RNA Sequencing

Elisha Roberson, Ph.D.

Depts. of Medicine & Genetics

eroberson@wustl.edu

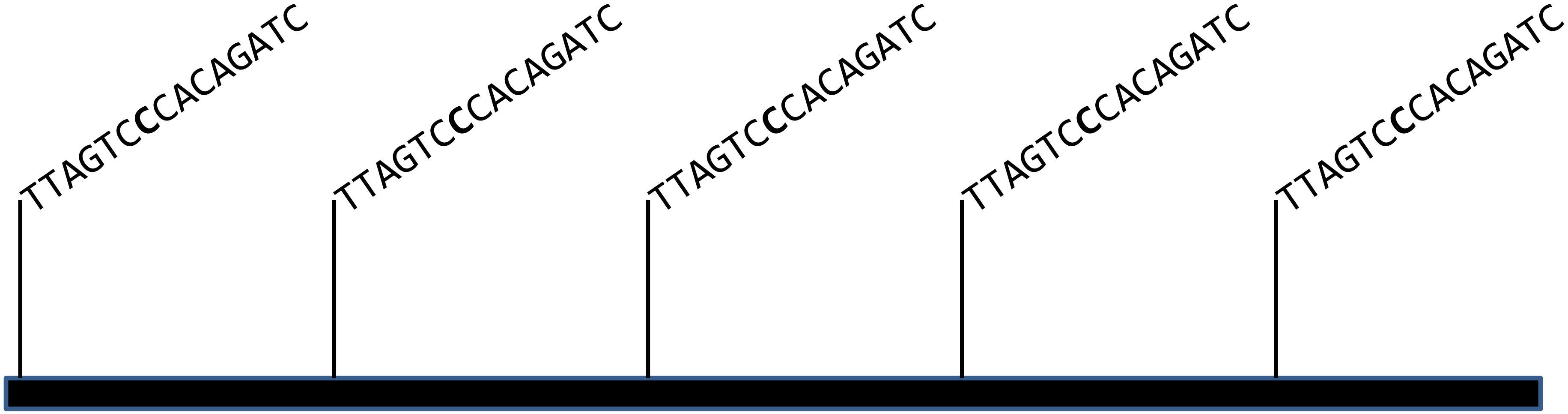
2016-December-06

Washington University in St. Louis

The Roberson Lab

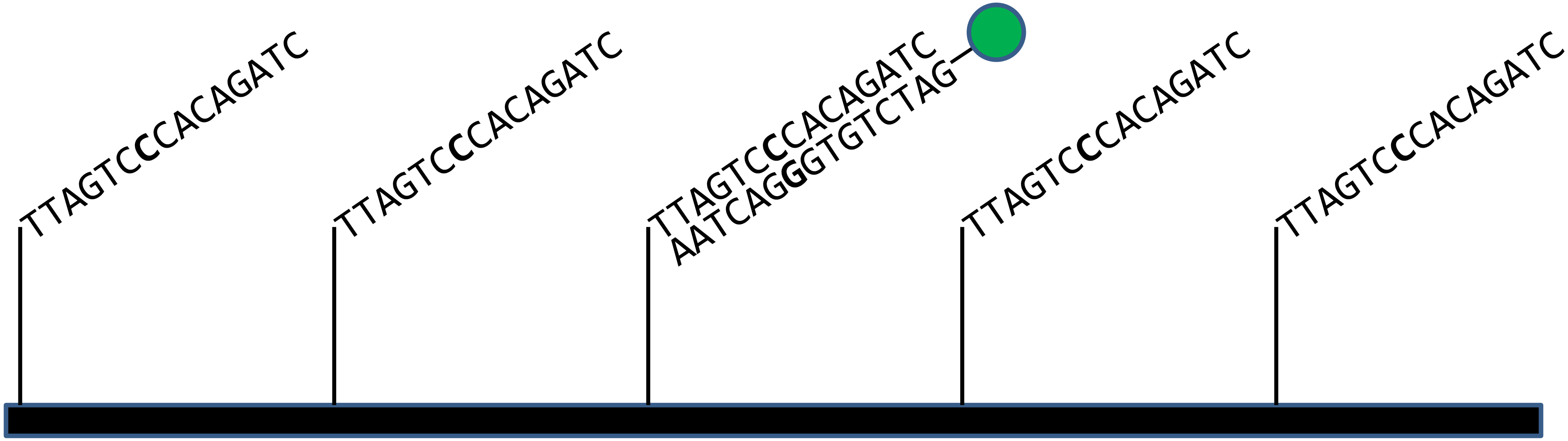
Comparing RNA-Seq and microarrays

RNA-Seq vs. microarrays



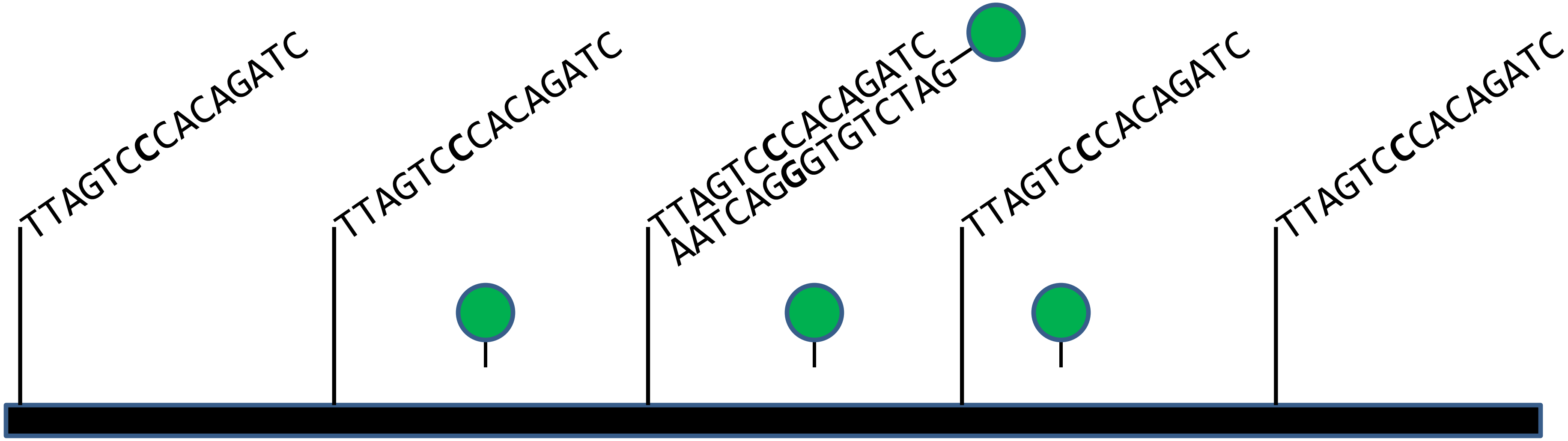
- Hybridization requires known targets

RNA-Seq vs. microarrays



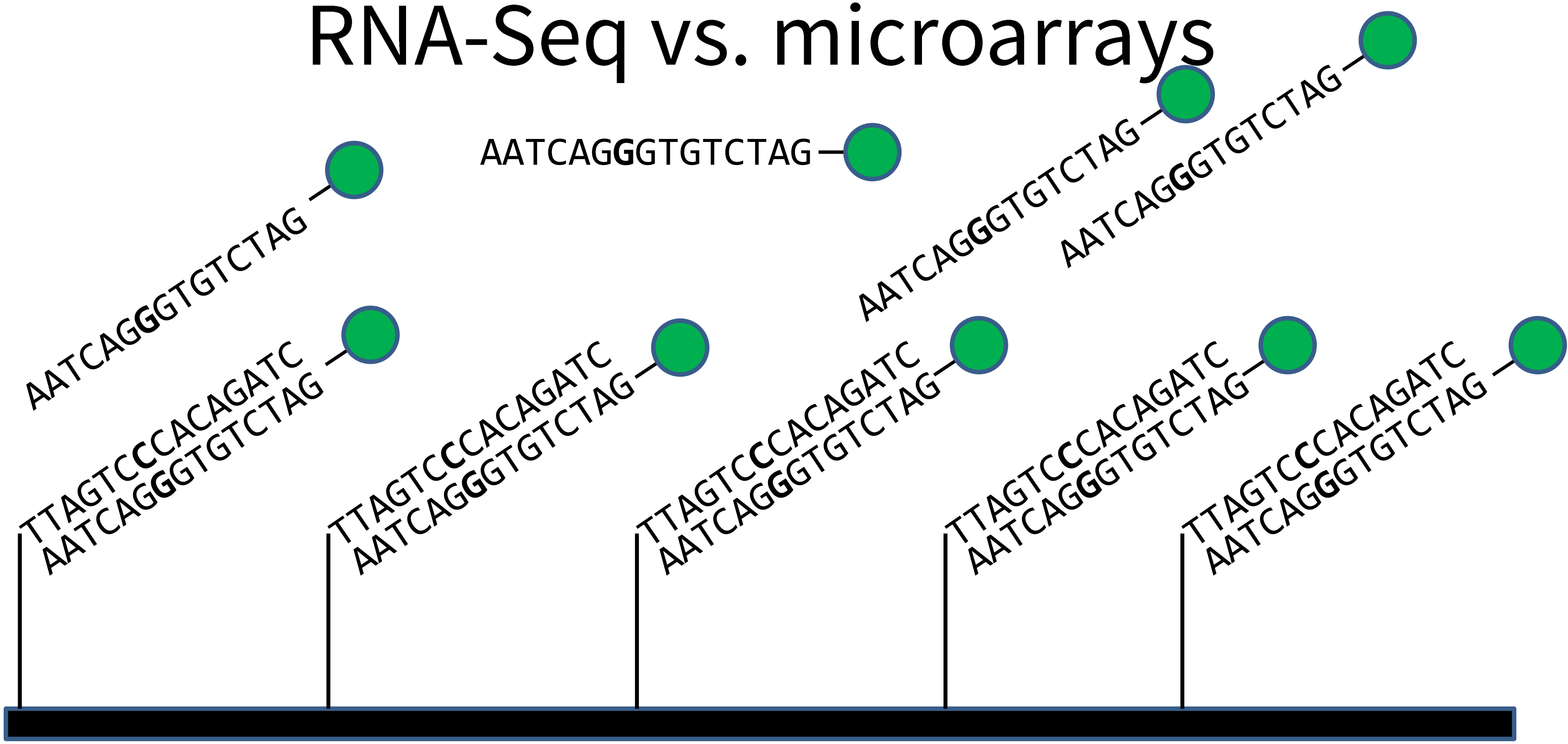
- Expression detected by fluorescence

RNA-Seq vs. microarrays



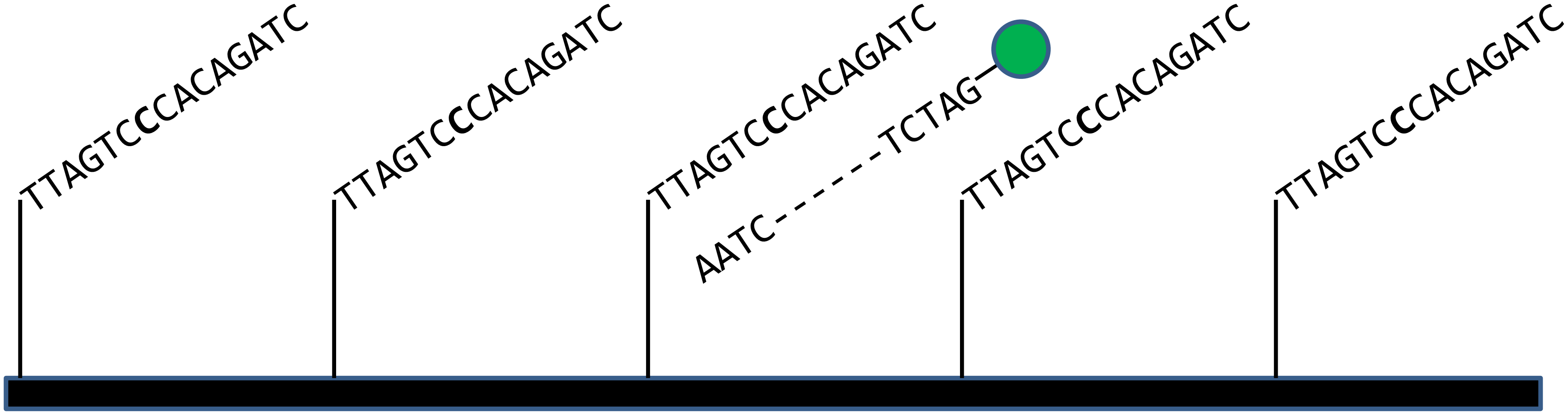
- Low-level expression can be difficult to detect compared to background

RNA-Seq vs. microarrays



- High-level expression can saturate probes

RNA-Seq vs. microarrays



- Variation in the subject's RNA sequence can affect binding kinetics

RNA-Seq vs. microarrays

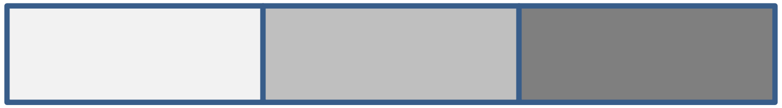
RNA-Seq differences

- Counting – no probes to saturate (though some sequences can predominate the library) and can always sequence more to get low-expressors

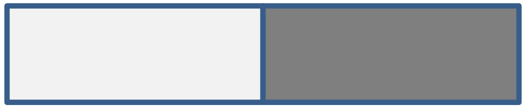
- Doesn’t *require* a known gene sequence

- Not affected by variation, as long as it doesn’t affect transcript stability

Isoform A



Isoform B

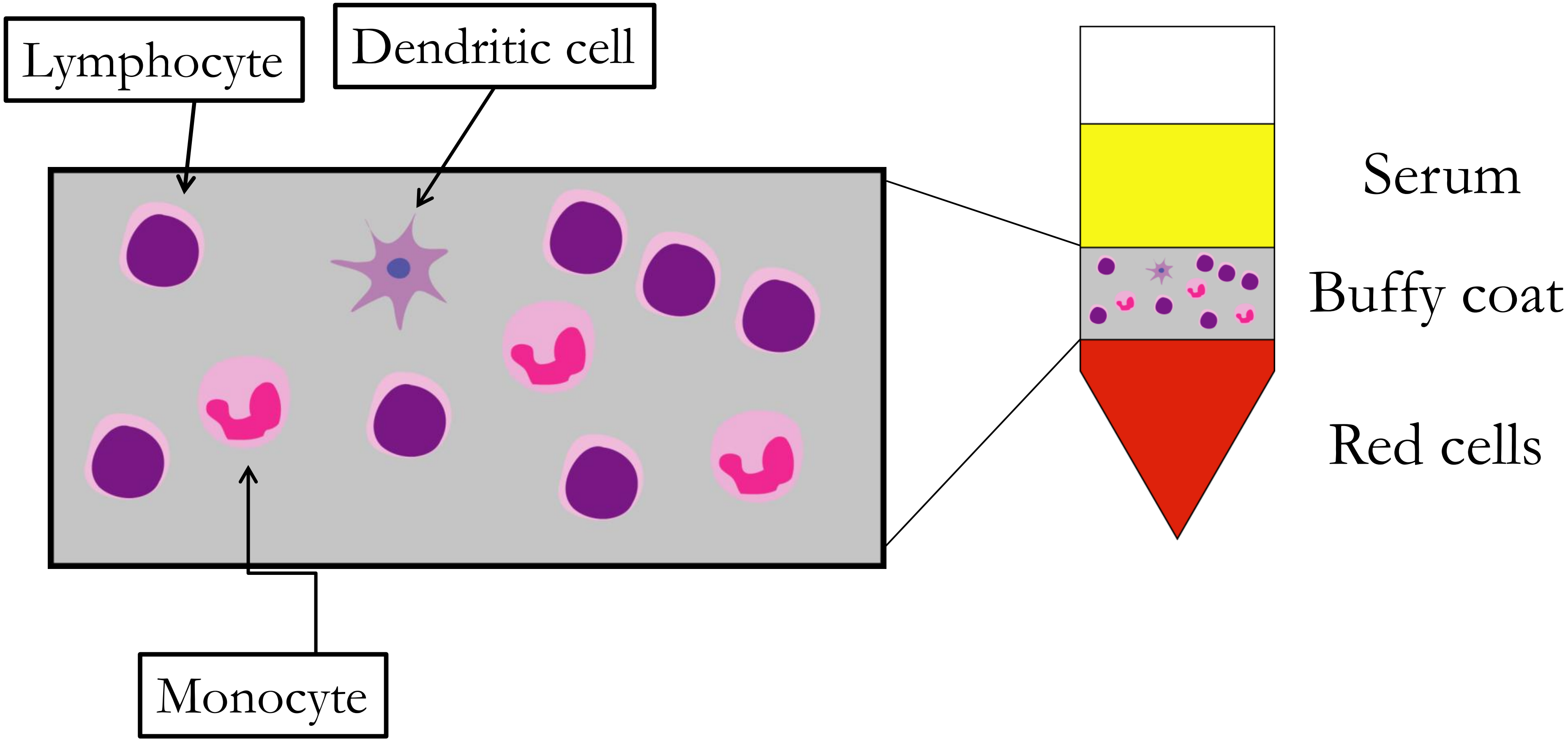


Sample prep is ***critical***

Collection & storage

- Blood is low-risk, easy to access, and frequently used
- ***But*** strongly enriched for red cells. What transcripts will predominate?

