University of Massachusetts Medical School

UMASS

RNA-Seq primer

Sequencing: applications

Counting applications

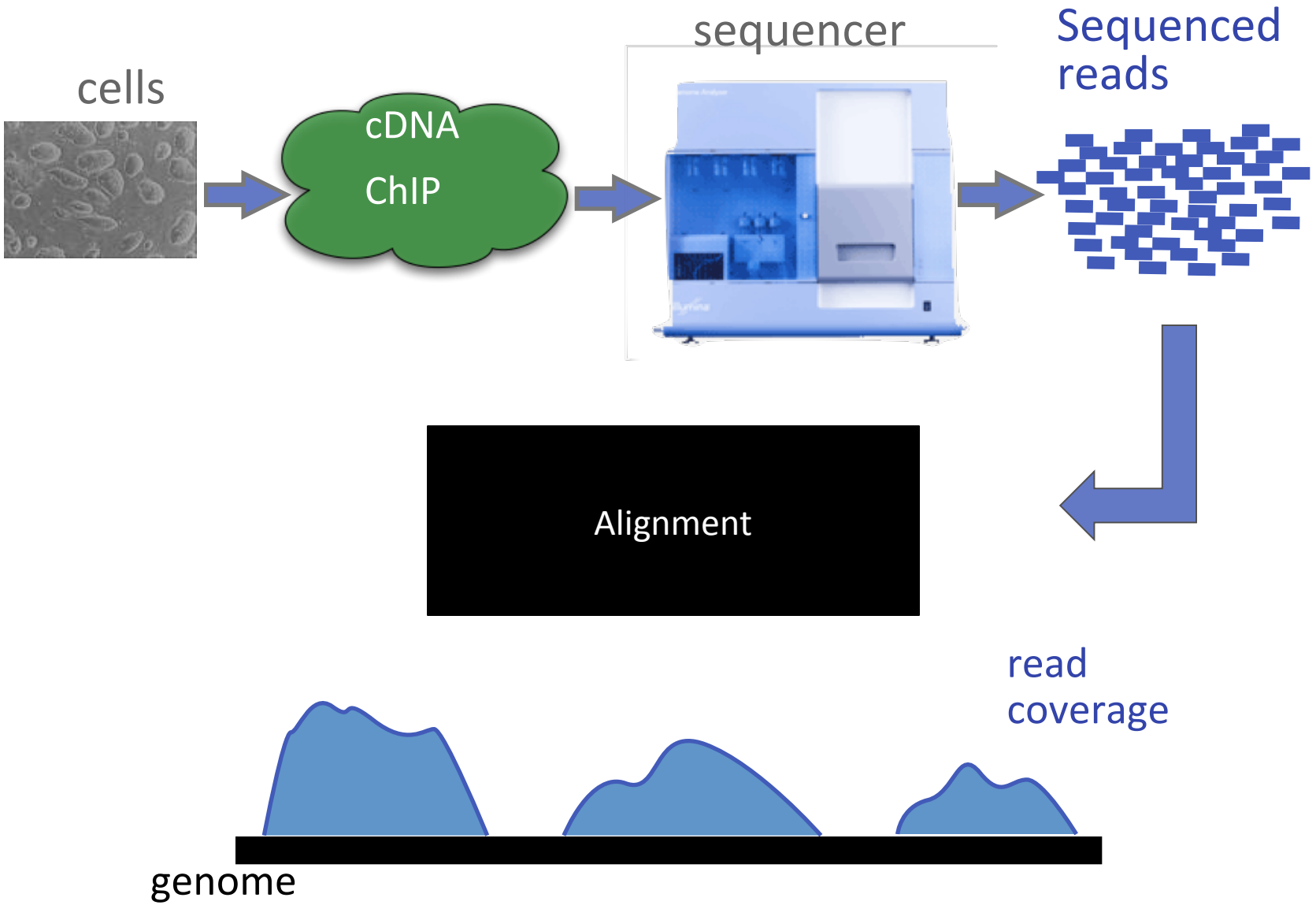
- Profiling
  - microRNAs
  - Immunogenomics
  - Transcriptomics
- Epigenomics
  - Map histone modifications
  - Map DNA methylation
  - 3D genome conformation
- Nucleic acid Interactions

Polymorphism/mutation discovery

  - Bacteria
  - Genome dynamics
  - Exon (and other target) sequencing
  - Disease gene sequencing
- Variation and association studies
- Genetics and gene discovery

- Cancer genomics
  - Map translocations, CNVs, structural changes
  - Profile somatic mutations
- Genome assembly
- Ancient DNA (Neanderthal)
- Pathogen discovery
- Metagenomics

Counting applications



Sequencing libraries to probe the genome

- RNA-Seq
  - Transcriptional output
  - Annotation
  - miRNA
  - Ribosomal profiling
- ChIP-Seq
  - Nucleosome positioning
  - Open/closed chromatin
  - Transcription factor binding
- CLIP-Seq
  - Protein-RNA interactions
- Hi-C
  - 3D genome conformation

RNA-Seq libraries I: "Standard" full-length

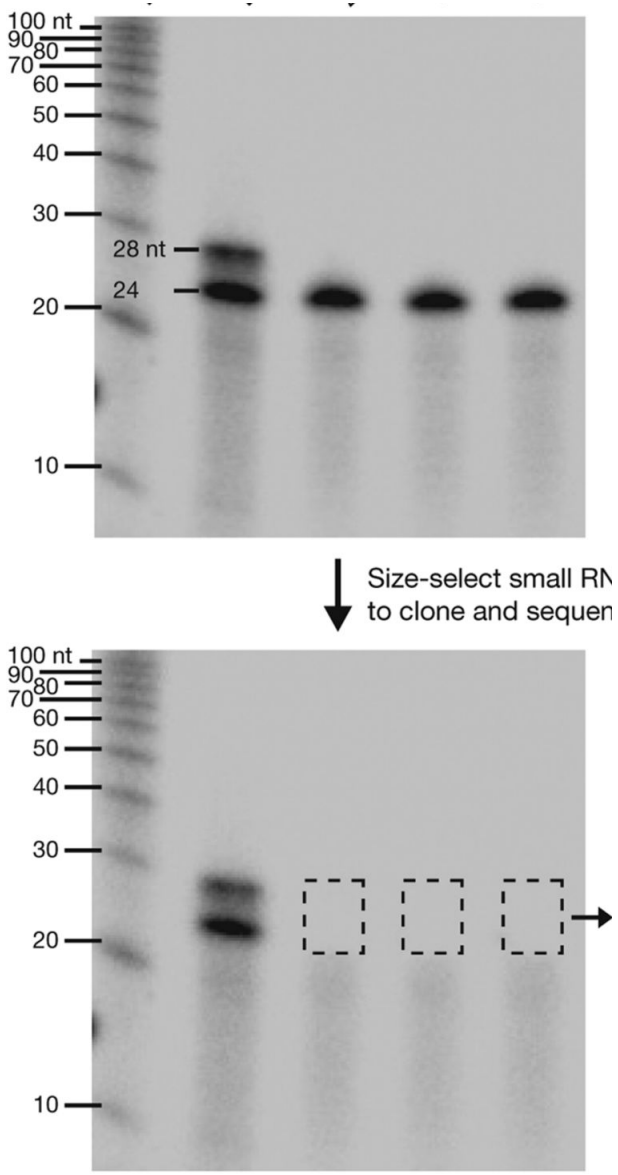
- "Source: intact, **high qual**. RNA (polyA selected or ribosomal depleted)
- RNA → cDNA → sequence
- Uses:
  - Annotation. Requires high depth, paired-end sequencing. ~50 mill
  - Gene expression. Requires low depth, single end sequence, ~ 5-10 mill
  - Differential Gene expression. Requires ~ 5-10 mill, at least 3 replicates, single end

RNA-Seq libraries II: 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

RNA-Seq libraries III: Small RNA libraries

- Source: size selected RNA
- Uses: miRNA, piRNA annotation and quantification
  - Short single end 30-50 bp reads
  - Require "clipping"
  - Depth: 5-10 mill reads

