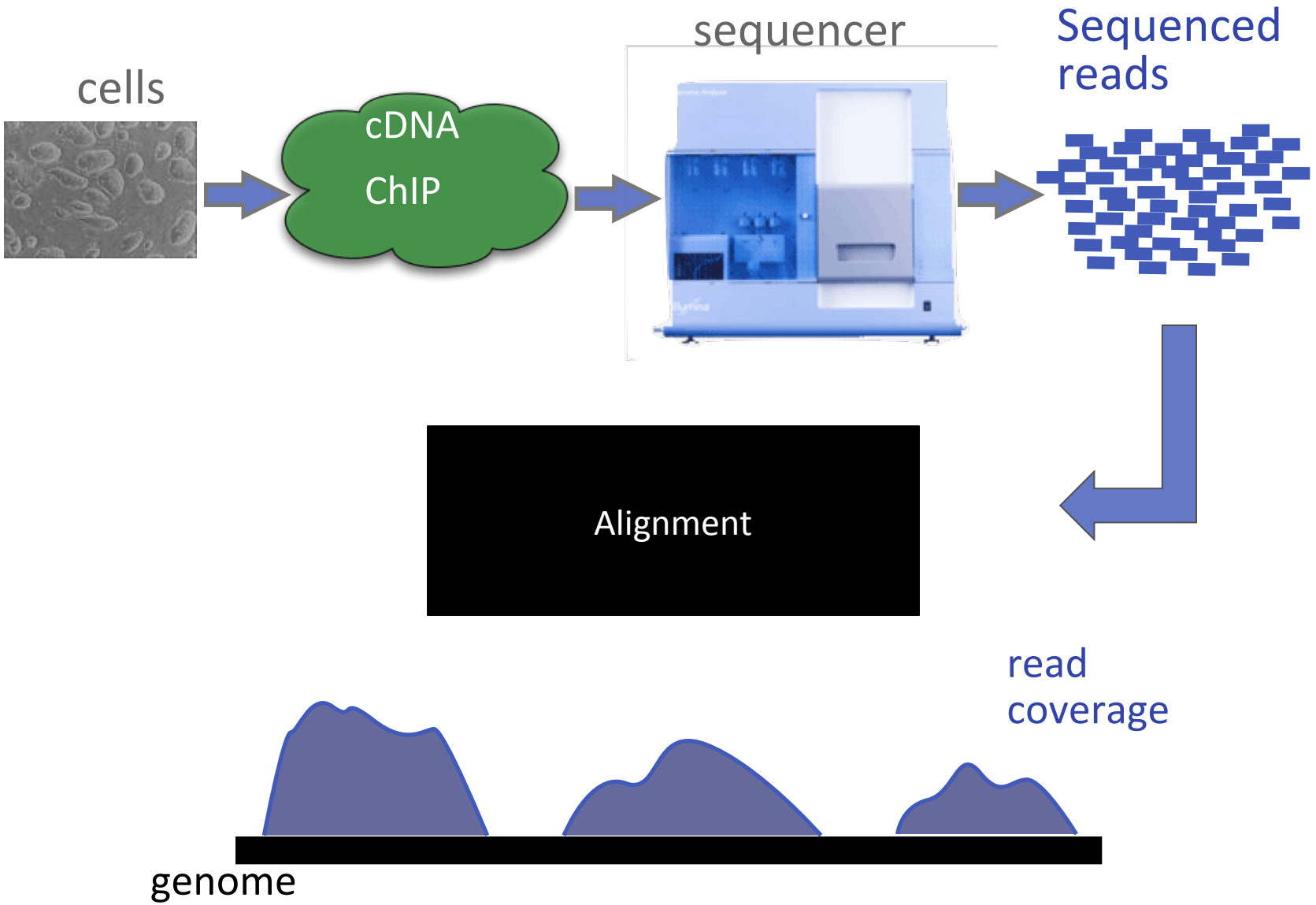
Gene expression from RNA-Seq

Once sequenced the problem becomes computational



Considerations and assumptions

- High library complexity
  - #molecules in library >> #sequenced molecules
- 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. 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)

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

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 process of RNA-Seq quantification

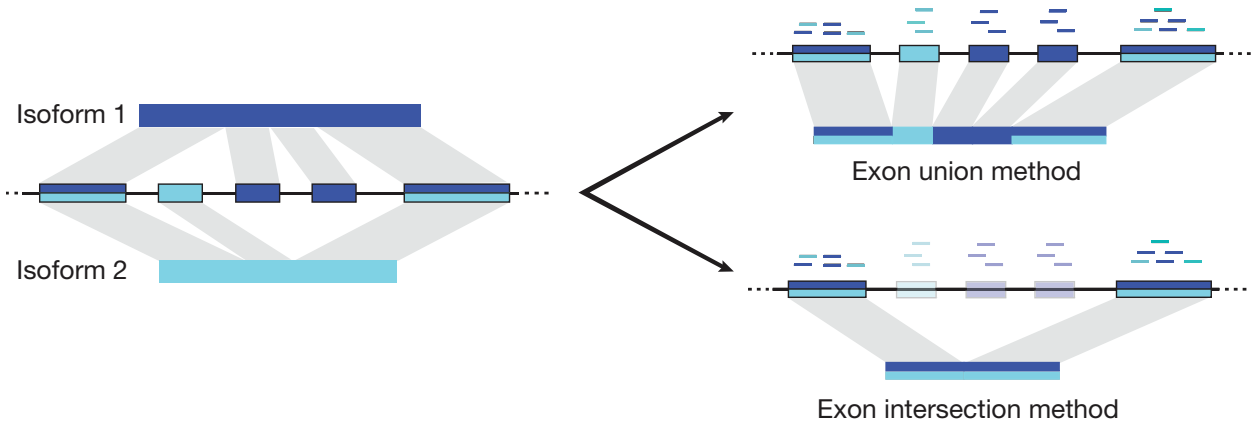
- Sequenced reads are aligned to a reference sequence
  - the species genome or
  - its transcriptome
- Transcript abundance is measured:
  - By counting reads mapped to each transcript (not accurate when multiple isoforms share sequence)
  - By solving a maximizing the likelihood of the observed mapping given transcript abundance
- To compare samples 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 vaue
- 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
  - Row j gives the expression of genes in sample j.