Collection & storage



- Freeze PBMCs directly
- Lyse directly in Trizol

Collection & storage

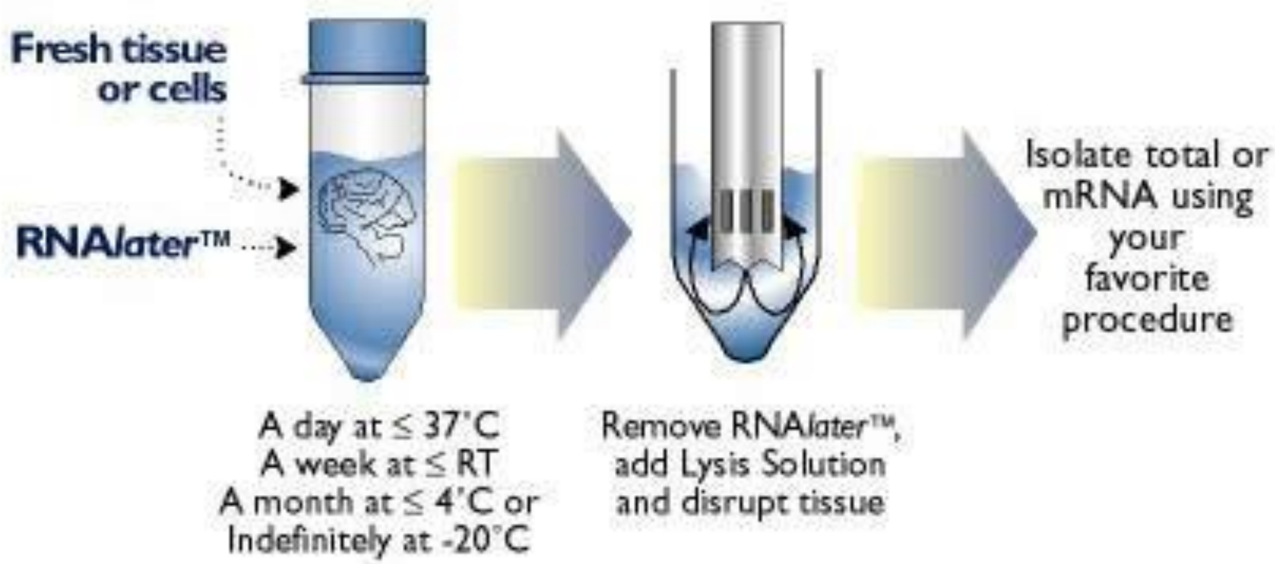
Tissues

- Biopsies, whole organs, etc.
- Unlike blood, the complex structure and embedded connective fibers make tissues more difficult to process and preserve.

*Nischal *et al*. J Cutan Aesthet Surg. 2008 Jul-Dec; 1(2): 107–111.

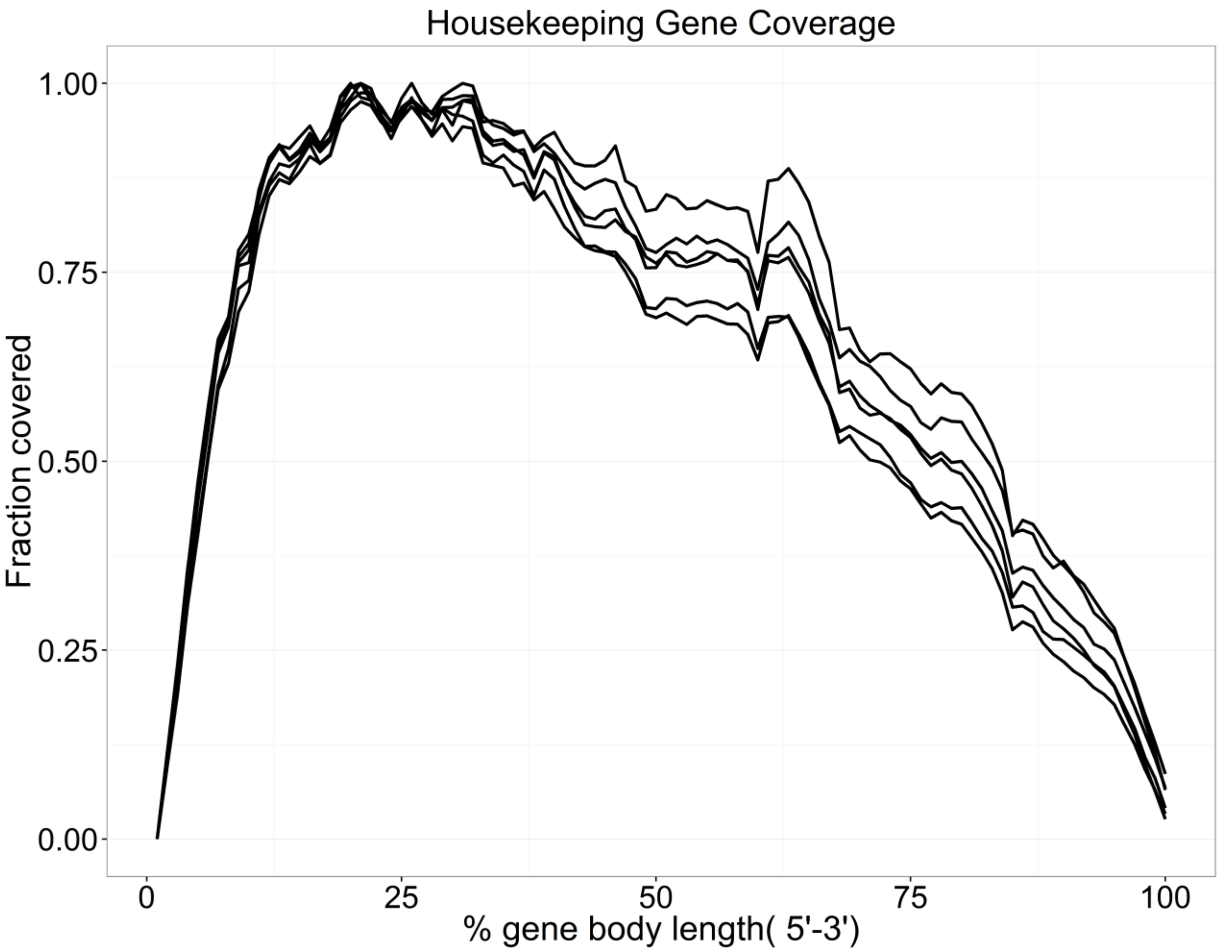
Collection & storage

Ambion
THE RNA COMPANY
Read SDS
P/N: AM7020
L/N: 1204009
MADE IN USA
RNA*later*® Solution
Store at RT
RNA Stabilization Solution
250 mL
AM7020$1204009
For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.


Fresh tissue or cells
RNA*later*™
A day at ≤ 37°C
A week at ≤ RT
A month at ≤ 4°C or
Indefinitely at -20°C
Remove RNA*later*™, add Lysis Solution and disrupt tissue
Isolate total or mRNA using your favorite procedure

- Some solutions, like RNAlater, can preserve tissue RNA for short-term storage at RT or long-term storage at -20°C

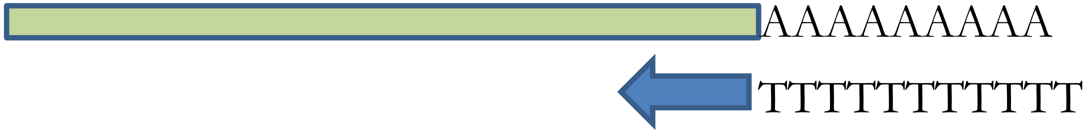
*Thermofisher



- PBMCs shipped on (too little) dry ice
- Degraded 3’-5’
- PolyA kit fails

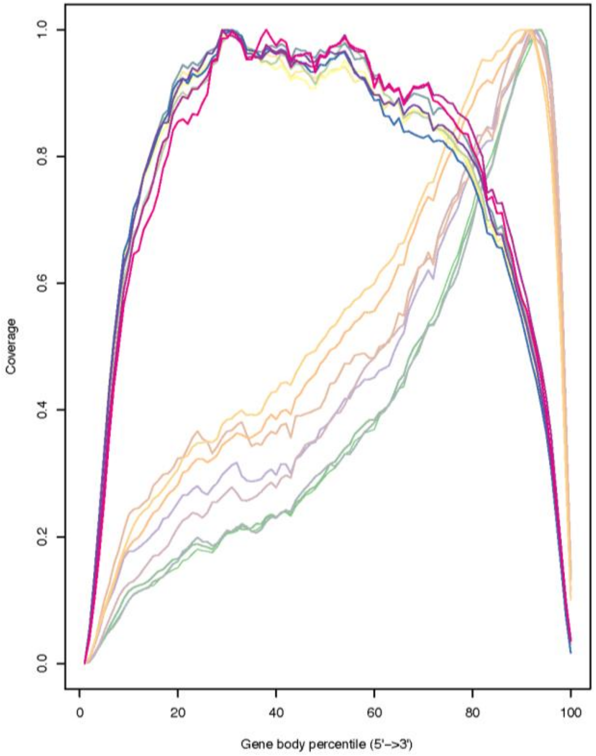
Library prep / cDNA methods

Polyadenylation (polyA) preps



AAAAAAAAA

TTTTTTTTTTT

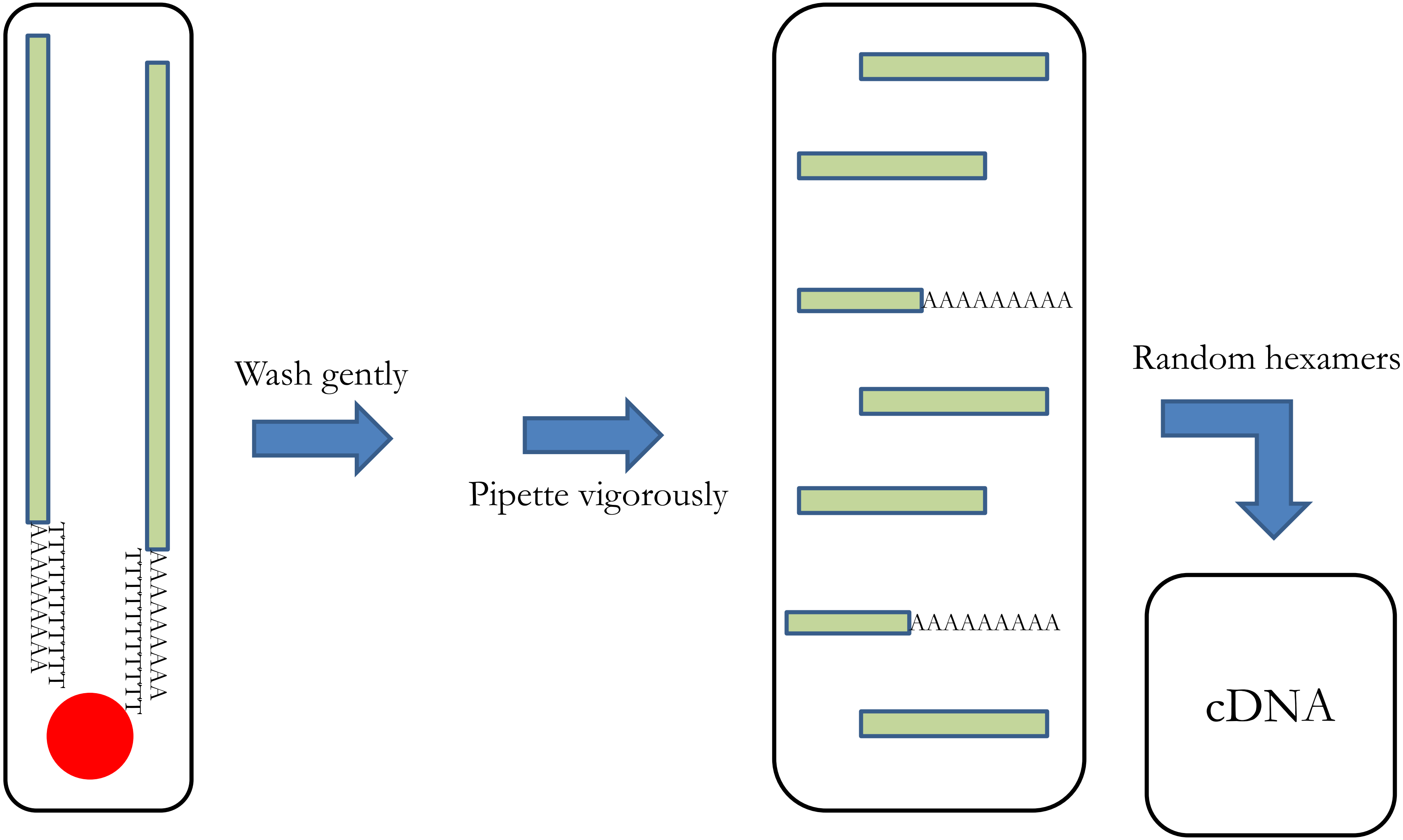


- polyT primed first-strand synthesis works

- **But** can lead to 3' bias and only captures polyadenylated transcripts.

