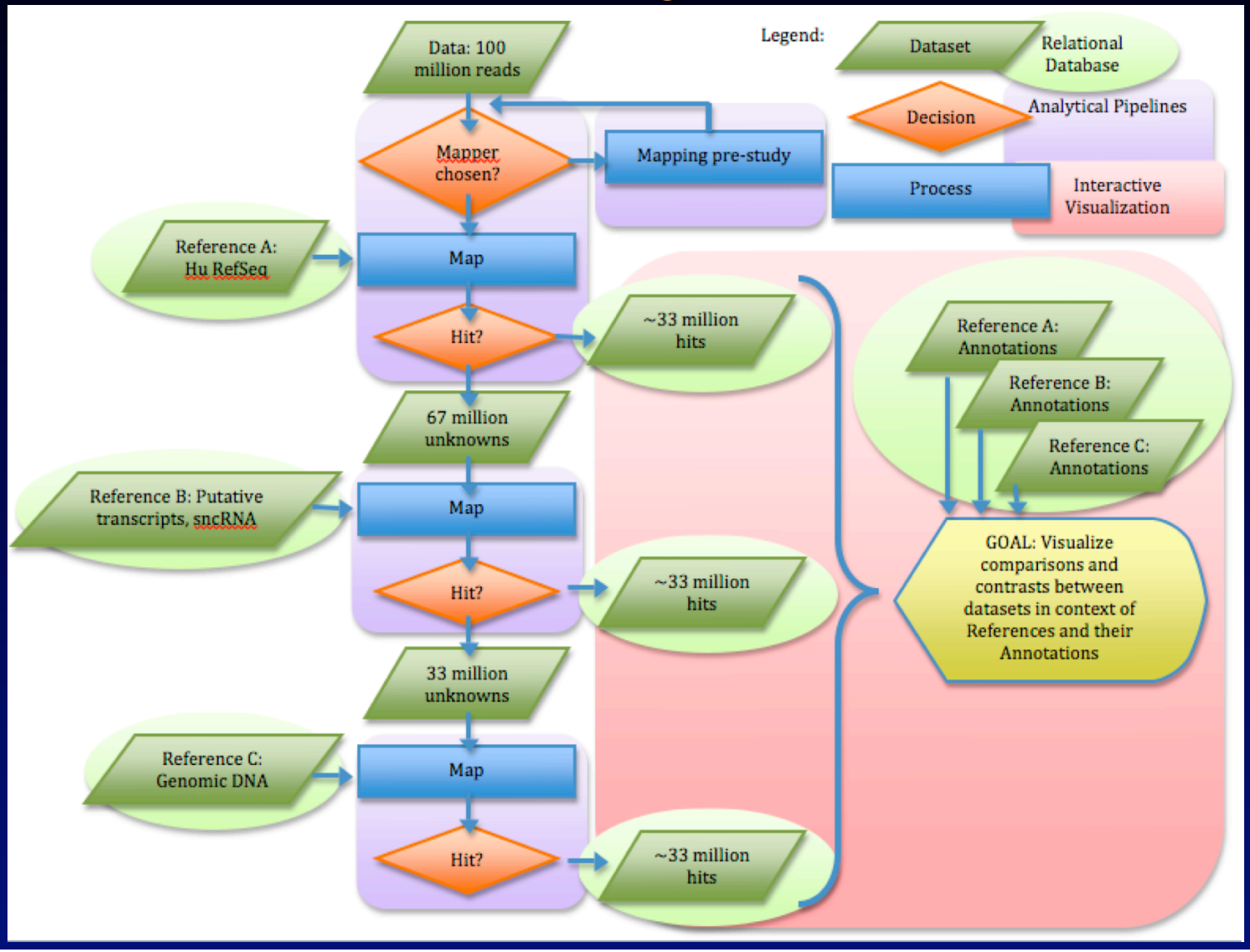
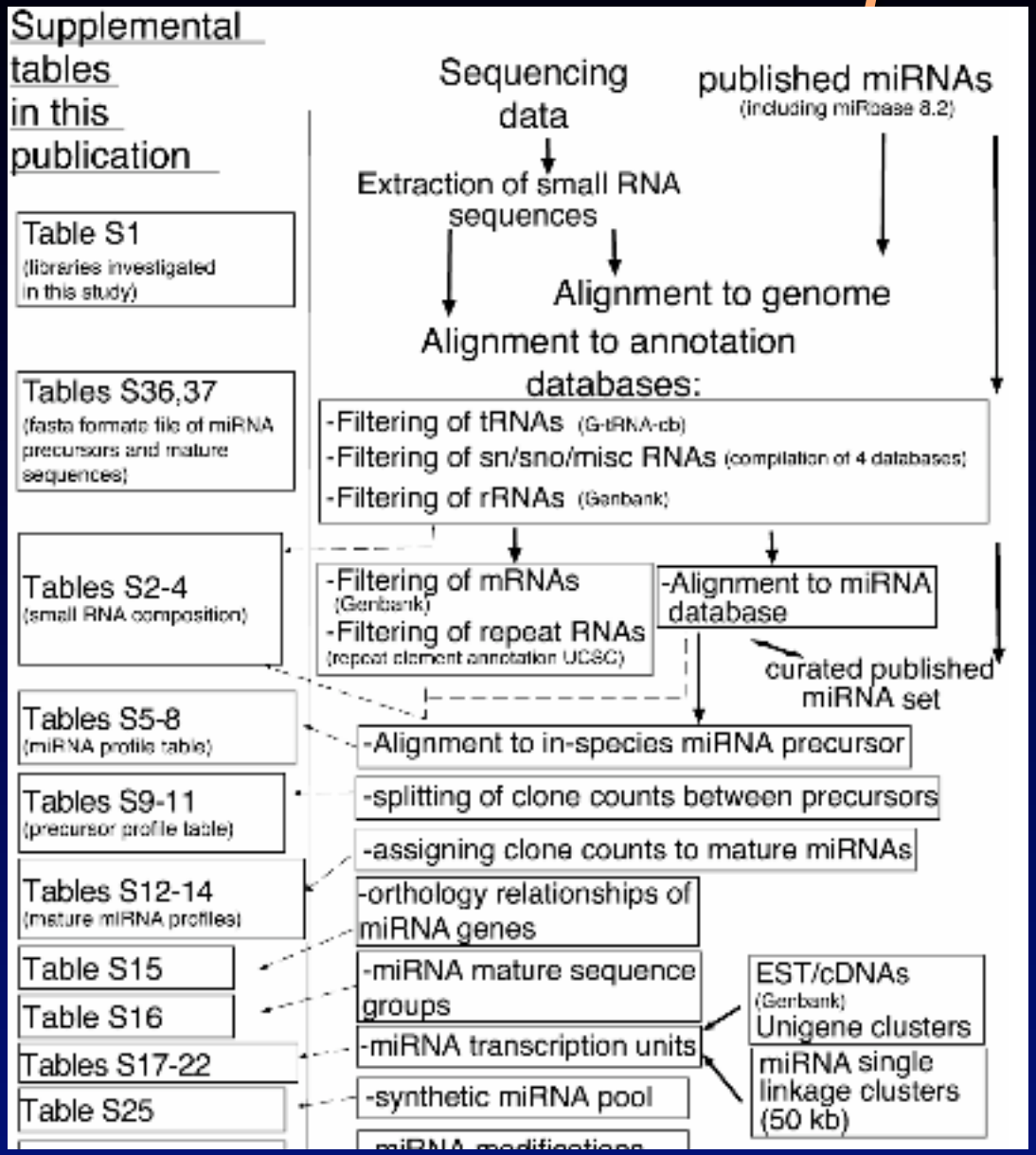
RNA-seq

