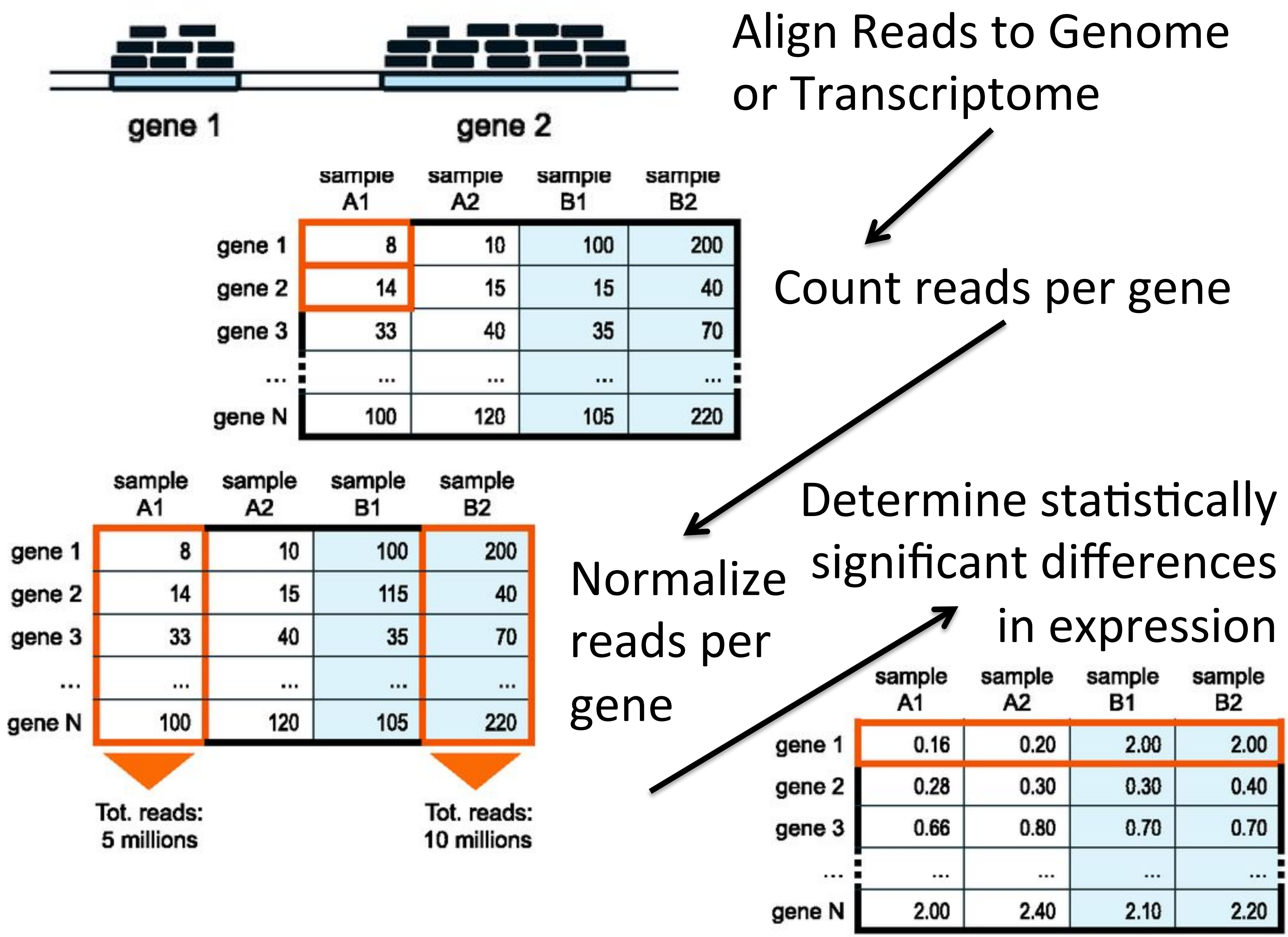
RNA-Seq Analysis

- Any questions from last class about RNA-Seq libraries?



Align Reads to Genome or Transcriptome

Count reads per gene

	sample A1	sample A2	sample B1	sample B2
gene 1	8	10	100	200
gene 2	14	15	15	40
gene 3	33	40	35	70
...	...	...	...	...
gene N	100	120	105	220

Normalize reads per gene

	sample A1	sample A2	sample B1	sample B2
gene 1	8	10	100	200
gene 2	14	15	115	40
gene 3	33	40	35	70
...	...	...	...	...
gene N	100	120	105	220

Tot. reads:
5 millions

Tot. reads:
10 millions

Determine statistically significant differences in expression

	sample A1	sample A2	sample B1	sample B2
gene 1	0.16	0.20	2.00	2.00
gene 2	0.28	0.30	0.30	0.40
gene 3	0.66	0.80	0.70	0.70
...	...	...	...	...
gene N	2.00	2.40	2.10	2.20

Francesca Finotello, and Barbara Di Camillo Briefings in Functional Genomics 2014;bfgp.elu035

Read Alignment

- Let’s try it

- What difficulties did you have with the alignment?
- What would you expect to find in a human genome that we mostly don't see here?
- Were there mistakes in these data? What were they?
  - How would you determine if they are features or mistakes?
- What did you do with "reads" that matched two places?
- Other thoughts on this alignment?

