

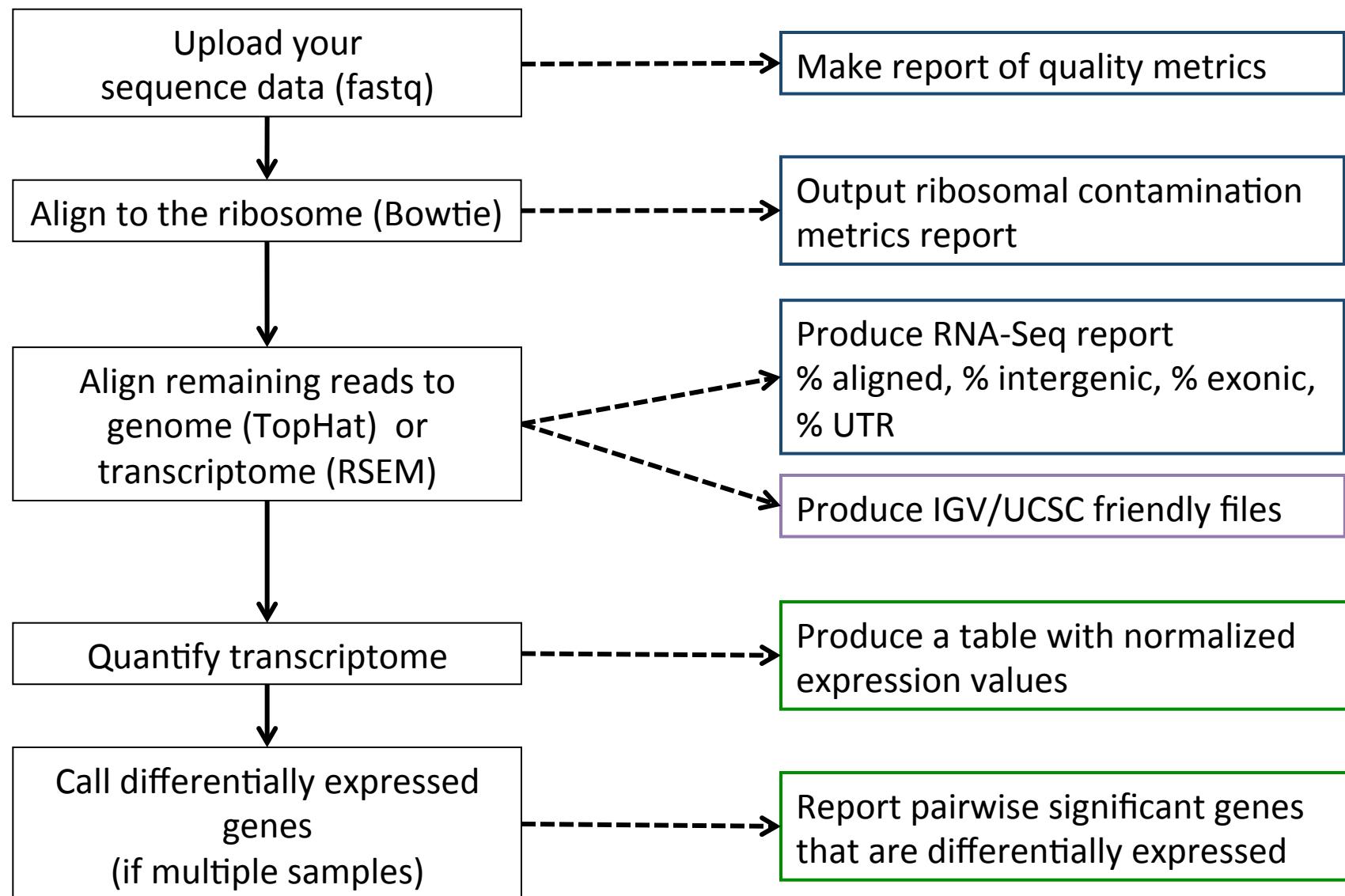


# **Week 5**

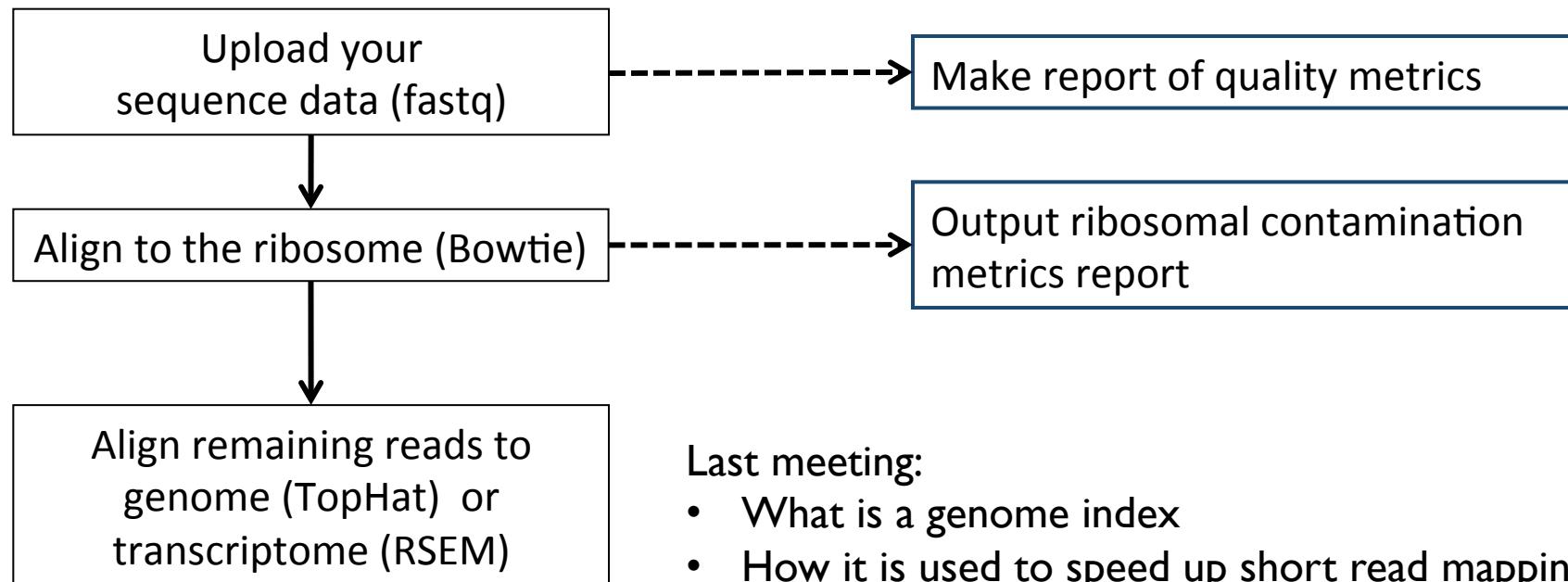
# **FDR**

# **Alignment**

# Our typical RNA quantification pipeline



# Alignment requires pre-processing



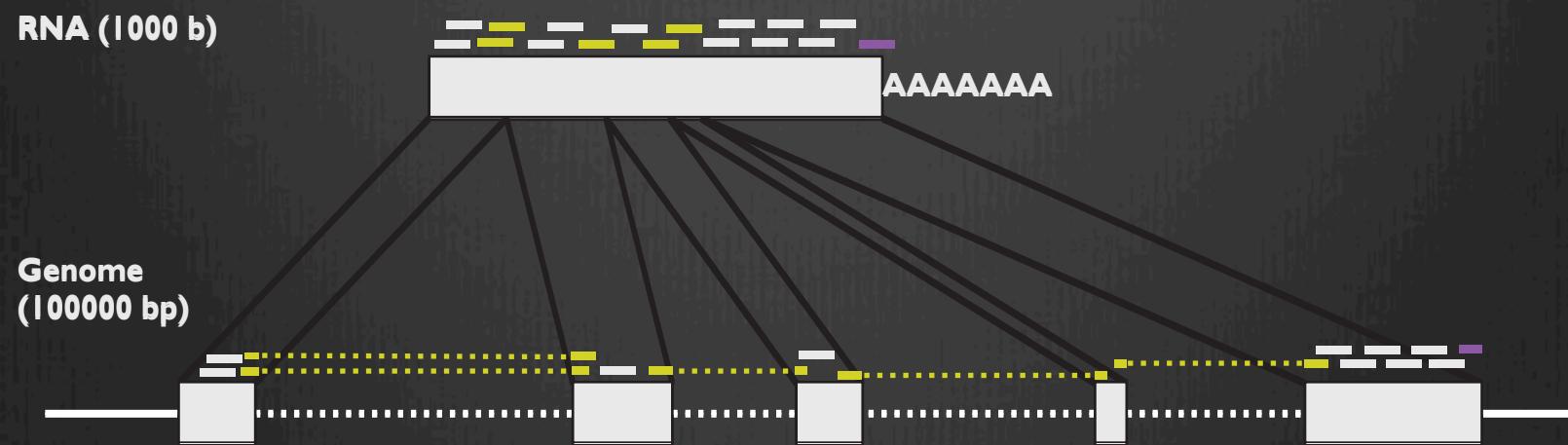
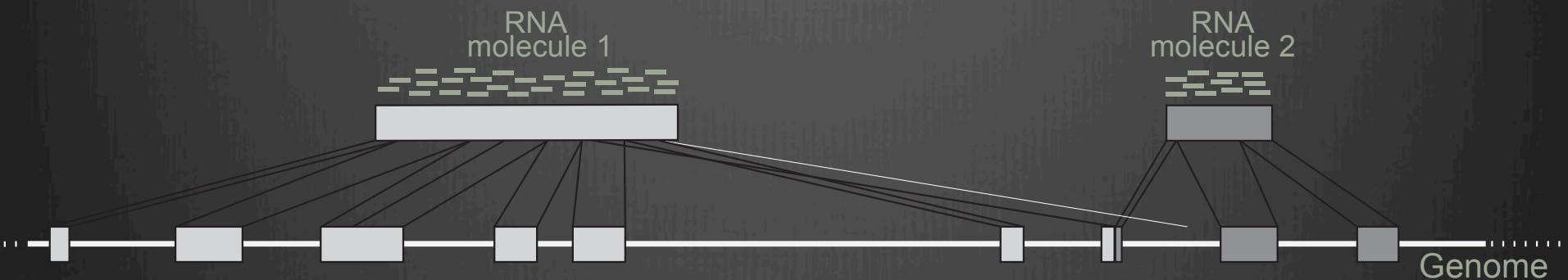
Last meeting:

- What is a genome index
- How it is used to speed up short read mapping
- What is mapping quality

```
bowtie2-build -f mm10.fa mm10
```

```
rsem-prepare-reference \  
--gtf ucsc.gtf --transcript-to-gene-map ucsc_into_genesymbol.rsem \  
mm10.fa mm10.rsem
```

