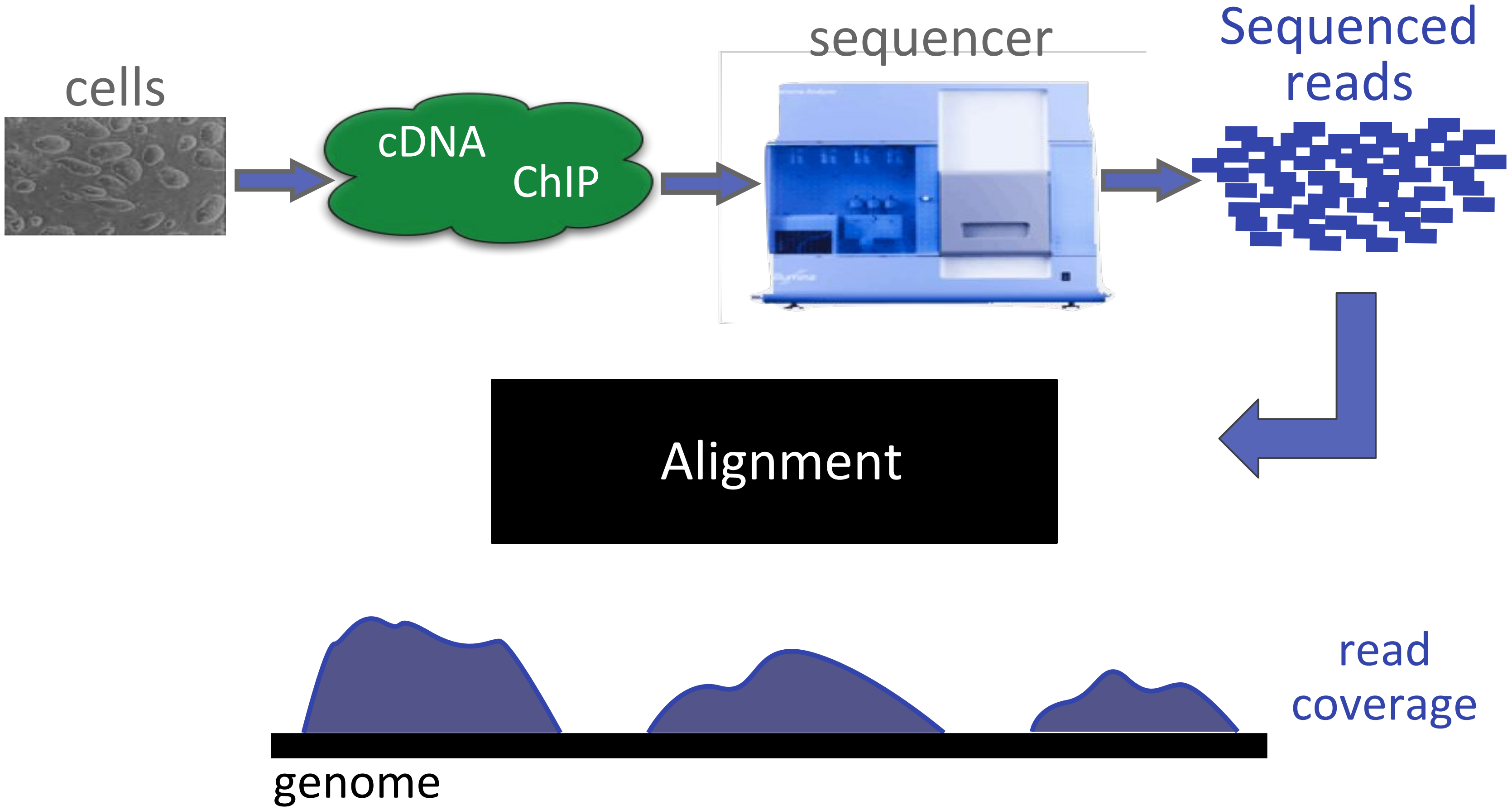
Gene expression from RNA-Seq

Once sequenced the problem becomes computational



cells

cDNA

ChIP

sequencer

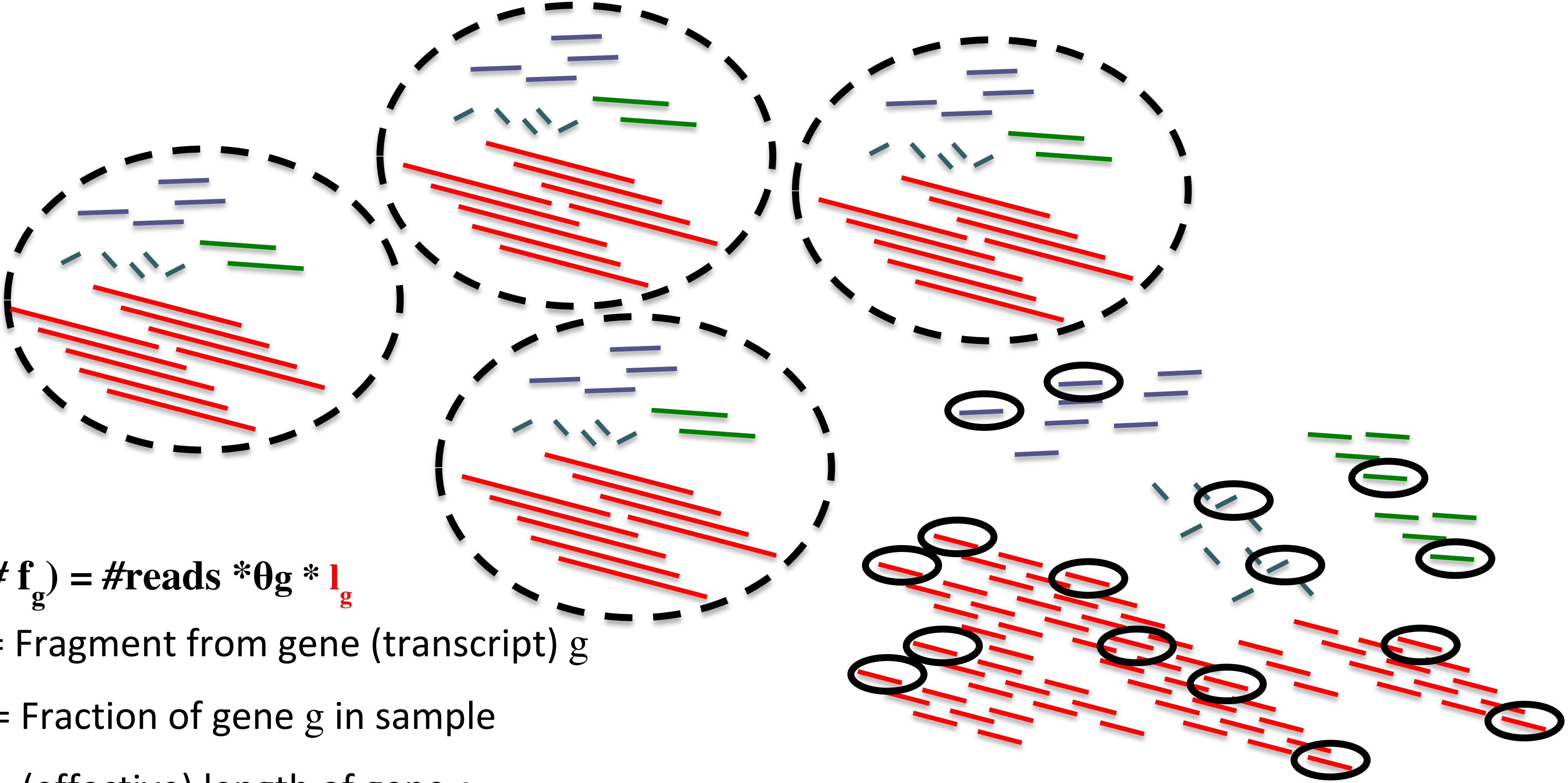
Sequenced reads

Alignment

read coverage

genome

Where we left….



$$E(\#\, f_g) = \#reads * \theta g * l_g$$

f_g = Fragment from gene (transcript) g

θ_g = Fraction of gene g in sample

l_g = (effective) length of gene g

Considerations and assumptions

1. **High library complexity**
   - #molecules in library >> #sequenced molecules
2. **Short reads**
   - Read length << sequenced molecule length

Not all applications satisfy this:

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

Corollaries

- Libraries satisfying assumptions 1 & 2 only measure relative abundance
- Key quantity: # fragments sequenced for each transcript.

  Data: *Aligned reads*

  Wanted*: transcript generated the observed read?