Were there mistakes in these data? What were they?

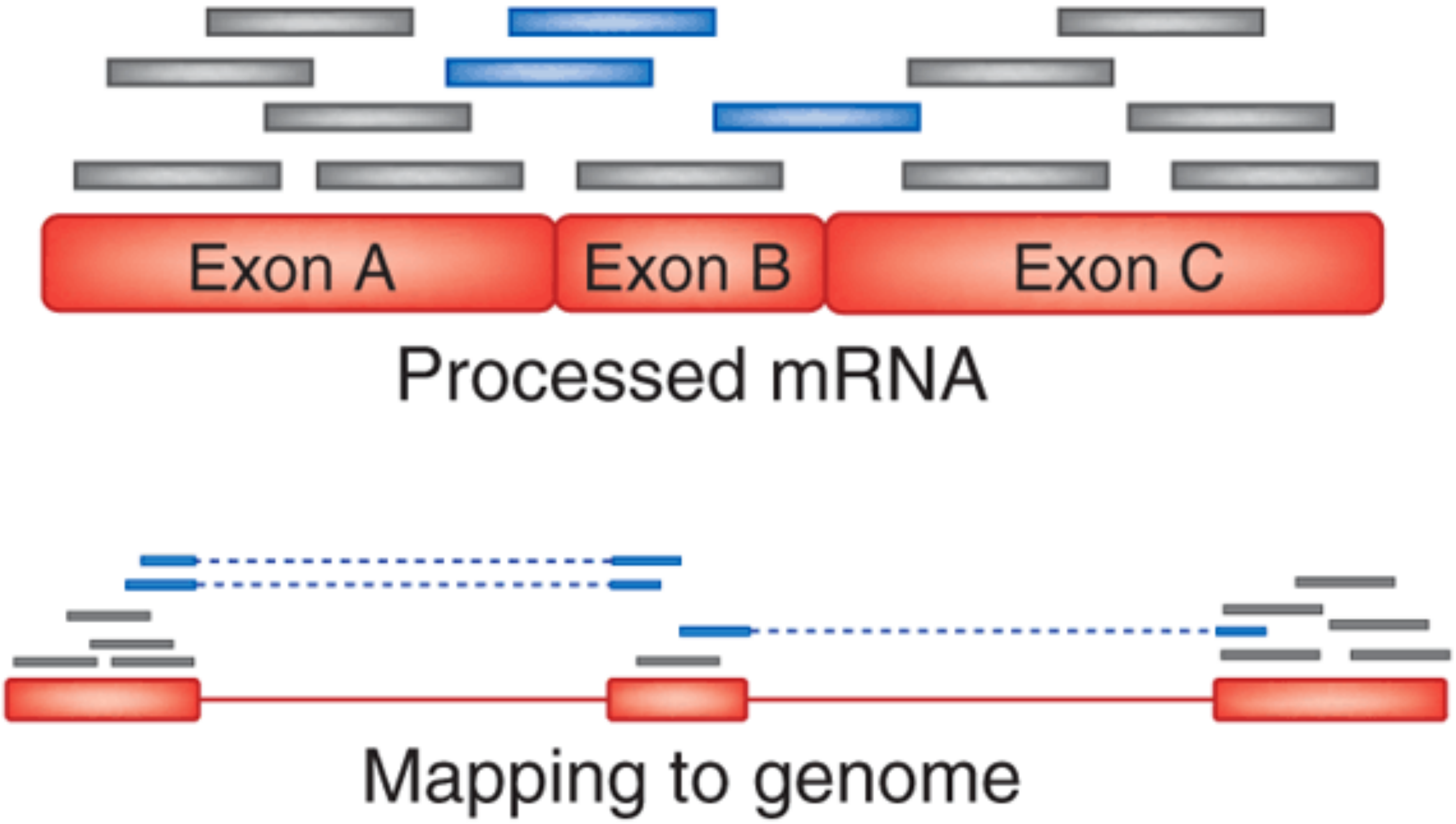
How would you determine if they are features or mistakes?

Biological Replicates!

Use of biological replicates helps identify artifacts in a particular RNA-seq experiment from something unusual going on in this sample.