# RNA-Seq Read mapping

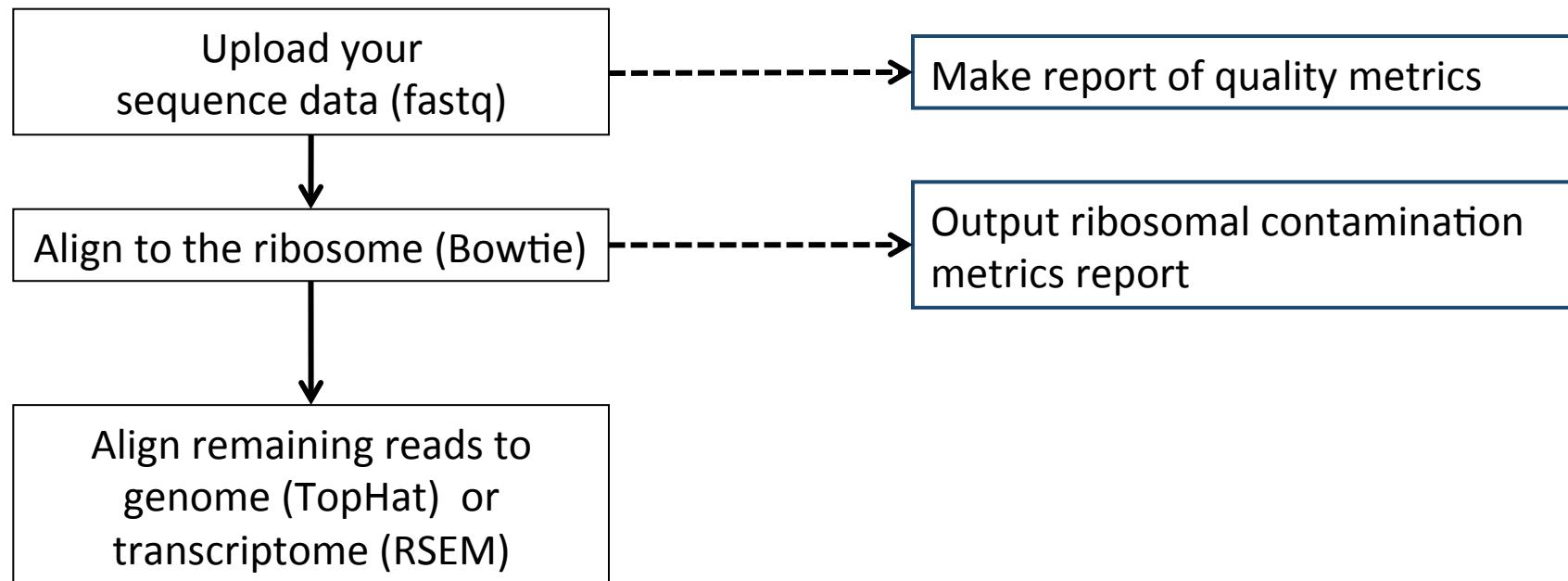


## Mapping RNA-Seq reads: Exon-first spliced alignment (e.g. TopHat)

---



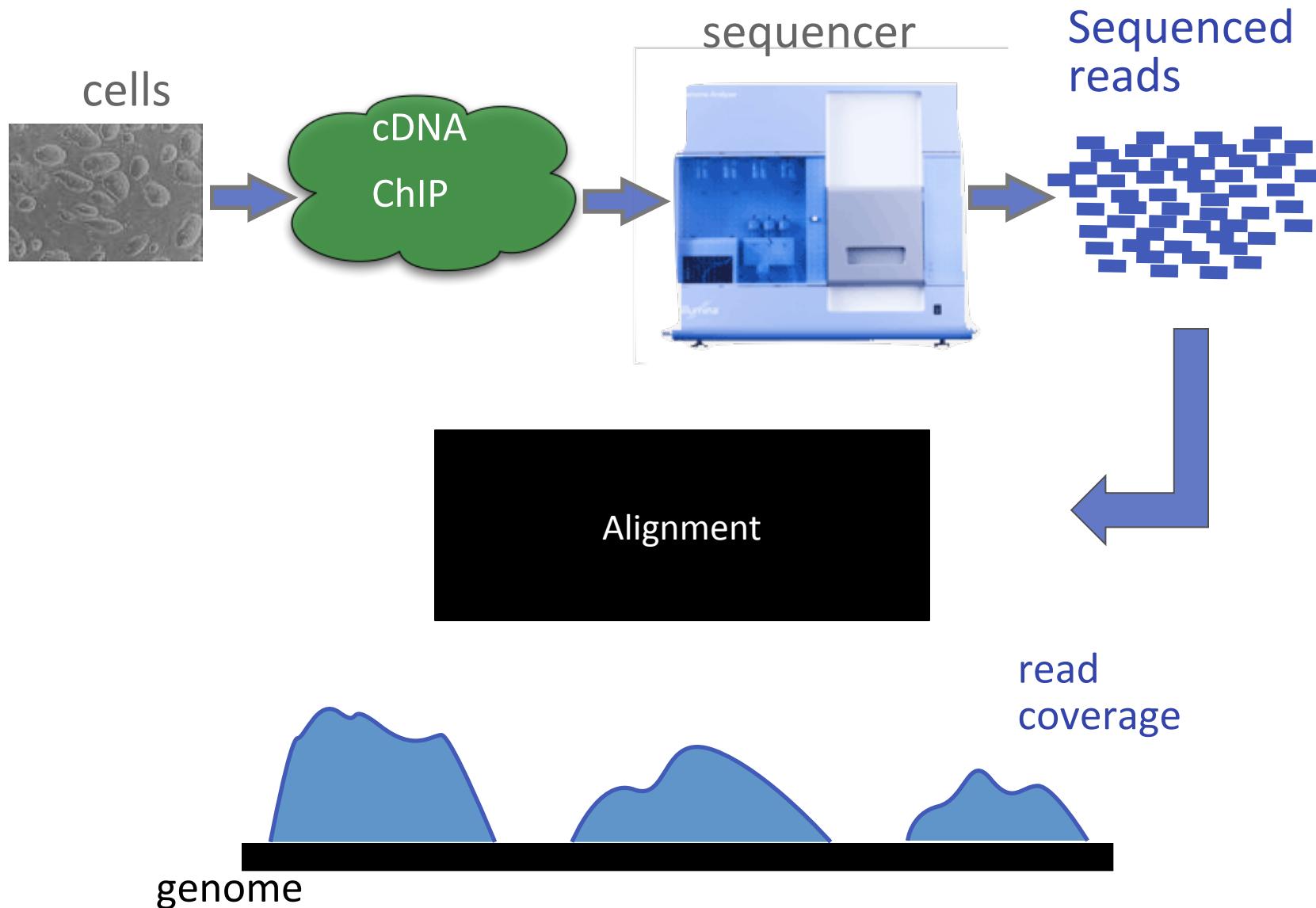
# Short read alignment



```
tophat2 --library-type fr-firststrand --segment-length 20 \  
-G genome.quantification/ucsc.gtf -o tophat/th.quant.ctrl1 \  
genome.quantification/mm10 fastq.quantification/control_rep1.1.fq \  
fastq.quantification/control_rep1.2.fq
```

```
/project/umw_biocore/bin/igvtools.sh count -w 5 tophat/th.quant.ctrl1.bam \  
tophat/th.quant.ctrl1.bam.tdf genome.quantification/mm10.fa
```