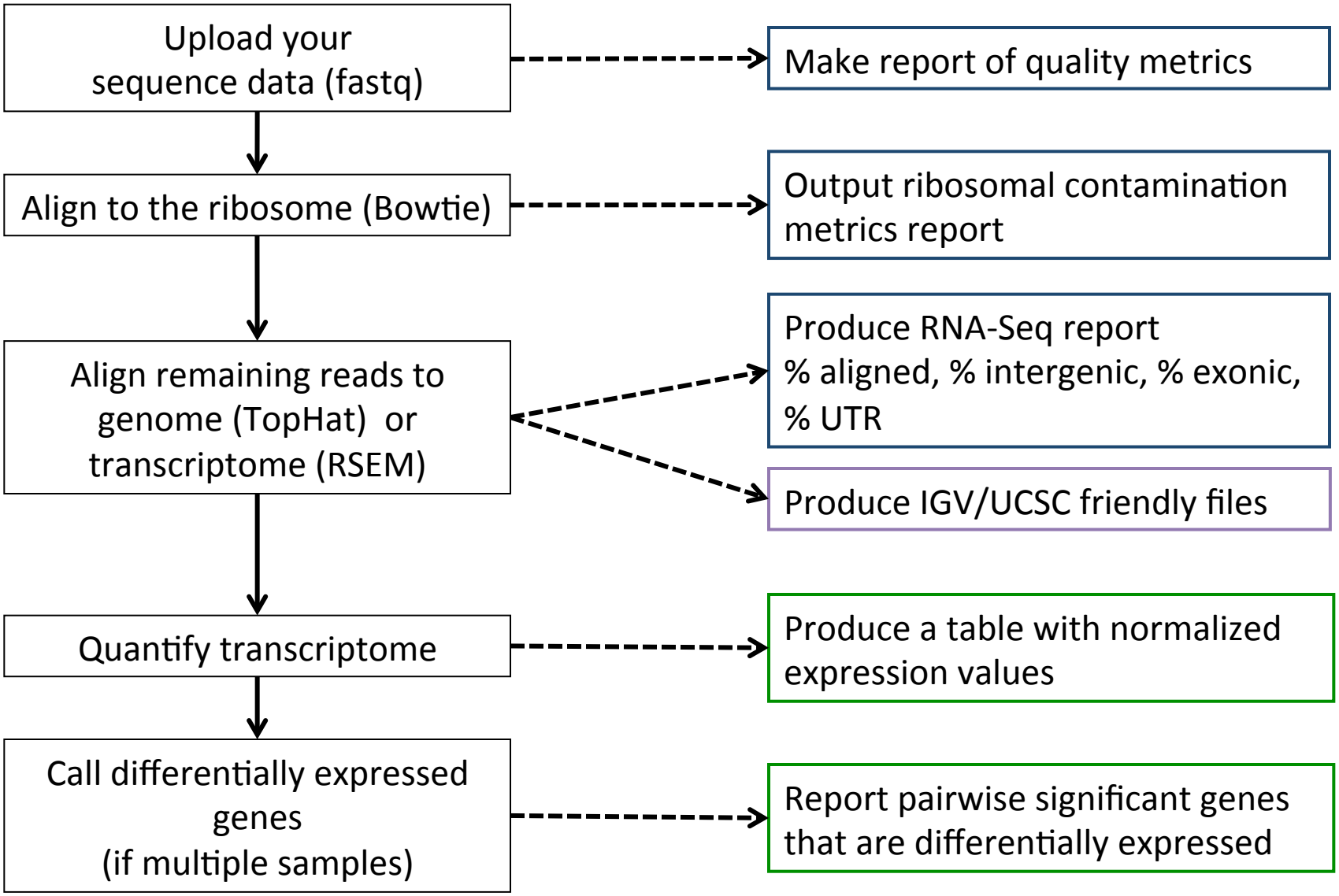


Malonne et al. CSHL protocols, 2011

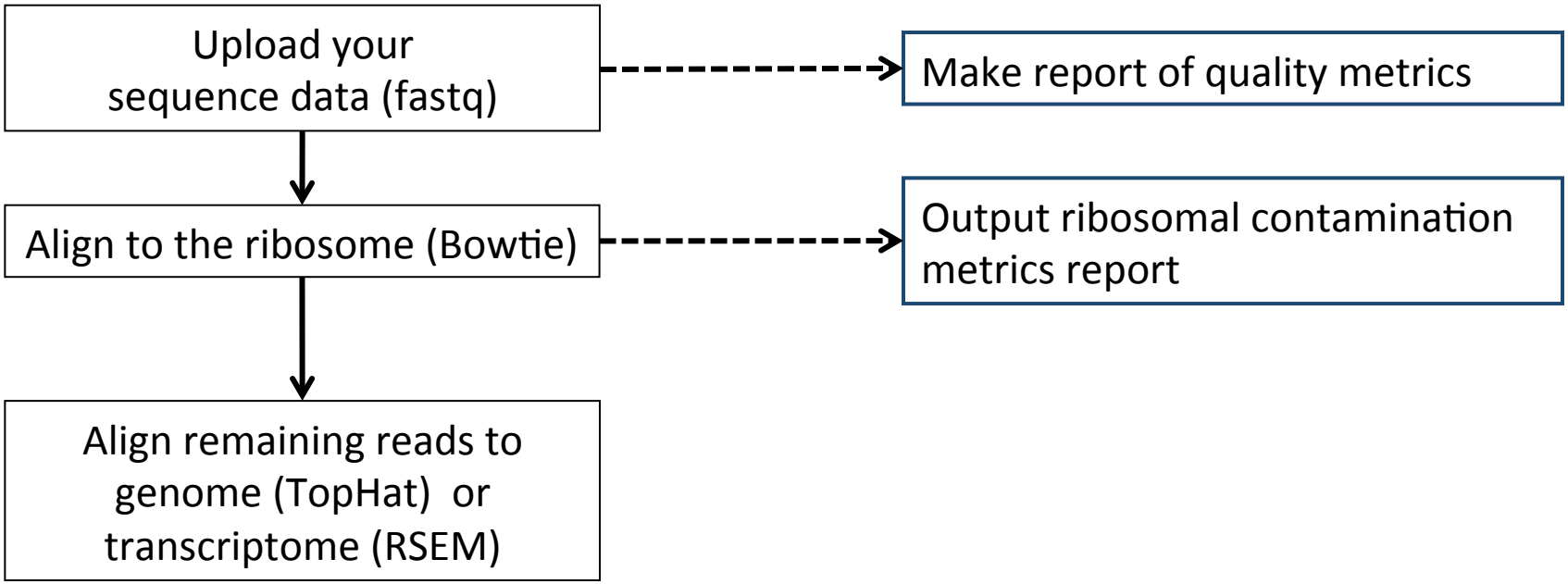
When you need both annotation and quantification

- Attempt three replicates per condition
- Pool libraries to obtain ~15 mill reads per replicate
- Sequence using paired ends
- Analysis:
  - Merge replicate alignments for annotation
  - Split alignments for differential expression analysis

Our typical RNA quantification pipeline

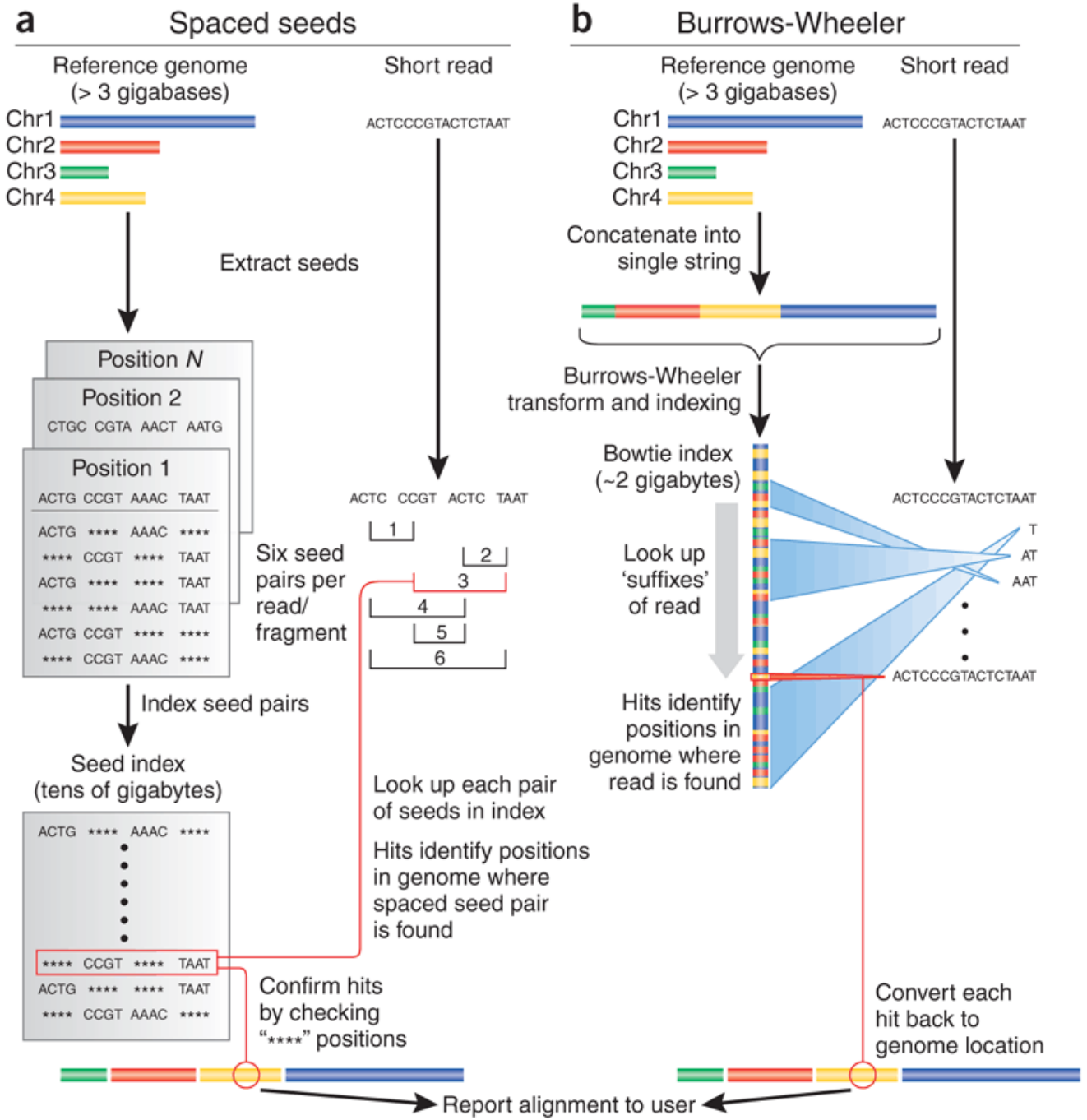


Alignment requires pre-processing




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

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



Trapnell, Salzberg, Nature Biotechnology 2009

Spaced seed alignment – Hashing the genome

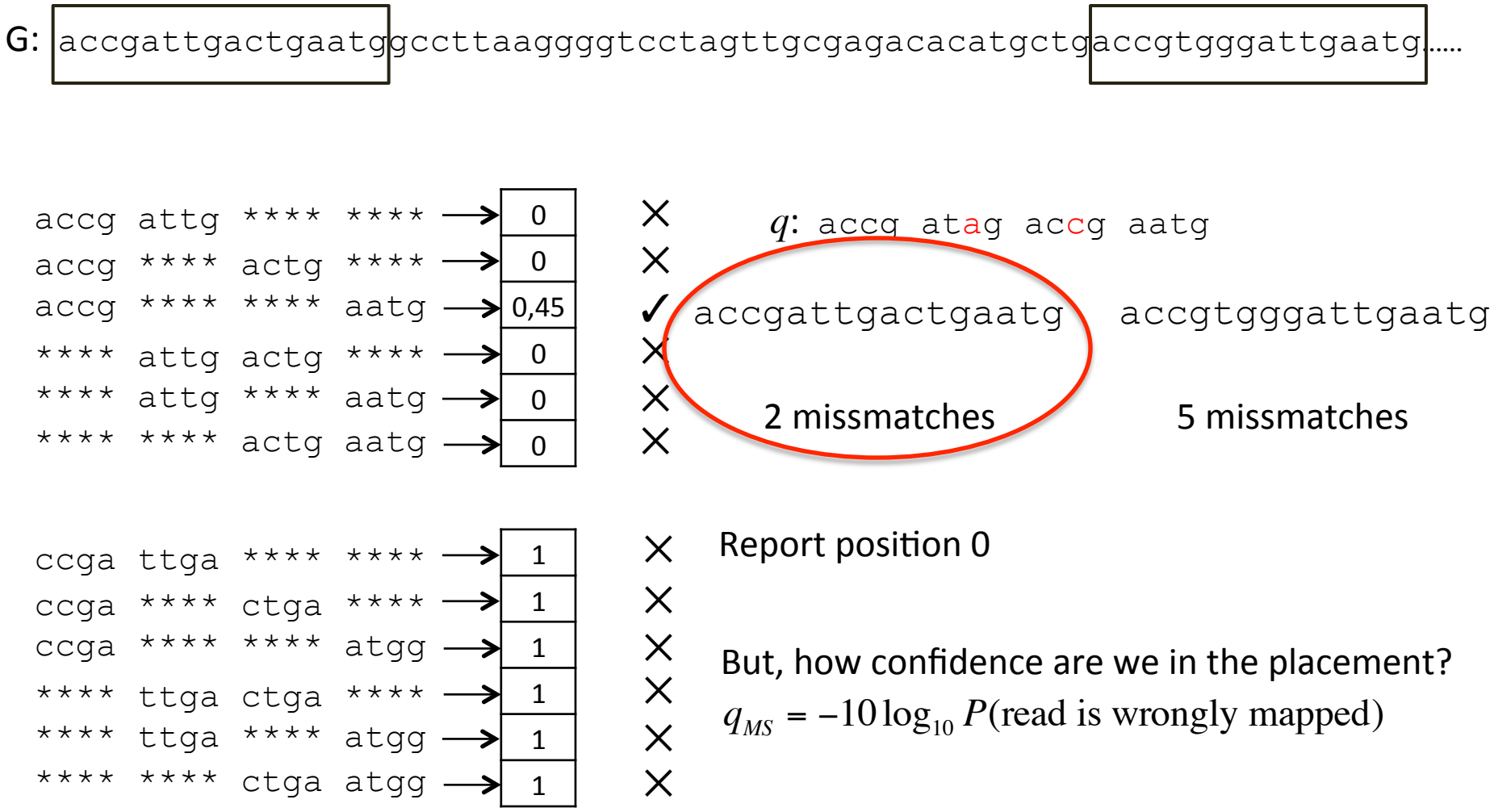
G: accgattgactgaatggccttaaggggtcctagttgcgagacacatgctgaccgtgggattgaatg……