- Run the experiment on a a separate sample and take data
  - 3 flasks of yeast, 3 dishes of cells, 3 different plants, etc.
- Always use at least 3x biological replicates for an RNA-seq study

Aligning Reads



Grey reads are easy to match (we’ll talk about how in a minute), Blue reads are hard

Cole Trapnell & Steven L Salzberg; *Nature Biotechnology* **27**, 455 - 457 (2009) doi:10.1038/nbt0509-455

Multi-aligning reads

1. Throw out
   - Leads to loss of data
2. Distribute evenly between matching sites
   - Biases data based on distribution
3. Distribute based on coverage of nearby sites
   - Requires keeping these in memory and calculating coverage of each site before finalizing alignment (computationally expensive)

Aligning Reads

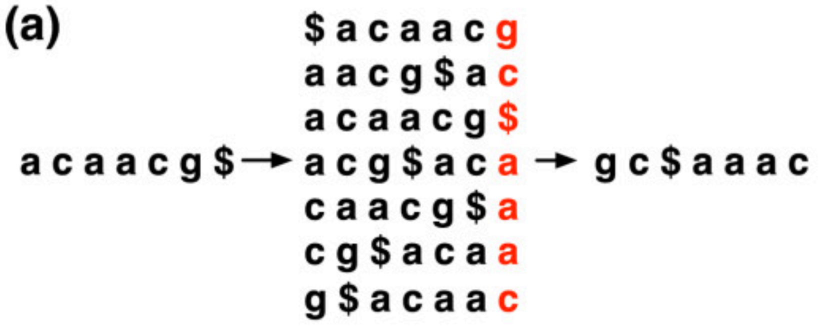
- Ways people align:
  - Tuxedo Suite (Bowtie and TopHat are the aligners)
  - Sailfish/Salmon

Bowtie/TopHat

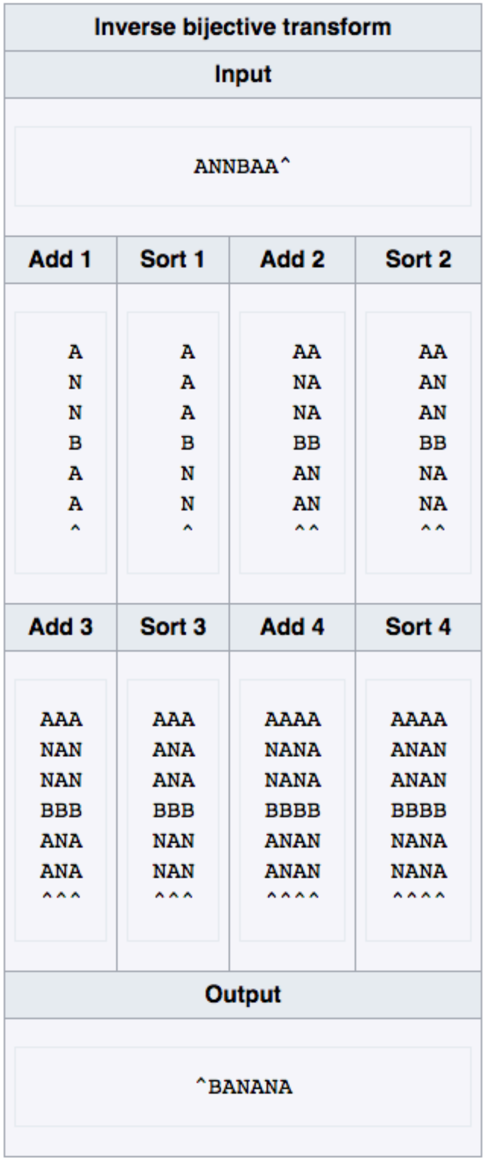
- Searching for matches/alignments though all 3 billion bases is computationally intractable
  - Too much memory used (2 bits*3 billion – 6GB)
  - Too many computation steps, so takes a LONG time
- Use a series of compression and sorting algorithms that allow for less memory needed and faster search
  - Burrows-Wheeler transform and FM-index

BW transform

- The genome sequence is broken into short strings
- These strings are permuted to a form that allows them to be sorted alphabetically, but retains the ability to backtrack to sequence

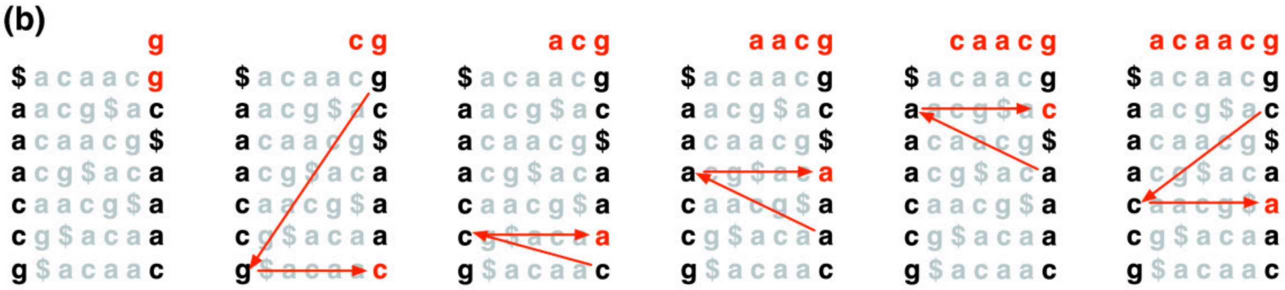


Langmead, B., et al; *Genome Biology* 200910:R25 DOI: 10.1186/gb-2009-10-3-r25



Inverse bijective transform			
Input			
ANNBAA^			
Add 1	Sort 1	Add 2	Sort 2
A N N B A A ^	A A A B N N ^	AA NA NA BB AN AN ^^	AA AN AN BB NA NA ^^
Add 3	Sort 3	Add 4	Sort 4
AAA NAN NAN BBB ANA ANA ^^^	AAA ANA ANA BBB NAN NAN ^^^	AAAA NANA NANA BBBB ANAN ANAN ^^^^	AAAA ANAN ANAN BBBB NANA NANA ^^^^
Output			
^BANANA			