# Once sequenced the problem becomes computational



## Considerations and assumptions

---

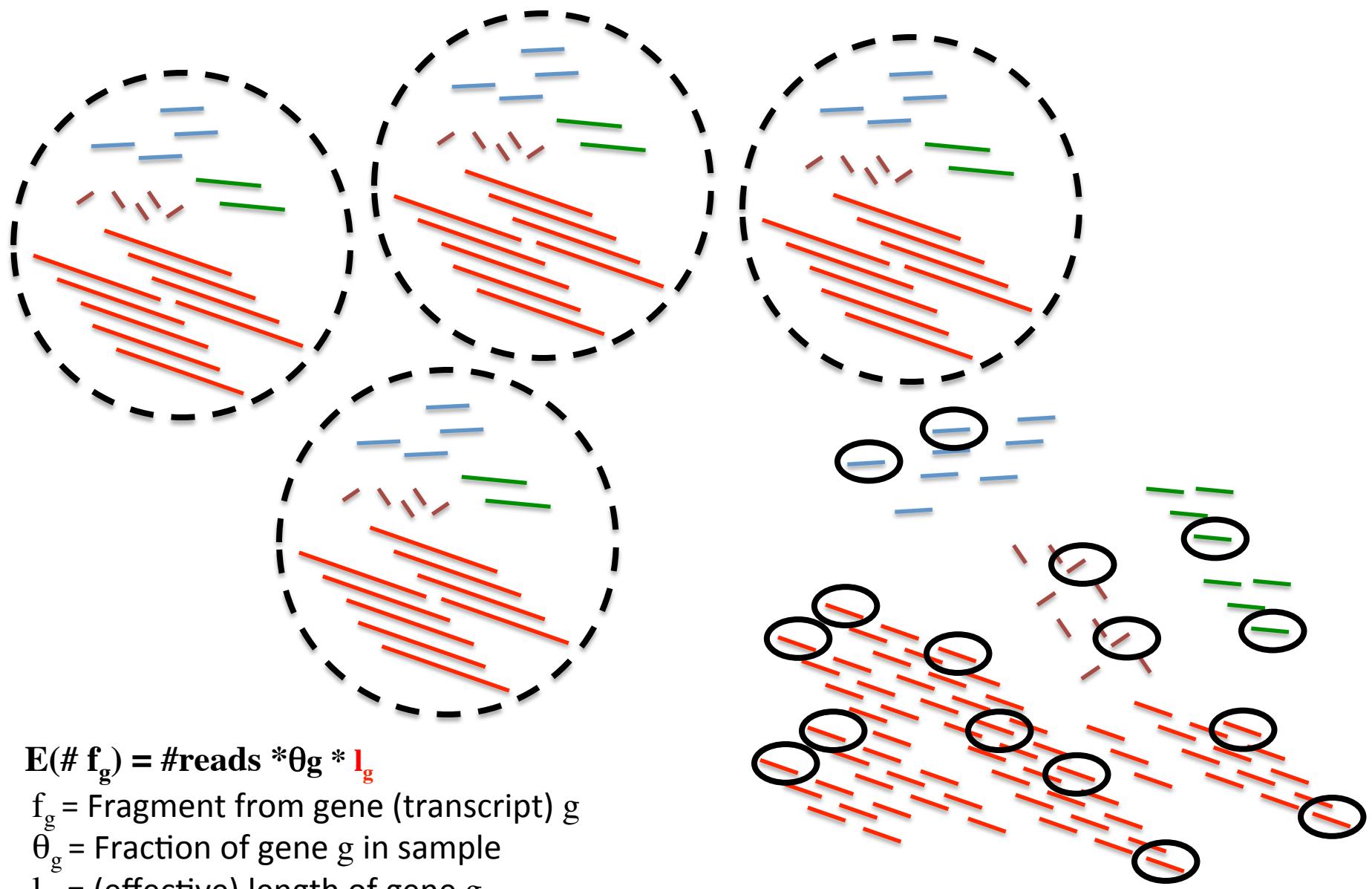
- High library complexity
  - #molecules in library  $\gg$  #sequenced molecules
- Short reads
  - Read length  $\ll$  sequenced molecule length

Not all applications satisfy this:

- miRNA sequencing
- Small input sequencing (e.g. single cell sequencing)

## A library satisfying assumptions 1 & 2

---



## Corollaries

---

- Libraries satisfying assumptions 1 & 2 only measure relative abundance
- Key quantity: # fragments sequenced for each transcript.  
Need to:
  - **Which transcript generated the observed read?**
- Isn't this easy?
  - Reads do not uniquely map
  - Transcripts or genes have different isoforms
  - Sequencing has a ~ 1% error rate
  - Transcripts are not uniformly sequenced

# The RNA-Seq quantification problem (simple case)

---

- Start with a set of previous gene/transcript annotations
- Assume only one isoform per gene
- Assume 1-1 read to transcript correspondence.

Let  $\Theta = \{\theta_g\}$  the relative abundance of each gene

let  $n_g$  the number of reads aligned to gene  $g$

$N = \sum n_g$  (Sequencing depth)

$$P(n_g | \theta_g) = \binom{N}{n_g} \theta_g^{n_g} (1 - \theta_g)^{N-n_g} \approx \frac{e^{-\theta_g N} (\theta_g N)^{n_g}}{n_g!}$$

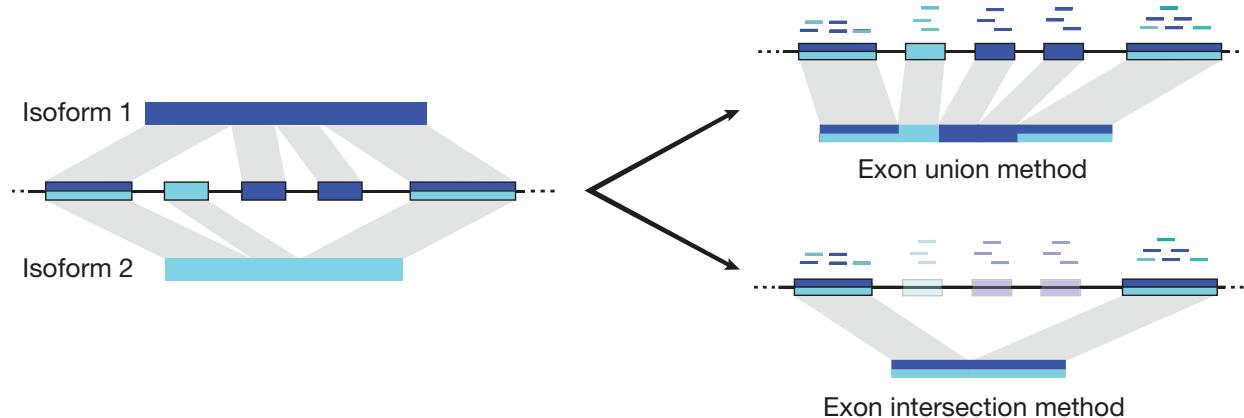
Using the Poisson  
approximation to the  
binomial

$$\mathcal{L}(\Theta) = \prod P(n_g)$$

We seek to maximize the likelihood of  
transcript frequencies given the data

Which, of course has MLE  $\theta_g = \frac{n_g}{\sum n_g}$

# The RNA-Seq quantification problem (more realistic case)



- Start with a set of previous gene/transcript annotations
- Define only one isoform per gene
- Assume 1-1 read to transcript correspondence. Reads (fragments) are now short, one transcript generates many fragments.