Store spaced seed positions

accg attg **** ****	0
accg **** actg ****	0
accg **** **** aatg	0,45
**** attg actg ****	0
**** attg **** aatg	0
**** **** actg aatg	0

ccga ttga **** ****	1
ccga **** ctga ****	1
ccga **** **** atgg	1
**** ttga ctga ****	1
**** ttga **** atgg	1
**** **** ctga atgg	1

Spaced seed alignment – Mapping reads



Mapping quality

What does $q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$ mean?

Lets compute the probability the read originated at genome position i

q: accg atag accg aatg

q_s: 30 40 25 30 30 20 10 20 40 30 20 30 40 40 30 25

$q_s[k] = -10\log_{10} P(\text{sequencing error at base k})$, the PHRED score. Equivalently:

$$P(\text{sequencing error at base k}) = 10^{-\frac{q_s[k]}{10}}$$

So the probability that a read originates from a given genome position i is:

$$P(q \mid G,i) = \prod_{j\ \text{match}} P(q_j \text{good call}) \prod_{j\ \text{missmatch}} P(q_j \text{bad call}) \approx \prod_{j\ \text{missmatch}} P(q_j \text{bad call})$$

In our example

$$P(q \mid G,0) = \left[(1-10^{-3})^6 (1-10^{-4})^4 (1-10^{-2.5})^2 (1-10^{-2})^2\right]\left[10^{-1}10^{-2}\right] = [0.97]*[0.001] \approx 0.001$$

Mapping quality

What we want to estimate is $q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$

That is, the posterior probability, the probability that the region starting at i was sequenced *given* that we observed the read *q*:

$$P(i \mid G,q) = \frac{P(q \mid G,i)P(i \mid G)}{P(q \mid G)} = \frac{P(q \mid G,i)P(i \mid G)}{\sum_j P(q \mid G,j)}$$

Fortunately, there are efficient ways to approximate this probability (see Li, H *genome Research* 2008, for example)

$$q_{MS} = -10\log_{10}(1 - P(i \mid G,q))$$

Considerations

- Trade-off between sensitivity, speed and memory
  - Smaller seeds allow for greater mismatches at the cost of more tries
  - Smaller seeds result in a smaller tables (table size is at most 4^k), larger seeds increase speed (less tries, but more seeds)

Short read mapping software

Seed-extend

	Short indels	Use base qual
Maq	**No**	**YES**
RMAP	Yes	**YES**
SeqMap	Yes	NO
SHRiMP	Yes	NO

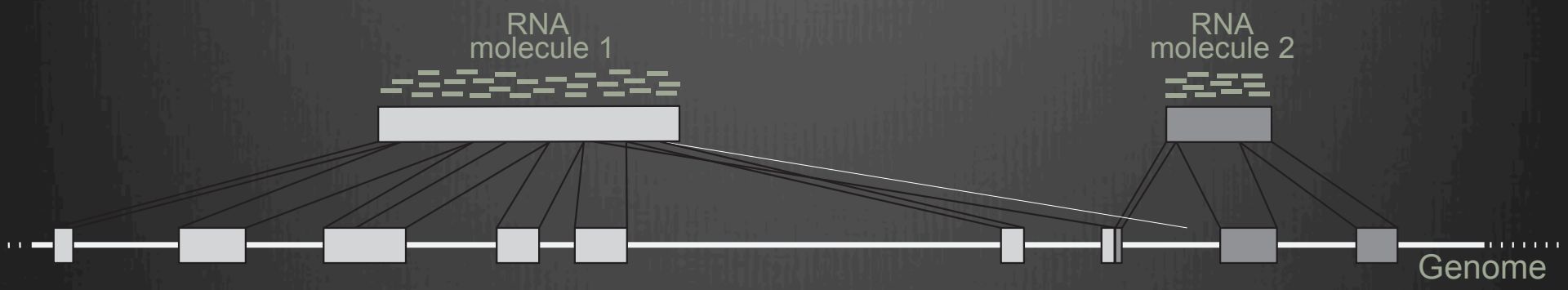
BWT

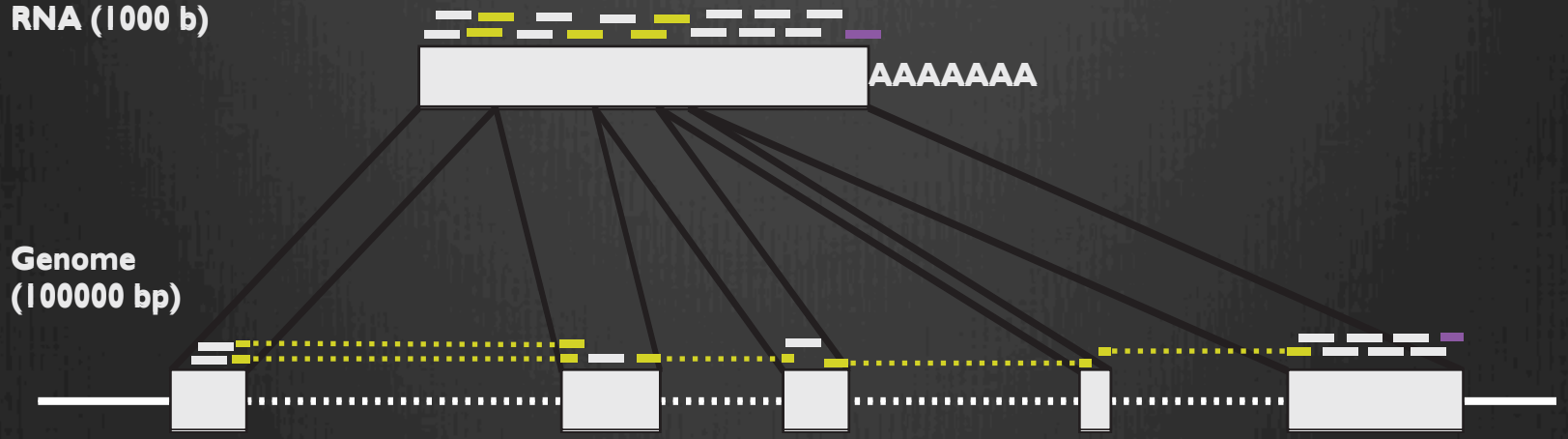
	Use Base qual
BWA	**YES**
Bowtie	NO
Stampy*	YES
Bowtie2*	(NO)

***Stampy is a hybrid approach which first uses BWA to map reads then uses seed-extend only to reads not mapped by BWA**

***Bowtie2 breaks reads into smaller pieces and maps these "seeds" using a BWT genome.**

RNA-Seq Read mapping





Mapping RNA-Seq reads: Seed-extend spliced alignment (e.g. GSNAP)

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

Short read mapping software for RNA-Seq

Seed-extend

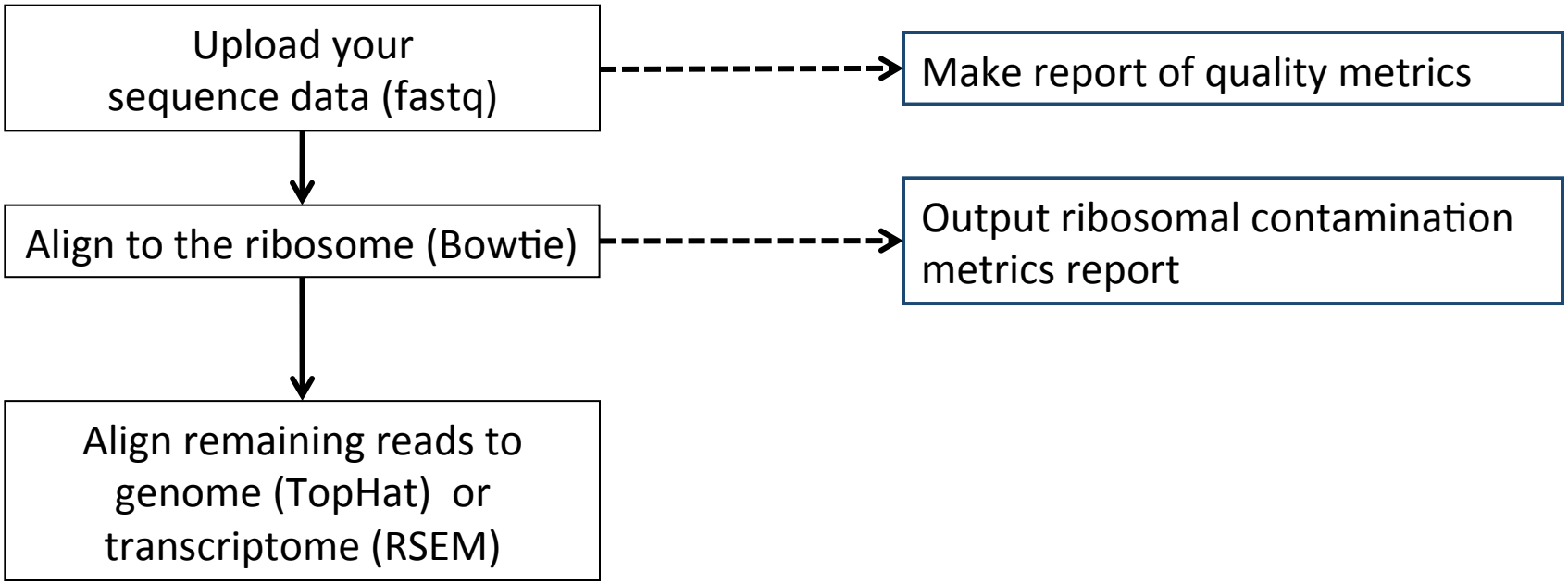
	Short indels	Use base qual
GSNAP	**Yes**	?
QPALMA	Yes	NO
BLAT	Yes	NO