gene	$L^{\Delta 1,2}$.rep1	$L^{\Delta 1,2}$.rep2	$L^{\Delta 1,2}$.rep3	$L^{\Delta 1}$.rep1	$L^{\Delta 1}$.rep2	$L^{\Delta 1}$.rep3	$L^{\Delta 2}$.rep1	$L^{\Delta 2}$.rep2	$L^{\Delta 2}$.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, how are these quantities computed?



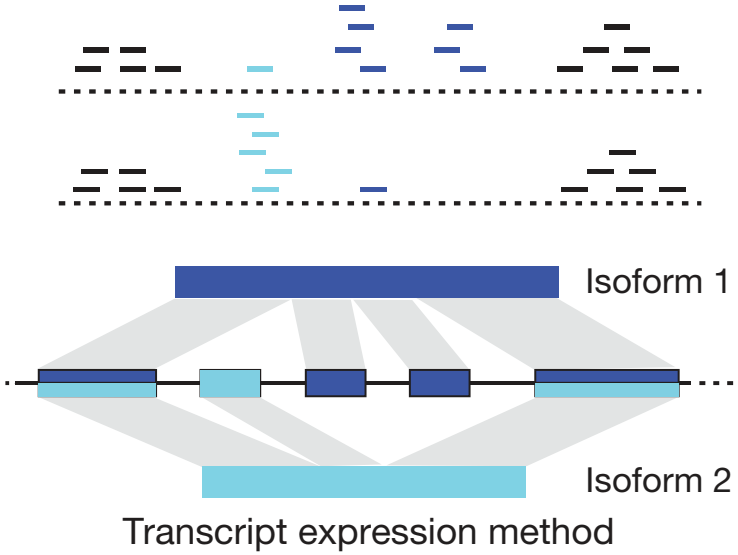
- Start with a set of previous gene/transcript annotations
- ~~Assume~~ Define only one isoform per gene
- ~~Assume 1-1 read to transcript correspondence.~~ Reads (fragments) are now short, one transcript generates many fragments.

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: $\theta_g = \frac{n_g}{\tilde{l}_g N}$

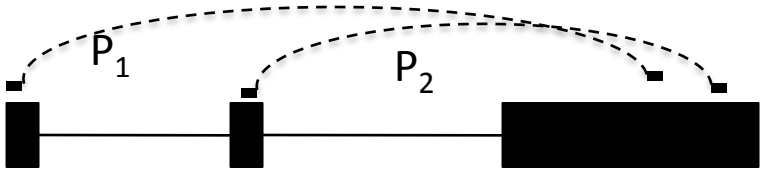
The RNA-Seq quantification problem. Isoform deconvolution



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

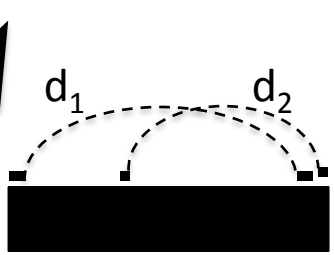
Parameters: Transcript relative abundance
Latent variables: Fragment alignment source
Observed variables: N fragment alignments, transcripts, *fragment length distribution*