- Isn’t this easy?
  - Reads do not uniquely map
  - Genes have different isoforms with overlapping exons
  - 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

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

$$\text{let } n_g \text{ the number of reads aligned to gene } g$$

$N = \sum n_g$ (Sequencing depth)

The RNA-Seq quantification problem (simple case)

When a success has probability **θ**, the probability of **n** successes in **N** tries can be calculated by:

$$P(n_g \mid \theta_g) = \binom{N}{n_g} \theta_g^{n_g} (1 - \theta_g)^{N-n_g}$$

Which, has maximum probability at $\theta_g = \frac{n_g}{\sum n_g}$

The process of RNA-Seq quantification

- Sequenced reads are aligned to a reference sequence
  - the species genome or
  - its transcriptome
- Transcript abundance is measured:
  - By counting reads mapped to each transcript (not accurate when multiple isoforms share sequence)
  - By solving a maximum likelihood of the observed mapping given transcript abundance
- To compare samples, the counts need to be normalized
  - Libraries have different sequencing depth
  - Sample composition may be different
  - Most standard normalization: counts → Transcripts per Million (TPM) units

The gene expression table

Genes are quantified, each gene or isoform has:

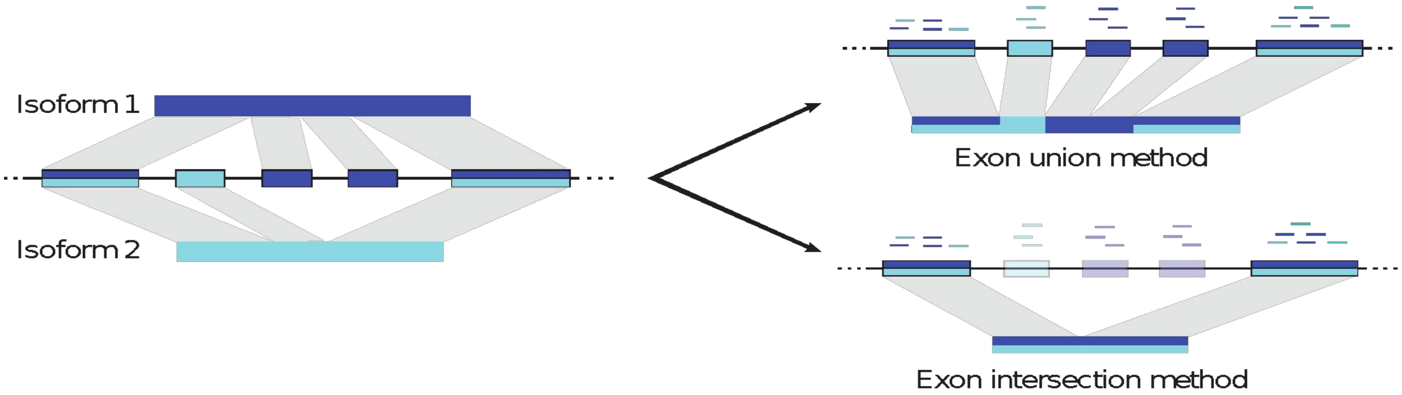
- A TPM value
- A (expected) fragment count value

All samples were quantified in the same fashion and arranged into a table of genes (22,000) x samples (24)

- Row i gives the expression of the gene i across all samples
- Column j gives the expression of genes in sample j

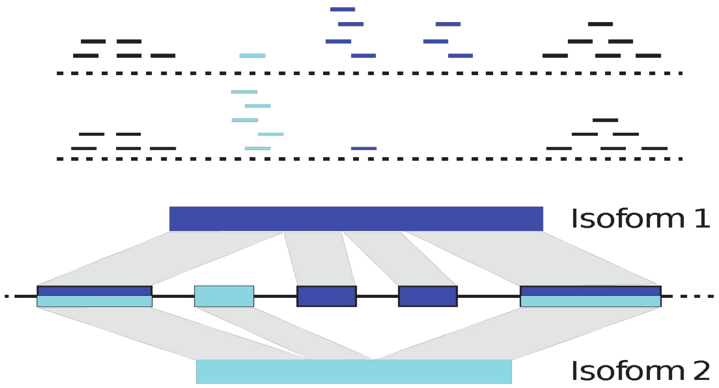
GENE	$L^{\Delta1,2}$rep1	$L^{\Delta1,2}$rep2	$L^{\Delta1,2}$rep3	$L^{\Delta1}$rep1	$L^{\Delta1}$rep2	$L^{\Delta1}$rep3	$L^{\Delta2}$rep1	$L^{\Delta2}$rep2	$L^{\Delta2}$rep3
Mir301	0	0	0	0	0	0	0	0	0
Cpne2	157	158.98	88.04	69	111.99	114.33	93	208	140
Capn5	36	65	46	46	69	42	33	58	59.01
Lage3	313.06	241.23	276.23	218.9	285.19	359.65	269.7	359.04	417.47
Brd7	379	358.58	390	336	357.26	368.08	264	564.07	476
Dimt1	77	68	58	54	62	60	54	76	97.03
AK017068	0	0	0	0	0	0	0	0	0

But, what happens when there are different isoforms?