*RSEQC

Polyadenylation (polyA) preps



Wash gently

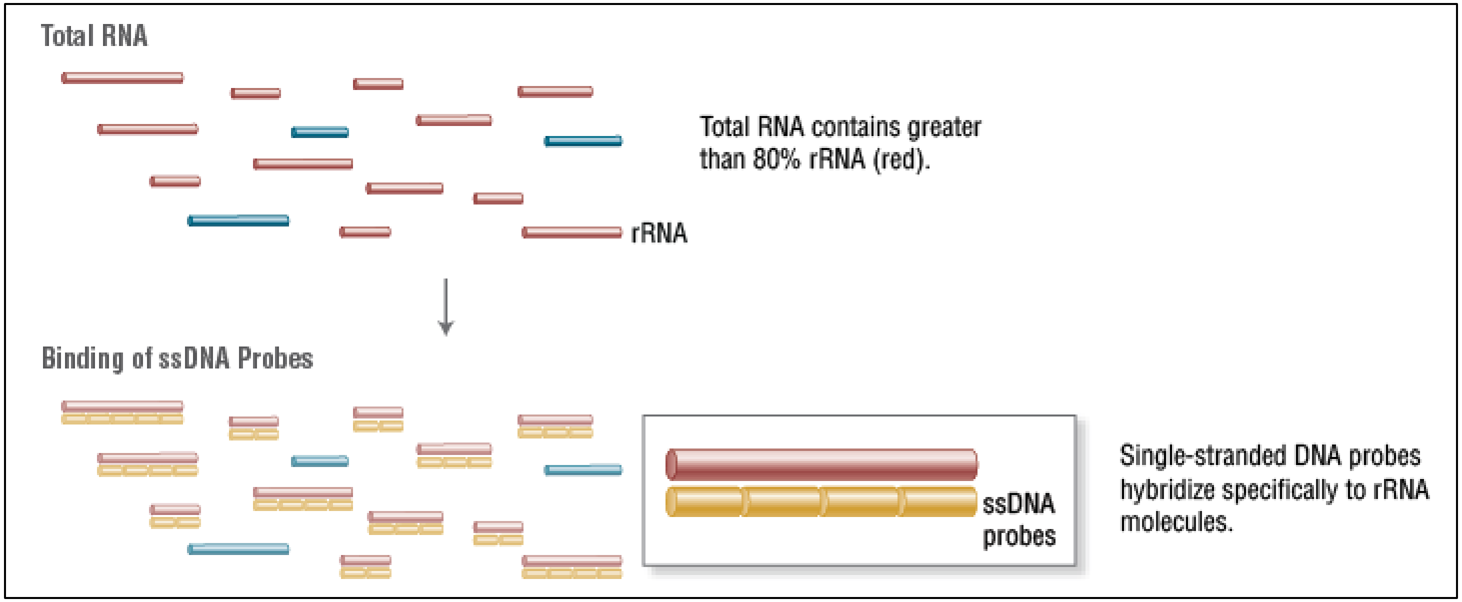
Pipette vigorously

Random hexamers

cDNA

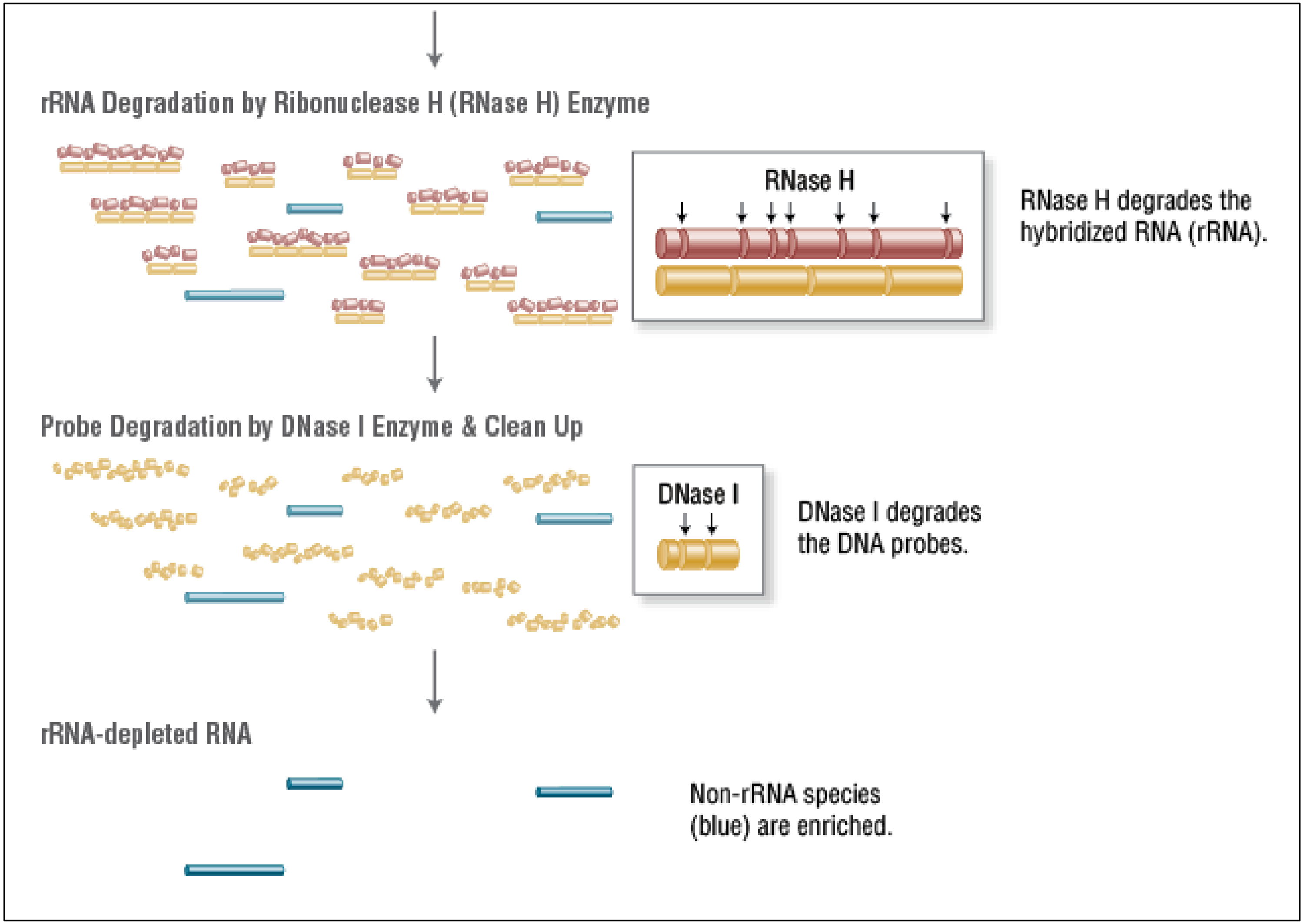
Ribosomal depletion preps

- Majority of cellular RNA is non-coding, particularly ribosomal RNA
- RNA polymerase I (28S, 18S, 5.8S rRNA) and Pol III (5S rRNA)
- **<u>Not polyadenylated</u>**


Total RNA
Total RNA contains greater than 80% rRNA (red).
rRNA
Binding of ssDNA Probes
ssDNA probes
Single-stranded DNA probes hybridize specifically to rRNA molecules.

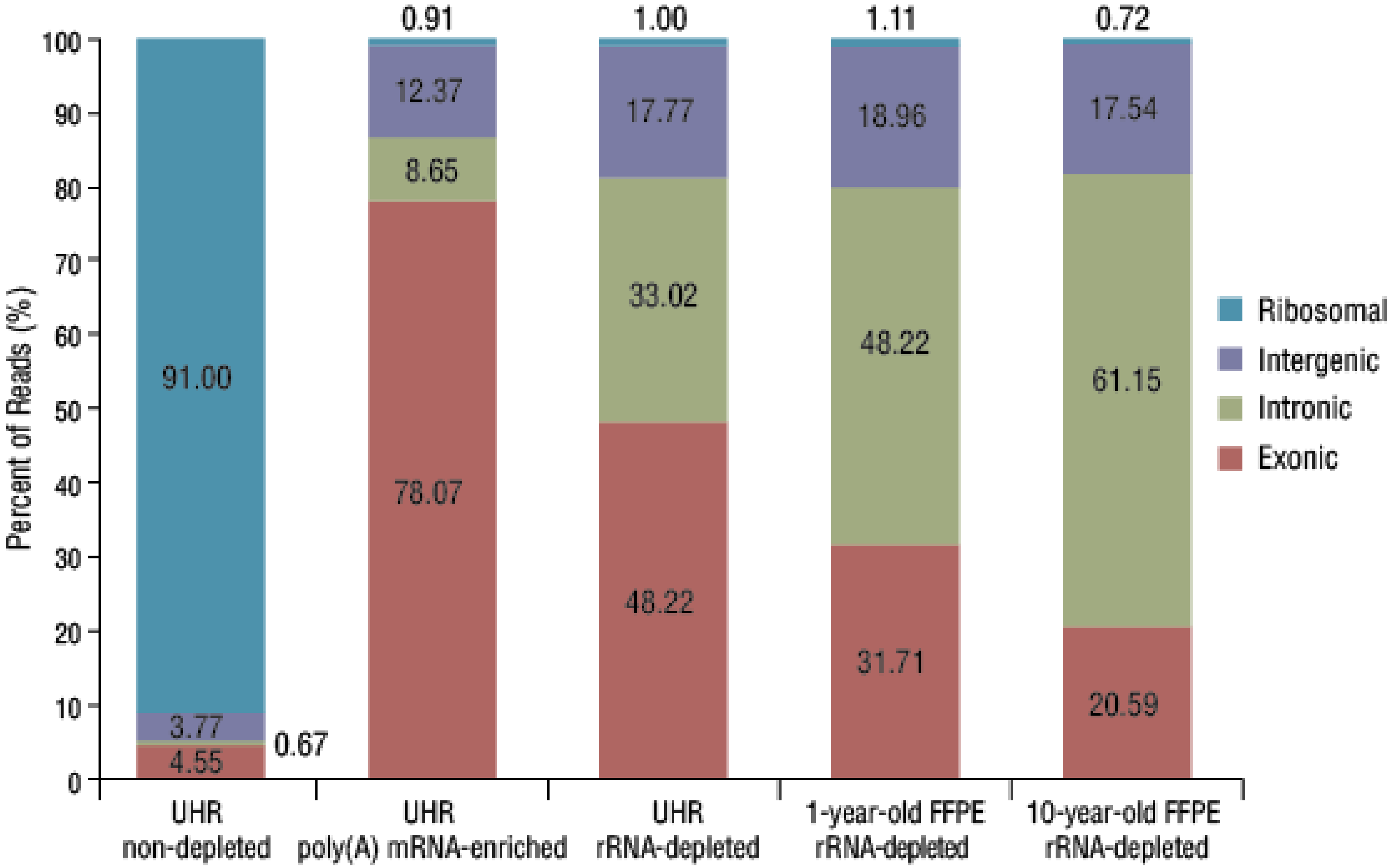
*NEB

Ribosomal depletion preps


rRNA Degradation by Ribonuclease H (RNase H) Enzyme
RNase H
RNase H degrades the hybridized RNA (rRNA).
Probe Degradation by DNase I Enzyme & Clean Up
DNase I
DNase I degrades the DNA probes.
rRNA-depleted RNA
Non-rRNA species (blue) are enriched.

*NEB

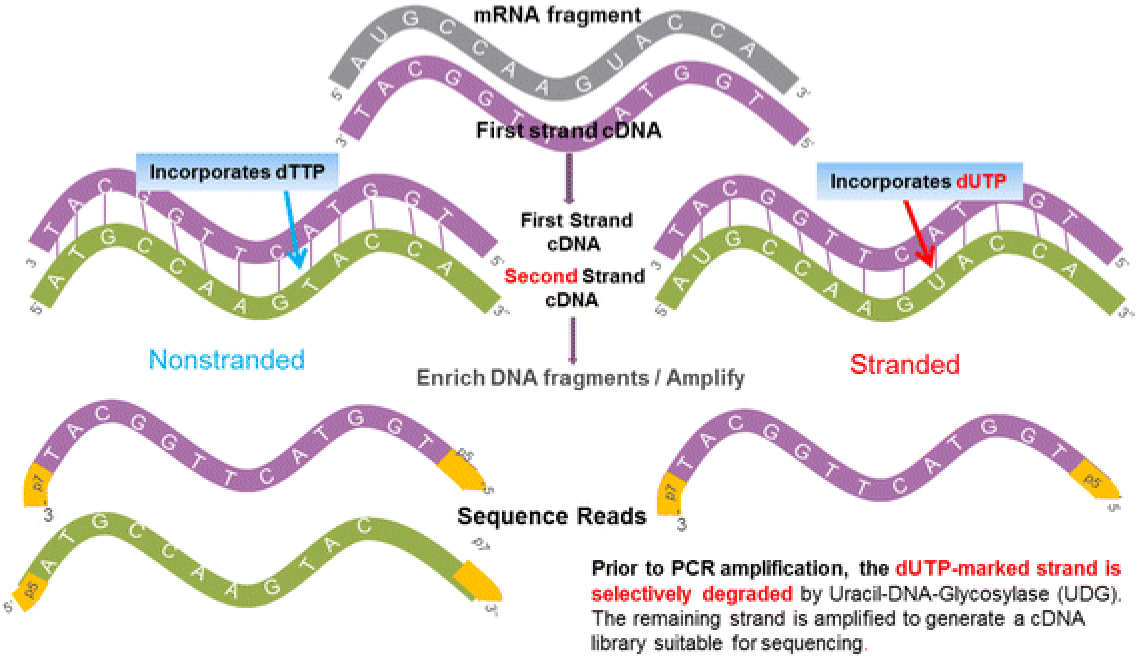
Ribosomal depletion preps



*NEB

Stranded preps

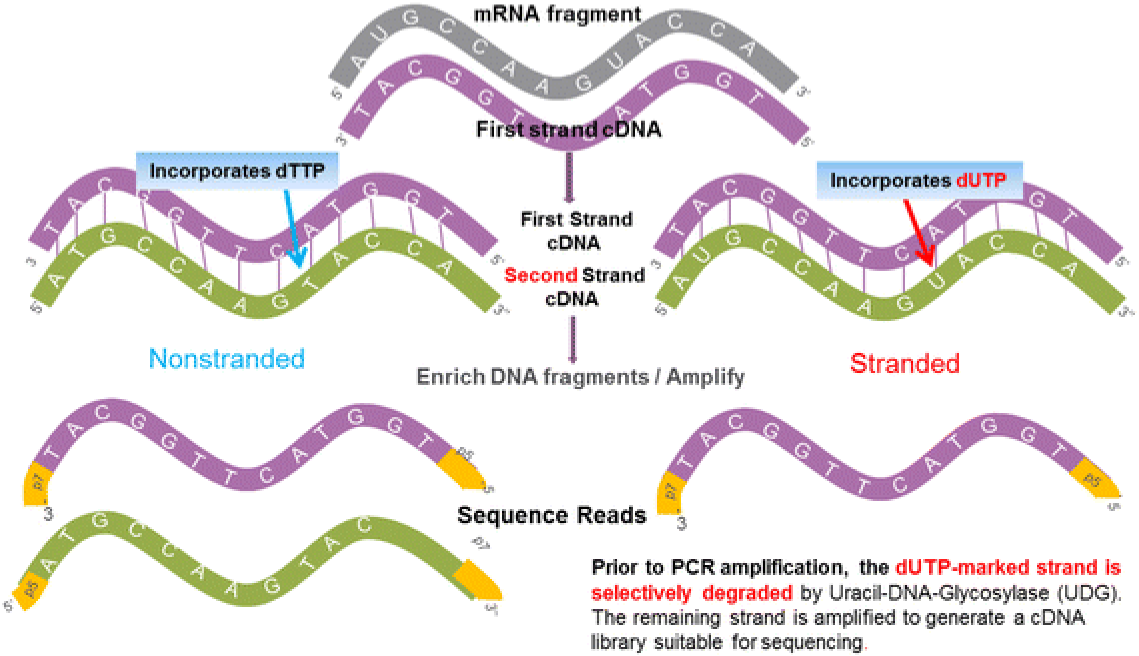
- Standard cDNA → library prep retains no information about transcript strand.
- Some loci have antisense transcripts



Prior to PCR amplification, the dUTP-marked strand is selectively degraded by Uracil-DNA-Glycosylase (UDG). The remaining strand is amplified to generate a cDNA library suitable for sequencing.

*2015 Zhao *et al*. PMCID: PMC4559181

Stranded preps



Prior to PCR amplification, the dUTP-marked strand is selectively degraded by Uracil-DNA-Glycosylase (UDG). The remaining strand is amplified to generate a cDNA library suitable for sequencing.