We can estimate the insert size distribution

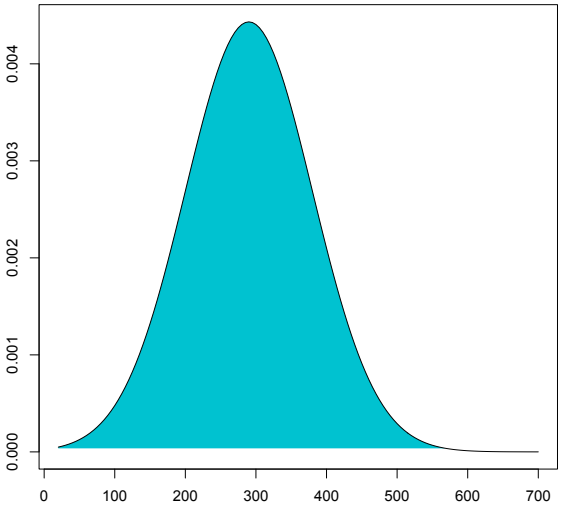


Get all single isoform reconstructions

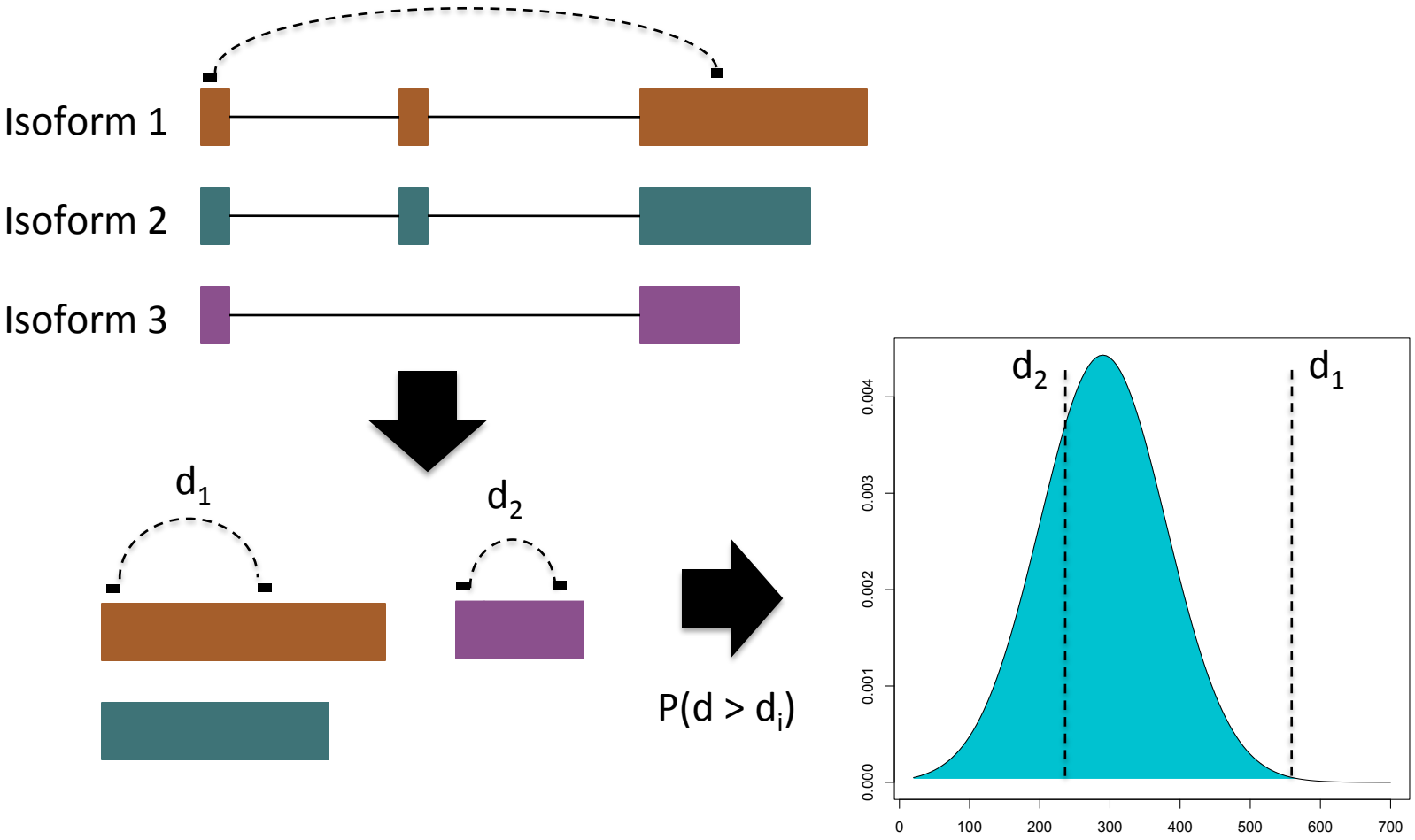
Splice and compute insert distance



Estimate insert size empirical distribution

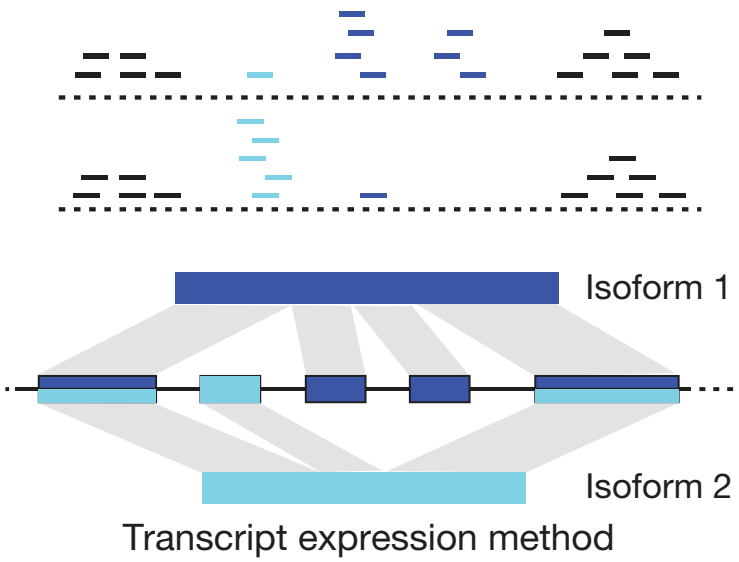


… and use it for probabilistic read assignment

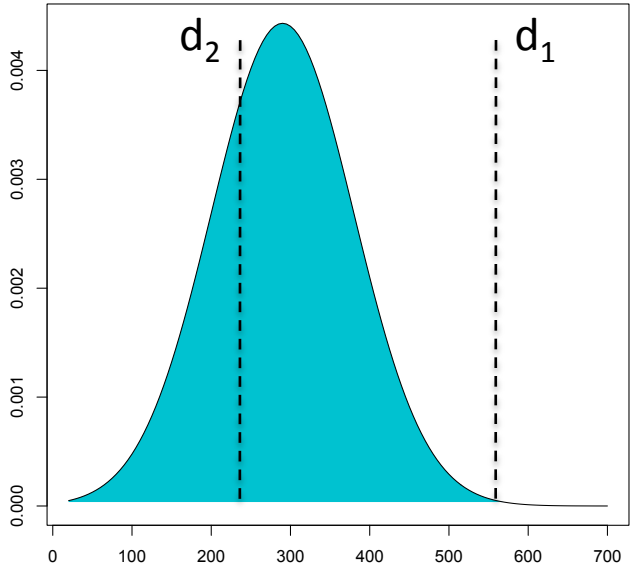


For methods such as MISO, Cufflinks and RSEM, it is critical to have paired-end data