- RNA-seq analysis usually requires a pipeline – let's look at two examples

Data Analysis: RNA-Seq



Pipeline example

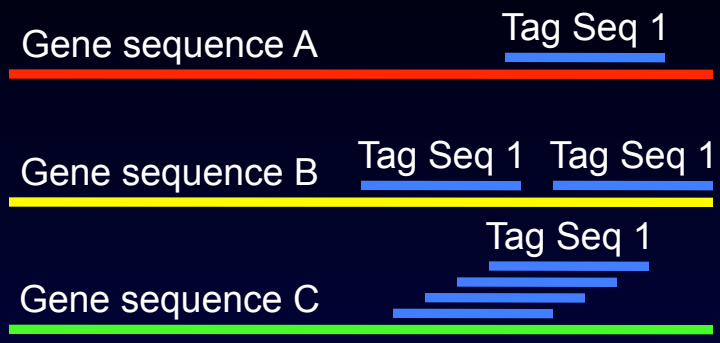


From: Landgraf, et. al., “A mammalian microRNA expression atlas based on small RNA library sequencing.”, Nat Biotechnol. 2007 Sep; 25(9):996-7, supplemental materials

RNA-seq pipelines

- Pipelines are based on rule sets:
  - What are you trying to find?
  - What do you need to rule-out?
  - What what matters most – sensitivity or specificity?