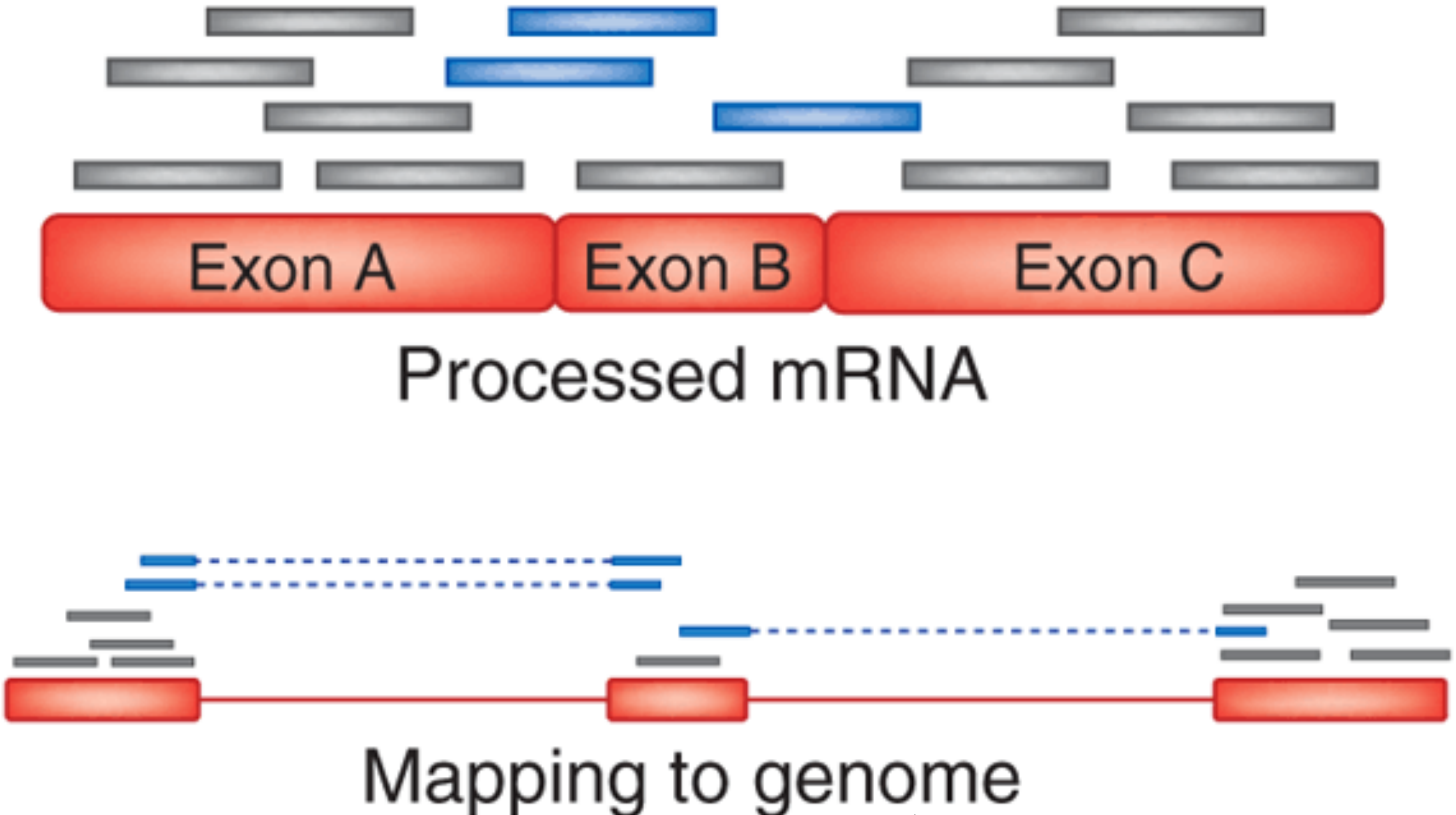
https://en.wikipedia.org/wiki/Burrows%E2%80%93Wheeler_transform



Langmead, B., et al; *Genome Biology* 200910:R25 DOI: 10.1186/gb-2009-10-3-r25

Using a similar back permutation, we can reconstruct original sequence

As these BW transforms of sequences are sorted alphabetically, rows that begin with the same sequence are ordered together, so a search can start at the section that matches the beginning making it faster