- Normal first strand synthesis. 2nd strand incorporates uracil
- Uracil-DNA glycosylase excises U-base from DNA
- Endonuclease VIII breaks backbone at those sites

*2015 Zhao *et al*. PMCID: PMC4559181

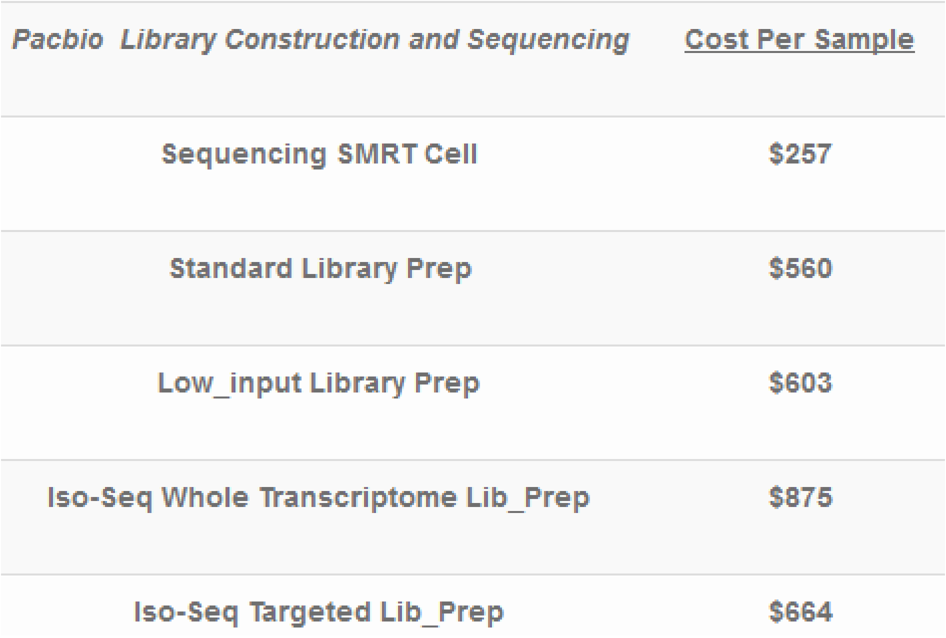
Sequencing choices

Short reads

- Illumina
  - Single-end vs. paired-end
    - Paired-end superior. Estimates insert size empirically
  - Read length
    - Greater cycle number preferred
    - 2x75 good compromise
  - Depth / coverage
    - Very different from DNA seq
    - Variable gene length and expression level
    - Several tools to estimate

Long reads – Pac Bio SMRT

- Full-length isoform sequencing ($$$$$)



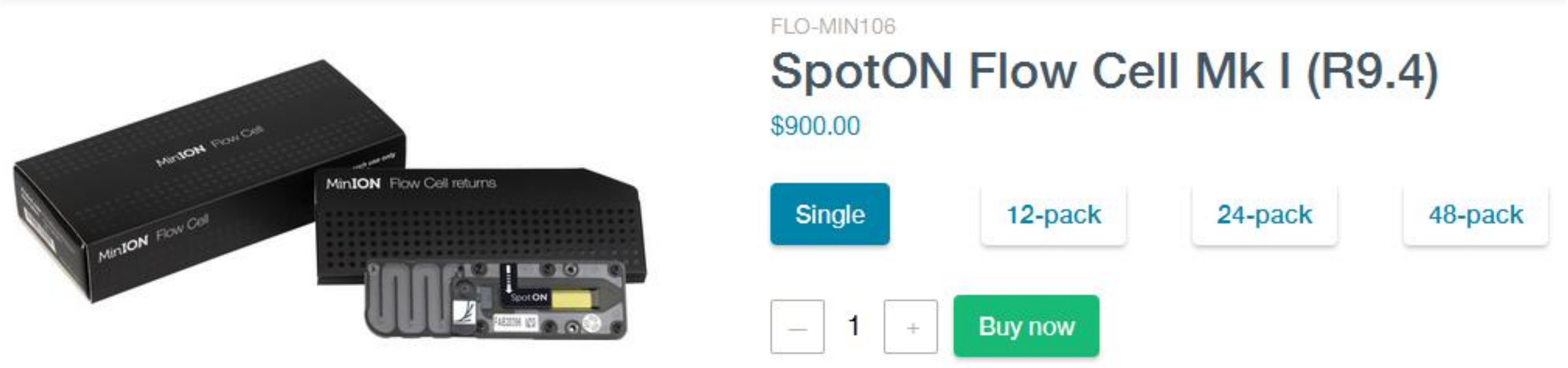
Pacbio Library Construction and Sequencing	*Cost Per Sample*
Sequencing SMRT Cell	$257
Standard Library Prep	$560
Low_input Library Prep	$603
Iso-Seq Whole Transcriptome Lib_Prep	$875
Iso-Seq Targeted Lib_Prep	$664

Min $1,100 / sample for whole transcriptome.

*NCI seq facility. Access 2016-12-05 https://ostr.cancer.gov/resources/fnl-cores/sequencing-facility/pricing

Long reads – Oxford nanopore

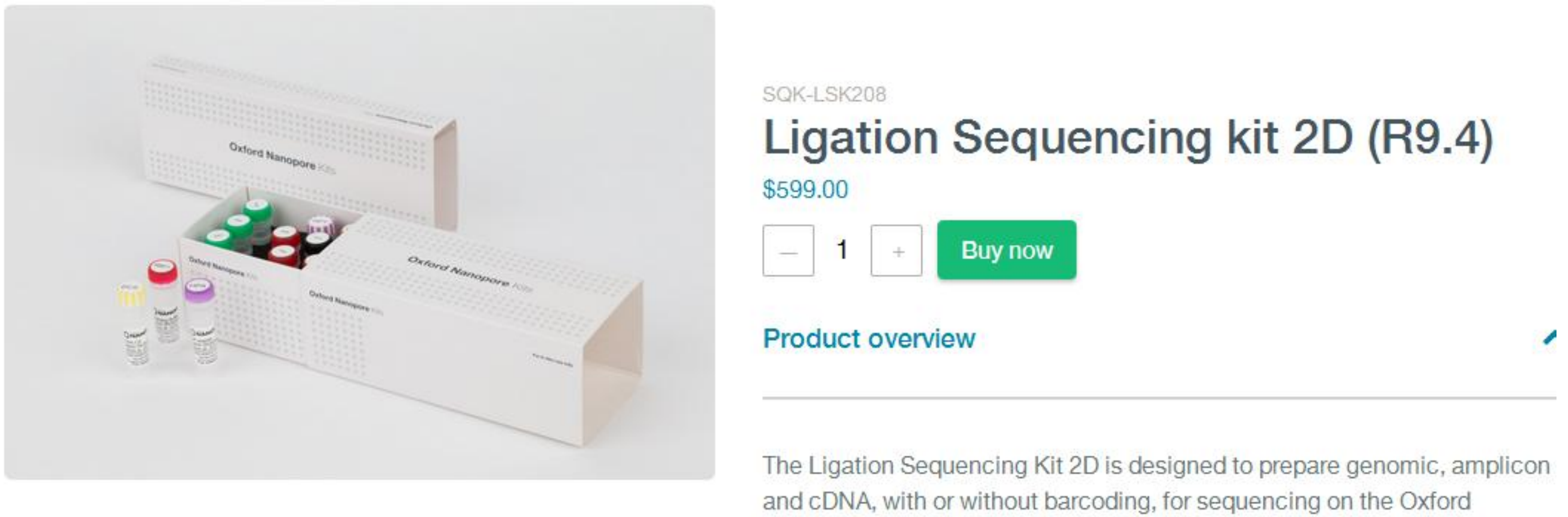
- MinION / PromethION (also $$$$$)



FLO-MIN106

SpotON Flow Cell Mk I (R9.4)

$900.00



SQK-LSK208

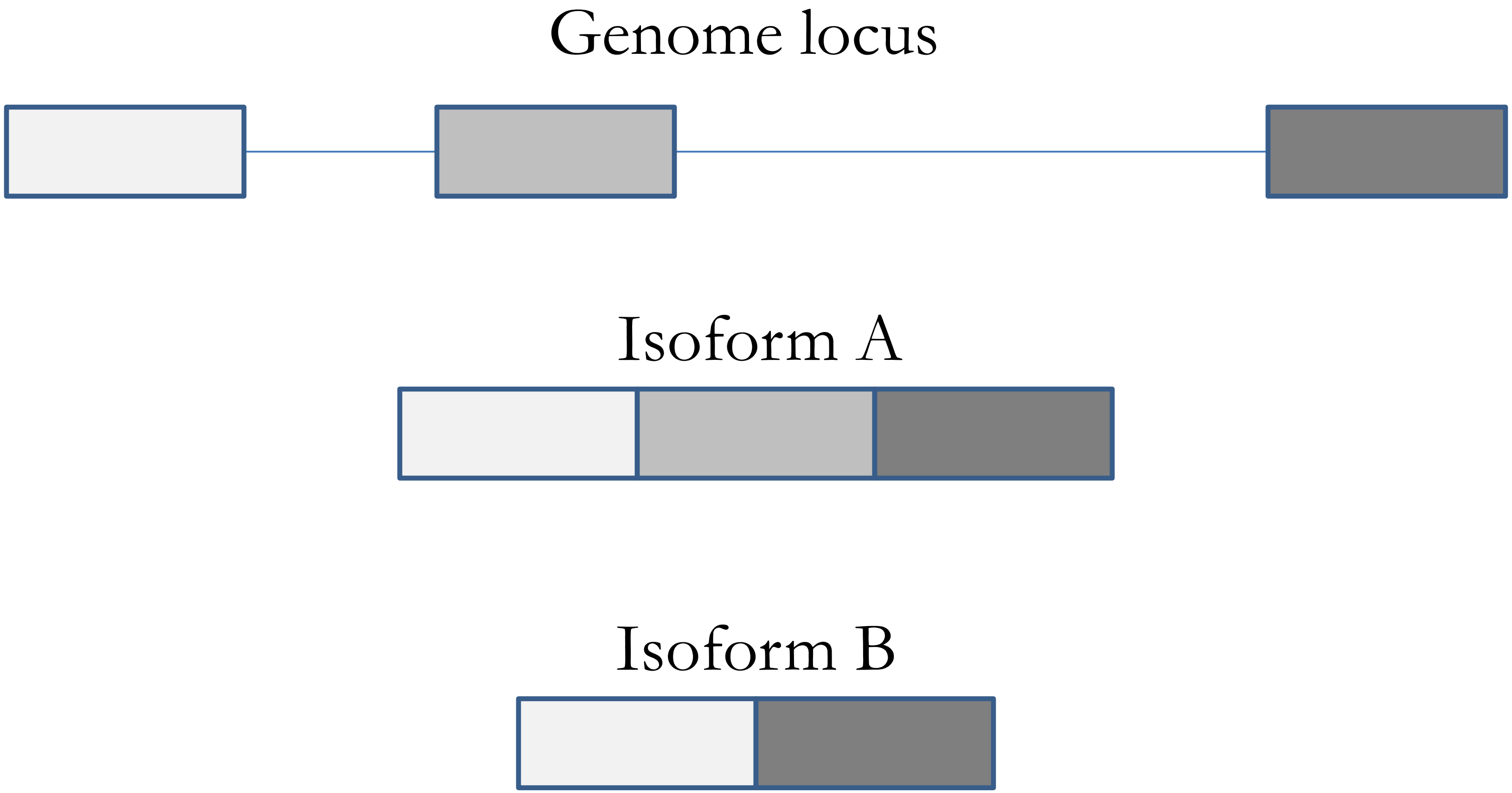
Ligation Sequencing kit 2D (R9.4)

$599.00

The Ligation Sequencing Kit 2D is designed to prepare genomic, amplicon and cDNA, with or without barcoding, for sequencing on the Oxford