The RNA-Seq quantification problem. Isoform deconvolution



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

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

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

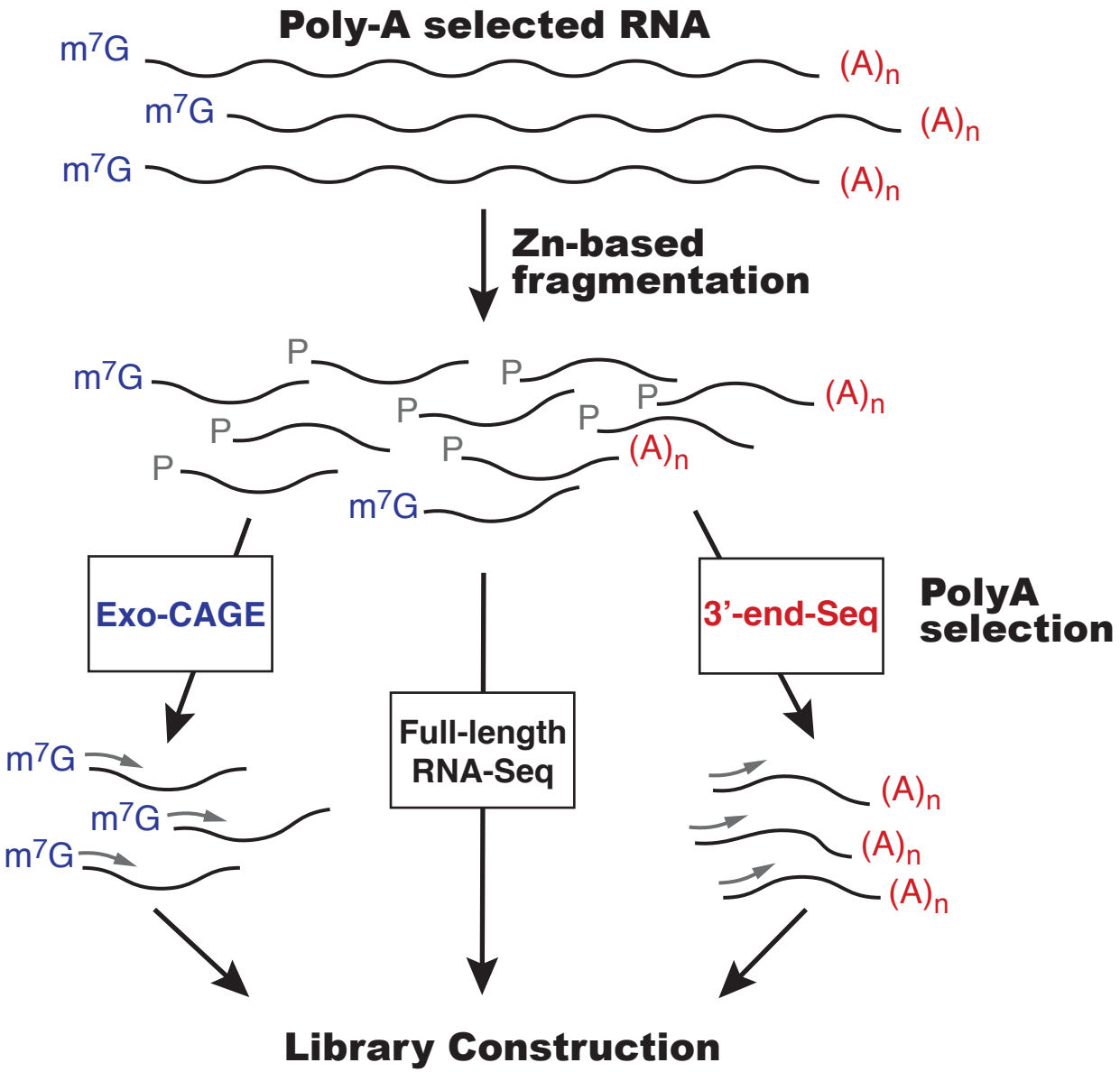
Summary: Current quantification models are complex

- 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*

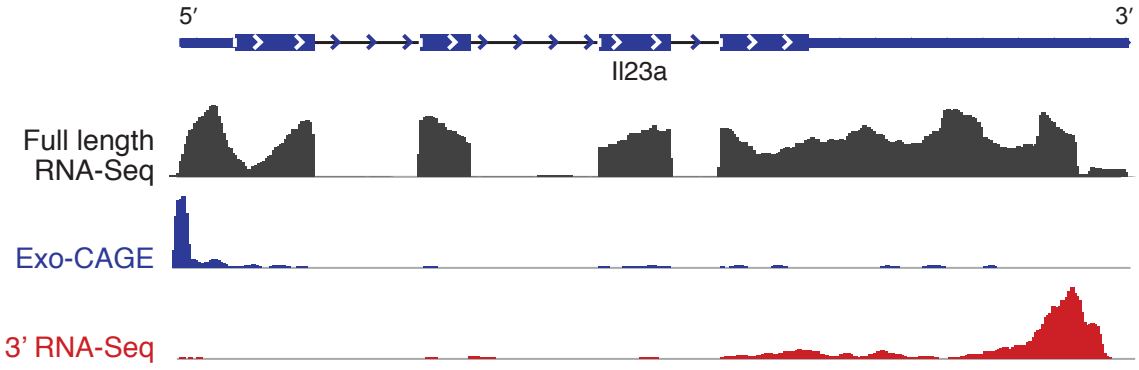
RNA-Seq libraries revisited: End-sequence libraries