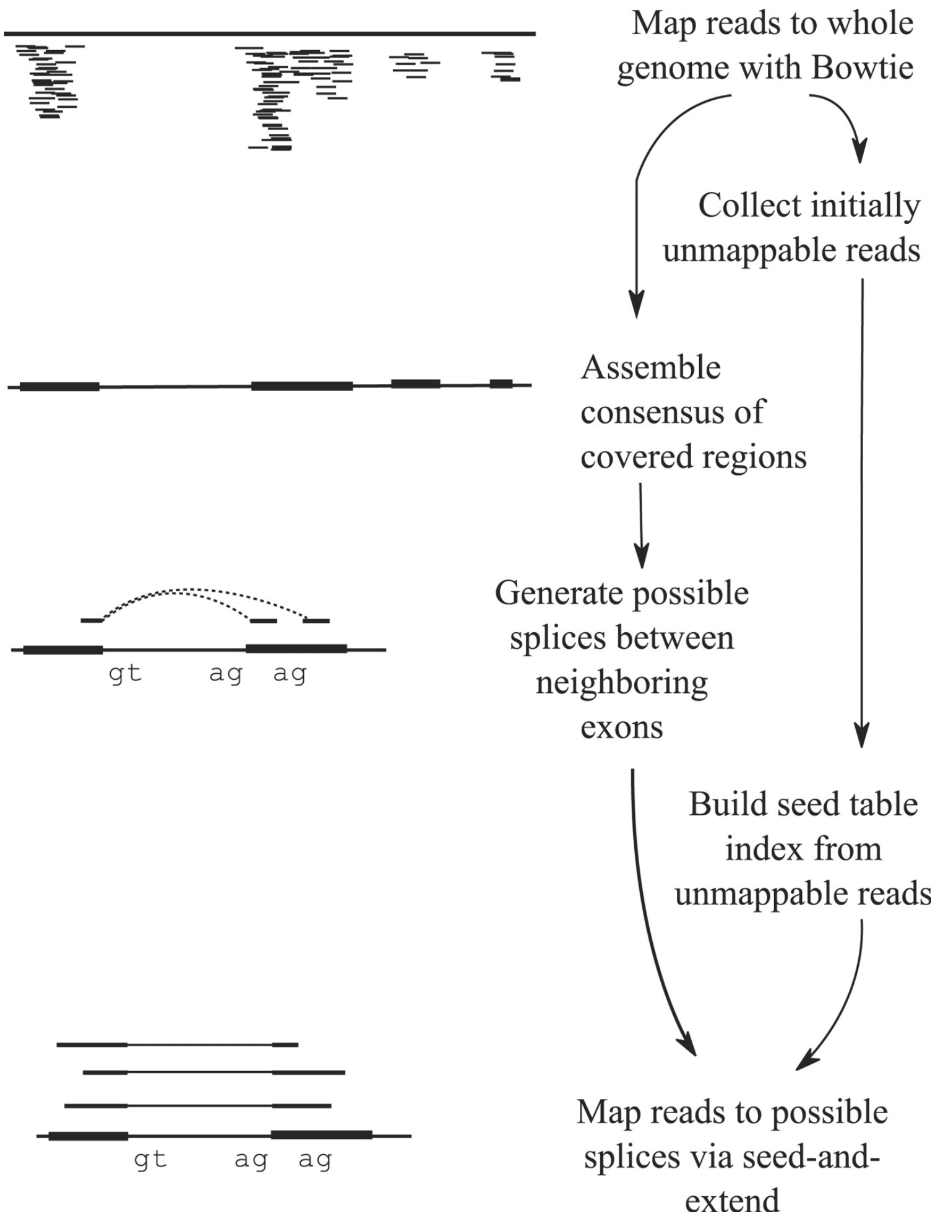
Cole Trapnell & Steven L Salzberg; *Nature Biotechnology* **27**, 455 - 457 (2009) doi:10.1038/nbt0509-455

Bowtie fails on Blue ones (with genome as reference)

- Aligns those that it can
- Pass those it can’t off to TopHat

How would you design an algorithm to map exon-exon splitting reads?