Alignment and counting

RNA-Seq does not look like the genome

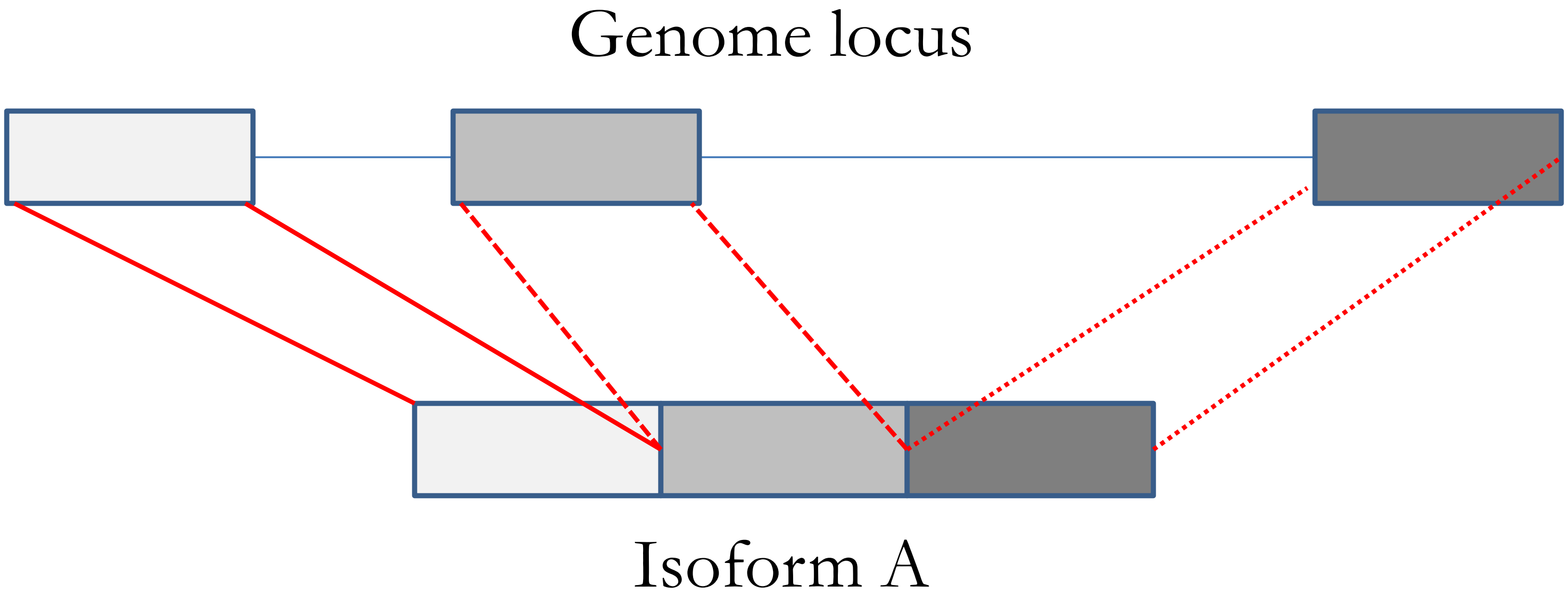


Genome locus

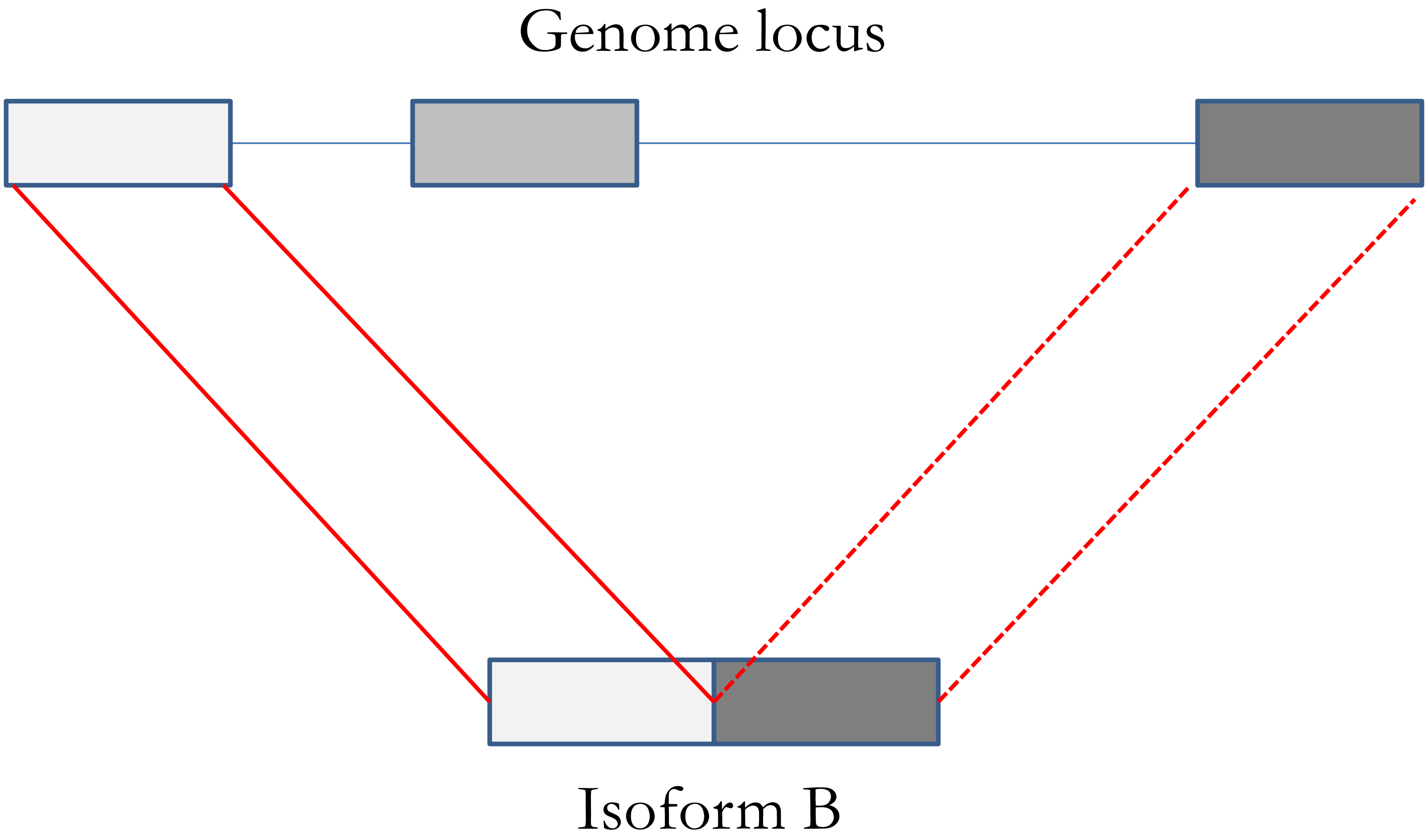
Isoform A

Isoform B

RNA-Seq does not look like the genome

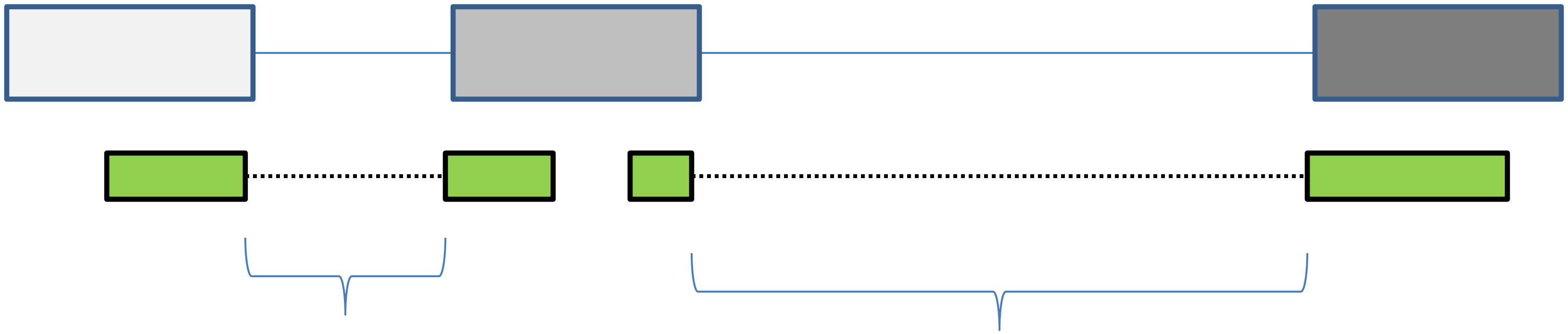


RNA-Seq does not look like the genome



RNA-Seq does not look like the genome

Genome locus



Deletion

Deletion

Genomic aligners expect the library to reflect **genome** architecture. Intron splicing looks like large deletions, and can confuse aligner.

One alternative is to align to transcript FASTA rather than whole-genome.

Splice-aware genome aligners

- Tophat2 (bowtie derived)
- Spliced Transcript Alignment to Reference (STAR)

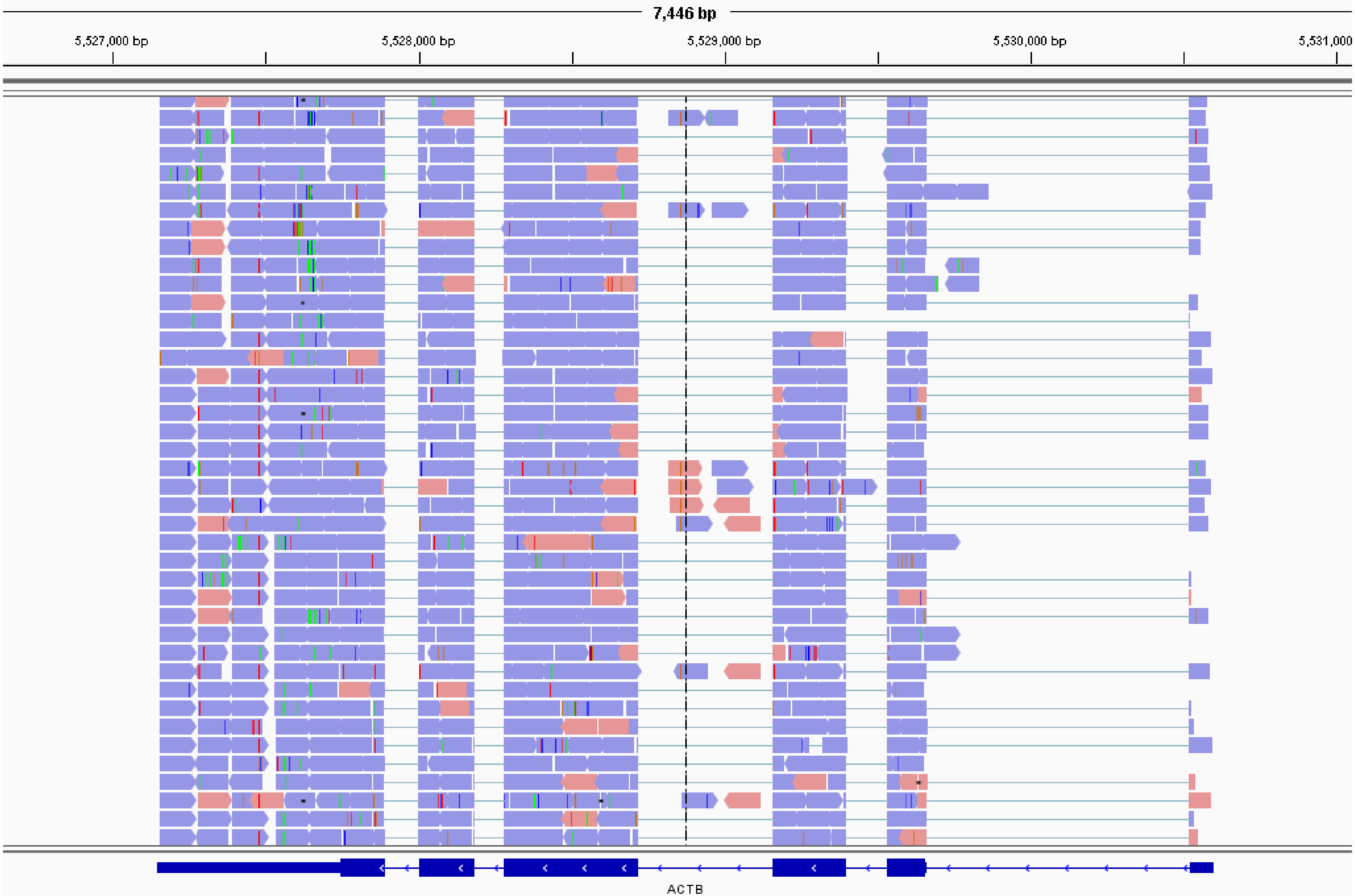
Transcript alignment

- Kallisto
- Sailfish / Salmon

De novo assembly

- ABySS / TransABySS
- Trinity
- SOAPdenovo-Trans

Ribosomal depletion alignment



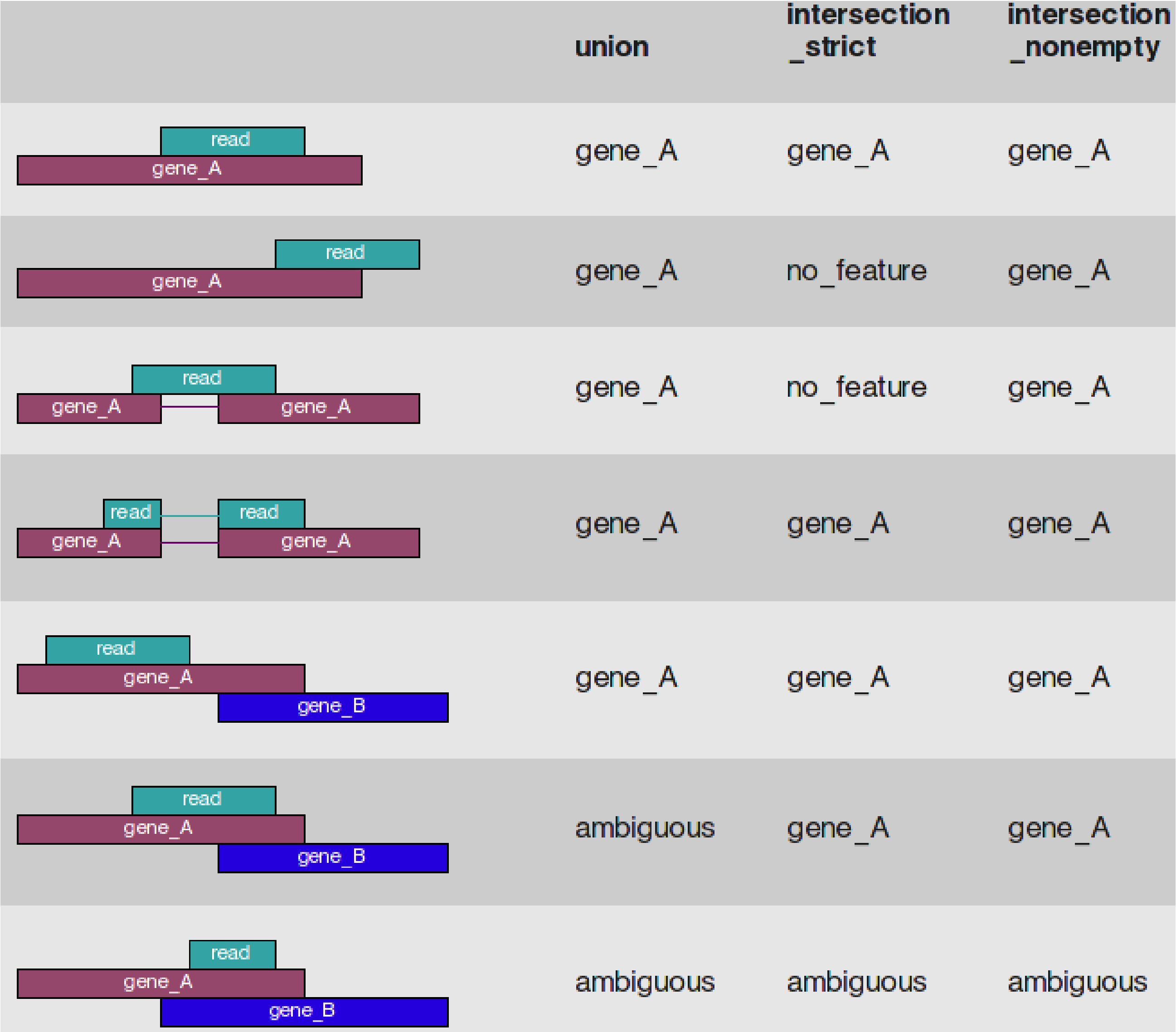
Counting reads

Counting tools

HTSeq count

subread featureCount

RSEM

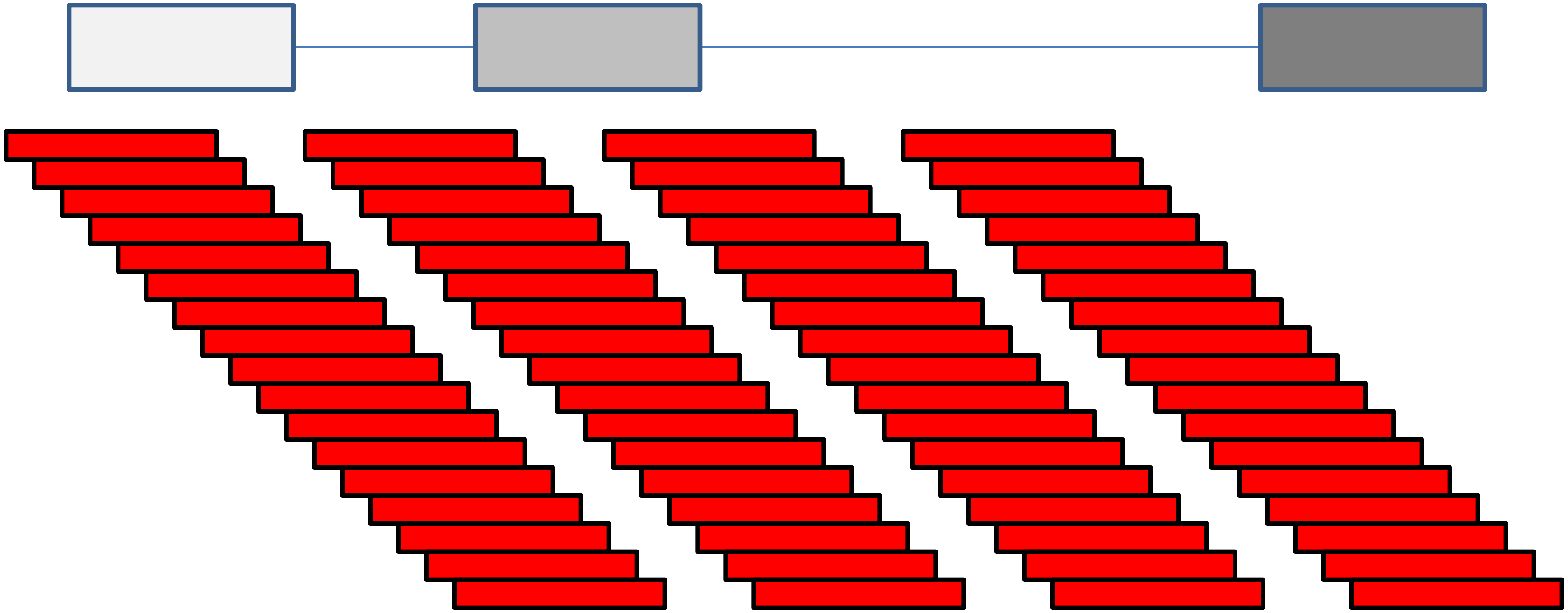


	union	intersection_strict	intersection_nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	ambiguous	ambiguous	ambiguous