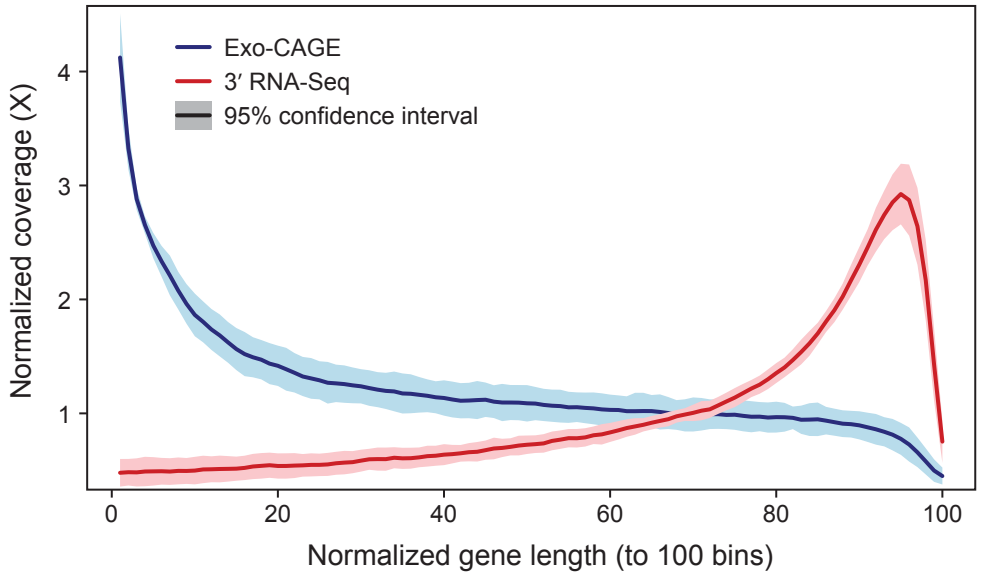
- Target the start or end of transcripts.
- Source: End-enriched RNA
  - Fragmented then selected
  - Fragmented then enzymatically purified
- Uses:
  - Annotation of transcriptional start sites
  - Annotation of 3’ UTRs
  - Quantification and gene expression
  - Depth required 3-8 mill reads
  - **Low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