TopHat



Map reads to whole genome with Bowtie

Collect initially unmappable reads

Assemble consensus of covered regions

Generate possible splices between neighboring exons

Build seed table index from unmappable reads

Map reads to possible splices via seed-and-extend

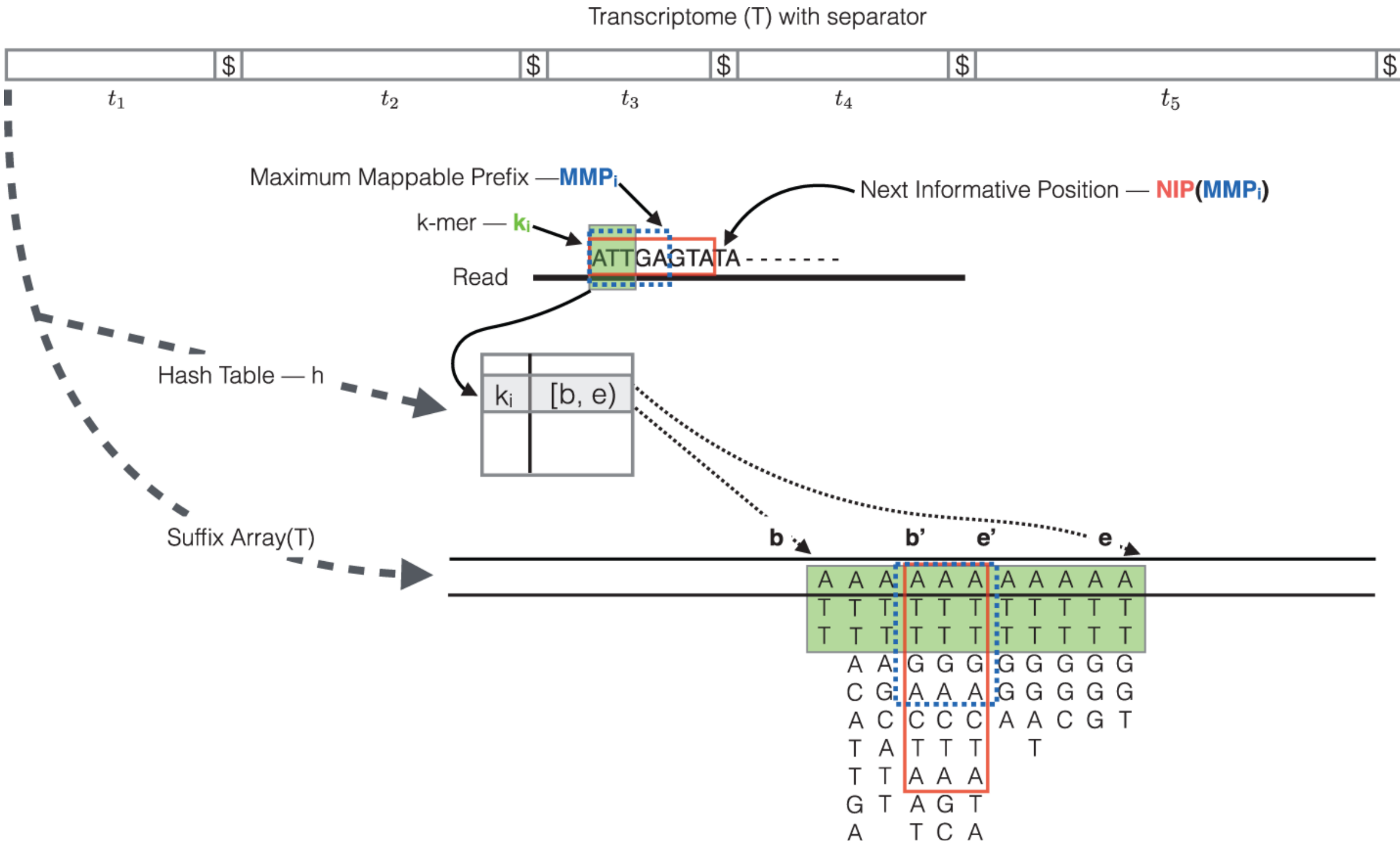
- Uses the coverage of the grey reads to identify exons that are present
- Predicts possible splicing events based on these exons and all potential splice donors and acceptors in those exons
- Searches in the un-mapped reads for matches to these potential splice junctions, and grown out any matches
- Allows for discovery of new splice sites

Trapnell, C. et al; *Bioinformatics*. 2009;**25(**9):1105-1111. doi:10.1093/bioinformatics/btp120

Sailfish/Salmon Mapping

- As more transcriptomes are available, or can be assembled *de novo* from RNA-Seq data, move away from aligning to genome to mapping to transcriptome
  - Quicker (by hours)
  - Doesn't give full alignment, but identifies which transcript read comes from for quantification