*HTSeq count

RNA-Seq complications - duplicates

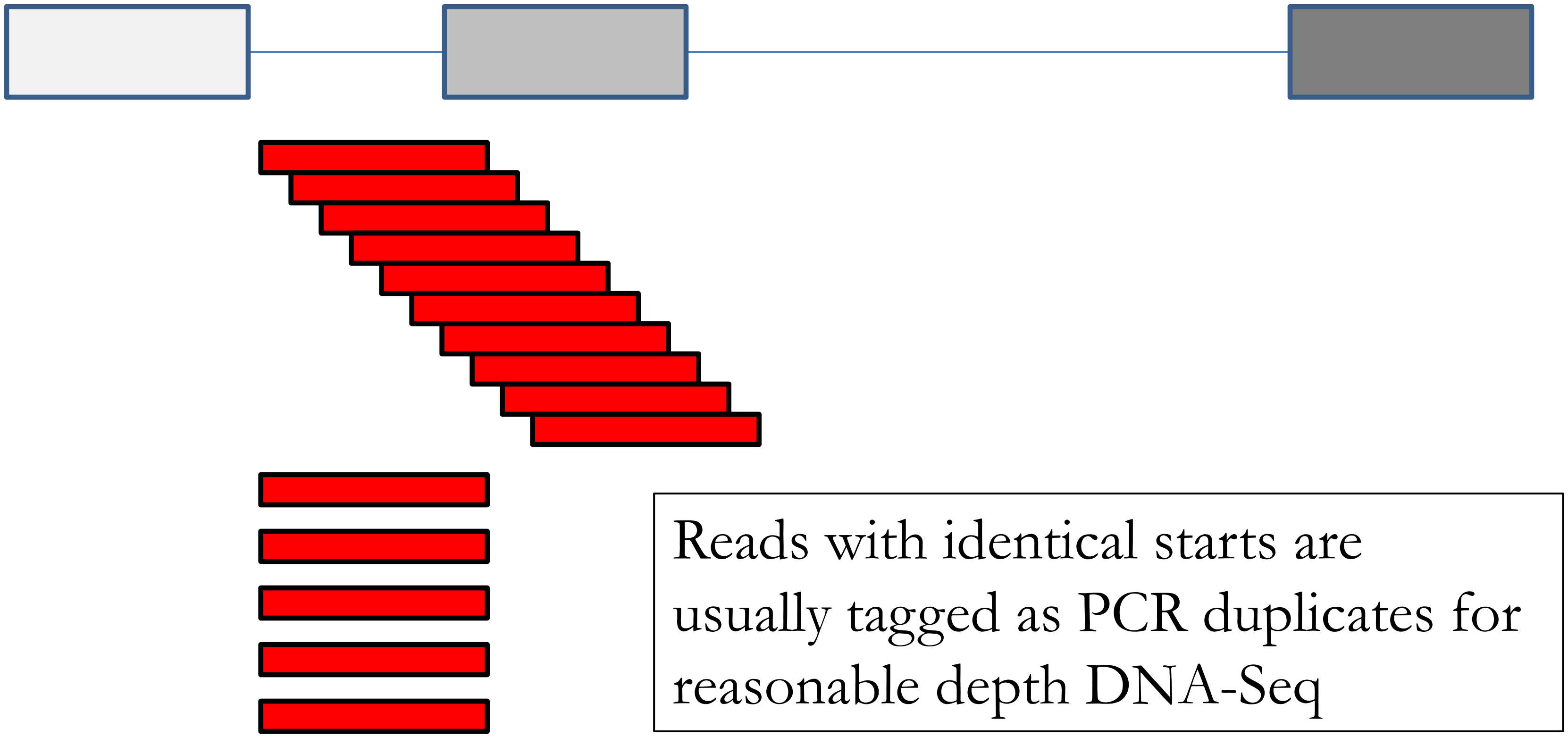
Genome locus



Highly covered genomic locus for DNA-Seq

RNA-Seq complications - duplicates

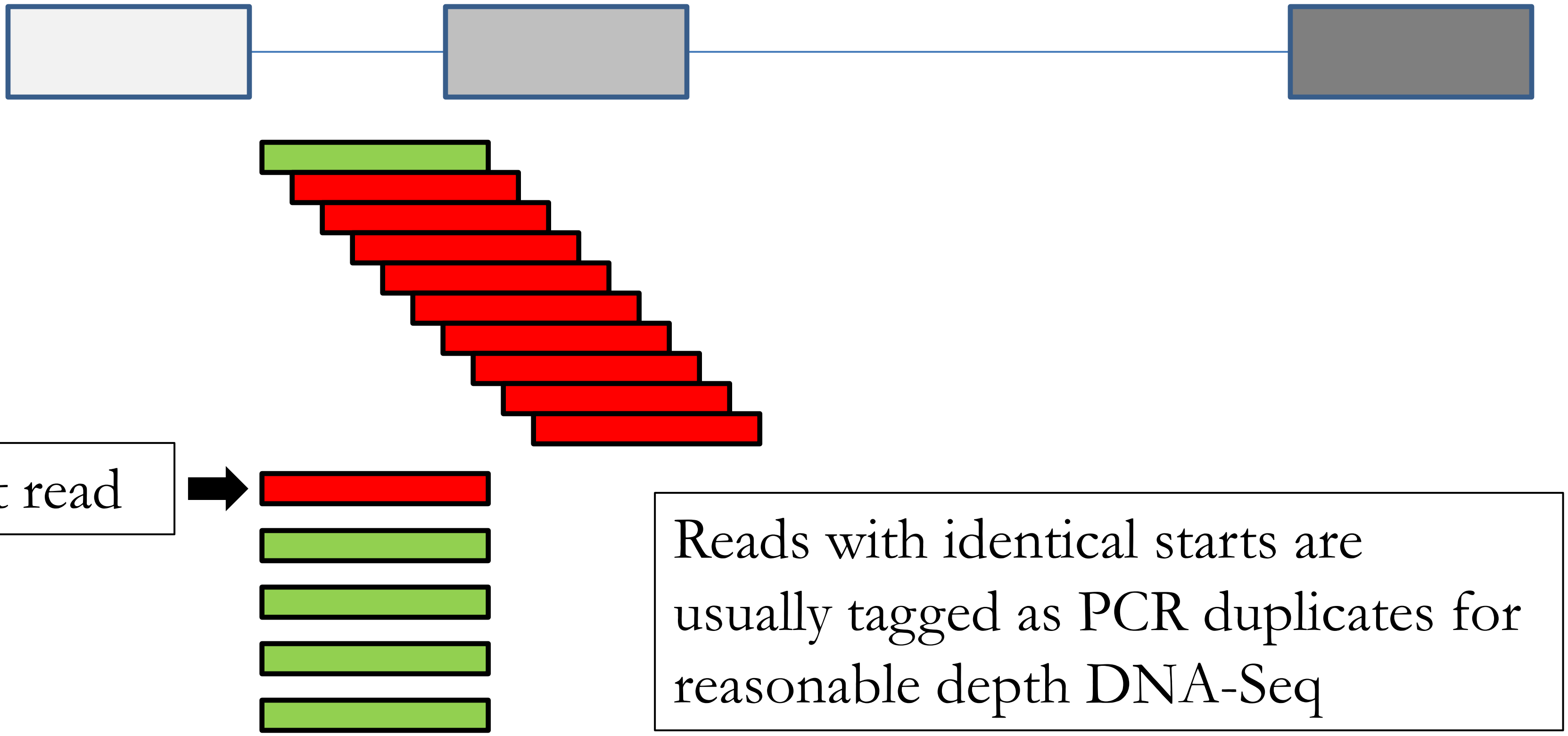
Genome locus



Reads with identical starts are usually tagged as PCR duplicates for reasonable depth DNA-Seq

RNA-Seq complications - duplicates

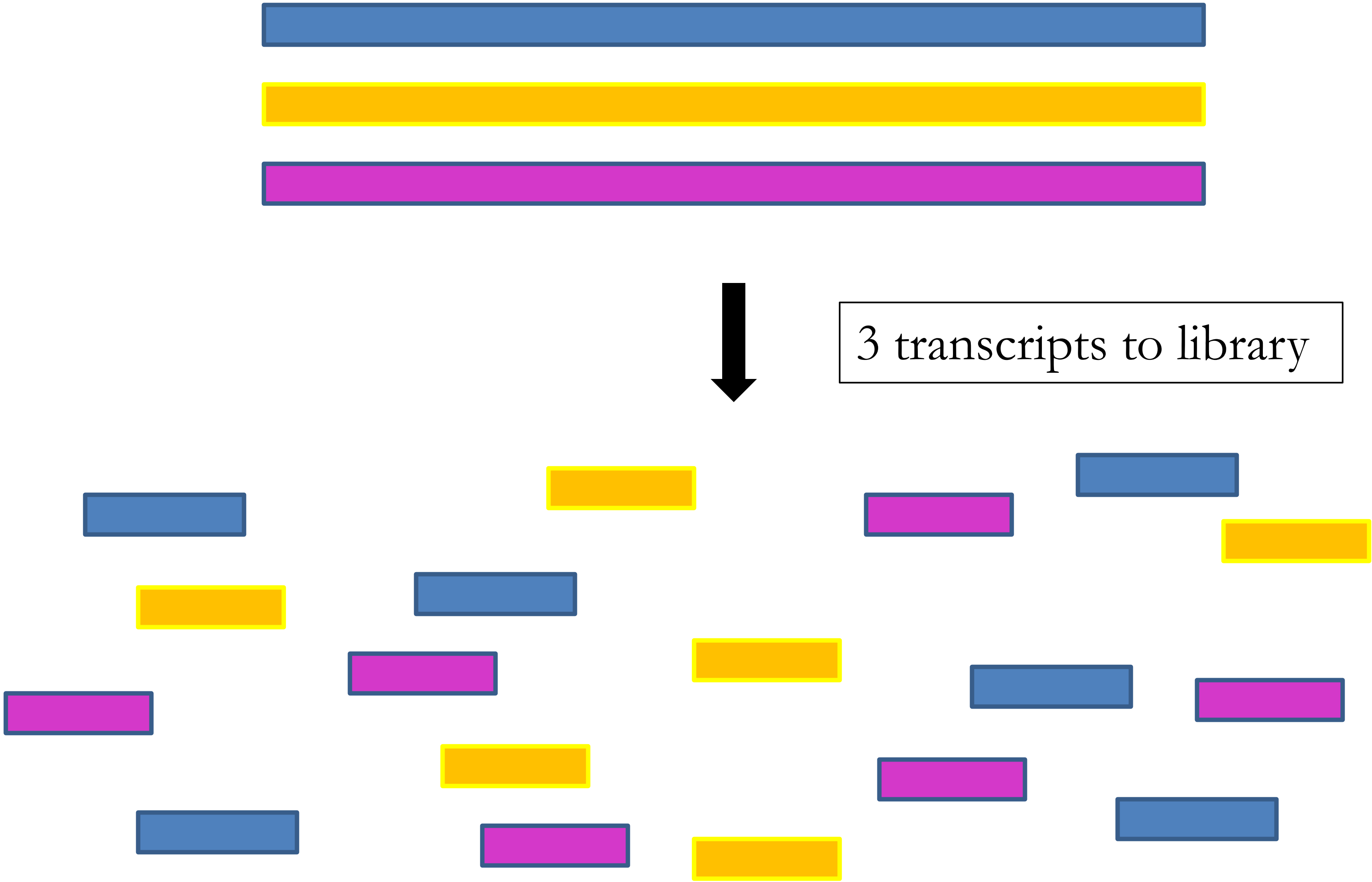
Genome locus



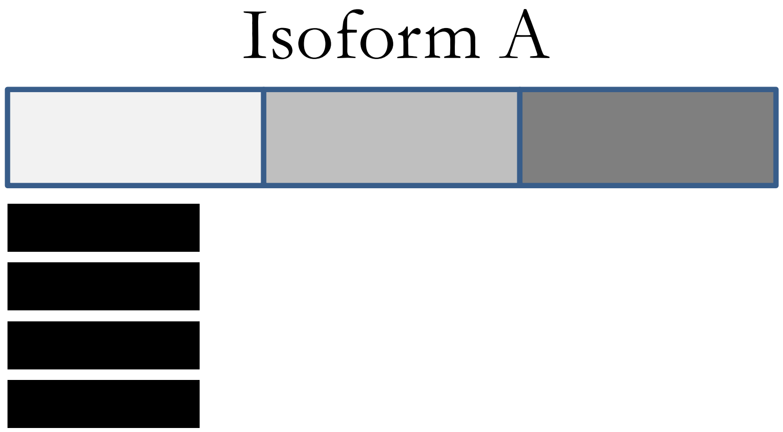
Best read

Reads with identical starts are usually tagged as PCR duplicates for reasonable depth DNA-Seq

RNA-Seq complications - duplicates


3 transcripts to library

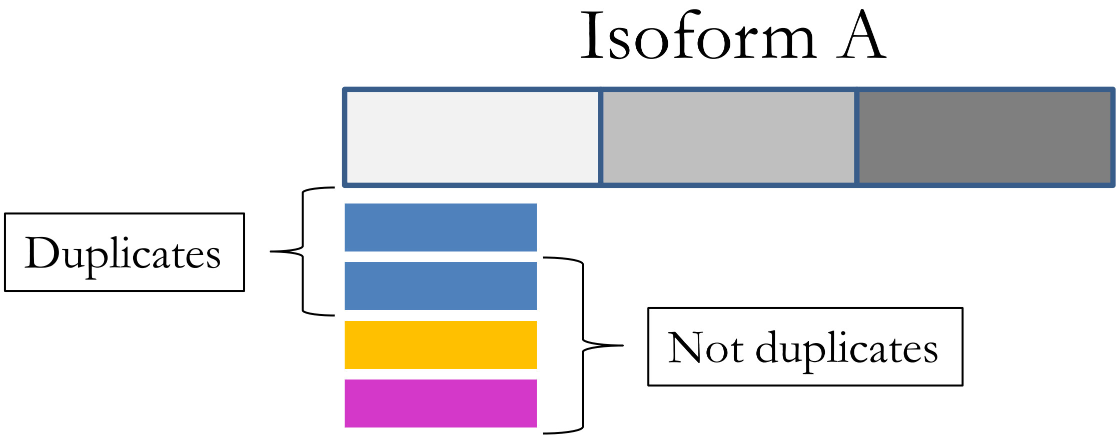
RNA-Seq complications - duplicates



- For DNA-Seq the target is equimolar, but RNA-Seq is more complicated

- Important considerations:
  - Sequencing depth
  - Gene relative expression level
  - Gene size

RNA-Seq complications - duplicates



- **<u>But</u>** this is a solvable problem

- Adding a short barcode to each fragment during PCR, called a unique molecular identifier (UMI) we know whether reads are truly unique, even with identical 5’ mapping

Estimating gene abundance & differential expression

Abundance metrics

- Fragments per Kb of exon per million mapped reads (FPKM)

- Transcripts per million (TPM)

- TPM preferable
  - Different total reads between experiments skew FPKM
  - TPM consistent, i.e. 1 TPM sample A and sample B really means similar abundance