RNA-Seq libraries: Summary



End-sequencing solution



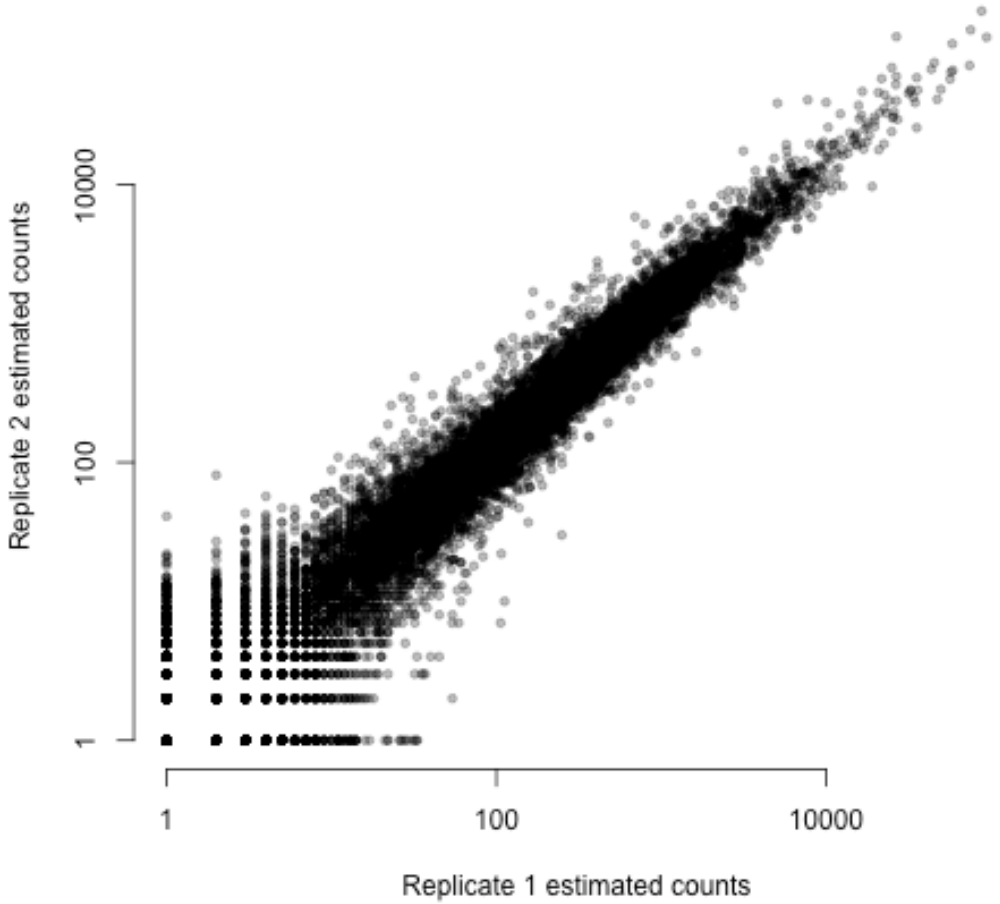


Analysis of counting data requires 3 broad tasks

- 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

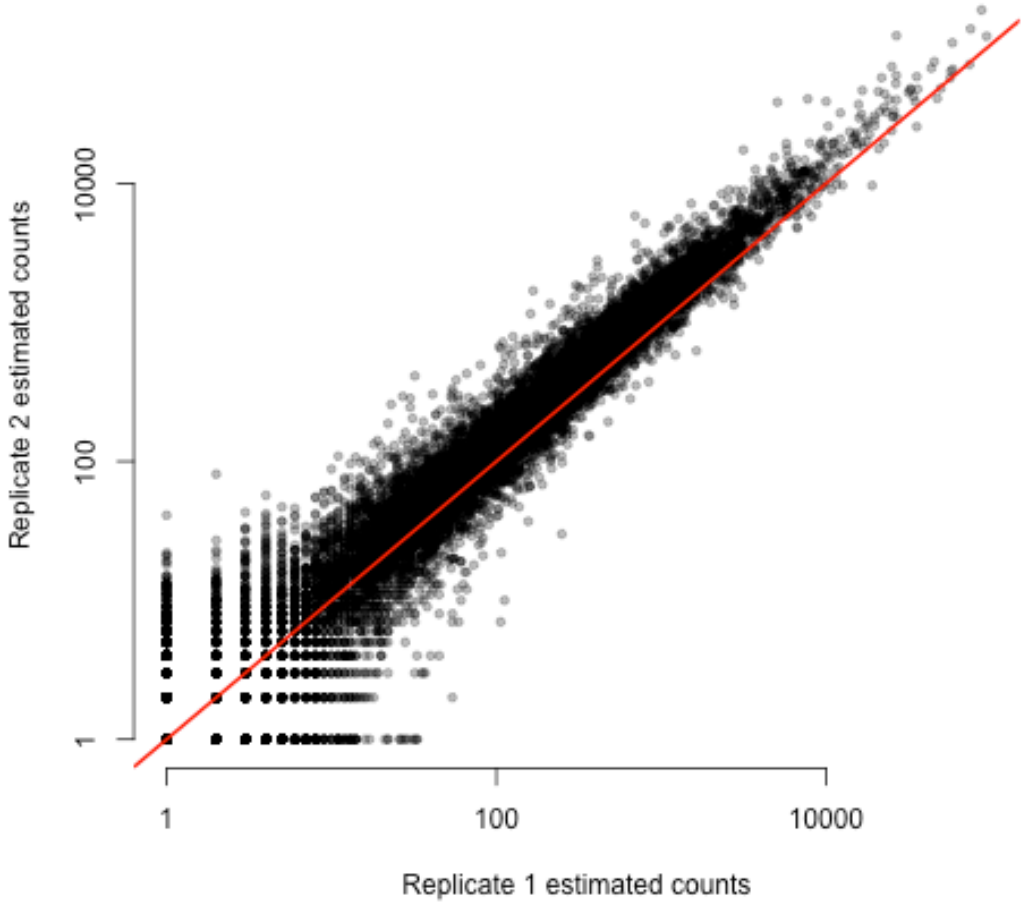
What are we normalizing?

A typical replicate scatter plot



What are we normalizing?

A typical replicate scatter plot



TPM normalization

- Accounts for:
  - Differences in sequencing depth
  - Differences in the number of reads generated by transcripts of different length

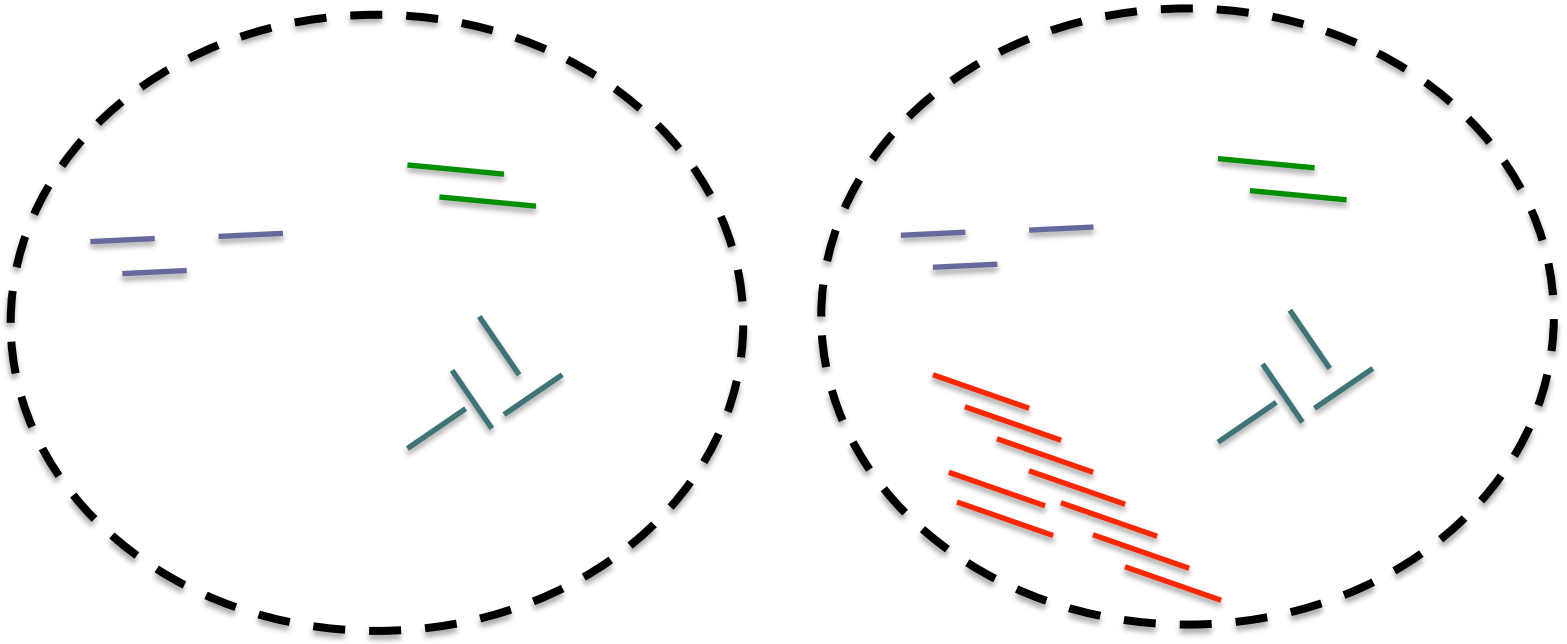
$$TPM = \frac{n_g}{N \cdot l_g}(10^3)(10^6)$$