Exon-first

	Use base qual
STAR	NO
TopHat	NO

Exon-first alignments will map contiguous first at the expense of spliced hits

Alignment requires pre-processing




```
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
```

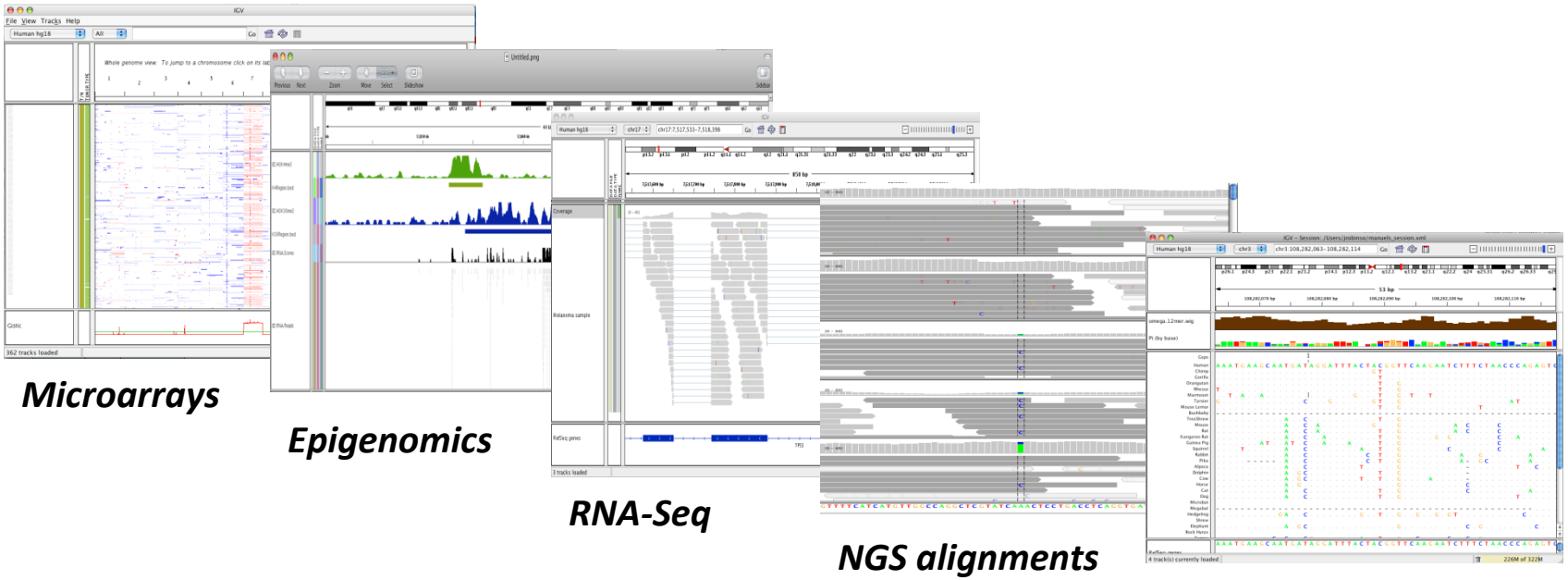
IGV: Integrative Genomics Viewer

Integrative Genomics Viewer

A desktop application

for the visualization and interactive exploration

of genomic data



Microarrays

Epigenomics

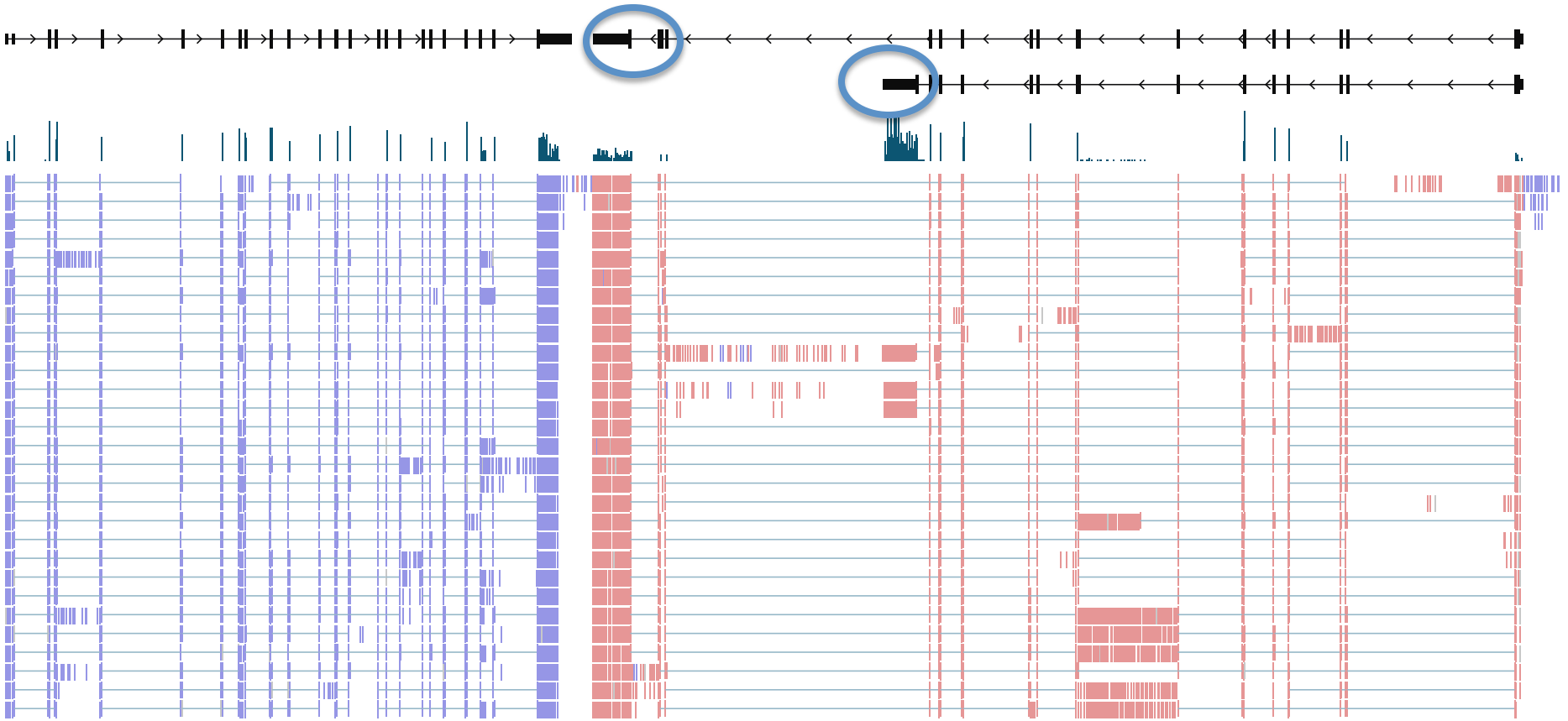
RNA-Seq

NGS alignments

Comparative genomics

BROAD INSTITUTE

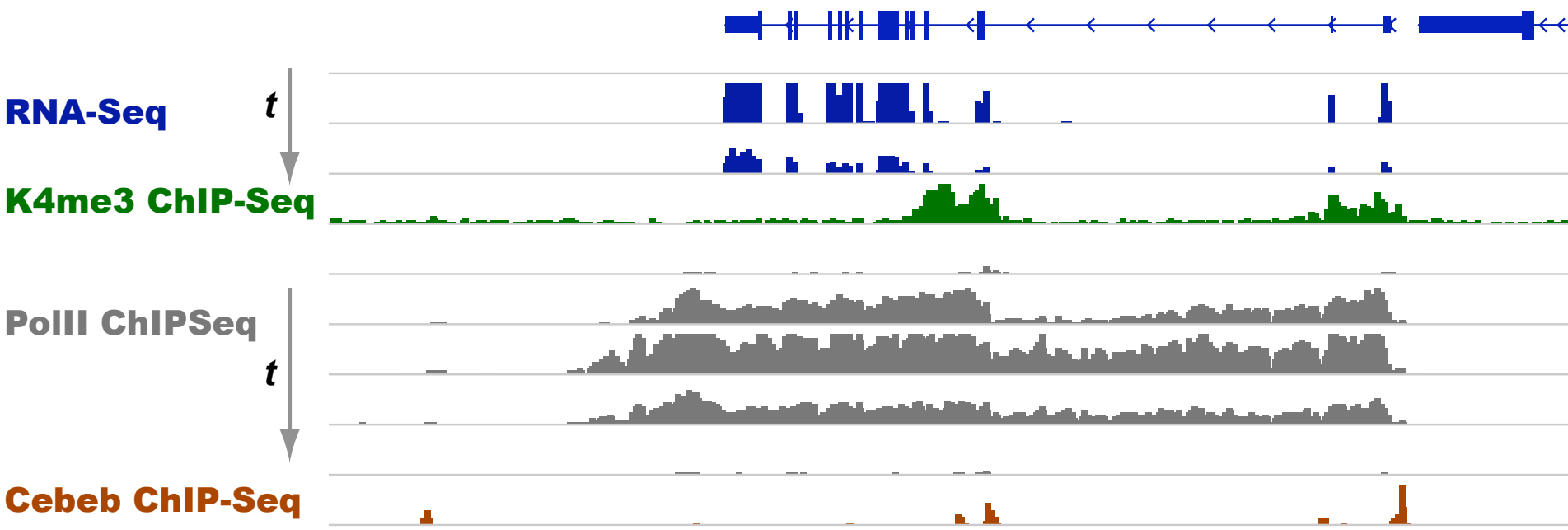
Visualizing read alignments with IGV — RNASeq



Strand specific library!