Rule Set Example



- Basis for definition of “hit”…
- Accept all hits
- Collapse intergenic non-unique
- Select random non-unique
- Select only unique
- Apply stat model to non-unique
- Summarize by gene, exon (gene model?)

Comparison of Short-Read Mappers & Filters

Mapping along normalized gene length – effects of post-mapping filters.

Fig 1a: Bowtie raw output,max.100 hits per tag (No filter)

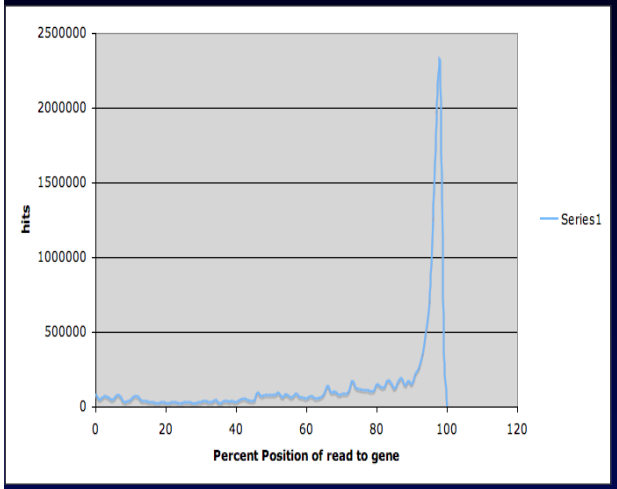


Fig 1b: Bowtie output, max.25 hits per tag, 3mis, nontiling, max. coverage of 1%

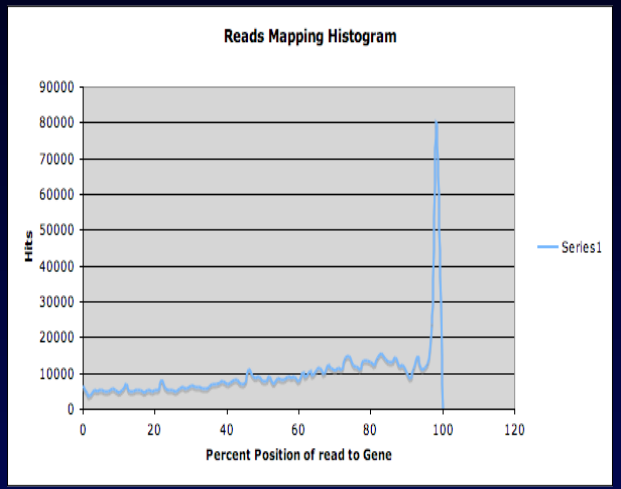


Fig 1c: Bowtie output,1 hit per tag, 3mis, nontiling, max.coverage of 1%, no polyA tails

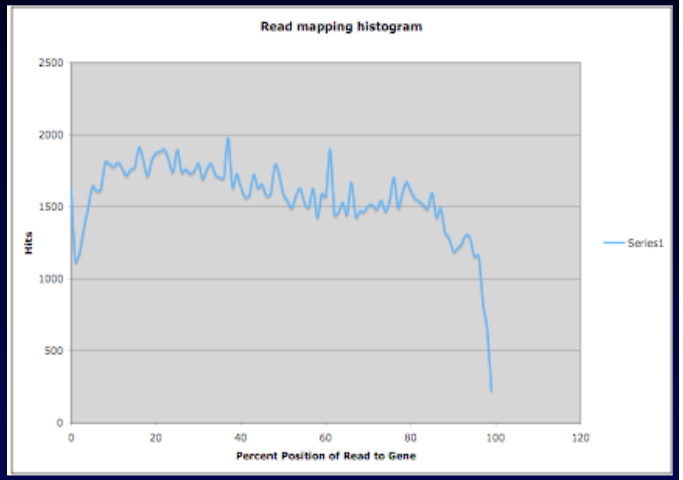


Fig 2a:SOAP2 raw output (No filter)