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

A probabilistic approach: Isoform deconvolution



Transcript expression method

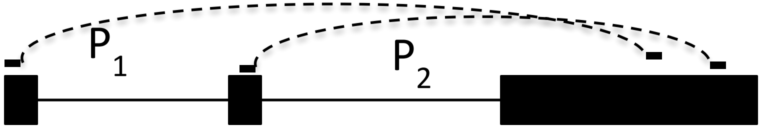
Main difference: quantification involves read assignment. Our model must capture read assignment uncertainty.

Objective: Transcript relative abundance

Unknown!: Fragment alignment source

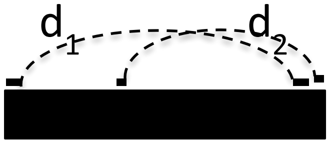
Observed variables: N fragment alignments, transcripts, *fragment length distribution*

We can estimate the insert size distribution

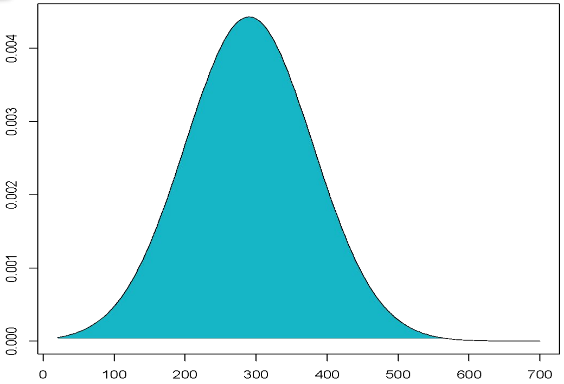


Get all single isoform reconstructions

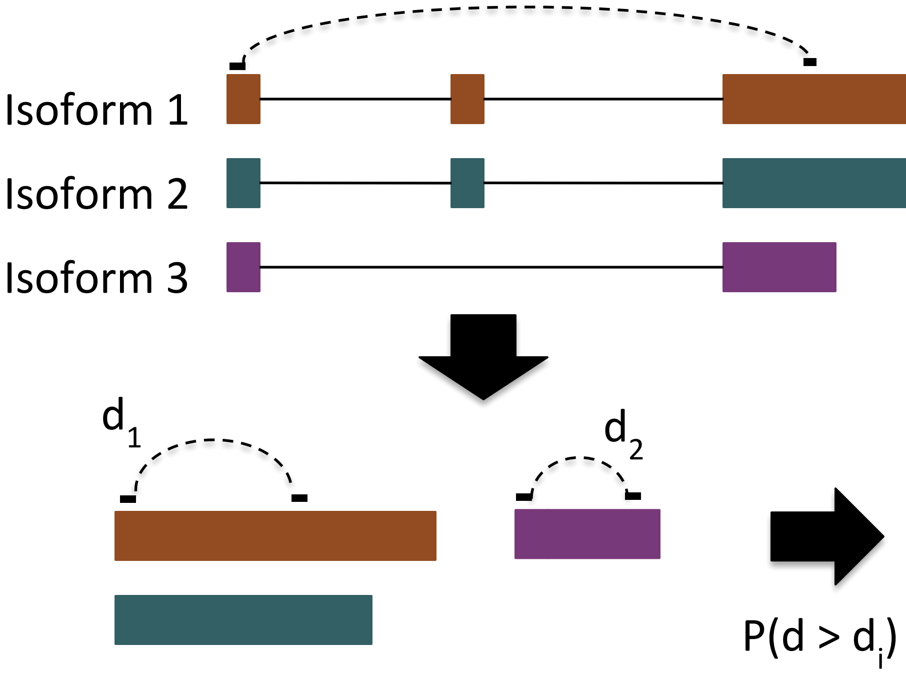
Splice and compute insert distance

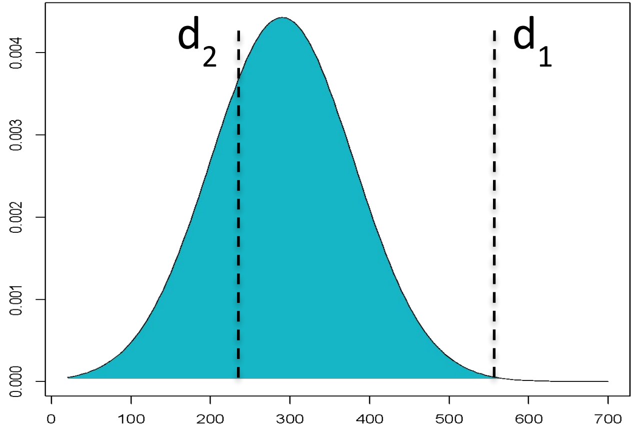


Estimate insert size empirical distribution