Modeling counts – edgeR, DESeq2

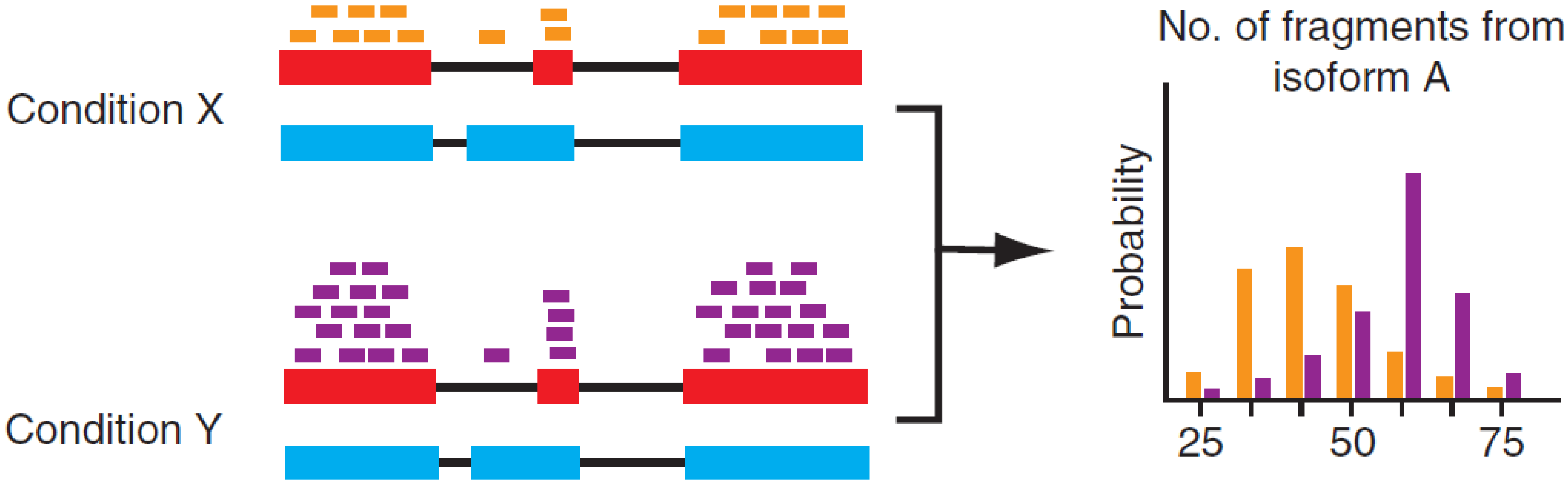
- **Counts are not normally distributed**

- What models counts?
  - Poisson distribution
    - But assumes variance & mean equal

  - Negative binomial
    - Mean ≠ variance
    - edgeR, DESeq2 R packages

Modeling counts – cuffdiff2

4) Combine uncertainty and overdispersion into a single model of fragment count variability (beta negative binomial)

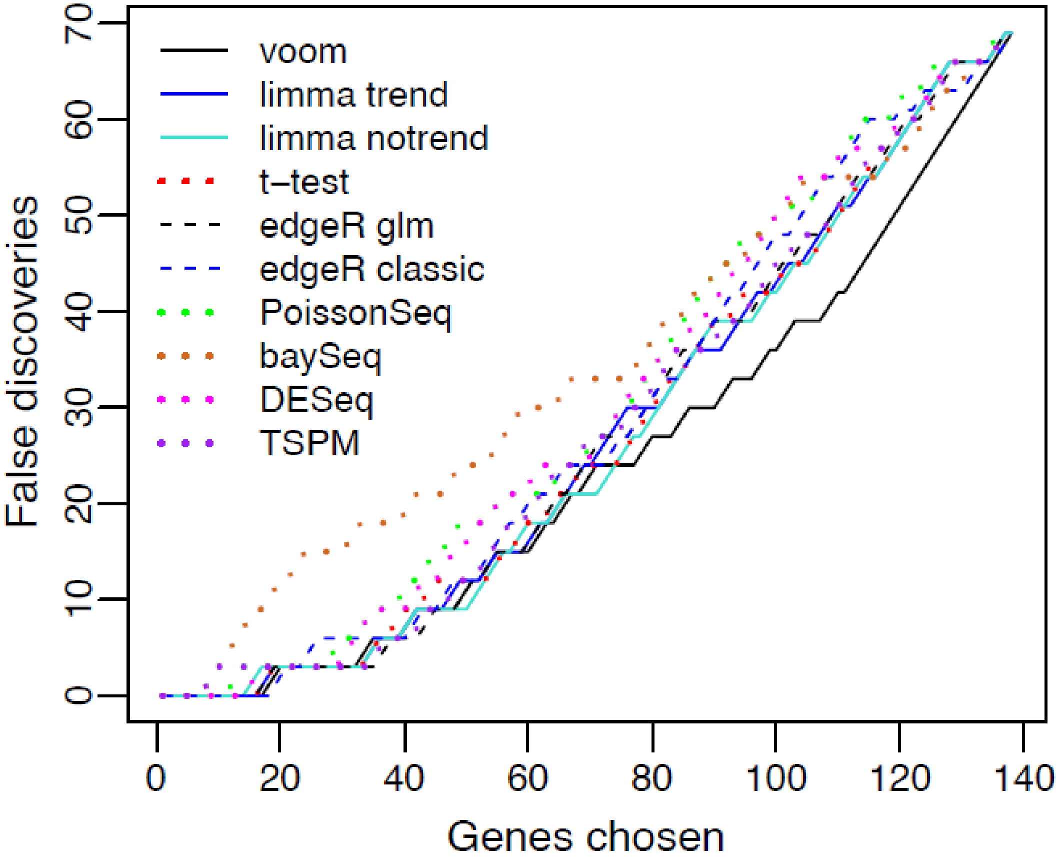


5) Test for signficance of changes between conditions in transcript-level counts

- edgeR & DESeq2 model **<u>gene-level</u>** differential expression
- cuffdiff2 tests for significant isoform-level DE

*2013 Trapnell *et al*. PMCID: PMC3869392

Modeling counts - VOOM

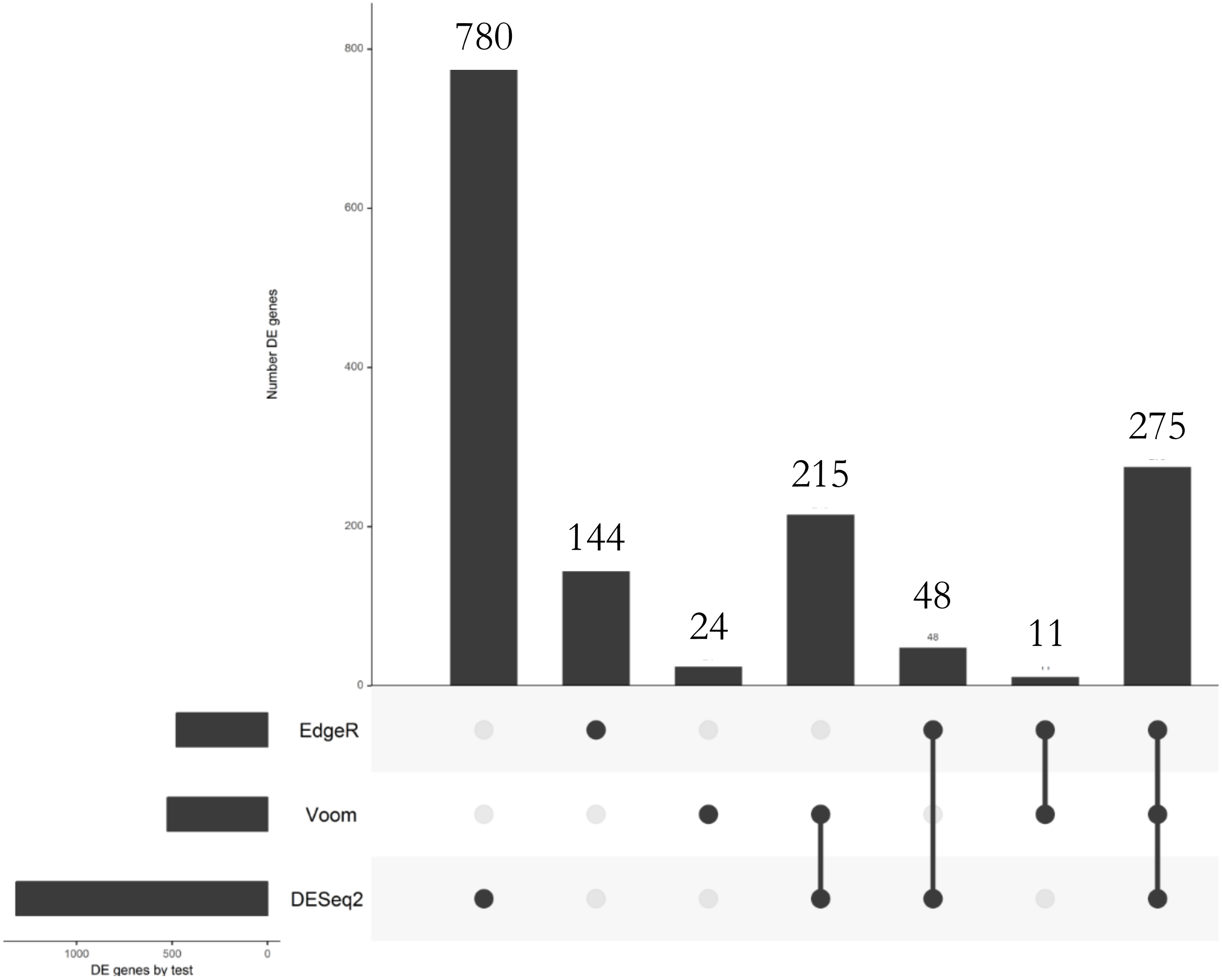


- VOOM uses log2 of counts per million normalization factor

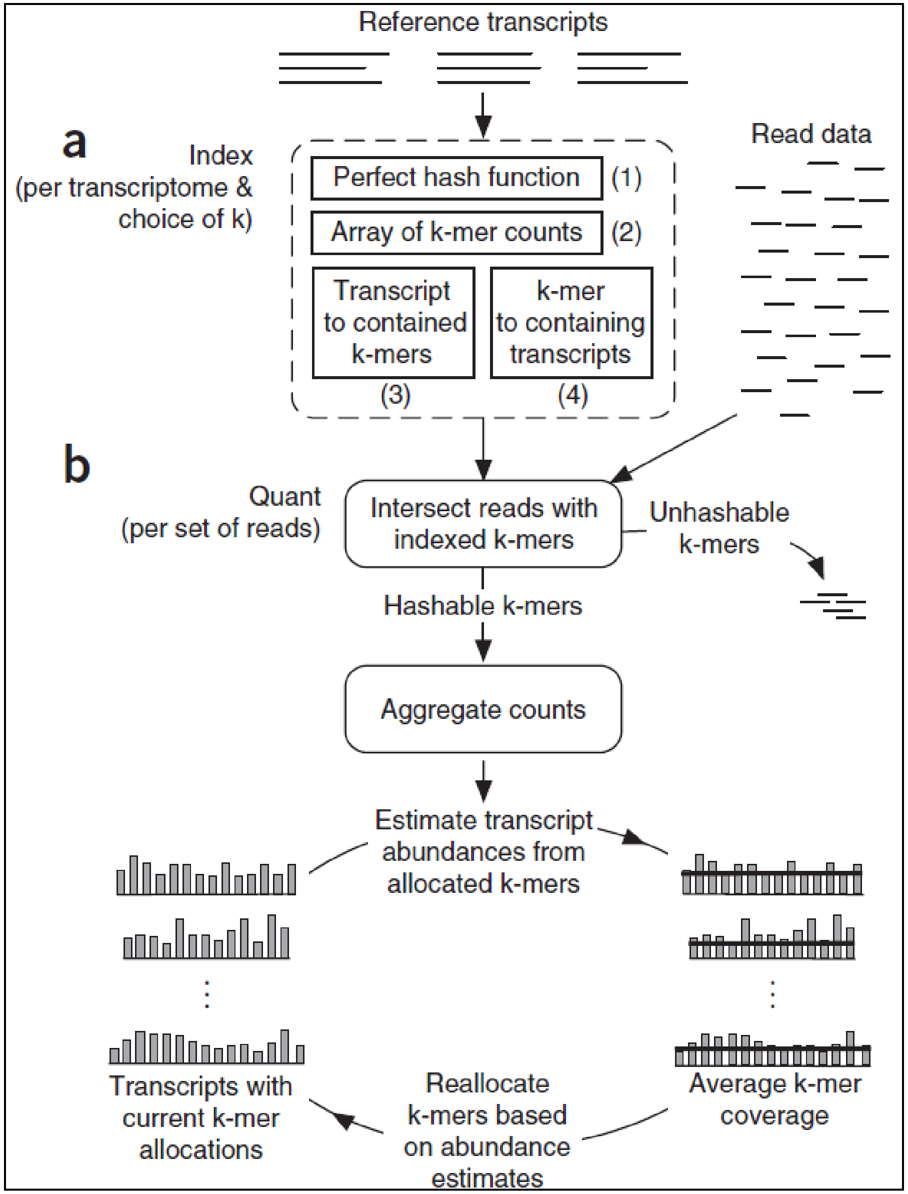
- Differential expression using the empirical Bayes limma pipeline

*2014 Law *et al*. PMCID: PMC4053721

How similar are gene DE algorithms?



Transcript k-mer modeling - sailfish


Reference transcripts
a
Index
(per transcriptome & choice of k)
Perfect hash function
(1)
Array of k-mer counts
(2)
Transcript to contained k-mers
(3)
k-mer to containing transcripts
(4)
Read data
b
Quant
(per set of reads)
Intersect reads with indexed k-mers
Unhashable k-mers
Hashable k-mers
Aggregate counts
Estimate transcript abundances from allocated k-mers
Transcripts with current k-mer allocations
Reallocate k-mers based on abundance estimates
Average k-mer coverage