Gap between reads spanning exons

Visualizing read alignments with IGV — zooming out

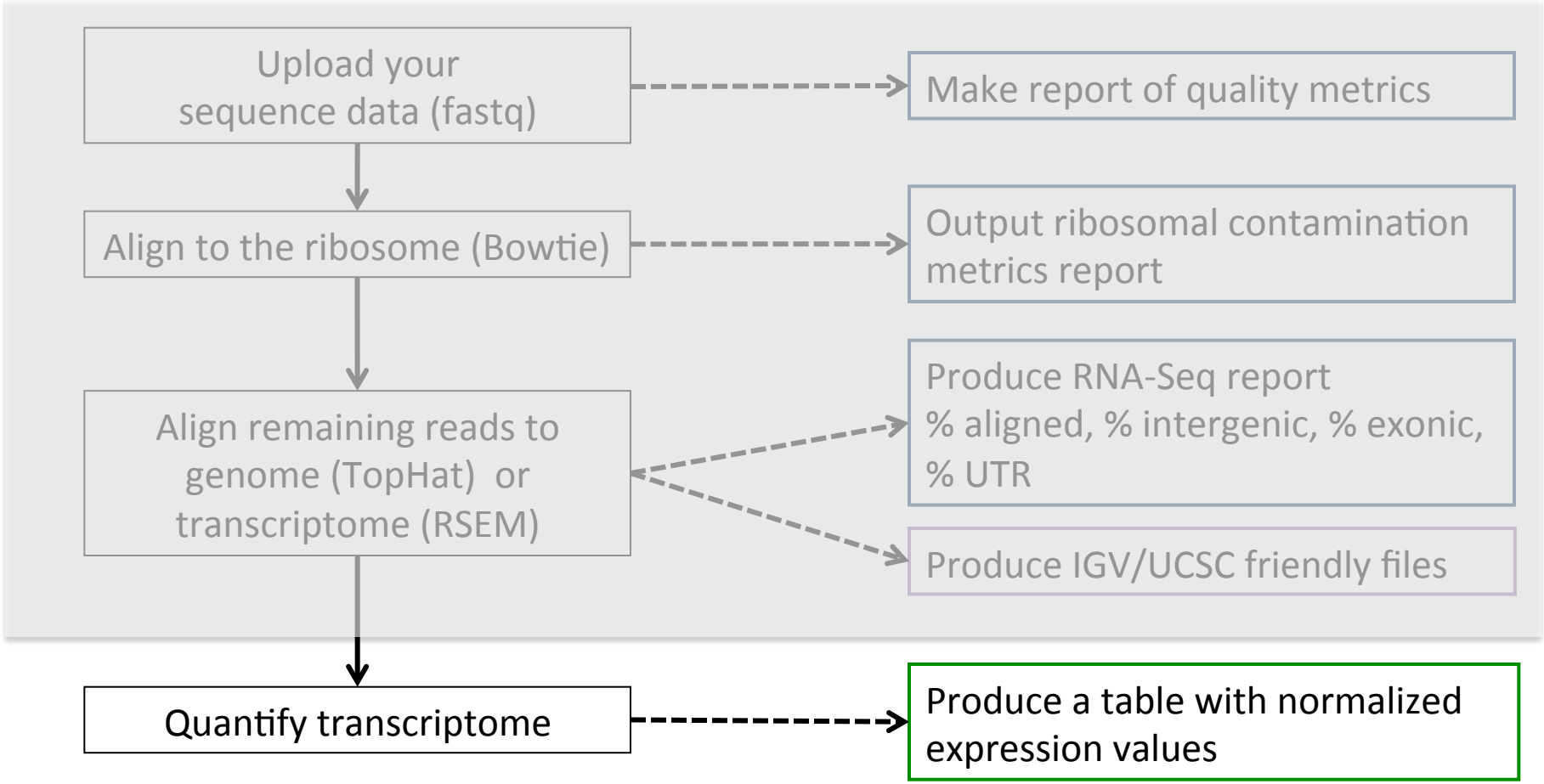


How do "short" read aligners responded to read increase?

- Break reads into seeds (e.g. 16nt every 10nt)
- Use BWT or HashTable to find candidate positions
- Prioritize candidates
- Extend top candidates using classical alignment techniques.

Aligner	Technique
TopHat2 (Bowtie2)	BWT
GSNAP	Hash Table
STAR	Suffix (similar to TopHat)

Computing gene expression

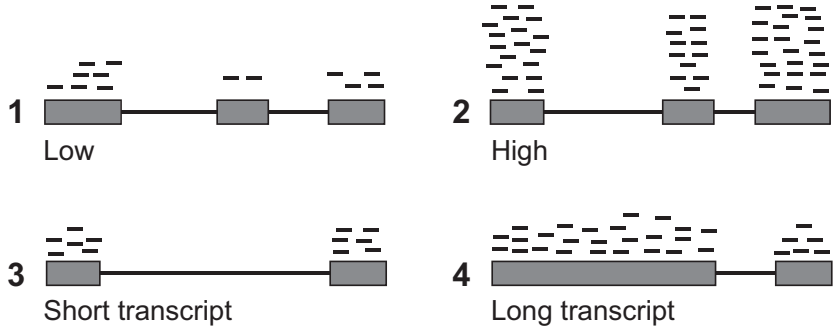



```
rsem-calculate-expression --paired-end --strand-specific -p 2 \
 --output-genome-bam fastq.quantification/control_rep1.1.fq \
fastq.quantification/control_rep1.2.fq genome.quantification/mm10.rsem \
rsem/ctrl1.rsem
```

RNA-Seq quantification

- Is a given gene (or isoform) expressed?
- Is expression gene A > gene B?
- Is expression of gene A isoform a_1 > gene A isoform a_2?
- Given two samples is expression of gene A in sample 1 > gene A in sample 2?

Quantification: only one isoform



$$RPKM = 10^9 \frac{\#reads}{length \times TotalReads}$$

Reads per kilobase of exonic sequence per million mapped reads
(*Mortazavi* et al Nature methods 2008)

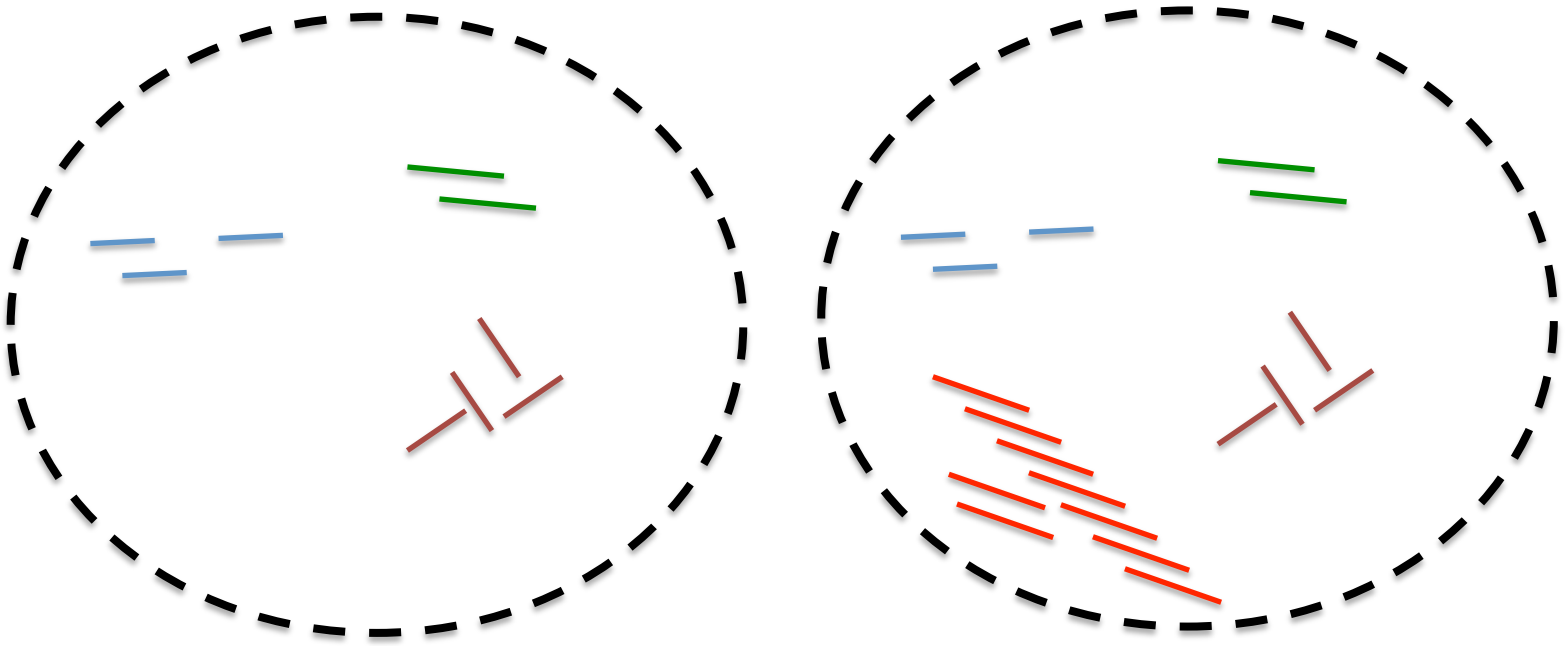
- Fragmentation of transcripts results in length bias: longer transcripts have higher counts
- Different experiments have different yields. Normalization is key for cross lane comparisons

Complexity increases when multiple isoforms exist

Normalization for comparing two different genes

- To compare within a sequence run (lane), RPKM accounts for length bias.
- RPKM is not optimal for cross experiment comparisons.
  - Different samples may have different compositions.
- FPKM superseded RPKM
- And later TPM = 10^6 x Fraction of transcript

Normalization for comparing a gene across samples

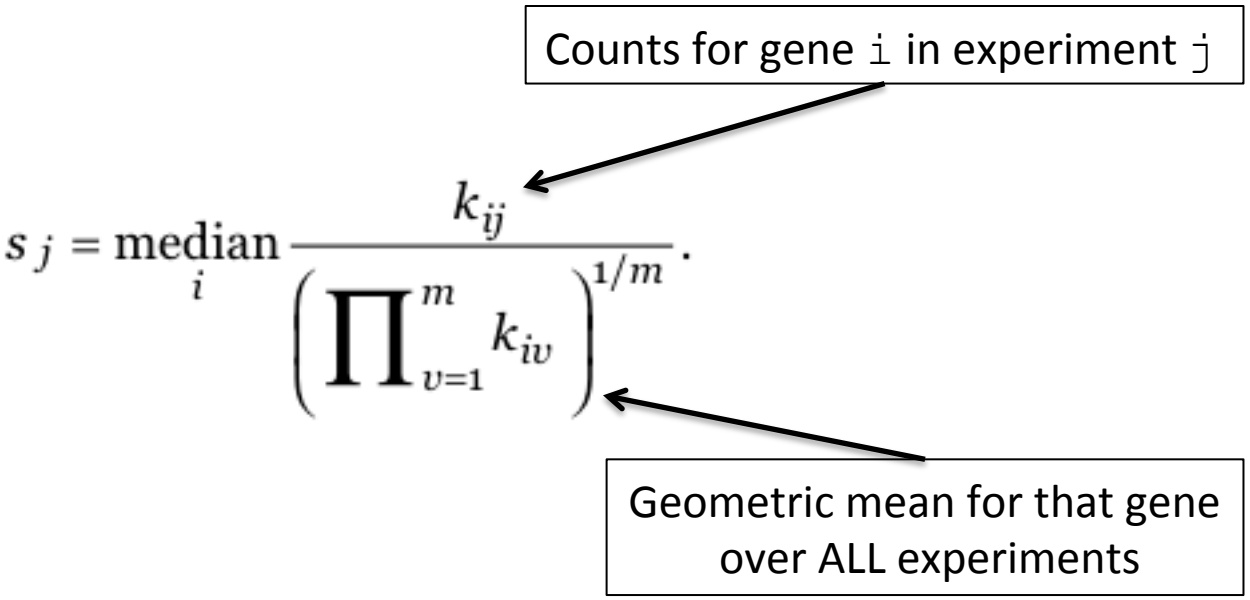


Cell type I

Cell type II

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

Step2: More robust normalization



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 (and do a short R practice)