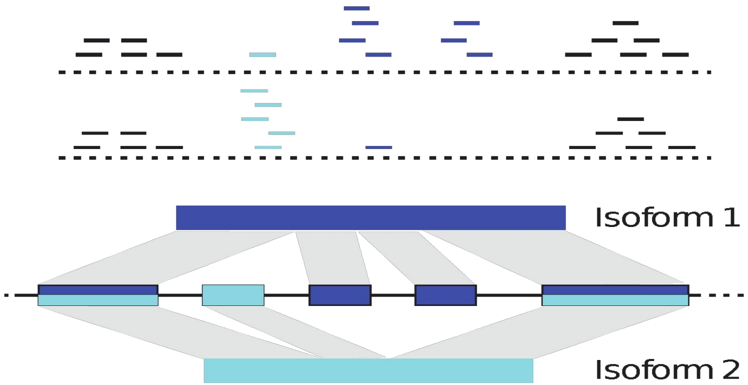
… and use it for probabilistic read assignment



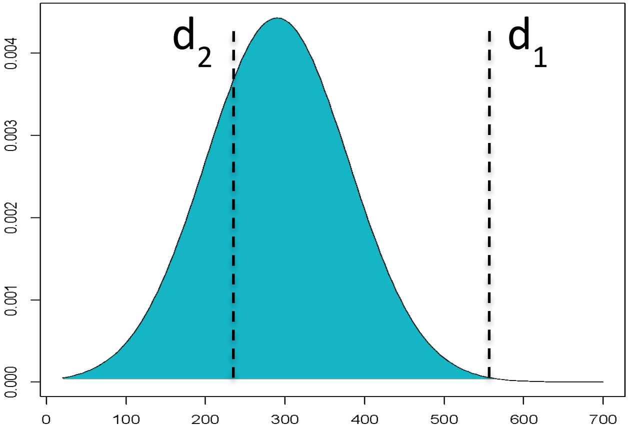


For methods such as MISO, Cufflinks and RSEM, it is critical to have paired-end data

The RNA-Seq quantification problem. Isoform deconvolution



Transcript expression method



Parameters: Transcript relative abundance

Latent variables: Fragment alignment source

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

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

Probability of the fragment alignment originating from t

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

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 mappability
  - 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

