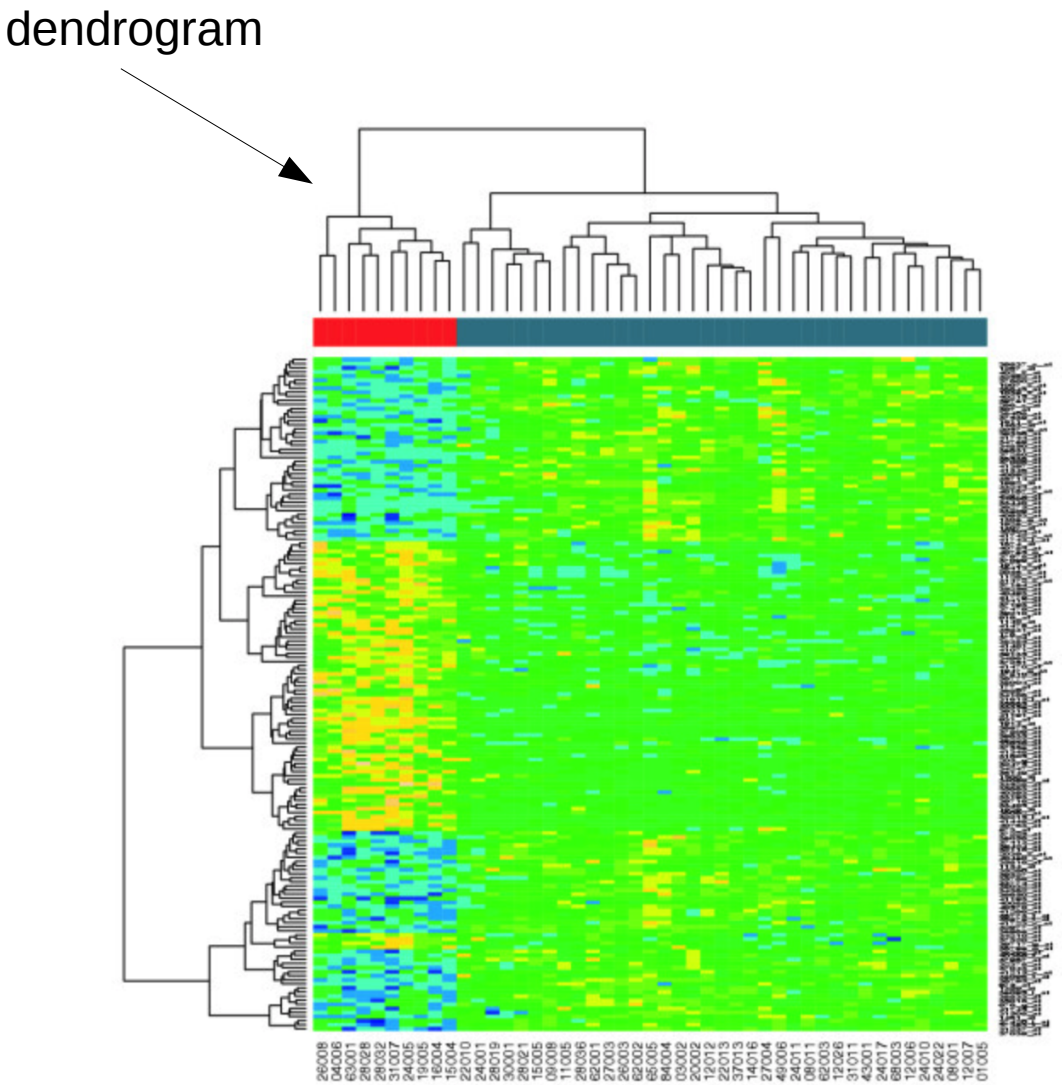


High expression

Low expression

A heatmap is a graphical representation of a table

Clustering



- Samples clusters indicate samples with similar phenotype
- Genes clusters indicated co-regulated genes