Quasi-Mapping

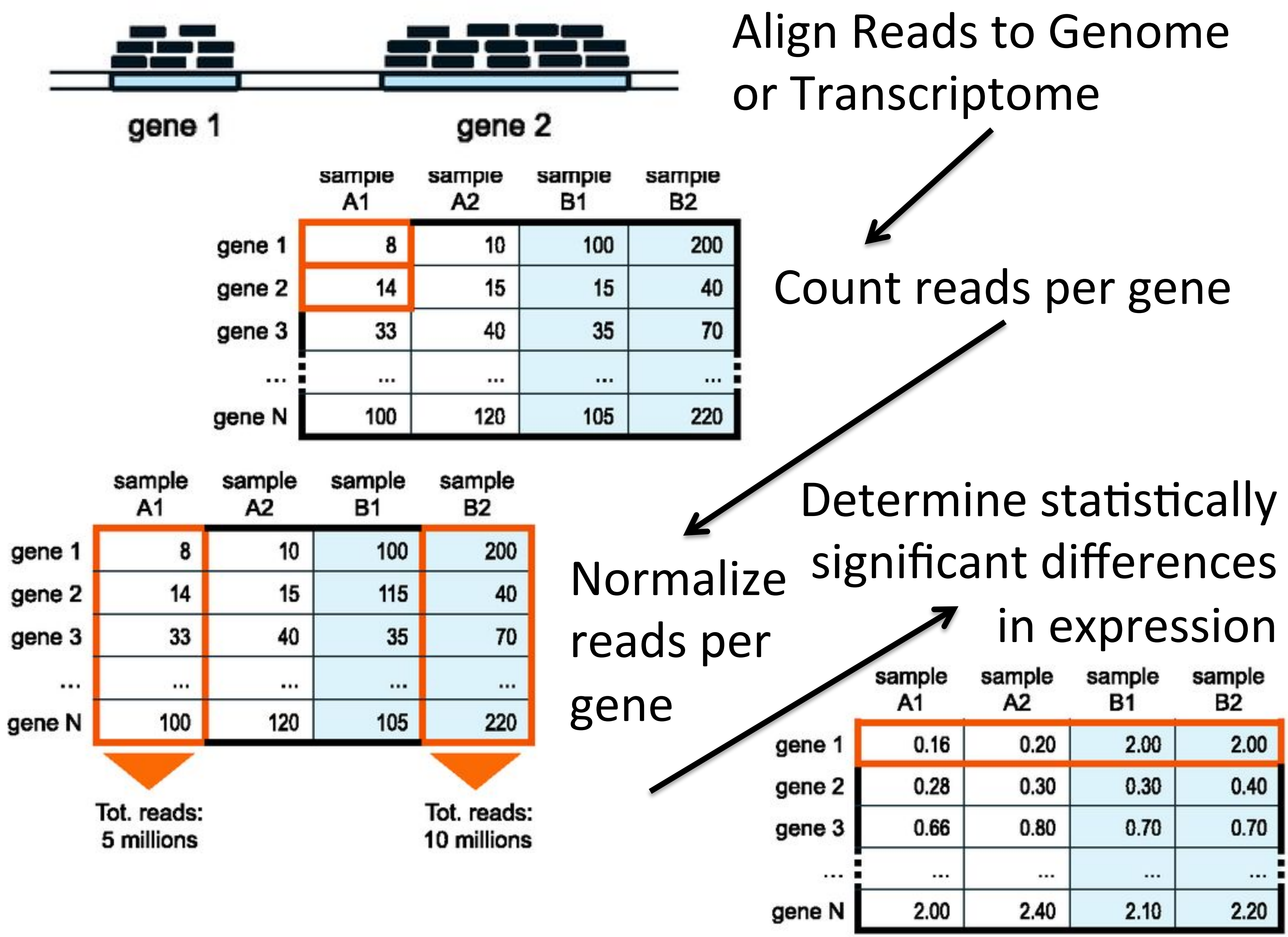


Srivastava, A. et al; *Bioinformatics*. 2016;**32**(12):i192-i200. doi:10.1093/bioinformatics/btw277

Faster and more accurate quantification than aligners

More memory needed

Requires indexing of transcriptome (NOT genome)



Align Reads to Genome or Transcriptome

Count reads per gene

	sample A1	sample A2	sample B1	sample B2
gene 1	8	10	100	200
gene 2	14	15	15	40
gene 3	33	40	35	70
...	...	...	...	...
gene N	100	120	105	220

Normalize reads per gene

	sample A1	sample A2	sample B1	sample B2
gene 1	8	10	100	200
gene 2	14	15	115	40
gene 3	33	40	35	70
...	...	...	...	...
gene N	100	120	105	220

Determine statistically significant differences in expression

	sample A1	sample A2	sample B1	sample B2
gene 1	0.16	0.20	2.00	2.00
gene 2	0.28	0.30	0.30	0.40
gene 3	0.66	0.80	0.70	0.70
...	...	...	...	...
gene N	2.00	2.40	2.10	2.20

Francesca Finotello, and Barbara Di Camillo Briefings in Functional Genomics 2014;bfgp.elu035

Determine statistical differences

What properties of the transcripts in a cell will determine the probability of finding a read from it in our RNA-seq library?

Determine statistical differences

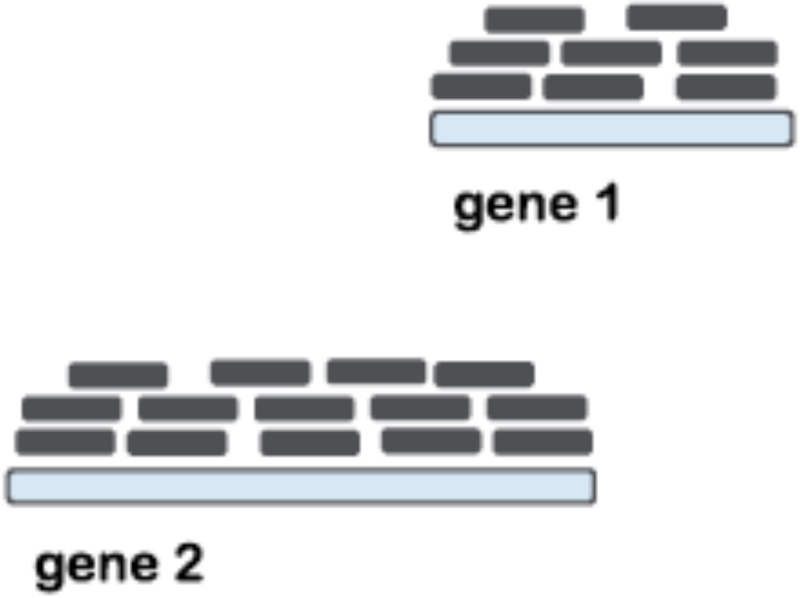
What properties of the transcripts in a cell will determine the probability of finding a read from it in our RNA-seq library?