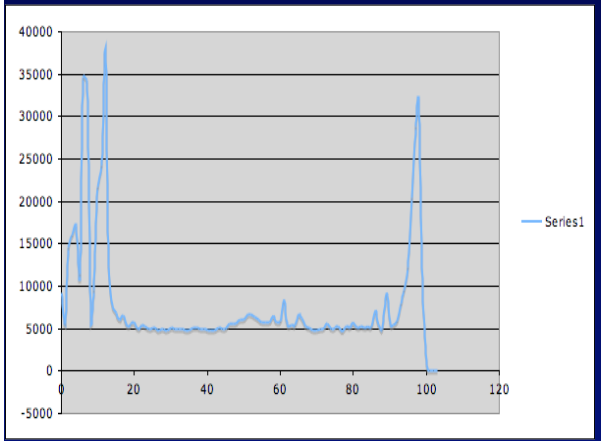


Fig 2b: SOAP2 output, 1 hit per tag, 3mis, nontiling, max. coverage of 1%

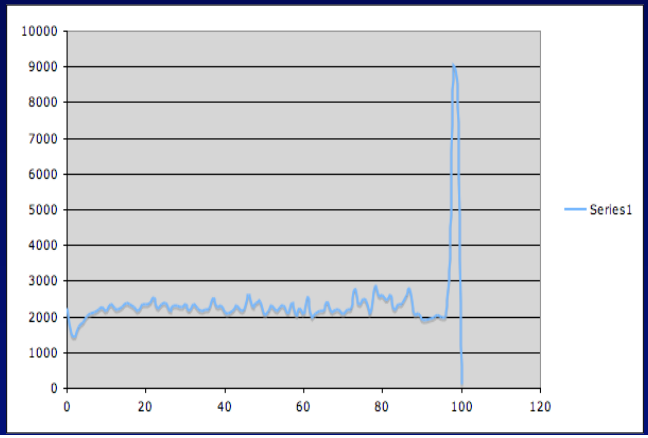
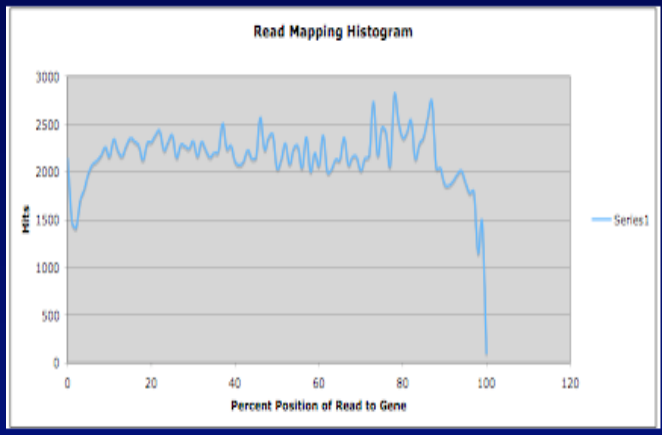
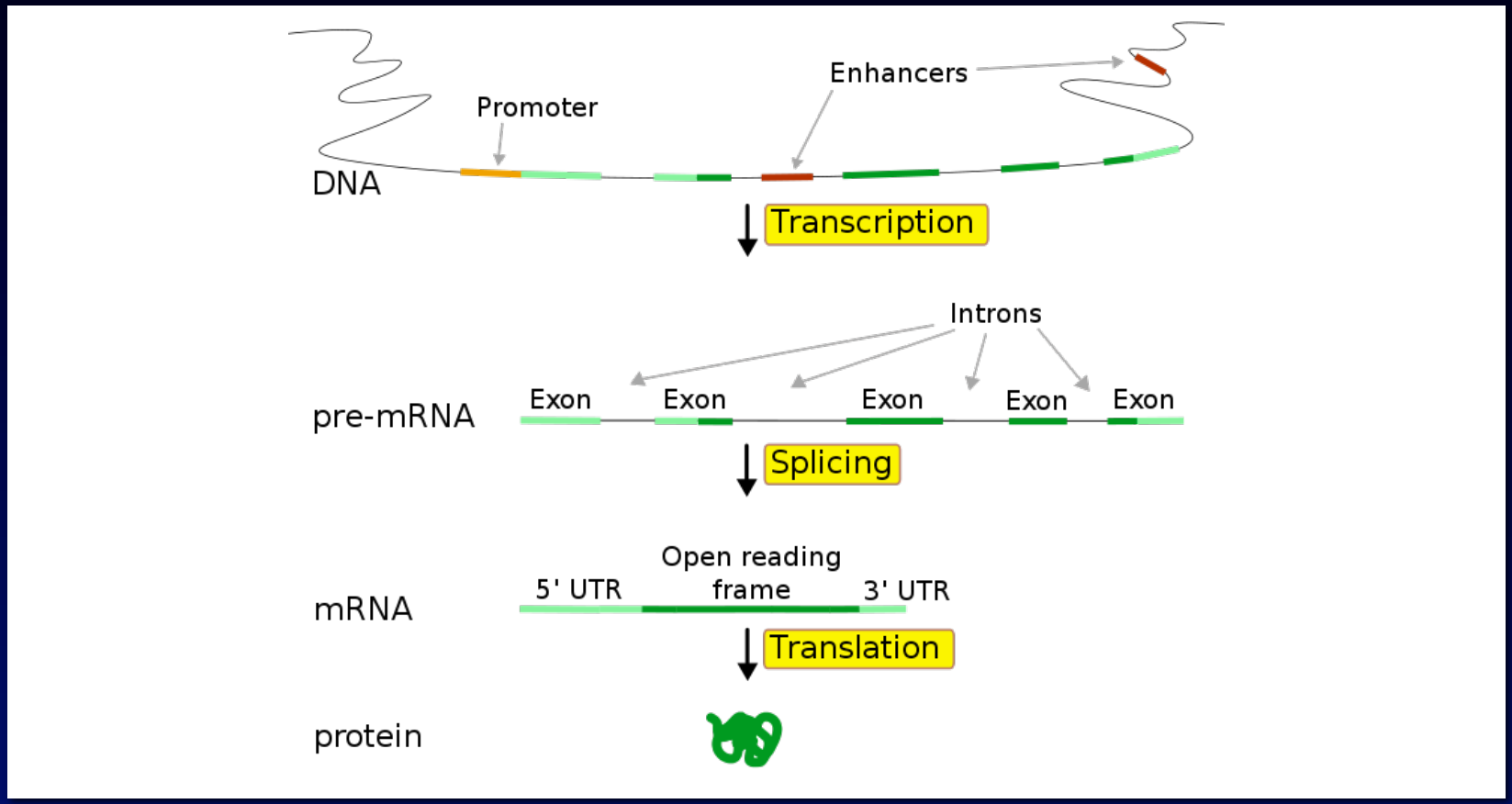