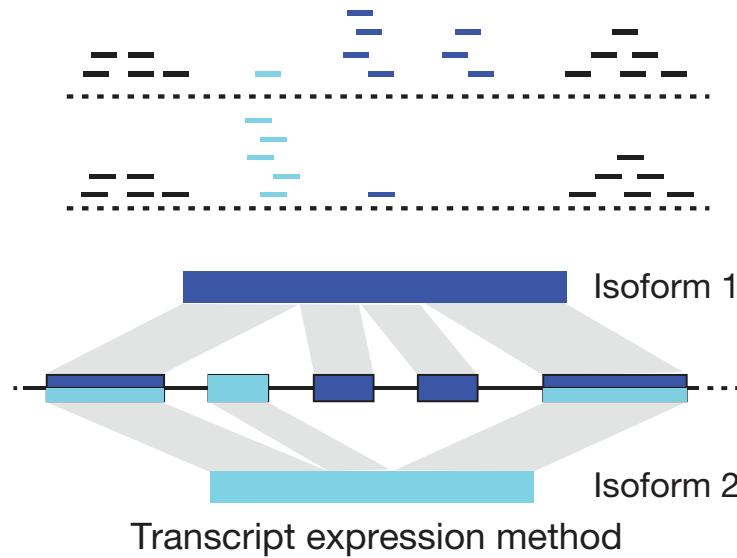
Change: Transcripts of different lengths generate  $f_g = \lfloor \frac{\tilde{l}_g}{w} \rfloor$  fragments

$\tilde{l}_g$  Transcript effective length

Model:  $P(n_g) = \frac{e^{-\theta_g \tilde{l}_g N} (\theta_g \tilde{l}_g N)^{n_g}}{n_g!}, N = \sum n_g$ , with MLE:  $\theta_g = \frac{n_g}{\tilde{l}_g N}$

# The RNA-Seq quantification problem. Isoform deconvolution

---



Main difference: quantification involves read assignment. Our model must capture read assignment uncertainty.

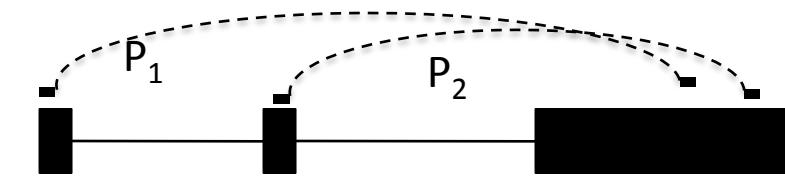
Parameters: Transcript relative abundance

**Latent variables:** Fragment alignment source

Observed variables: N fragment alignments, transcripts, *fragment length distribution*

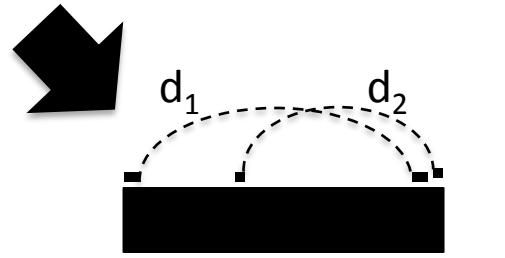
## We can estimate the insert size distribution