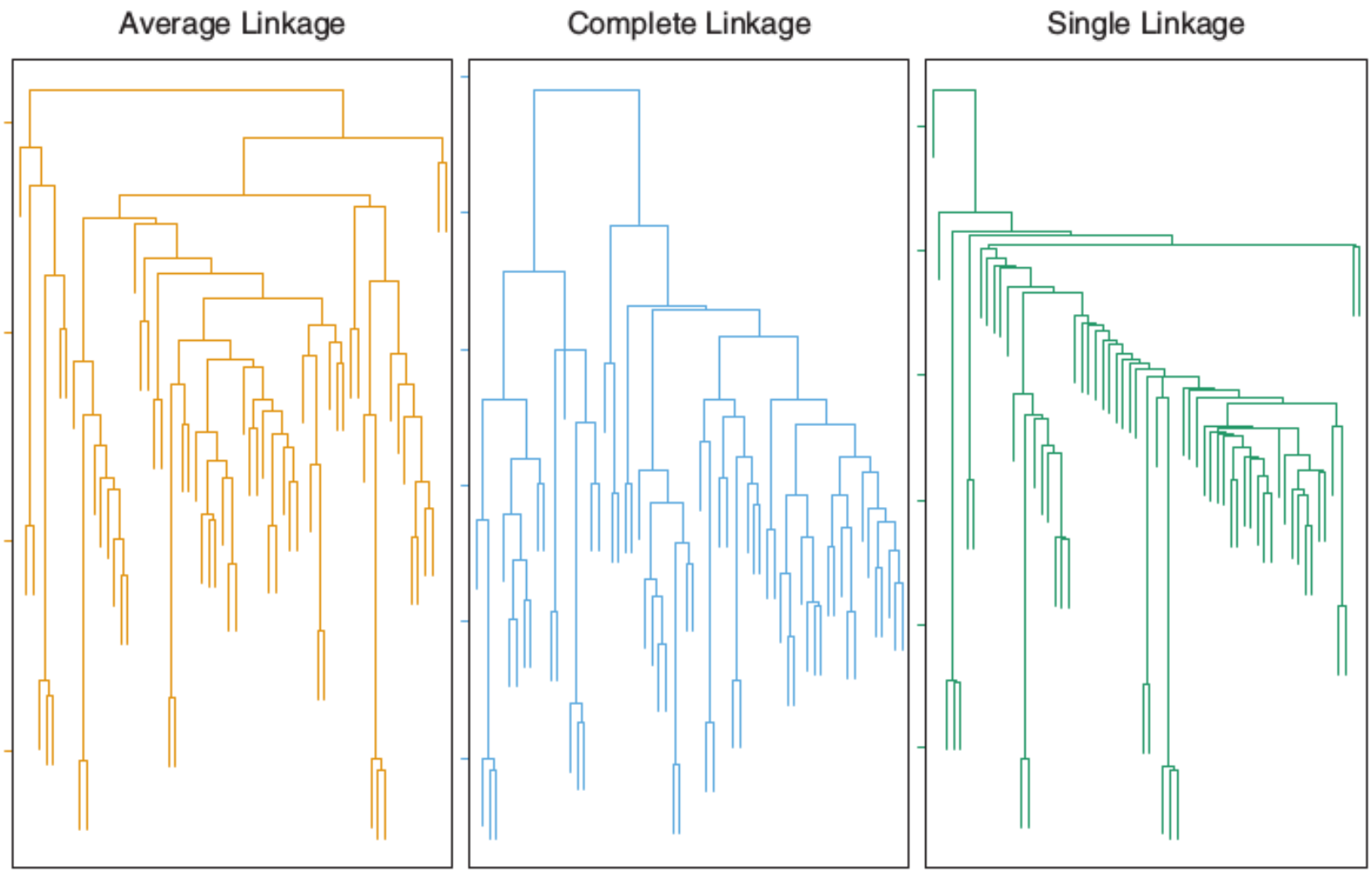
Hierarchical clustering

- The algorithm that creates the dendrogram is based on the definition of a *distance (dissimilarity)* measure between observation pairs
- The algorithm works as follow:
  - Start with N observation and a distance measure (e.g. Euclidean distance) for all N*(N-1)/2 pairs. Each observation is a cluster
  - For i=n, n-1, n-2, …, 2
    - Examine all inter-clusters distances and fuse the clusters with lower distance
      - The distance between the fused clusters represents the height of the bar in the dendrogram
    - Calculate new inter-cluster distances between the remaining I-1 clusters
- With the linkage we define a notion of distance between clusters (containing many observations)