Fig 2c: SOAP2 output,1 hit per tag, 3mis, nontiling, max.coverage of 1%, no polyA tails



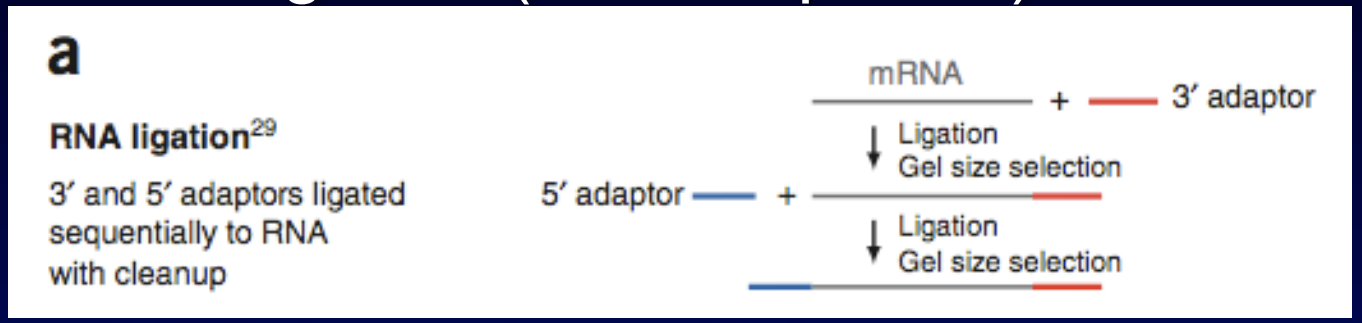
What is a gene?