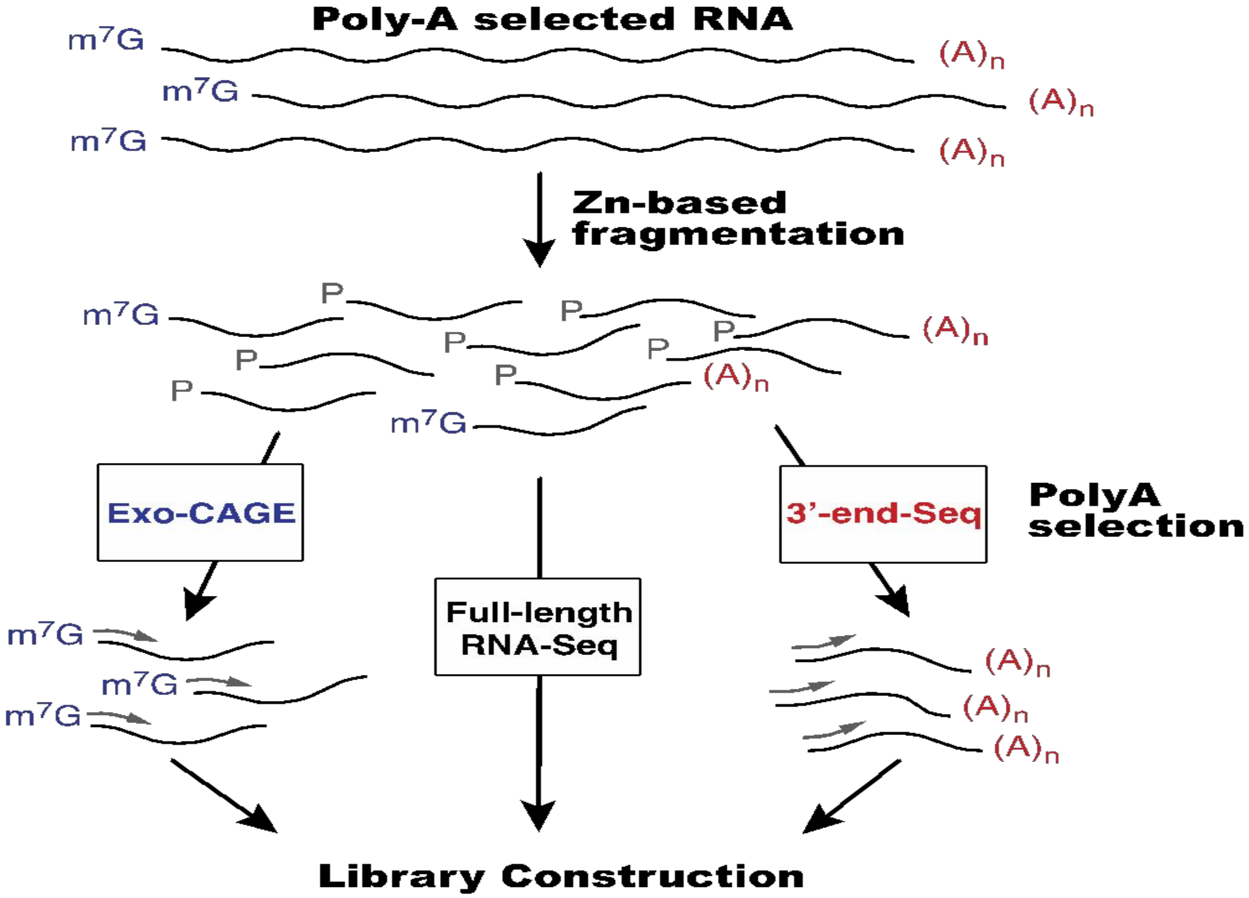
Source: End-enriched RNA

- Fragmented then selected
- Fragmented then enzymatically purified

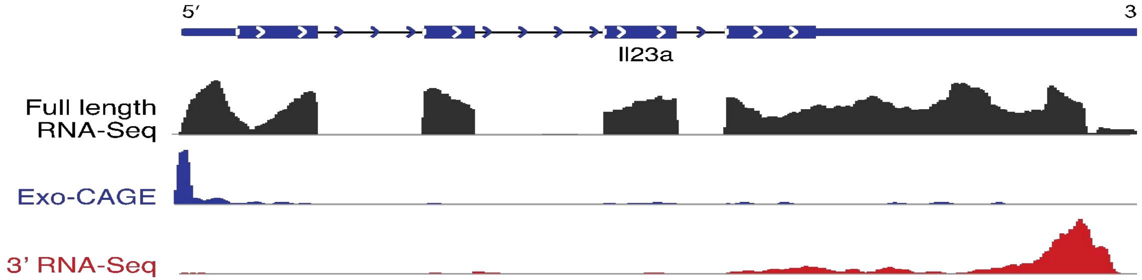
Uses:

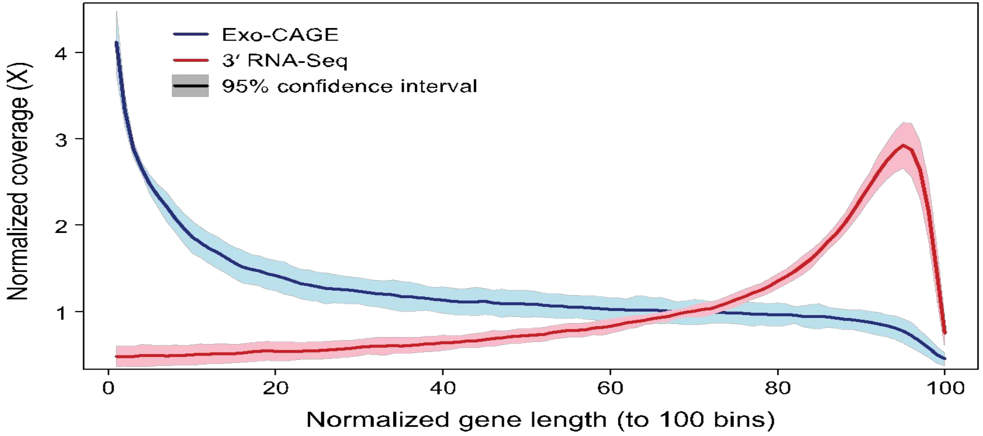
- Annotation of transcriptional start sites
- Annotation of 3’ UTRs
- Quantification and gene expression
- Depth required 3-8 million reads
- **Low quality RNA samples**
- **Single cell RNA sequencing**