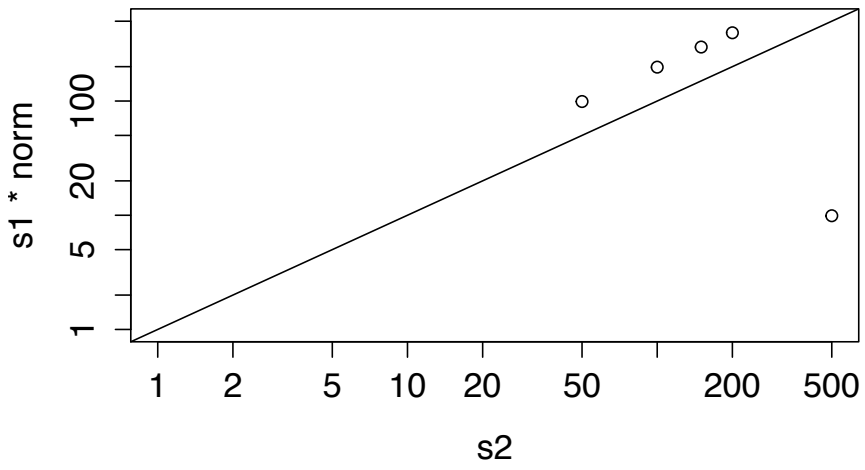
$> s1 = c(100, 200, 300, 400, 10)$
$> s2 = c(50, 100, 150, 200, 500)$

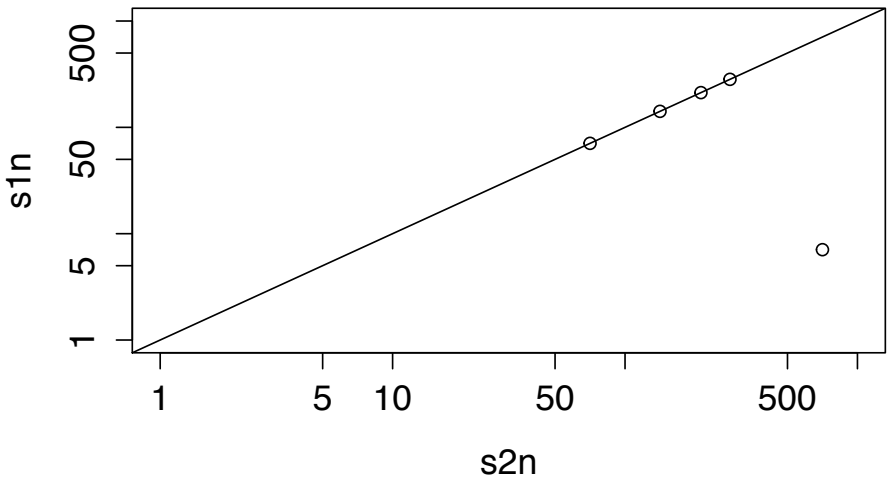
Similar read number,
one transcript many fold changed

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

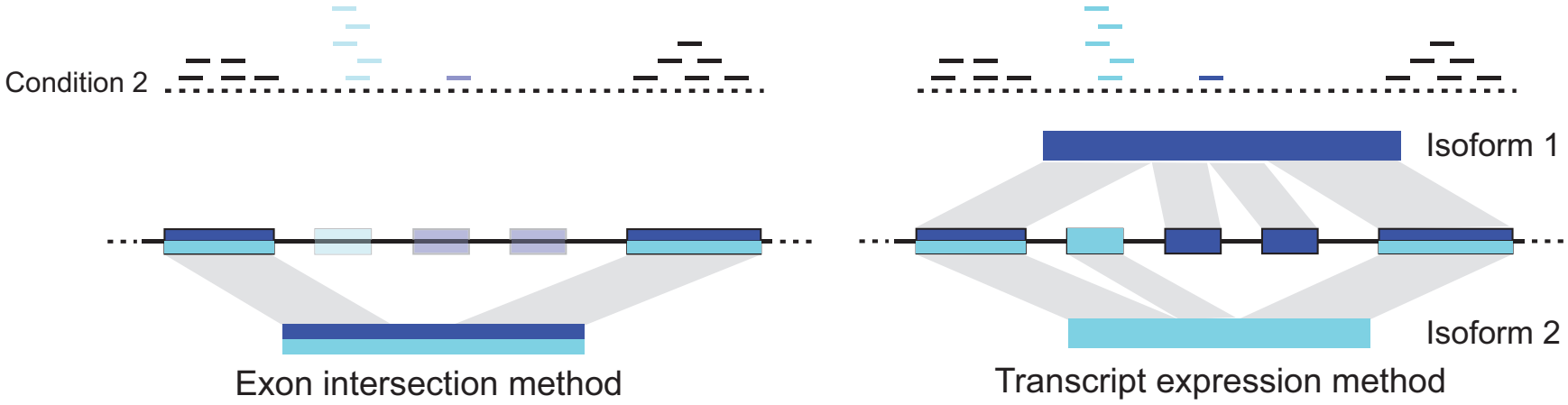
Size normalization results in 2-fold changes in *all* transcripts

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





But, how to compute counts for complex gene structures?



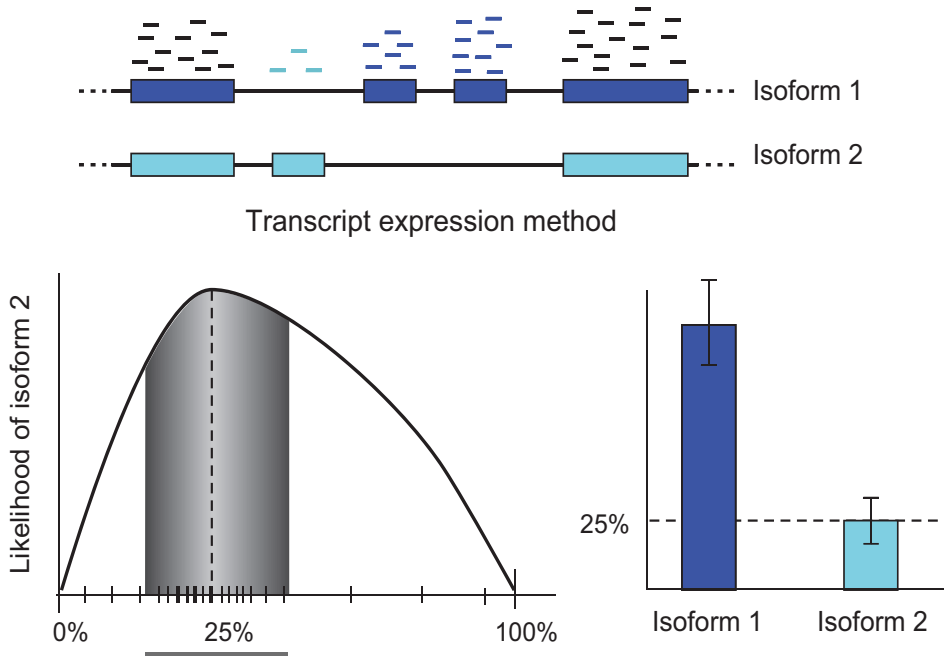
Three popular options:

Exon *intersection* model: Score constituent exons

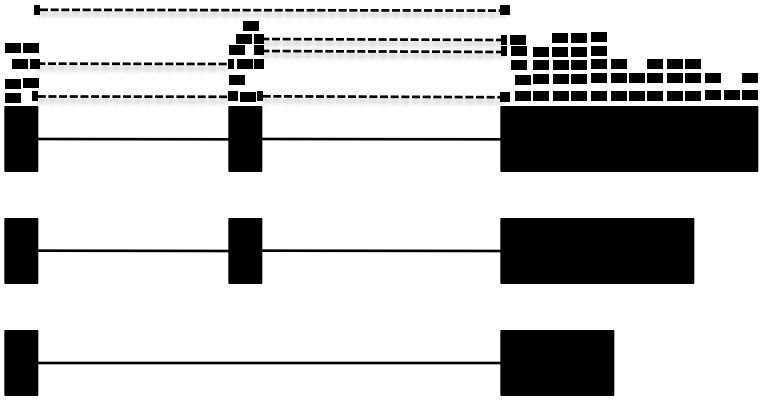
Exon *union* model: Score the the "merged" transcript

Transcript expression model: Assign reads uniquely to different isoforms. *Not a trivial problem!*

Quantification: read assignment method

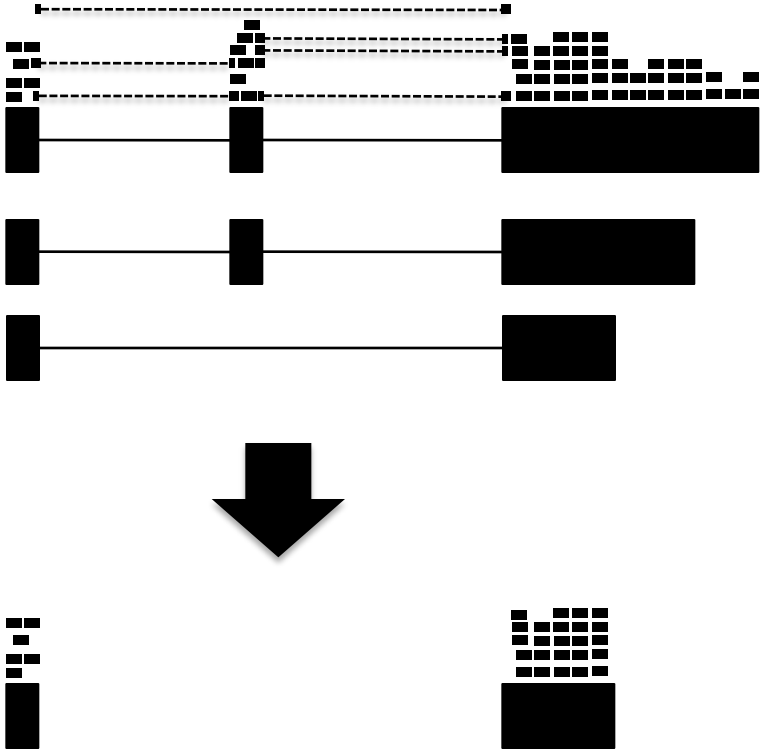


Quantification with multiple isoforms



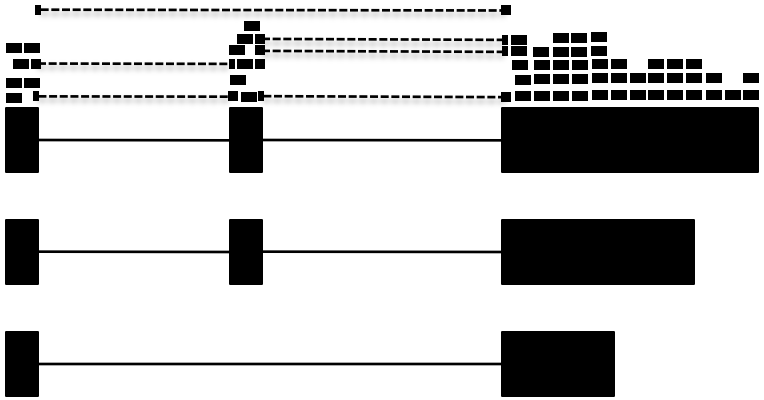
How do we define the gene expression?
How do we compute the expression of each isoform?