---

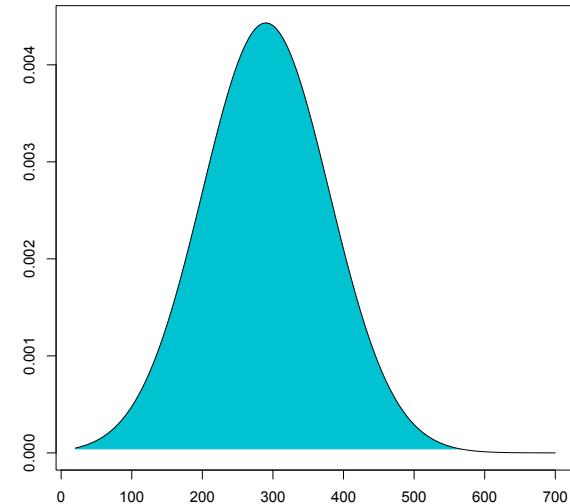


Get all single isoform reconstructions

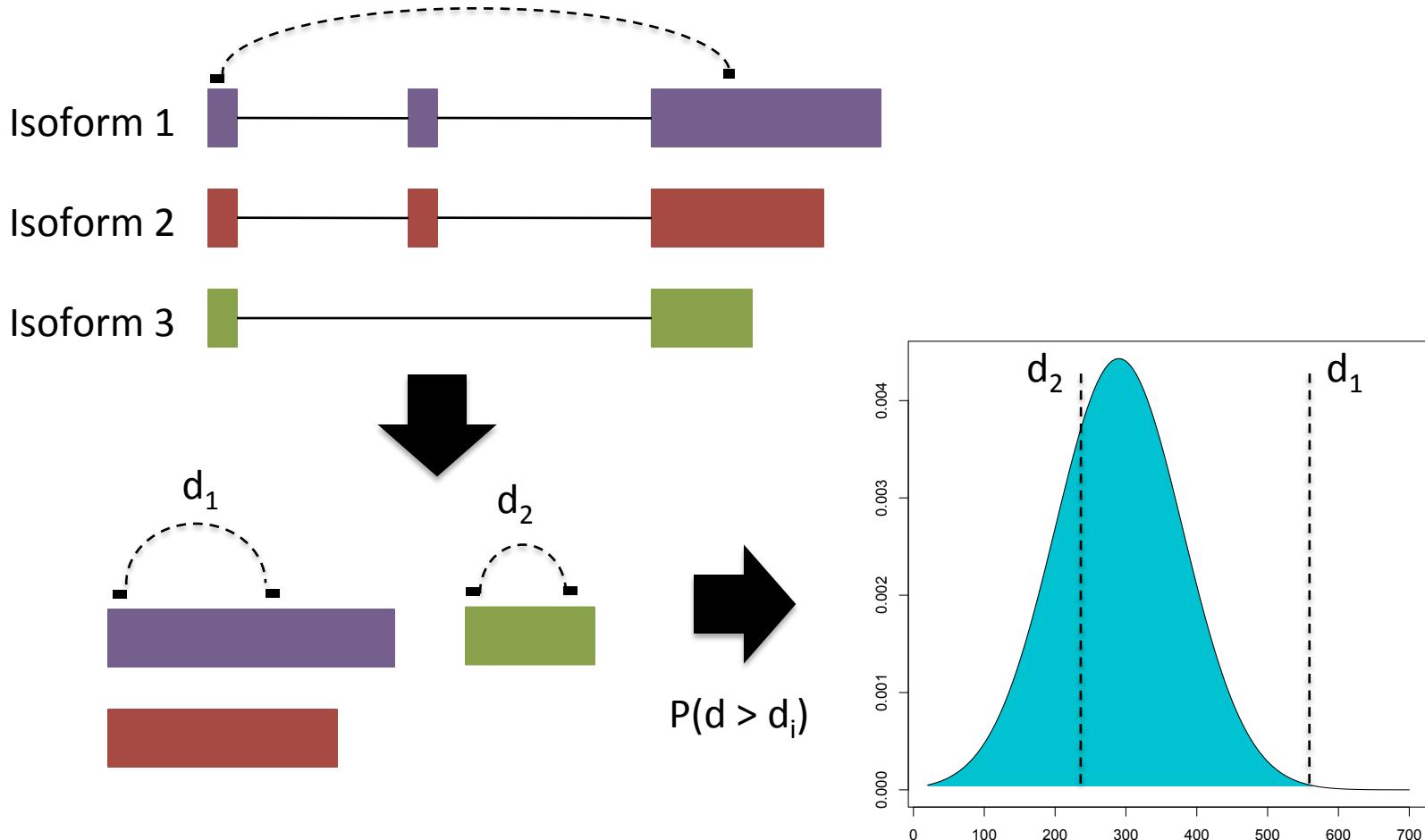
Splice and compute  
insert distance



Estimate insert size  
empirical distribution

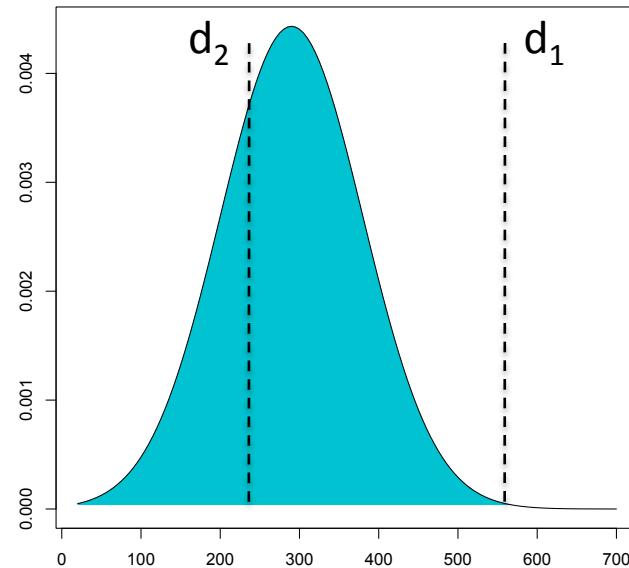
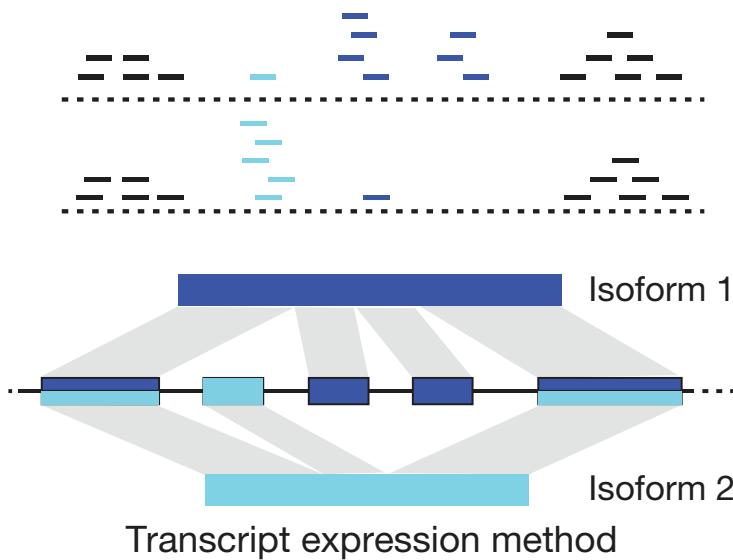


... and use it for probabilistic read assignment



For methods such as MISO, Cufflinks and RSEM, it is critical to have paired-end data

# The RNA-Seq quantification problem. Isoform deconvolution



Parameters: Transcript relative abundance

Latent variables: Fragment alignment source

Observed variables: N fragment alignments, transcripts, **fragment length distribution**

$$P(a \in t | D, \theta_t) = \frac{\theta_t \tilde{l}_t}{\sum_{a \in s} \theta_s \tilde{l}_s} P(l(a) | t, D)$$