n_g Estimated reads/fragments for the gene

N Total reads/fragments

l_g Length of the transcript

Sample composition impacts transcript ***relative*** abundance

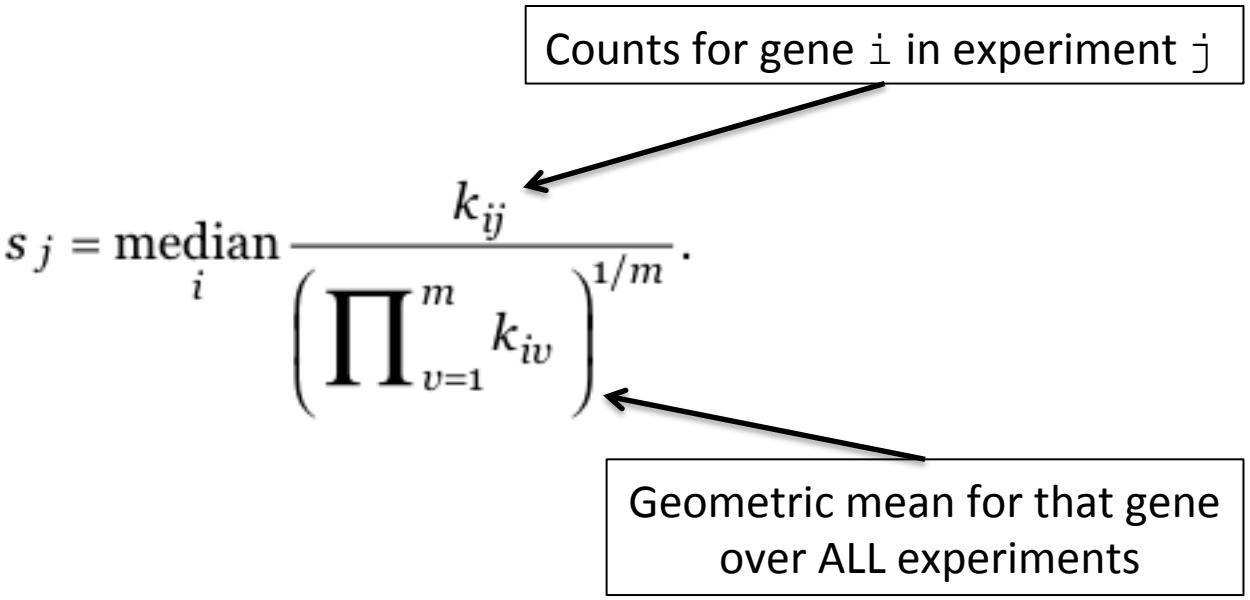


Cell type I

Cell type II

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

Example normalization techniques



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

Alders and Huber, 2010

Lets do an experiment

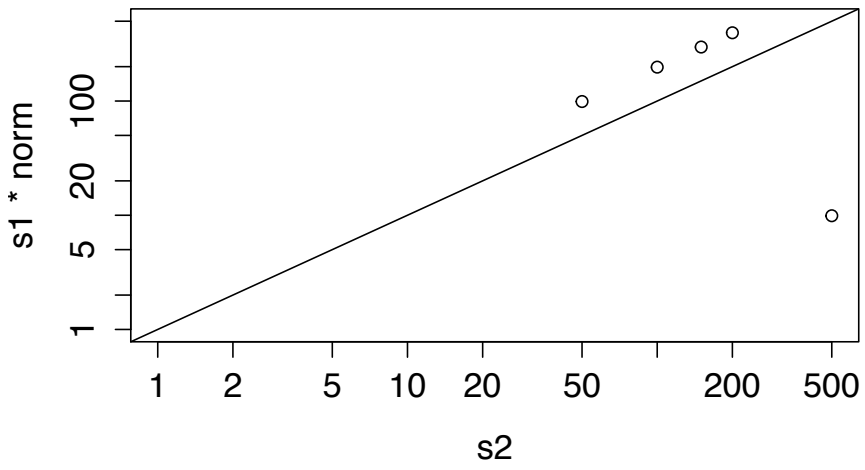
$> s1 = c(100, 200, 300, 400, 10)$
$> s2 = c(50, 100, 150, 200, 500)$