Computing gene expression

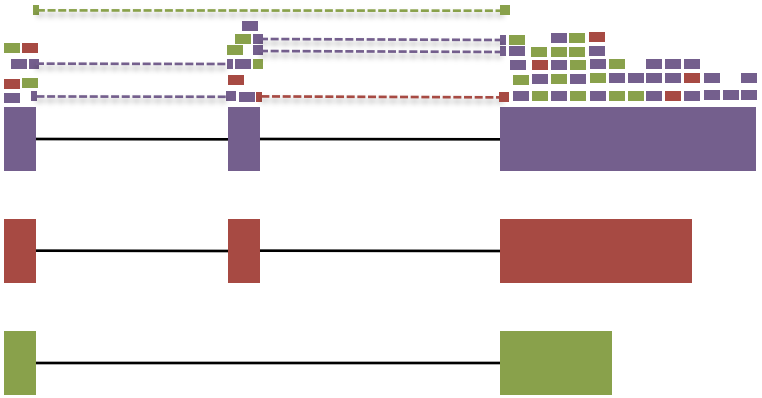


Idea1: RPKM of the constitutive reads (Neuma, Alexa-Seq, Scripture)

Computing gene expression — isoform deconvolution



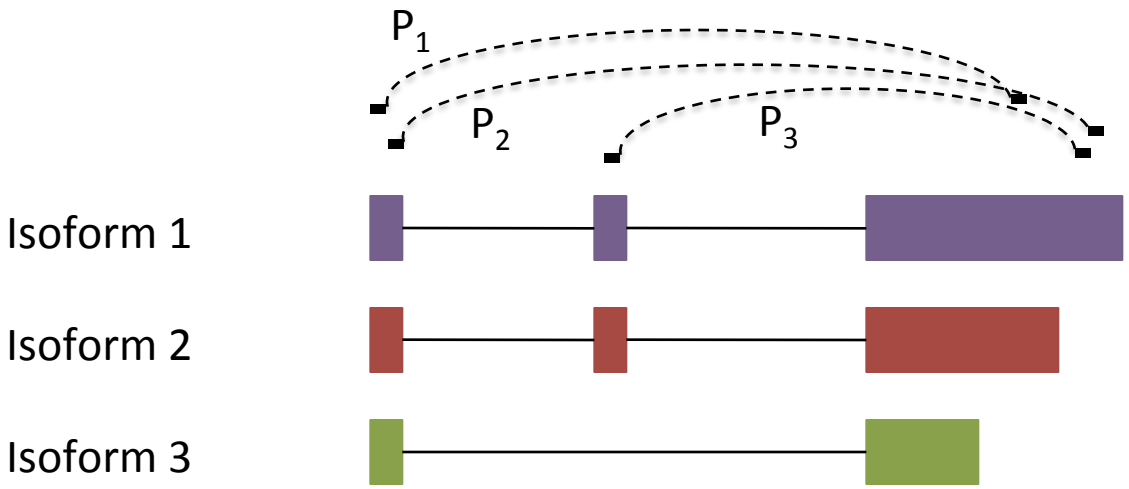
Computing gene expression — isoform deconvolution



If we knew the origin of the reads we could compute each isoform's expression. The gene's expression would be the sum of the expression of all its isoforms.

$$E = RPKM_1 + RPKM_2 + RPKM_3$$

Paired-end reads are easier to associate to isoforms

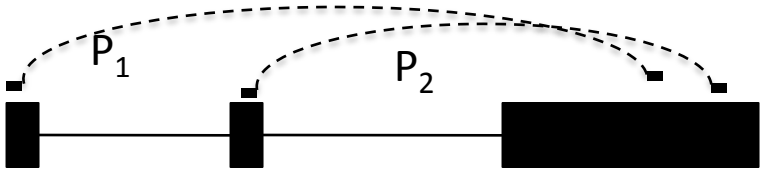


Paired ends increase isoform deconvolution confidence

- P_1 originates from isoform 1 or 2 but not 3.
- P_2 and P_3 originate from isoform 1

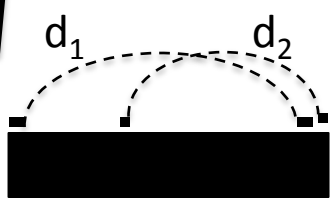
Do paired-end reads also help identifying reads originating in isoform 3?