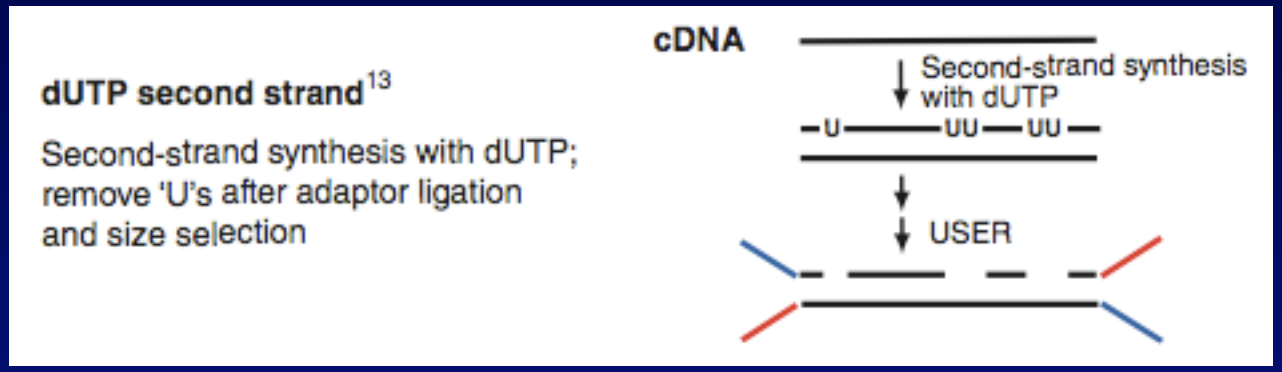
From Wikipedia's entry on "gene", which neglects UTR's

RNA-seq methods

- 3 methods:
  - ds cDNA (not strand specific)
  - RNA ligation (strand specific)



  - dUTP (strand specific)



Figures from Levin, et. al, Nat. Methods, 2010

Artifacts (?)