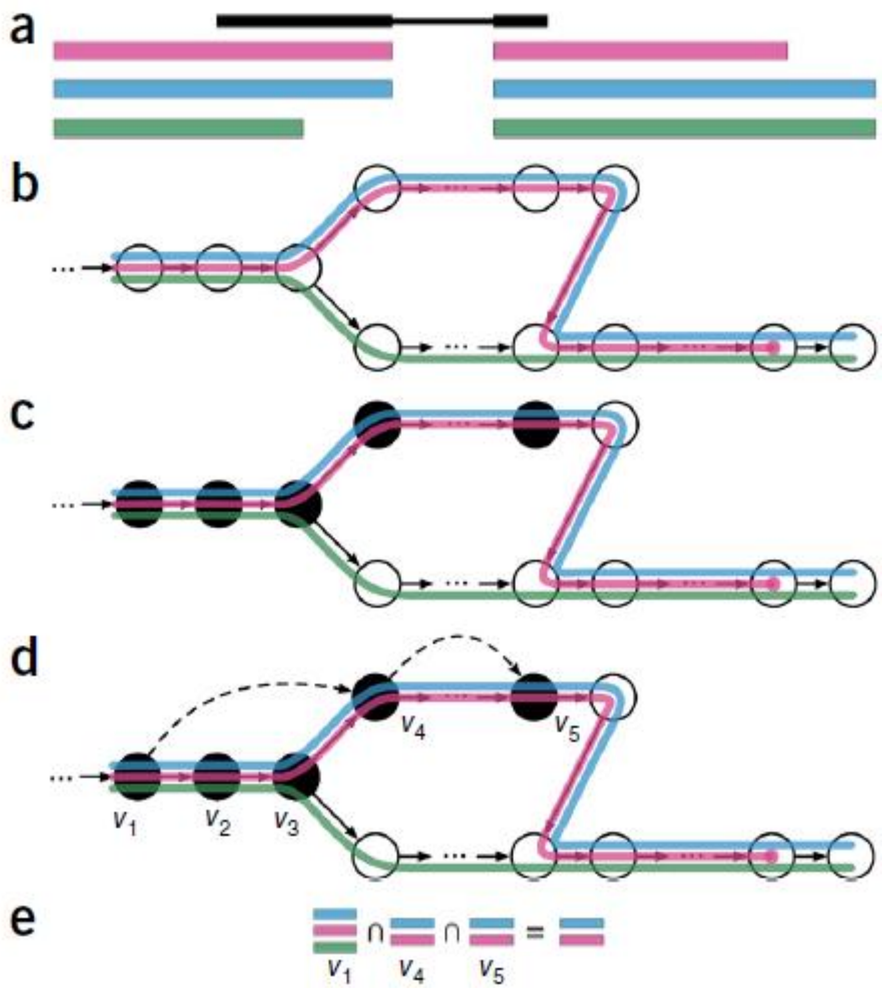
- Hashing uses a large amount of memory

- But lookups are blazingly fast

- Calculates TPMs

*2014 Patro *et al.* PMCID: PMC4077321

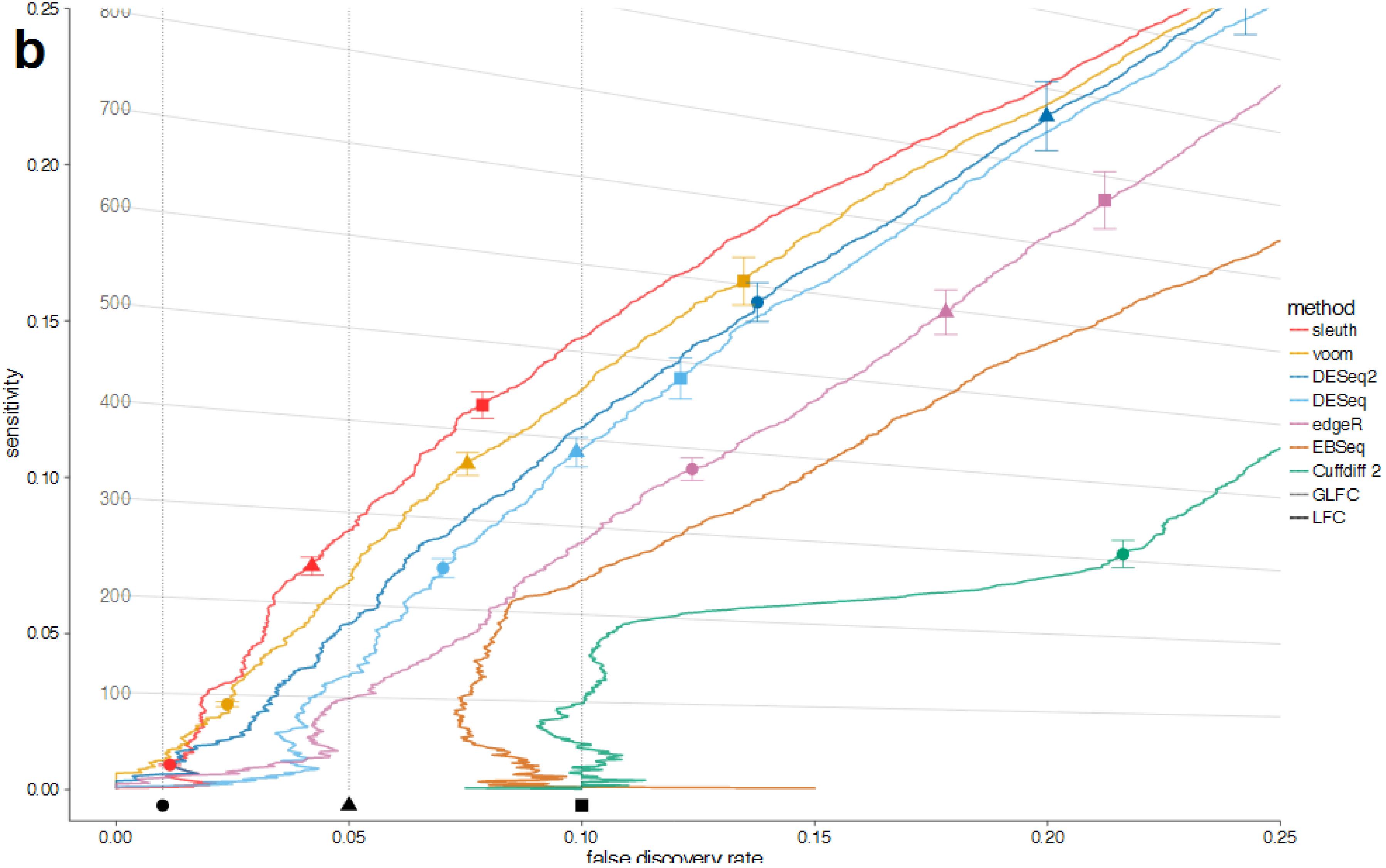
Transcript pseudoalignment - kallisto



- Builds a de Bruijn graph of transcript sequences
- Pseudoalignment – compatible transcripts, not where in transcript
- Very fast and efficient
- Allows for bootstrapping

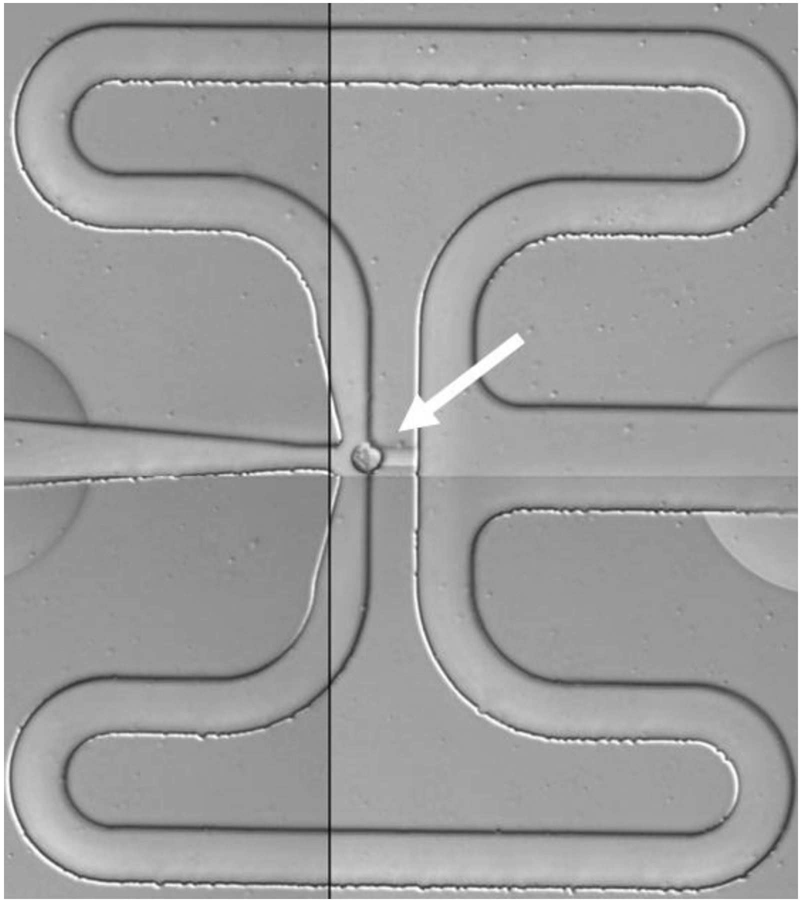
*2016 Bray *et al*. PMID: 27043002

Transcript DE from kallisto - Sleuth



*2016 Pimentel *et al.* http://dx.doi.org/10.1101/058164

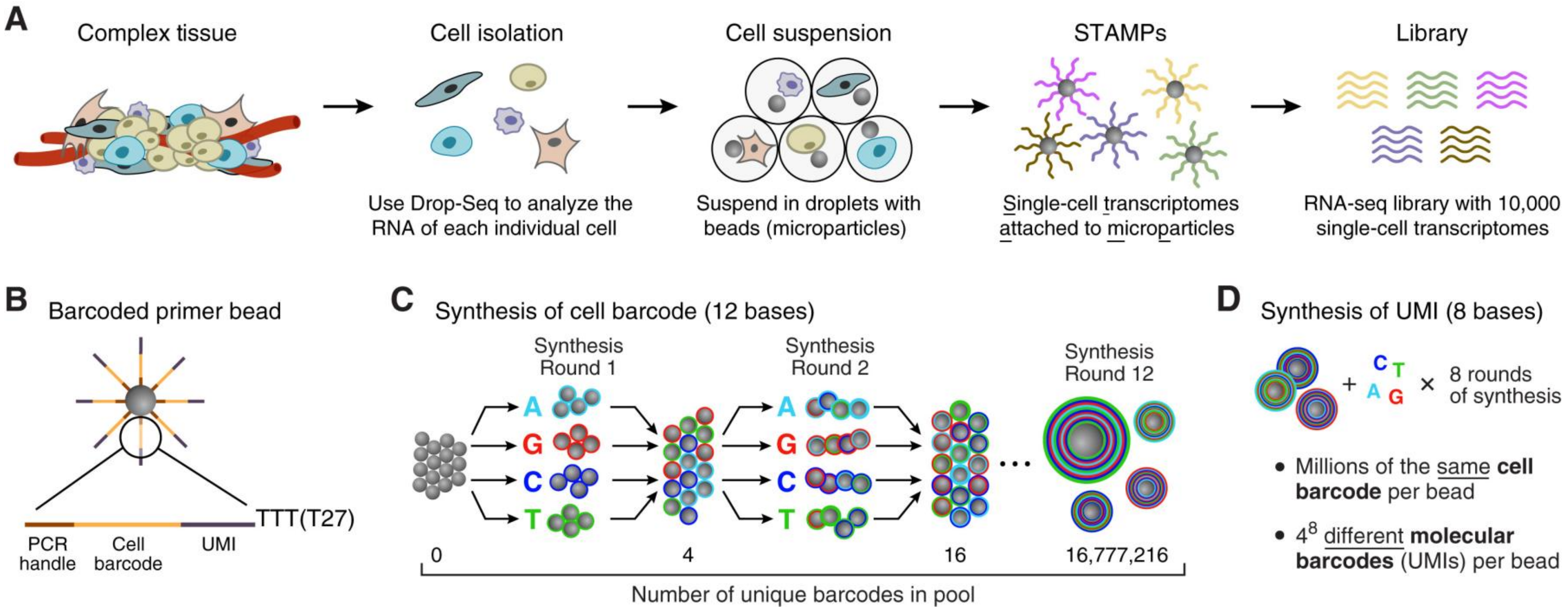
Single cell RNASeq – Fluidigm C1



- Microfluidics capture single cells
- Lyse cells and generate cDNA
- Requires live cells
- 96-800 cells / chip

http://biotech.illinois.edu/functionalgenomics/services-equipment/fluidigmsystem/singlecellautoprepsystem

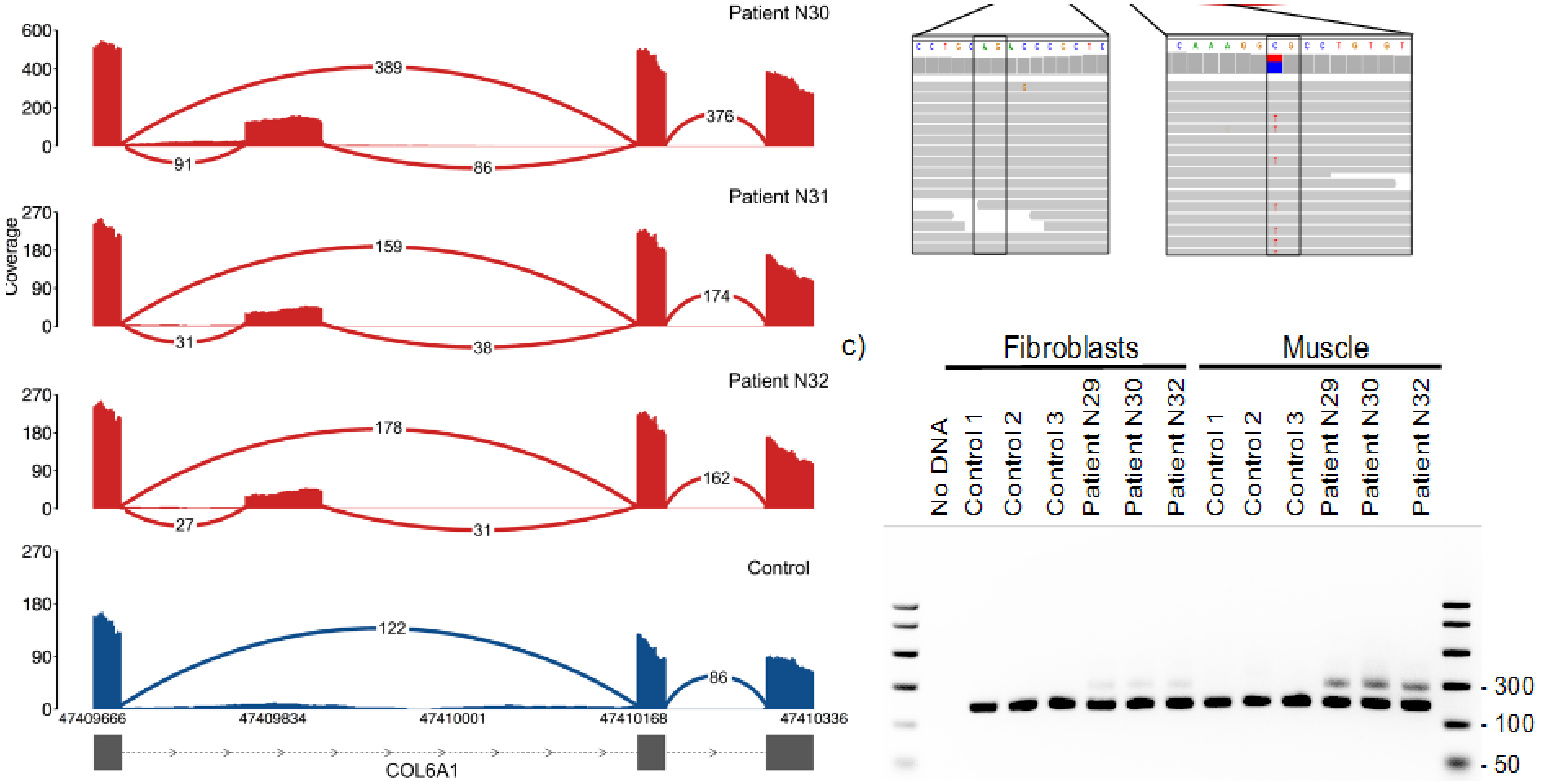
Single cell RNASeq – DropSeq


A
Complex tissue
Cell isolation
Use Drop-Seq to analyze the RNA of each individual cell
Cell suspension
Suspend in droplets with beads (microparticles)
STAMPs
Single-cell transcriptomes attached to microparticles
Library
RNA-seq library with 10,000 single-cell transcriptomes
B
Barcoded primer bead
TTT(T27)
PCR handle
Cell barcode
UMI
C
Synthesis of cell barcode (12 bases)
Synthesis Round 1
Synthesis Round 2
Synthesis Round 12
A
G
C
T
0
4
16
16,777,216
Number of unique barcodes in pool
D
Synthesis of UMI (8 bases)
C T A G
× 8 rounds of synthesis
Millions of the same cell barcode per bead
4^8 different molecular barcodes (UMIs) per bead

- Flows beads in a droplet.
- Cells, usu singles, merge into a droplet for library prep.
- 10s of thousands of cells.

* Macosko *et al*. 2015. PMCID: PMC4481139

RNA-Seq to diagnose Mendelian disease



* Cummings *et al*. 2015. http://biorxiv.org/content/biorxiv/early/2016/09/08/074153.full.pdf

Summary

- RNA-Seq vs. microarrays
  - Microarray requires knowing target sequence
  - Poor dynamic range (hard to detect low-level expressors, saturate at high-levels)

- Sample collection & storage
  - Blood should be spun down to PBMCs
    - Can be directly lysed
  - Tissue should have RNAlater applied soon as possible, followed by disruption.

Summary

- cDNA / library prep
  - Polyadenylation library methods selectively capture polyA transcripts
  - Ribosomal depletion methods degrade ribosomal RNA, but leave non-polyadenylated
  - Strandedness
    - There are an appreciable number of genes with antisense transcripts.
    - Also useful for identifying genes in species without a reference genome

Summary

- Sequencing technologies
  - Long reads (expensive), but sequence full isoform
  - Short read. Reasonable price.
- Aligner
  - Must use an aligner that is aware of introns
  - May align to either genes (STAR, etc) or transcriptome (Tophat 2 and kallisto)
- Counting
  - Subread featureCounts, HTSeq count, RSEM

Summary

- Differential expression
  - Gene, negative binomial: edgeR, DESeq2
  - Gene, log2 counts per million: VOOM
  - Transcript, TPM, kallisto.
  - Transcript, TPM, Sailfish

- Single-cell
  - Isolate a single cell and make a libraries (low-output)
  - Spike-ins help
  - UMIs help also