$$\mathcal{L}(\Theta | D, A, G) = \prod_{t \in G} \prod_{a \in t} P(a \in t | D, \theta_t)$$

Probability of the fragment alignment originating from t

Can be shown it is concave, and hence solvable by expectation maximization

## Summary: Current quantification models are complex

---

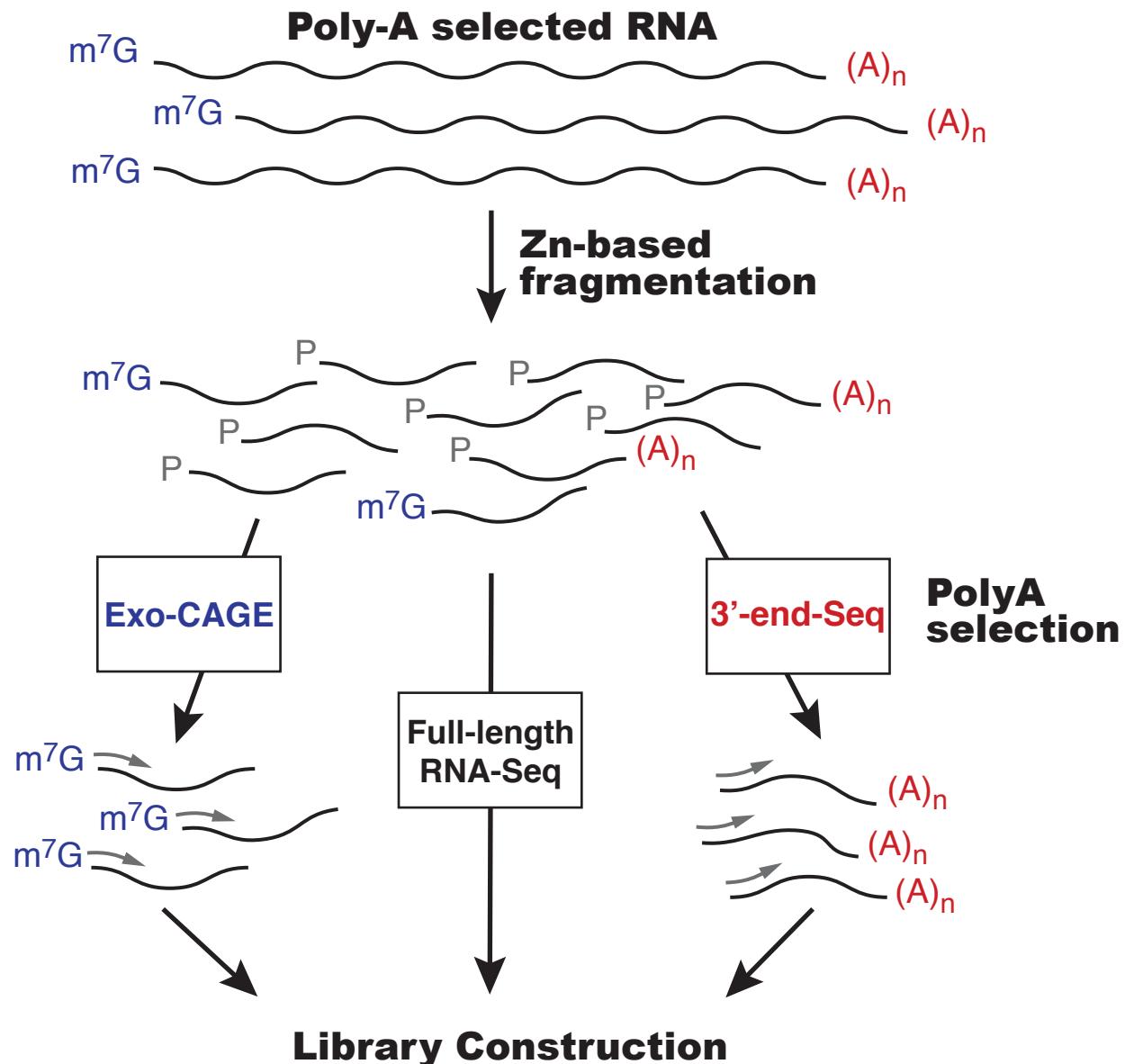
- In its simplest form we assume that reads can be unequivocally mapped. This allows:
  - Read counts distribute multinomial with rate estimated from the observed counts
- When this assumption breaks, multinomial is no longer appropriate.
- More general models use:
  - Base quality scores
  - Sequence mapability
  - Protocol biases (e.g. 3' bias)
  - Sequence biases (e.g. GC)
- Handling each of these involves a more complex model where reads are assigned probabilistically not only to an isoform but to a *different loci*