RNA-Seq libraries: Summary



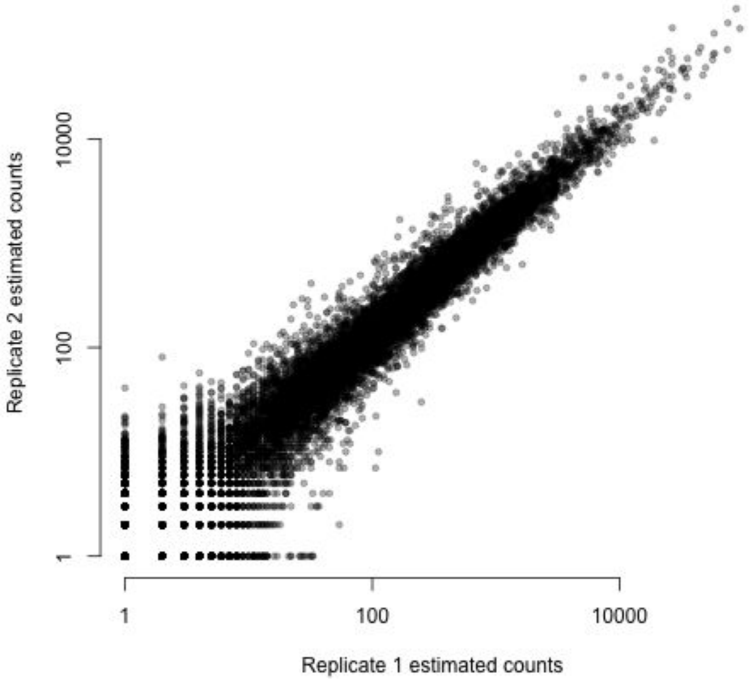
End-sequencing solution





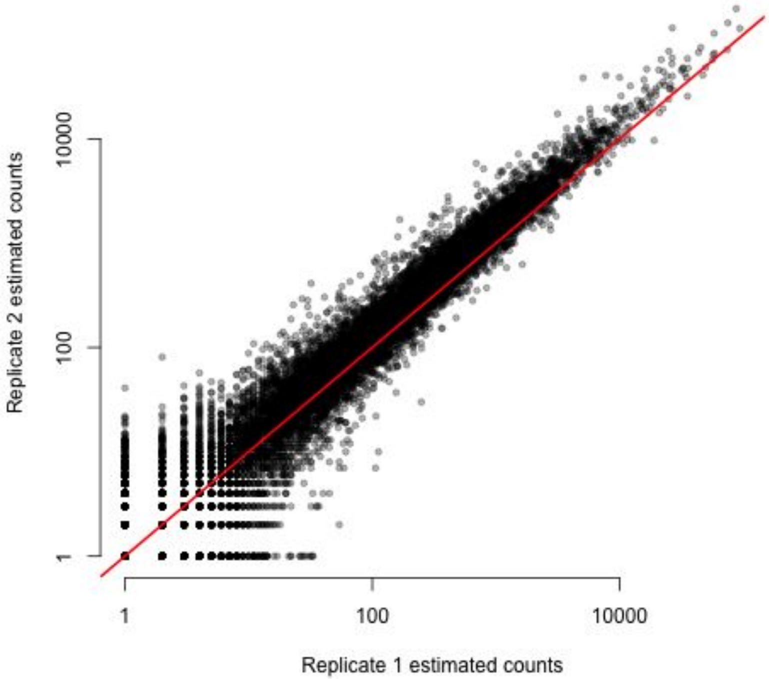
What are we normalizing?

A typical replicate scatter plot



What are we normalizing?