Linkage methods

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

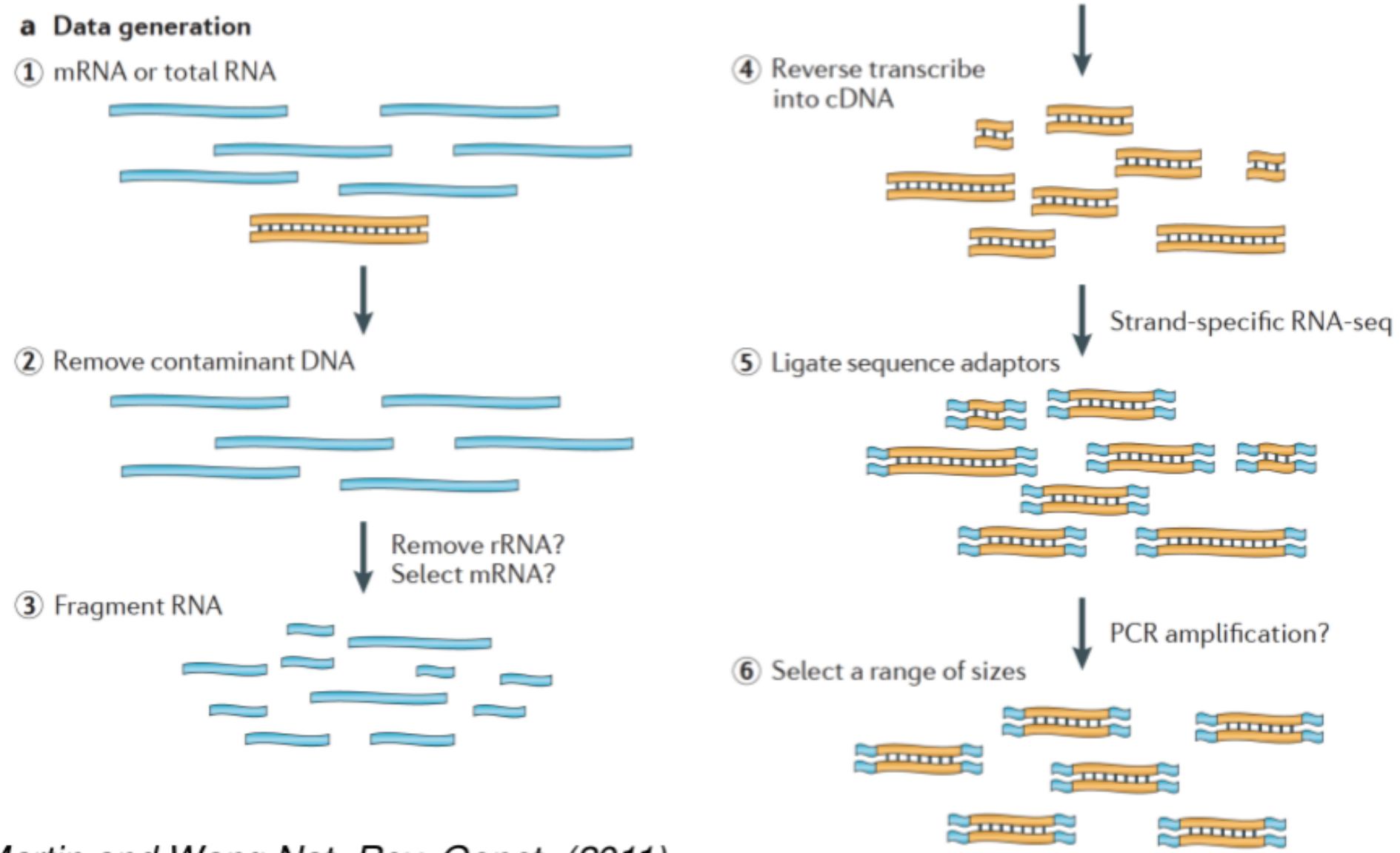
Linkage methods



RNA-seq

- Next-generation sequencing approach
- What is being sequenced is the cDNA from the mRNA component
- Sequencing of whole transcriptome of a sample (NGS) comparing it against the whole transcriptome of another sample