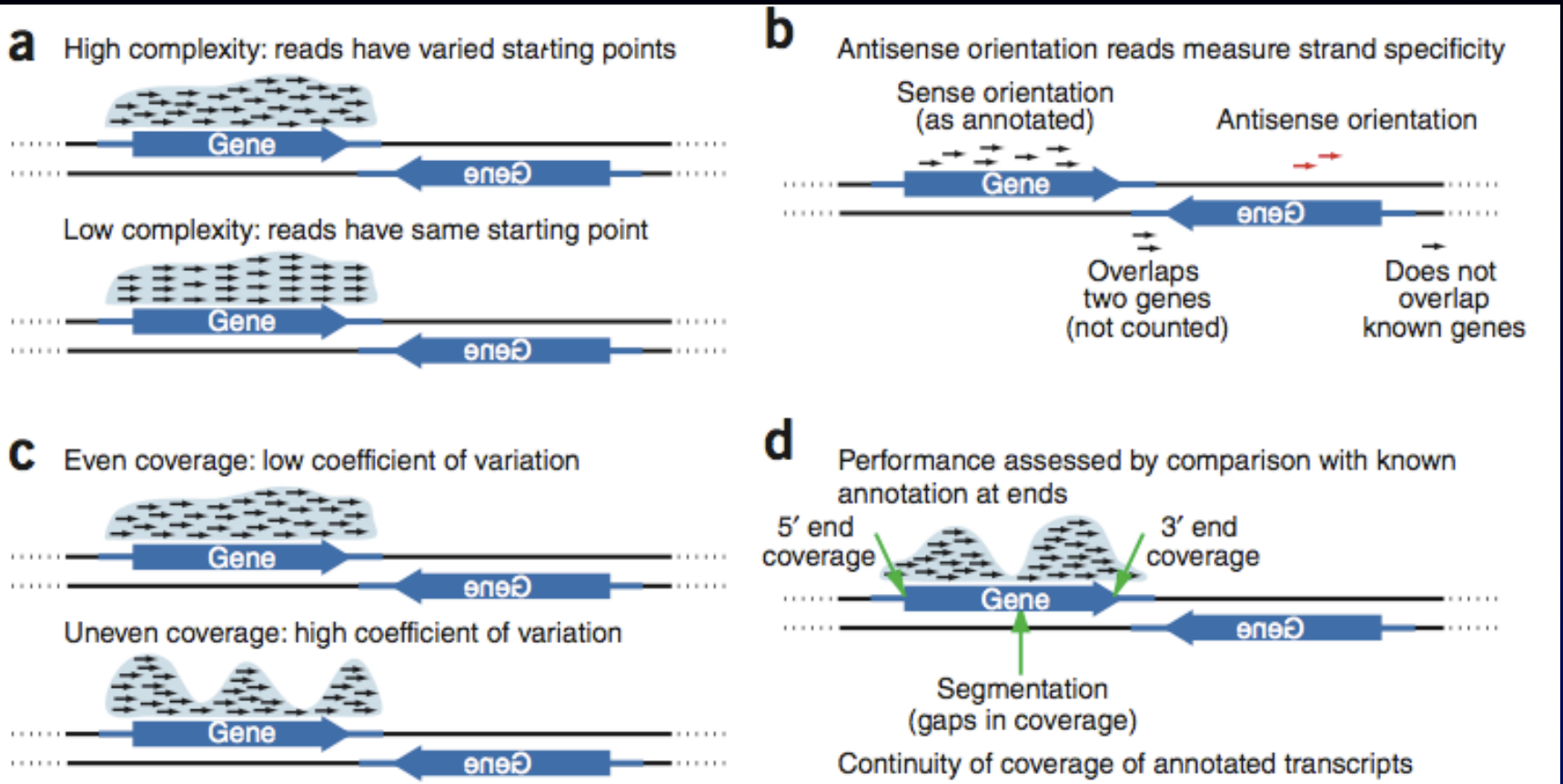


Figure 2 | Key criteria for evaluation of strand-specific RNA-seq libraries. (**a**–**d**) Categories of quality assessment were complexity (**a**), strand specificity (**b**), evenness of coverage (**c**) and comparison to known transcript structure (**d**). Double-stranded genome with gene ORF orientation (blue arrows) and UTRs (blue lines) are shown along with mapped reads (black and red arrows, reads mapped to sense and antisense strands, respectively).

From “Comprehensive comparative analysis of strand-specific RNA sequencing methods”, Levin, et. al.
Nature Methods 7, 709–715 (2010) doi:10.1038/nmeth.1491

Counting & normalization

- Example:
  - RNA Sample 1 has 1e7 reads
  - RNA Sample 2 has 1e8 reads
    - (Worse: Sample 2 has 60% mapping, Sample 1 has 80% mapping…)
- How do you normalize?
  - Mean? Median centering? Quantile? Variance stabilizing normalization?

Counting & normalization

- Example:
  - Gene A is 1000 bp, Gene B is 40,000 bp.
  - If Gene A has 10,000 reads and Gene B has 400,000 reads, is their expression equivalent?
- Practical normalization:
  - FPKM: Fragments per kilobase “gene/exon” length per million mapped reads.
  - BUT! This confounds VARIANCE.

Differential Expression

The *t* statistic to test whether the means are different

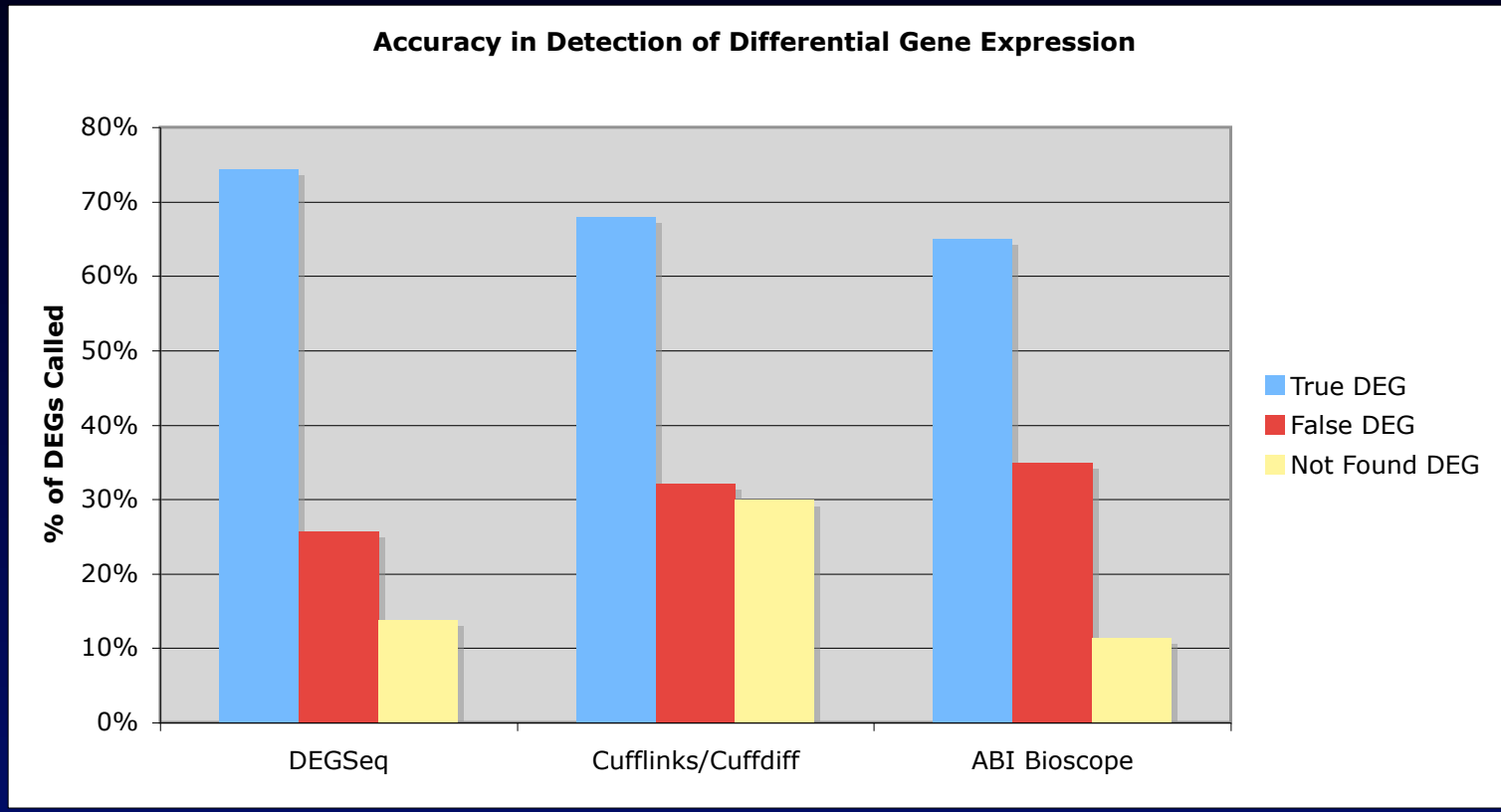
$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