A typical replicate scatter plot



Let’s do an experiment

$> s1 = c(100, 200, 300, 400, 10)$

$> s2 = c(50, 100, 150, 200, 500)$

Similar read number,
one transcript many fold changed

$>\mathrm{norm}{=}\mathrm{sum}(s2)/\mathrm{sum}(s1)$

$>\mathrm{plot}(s2, s1{*}\mathrm{norm},\mathrm{log}{=}\text{”xy”})$

$>\mathrm{abline}(a = 0, b = 1)$

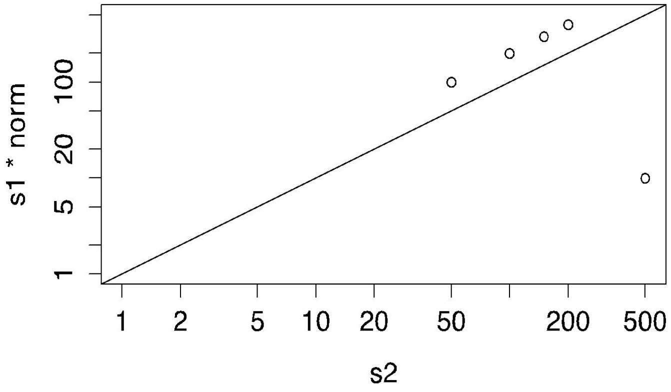
$>\mathrm{g} = \mathrm{sqrt}(s1 * s2t)$

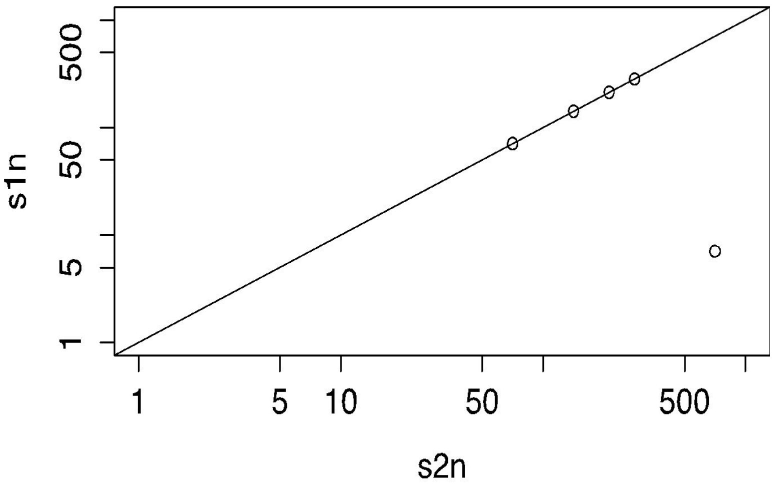
$>\mathrm{s1n} = s1/\mathrm{median}(s1/g);\ \ \mathrm{s2n} = s2/\mathrm{median}(s2/g)$

$>\mathrm{plot}(s2n, s1n,\mathrm{log}{=}\text{”xy”})$

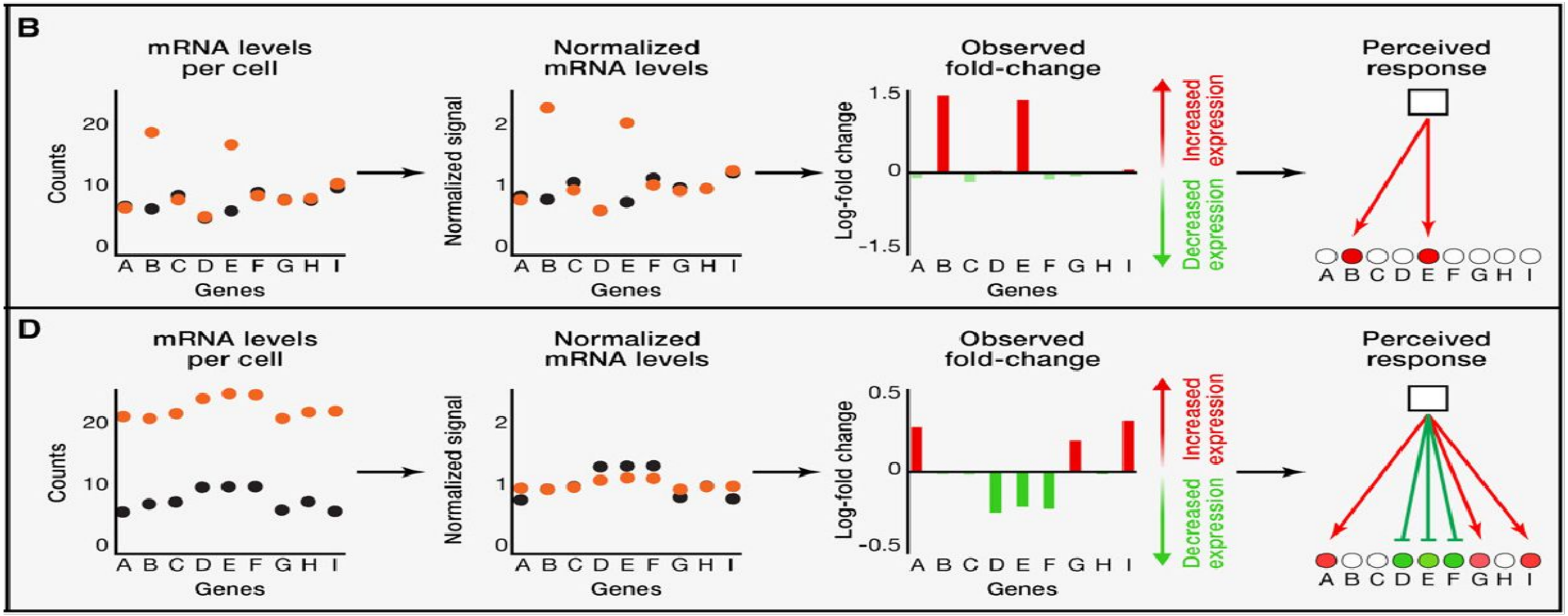
$>\mathrm{abline}(a = 0, b = 1)$

Library size normalization results
in 2-fold changes in *all* transcripts





When everything changes: Spike-ins



Lovén et al, Cell 2012

Finding DE genes

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