RNA-seq protocol



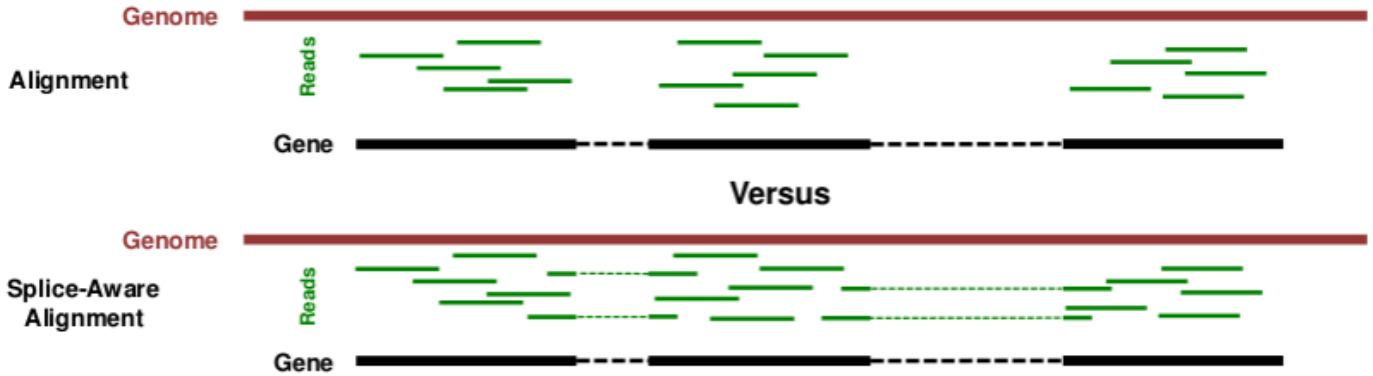
Martin and Wang Nat. Rev. Genet. (2011)

RNA-seq

- Biological replicates:
  - biological variation (at least 3 if possible, or more)
  - batch effects (or lane effects)
- Sequencing protocol
  - Library
  - Sequencing
    - Paired end (PE): duplicates, splicing analysis and discovery of novel isoforms
    - Single end (SE): enough for gene expression analysis
  - Sequencing depth:
    - 20-50M per sample for differential expression
  - Read length:
    - 100bp/150bp

RNA-seq data alignment

- Align your reads to a reference genome or transcriptome or to a genomic region
- Better to use splice-aware aligner, such as TopHat or STAR
  - refine the aligments according to coding sequences (exons) using known and/or predicted splice junctions



- Or assemble the reads *de-novo* (e.g. Trinity)

FASTQ files

- Output of the sequencing machine

Quantifying reads per gene

- Your aim is to count sequence reads per gene
- When mapping reads to genome:
  - Filter out rRNA, tRNA, mitRNA, etc
  - Filtering out (or in!) non-coding RNA
  - Deal with alternative splicing
  - Deal with overlapping genes, pseudogenes
  - Small reads mean many short overlaps at one end or the other of intron gaps
  - Allele specific gene expression

Counts

- Standard counts
  - Number of reads for transcript *i*

$$X_i$$

- Counts per million (CPM)
  - counts scaled by the number of fragments you sequenced (N) times one million.

$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

Number of reads sequenced

- used by differential expression methods

Intra sample normalization

- Do not use for direct comparison across samples
- Transcripts per million (TPM) is a measurement of the proportion of transcripts in your pool of RNA

$$\mathrm{TPM}_i = \frac{X_i}{\widetilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j}} \right) \cdot 10^6$$