where

$$S_{X_1X_2} = \sqrt{\frac{1}{2}(S_{X_1}^2 + S_{X_2}^2)}$$

- From Wikipedia: “Student’s t-test”

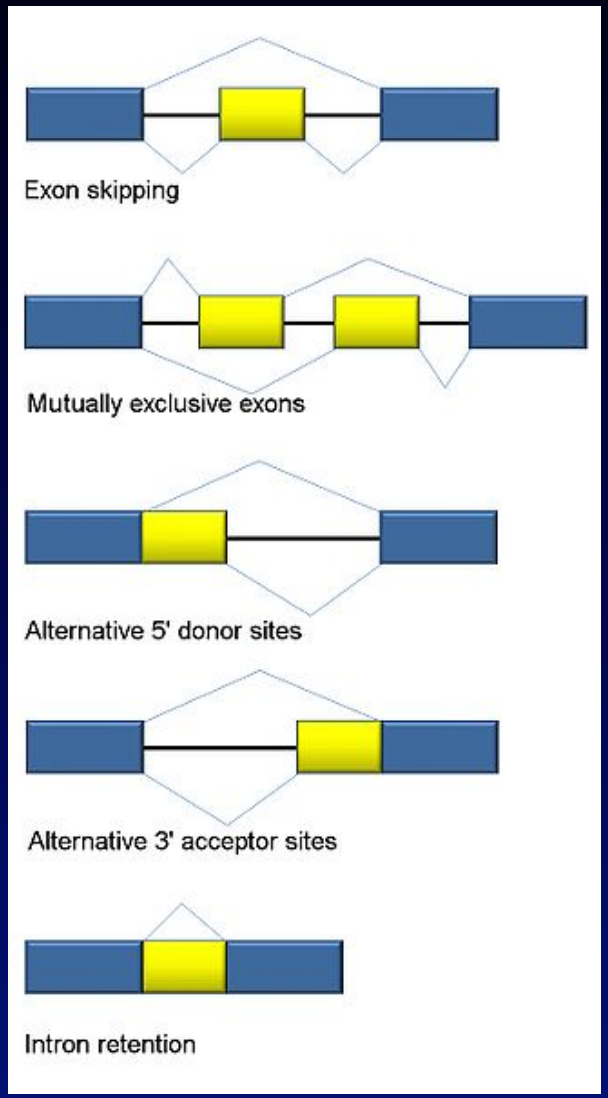
Differential Expression



Issues

- We didn’t even discuss what it means for a gene to have differential expression amongst it’s isoforms…

The trouble with measuring genes



- From Wikipedia's entry on "alternative splicing"