How highly expressed

How long it is

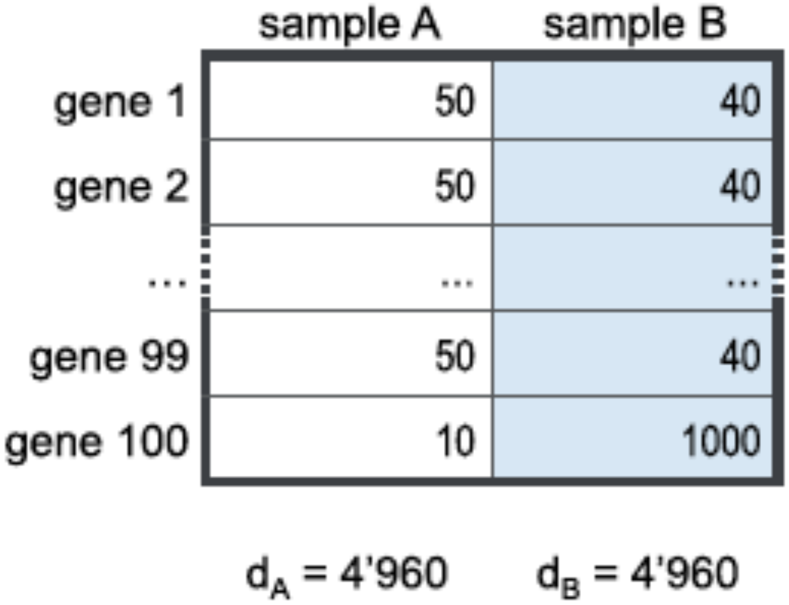
Normalizations

Length of a gene



‘Reads Per Kilobase of exon model per Million mapped reads’ (RPKM)

Number of reads



	sample A	sample B
gene 1	50	40
gene 2	50	40
...	...	...
gene 99	50	40
gene 100	10	1000

$d_A = 4'960$ $d_B = 4'960$

Different ways to account for this:

- Total counts
- 3rd quantile of reads
- Bottom 70% of reads

Francesca Finotello, and Barbara Di Camillo Briefings in Functional Genomics 2014;bfgp.elu035

f is the number of transcripts in the sample. It is made up of transcripts 1,…,F.

Each transcript in *f* has a length of l_f and an expression level of θ_{fj}.

If the library doesn't have bias, a read can start at any of the bases in the transcript, so each transcript has the number of possible reads of $\theta_{fj} l_f$.

The probability of a read coming from a specific transcript is the number of possible read starts in a given transcript divided by all possible read starts in the entire population of transcripts:

$$\pi_{fj} = \frac{\theta_{fj} l_f}{\sum_{f=1}^{F} \theta_{fj} l_f}$$

Assuming the reads are randomly sampling all possible start sites in all transcripts, we can model the probability of a read coming from transcript *f* as a Bernoulli process where the read either does from from *f* (P) or does not (1-P).

The Bernoulli process is defined as:

$$Y = \sum_{k-1}^{n} X_k \sim B(n, p)$$

The number of reads from *f* (N_{fj}) can be defined as a Bernoulli process within the population of total reads (R_j)

$$N_{fj} \sim B(R_j, \pi_{fj})$$

As R_j is much MUCH MUCH larger than π_{fj}, this can be modeled as a Poisson distribution with parameter λ_{fj}.

$$\lambda_{fj} = R_j \cdot \pi_{fj}$$

$$N_{fj} \sim P(\lambda_{fj})$$

Technical replicates have shown that this Poisson distribution holds for RNA-Seq data

HOWEVER, as the distribution of transcripts is different for different cells (*j*), the Poisson parameter is different for each sample - we are reading error AND biological difference

Need a different distribution to model

Use a Negative Binomial (NB) instead

$$\mathrm{var}(N_{fj}) = \mu_f(1 + \phi\mu_f^{\alpha-1})$$

This leads to the following:

$$N_{fj} = \mathrm{N}B(\mu_f, \phi)$$

When there is no biological difference, Φ will be zero, and the NB reverts to the Poisson

These allow us to determine which transcripts (*f*), have a different distribution in different samples (*j*). These can then be tested for statistical significance.

As the amount of data is rather limited in RNA-Seq, we have to estimate these parameters instead of calculating for each gene.

Summary

Questions?