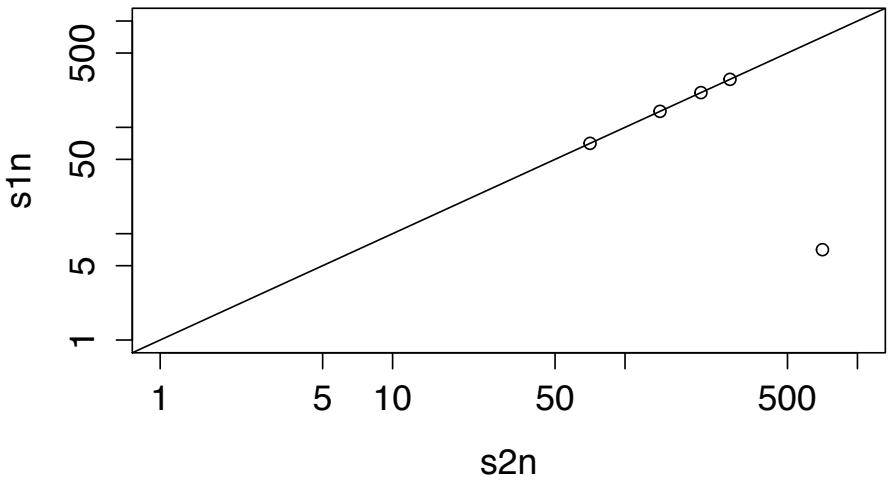
Similar read number,
one transcript many fold changed

$>\text{norm}=\text{sum}(s2)/\text{sum}(s1)$
$>\text{plot}(s2, s1*\text{norm},\text{log}=\text{"xy"})$
$>\text{abline}(a = 0, b = 1)$

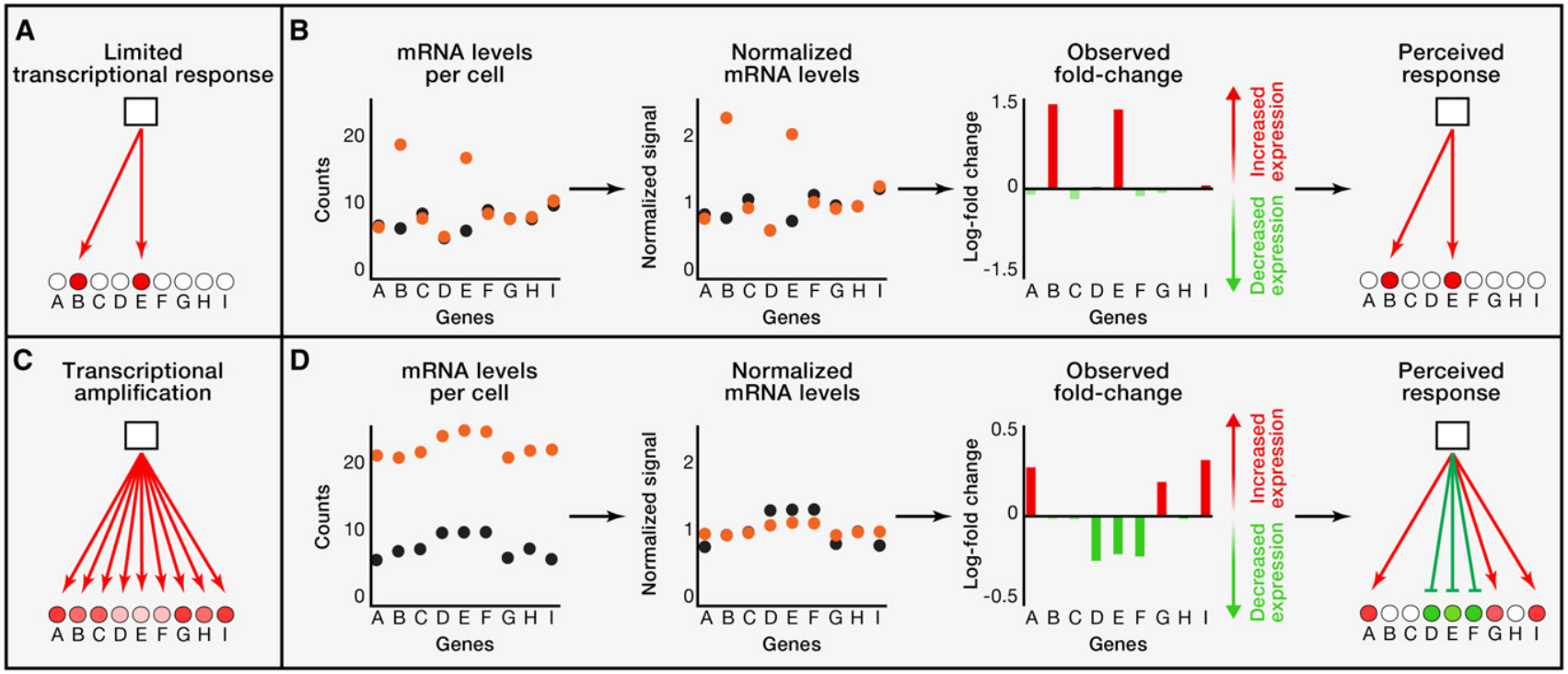
Size normalization results in 2-fold changes in *all* transcripts

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





When everything changes: Spike-ins



Lovén et al, Cell 2012