Length of transcript *i*

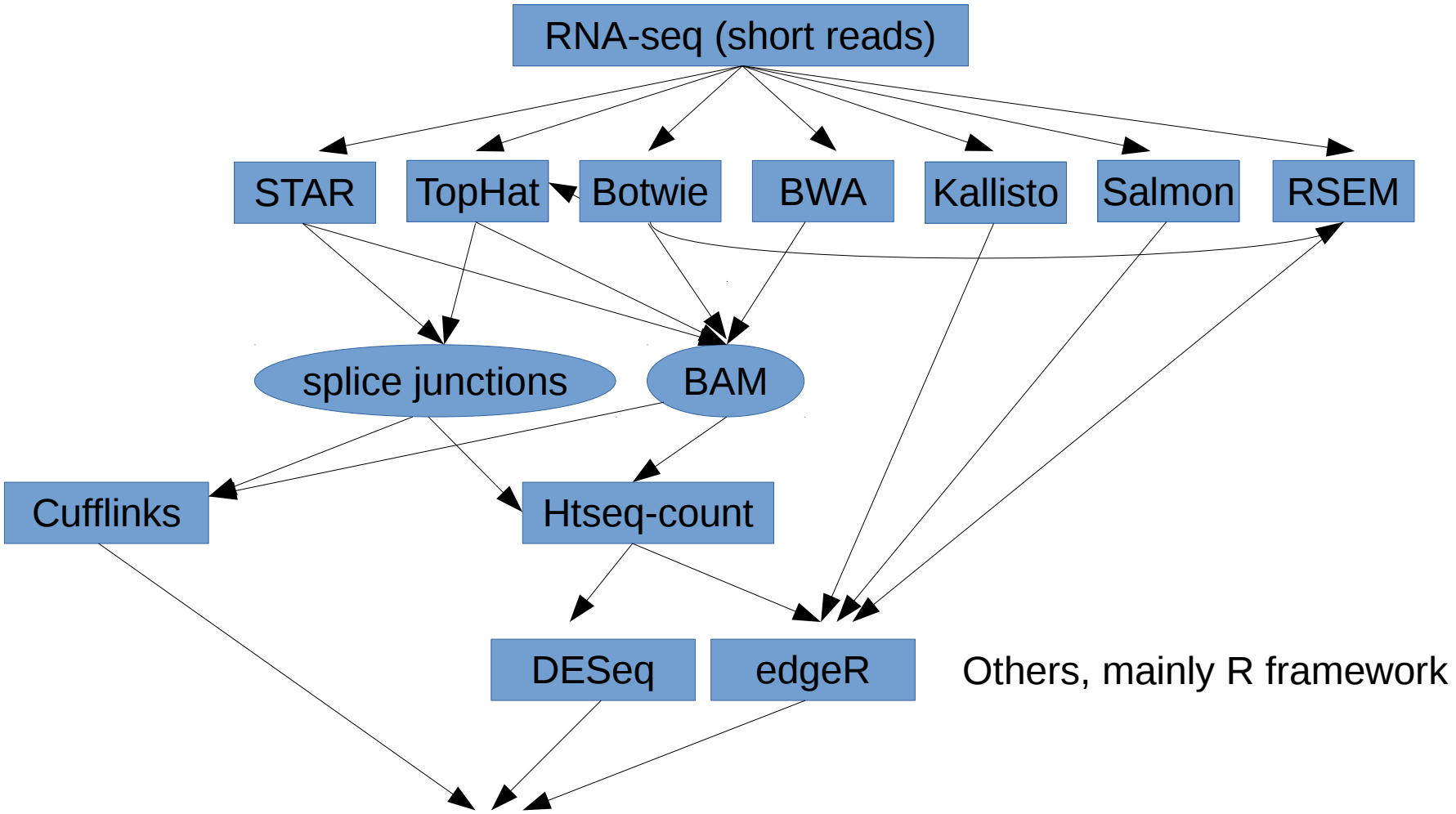
- Reads per kilobase of exon per million reads mapped (RPKM), or the more generic FPKM (substitute reads with fragments) are essentially the same thing

$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\widetilde{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\widetilde{l}_i N} \cdot 10^9$$

Inter sample normalization

- Many use “quantile” normalization
  - making distributions identical in statistical properties
- New normalization methods currently being published
- Different normalization methods give different results

Differential gene expression



Differential gene expression

- Best methods to discover DE are coupled with sophisticated approaches to normalization
- Very low expressing genes are tricky
  - FPKM<1