## RNA-Seq libraries revisited: End-sequence libraries

---

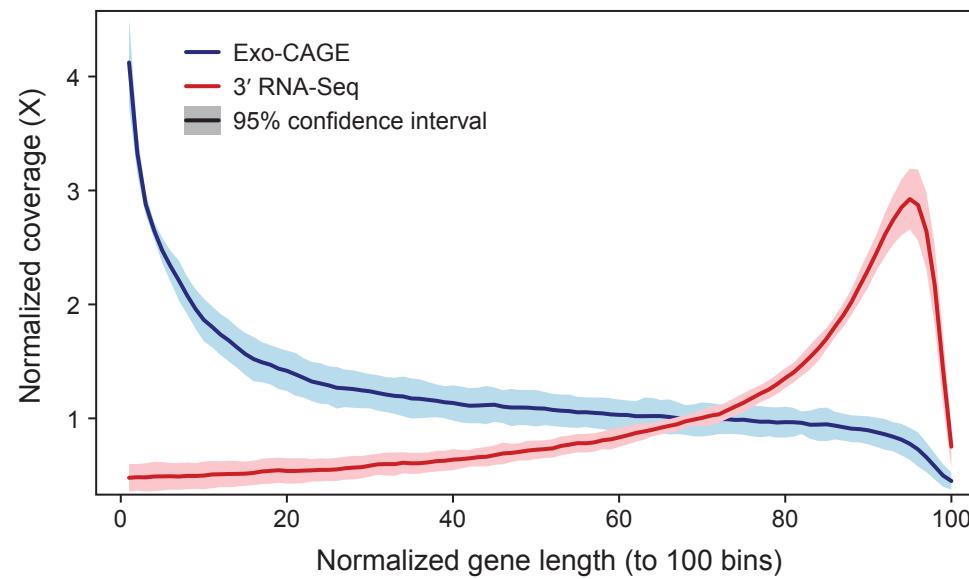
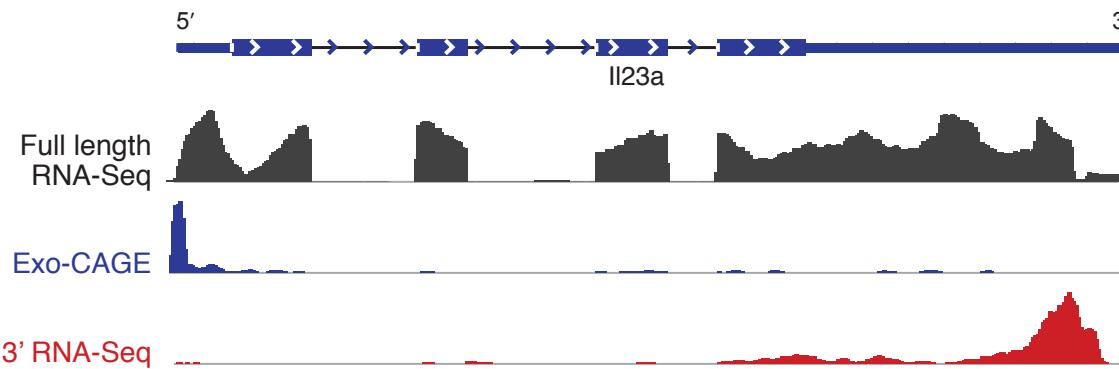
- Target the start or end of transcripts.
- Source: End-enriched RNA
  - Fragmented then selected
  - Fragmented then enzymatically purified
- Uses:
  - Annotation of transcriptional start sites
  - Annotation of 3' UTRs
  - Quantification and gene expression
  - Depth required 3-8 mill reads
  - **Low quality RNA samples**
  - **Single cell RNA sequencing**

# RNA-Seq libraries: Summary



# End-sequencing solution

---



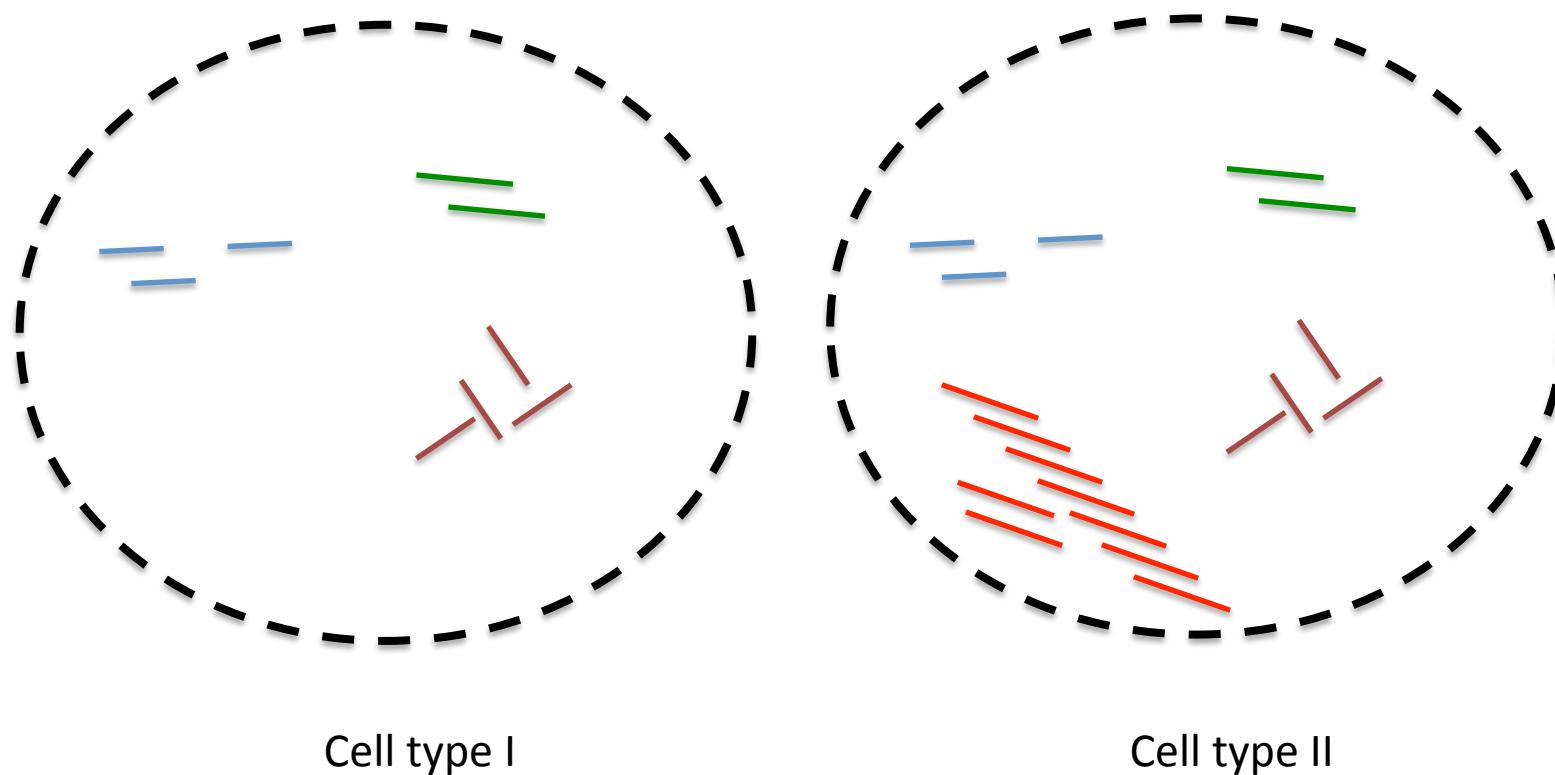
# Analysis of counting data requires 3 broad tasks

---

- Read mapping (alignment): Placing short reads in the genome
- Quantification:
  - Transcript relative abundance estimation
  - Determining whether a gene is expressed
  - Normalization
  - Finding genes/transcripts that are differentially represented between two or more samples.
- Reconstruction: Finding the regions that originated the reads

## Sample composition impacts transcript *relative* abundance

---



Normalizing by total reads does not work well for samples with very different RNA composition

## Example normalization techniques

$$s_j = \text{median}_i \frac{k_{ij}}{\left( \prod_{v=1}^m k_{iv} \right)^{1/m}}.$$

Counts for gene  $i$  in experiment  $j$

Geometric mean for that gene over ALL experiments

$i$  runs through all  $n$  genes

$j$  through all  $m$  samples

$k_{ij}$  is the observed counts for gene  $i$  in sample  $j$

$s_j$  is the normalization constant

# Lets do an experiment

---

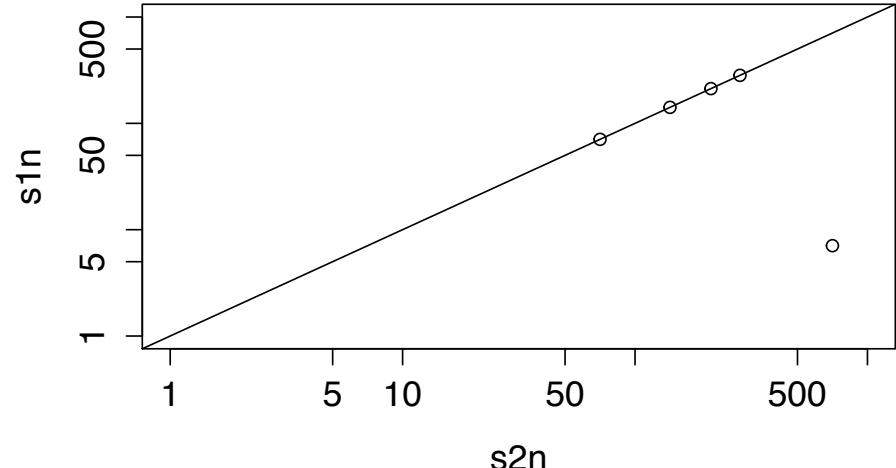
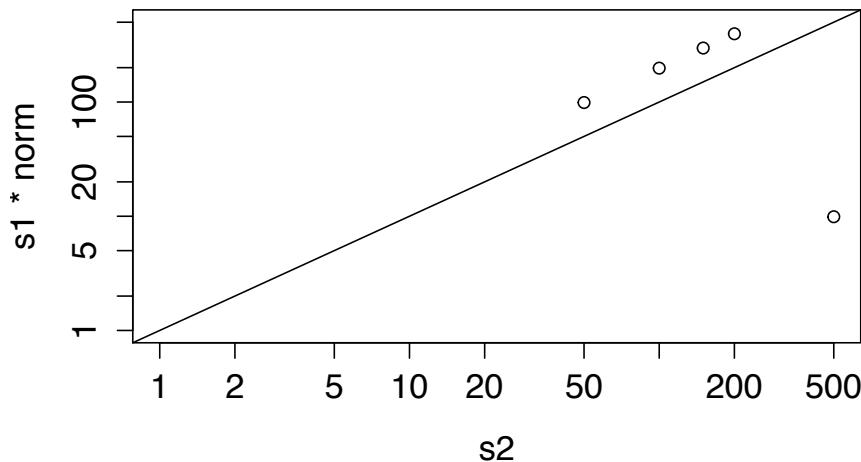
```
> s1 = c(100, 200, 300, 400, 10)  
> s2 = c(50, 100, 150, 200, 500)
```

```
> norm=sum(s2)/sum(s1)  
> plot(s2, s1*norm,log="xy")  
> abline(a = 0, b = 1)
```

```
> g = sqrt(s1 * s2t)  
> s1n = s1/median(s1/g); s2n = s2/median(s2/g)  
> plot(s2n, s1n,log="xy")  
> abline(a = 0, b = 1)
```

Similar read number,  
one transcript many fold changed

Size normalization results in 2-fold  
changes in *all* transcripts



# When everything changes: Spike-ins