We can estimate the insert size distribution

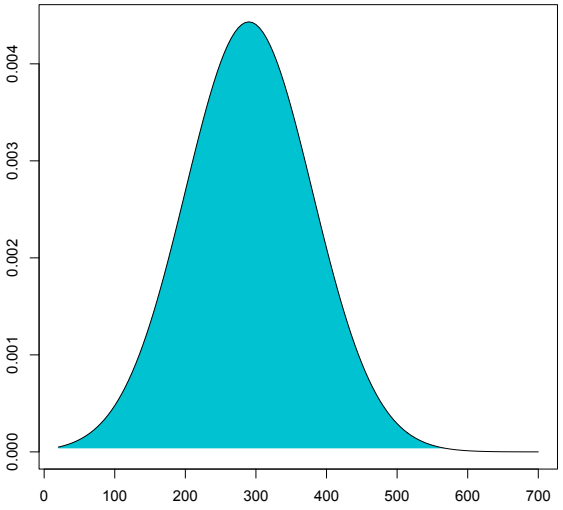


Get all single isoform reconstructions

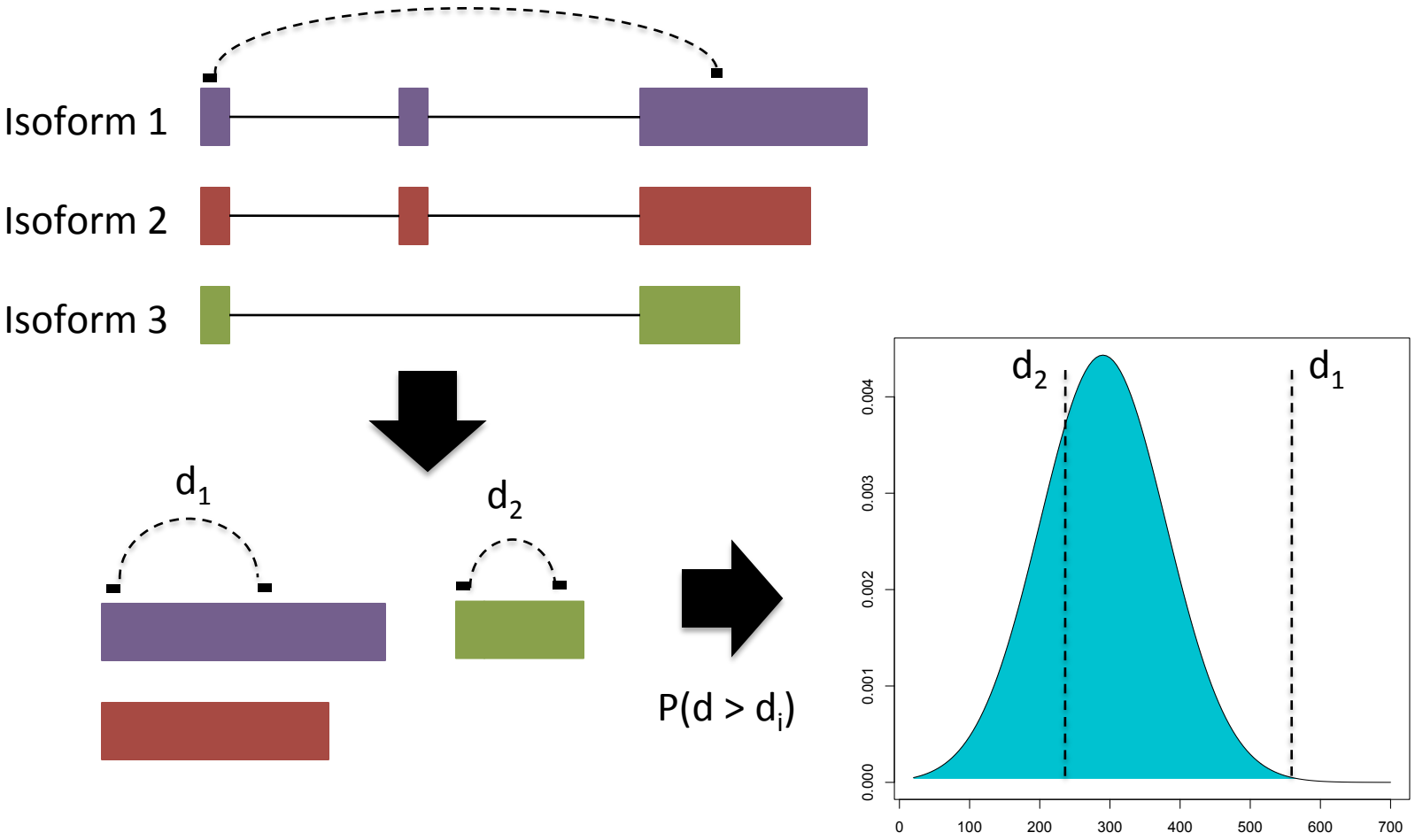
Splice and compute insert distance



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

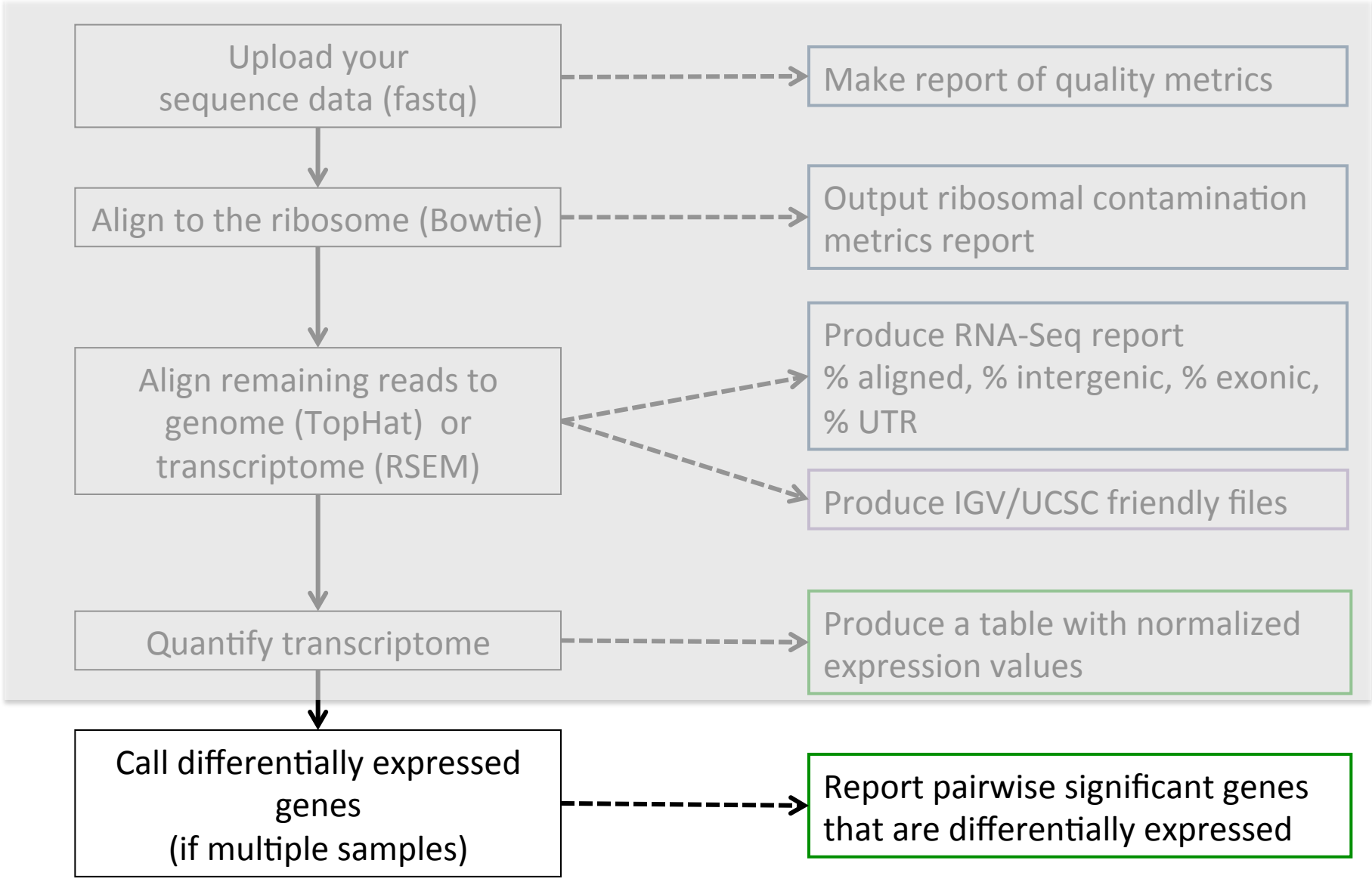
RNA-Seq quantification summary

- Counts must be estimated from ambiguous read/ transcript assignment.
  - Using simplified gene models (intersection)
  - Probabilistic read assignment
- Counts must be normalized
  - RPKM is sufficient for intra-library comparisons
  - More sophisticated normalizations to account for differences in library composition for inter-library comparisons.

Programs to measure transcript expression

	Implemented method
Alexa-seq	Gene expression using intersection model
ERANGE	Gene expression using union model
Scripture	Gene expression using intersection model
Cufflinks	**Transcript deconvolution by solving the maximum likelihood problem**
MISO	**Transcript deconvolution by solving the maximum likelihood problem**
RSEM	**Transcript deconvolution by solving the maximum likelihood problem**

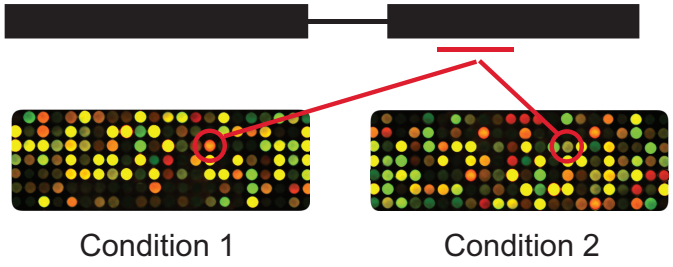
Differential gene expression

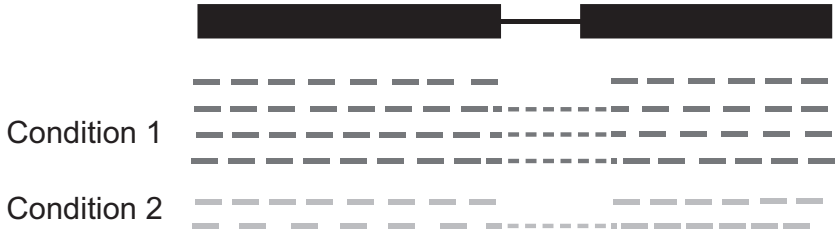


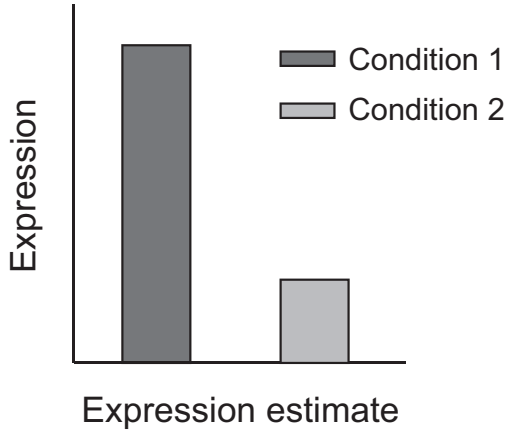
Differential Gene Expression Questions

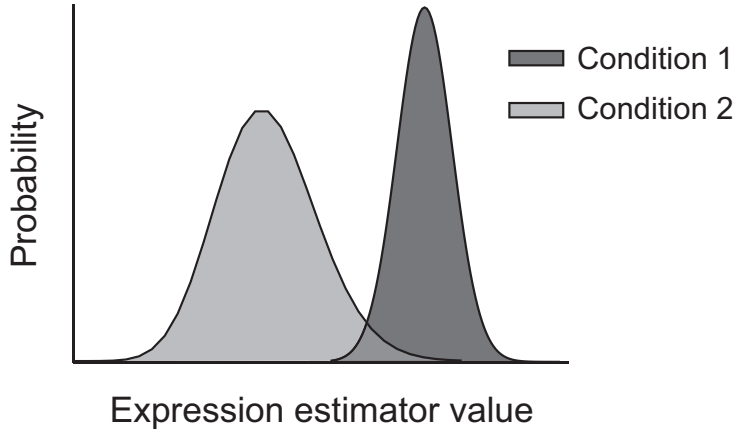
- Finding genes that have different expression between two or more conditions.
- Find gene with isoforms expressed at different levels between two or more conditions.
  - Find differentially used slicing events
  - Find alternatively used transcription start sites
  - Find alternatively used 3’ UTRs

Differential gene expression using RNA-Seq









•(Normalized) read counts ←→ Hybridization intensity

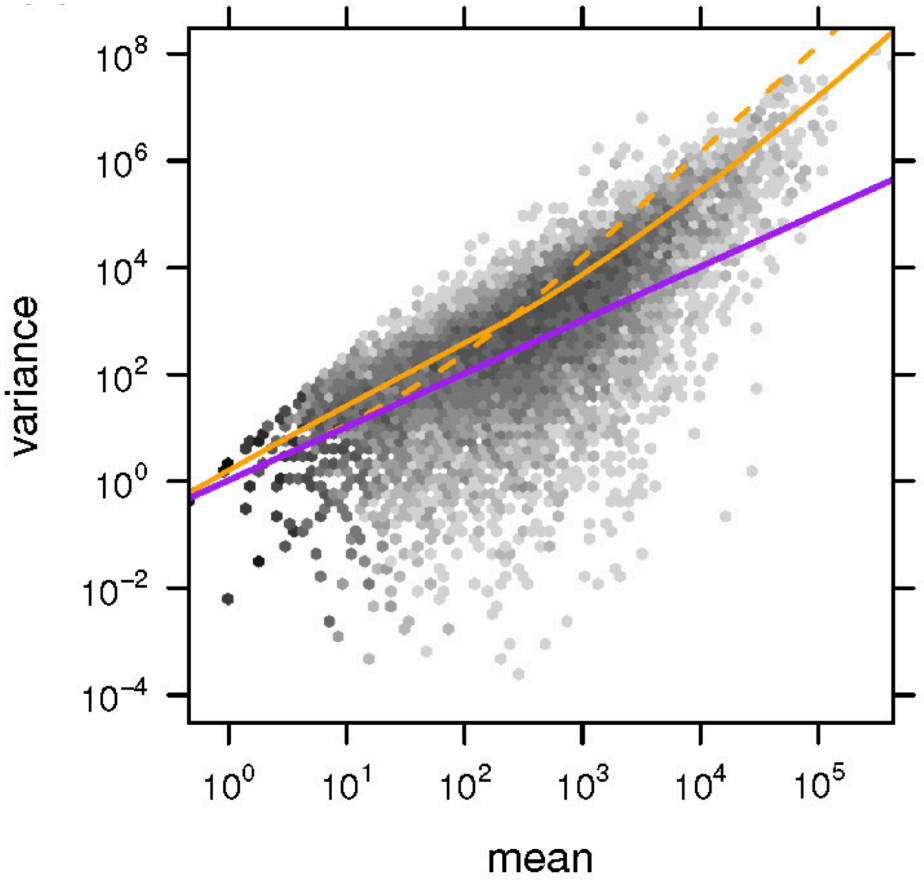
Differential analysis strategies

- Use read counts
  - Standard Fisher exact test

	Condition A	Condition B
Gene A reads	n_a	n_b
Rest of reads	N_a	N_b

  - Model read counts (Poisson, negative binomial) and test whether models are distinct
  - Use empirical approaches that do not rely on parametric assumptions (more on this later)

Poisson model does not work



Adapted from Anders, 2010

Biological variance does not follow a Poisson model

Using a parametric model (DESeq, Cufflinks)

Because of overdisperssion DESeq and Cufflinks uses a Negative binomial to model read counts

$$K_{g,s} \sim \mathcal{N}(K_{g,s}, \sigma_{g,s});\ \sigma_{g,s} = K_{g,s} + \nu_{g,s}$$

Given observed counts for two samples in replicates

$$k_{g,s_1} \ldots k_{g,s_n};\ k_{g,t_1} \ldots k_{g,t_m}$$

DESeq tests the null hypothesis that all counts are sampled from the same distribution

$$P(\sum_i k_{g,s_i} + \sum_j k_{g,t_j} | \mu_s = \mu_t)$$

Cufflinks differential isoform usage

Let a gene G have *n* isoforms and let $p_1, \ldots, p_n$ the estimated fraction of expression of each isoform.

Call this a the isoform expression distribution *P* for G

Given two samples the differential isoform usage amounts to determine whether H_0: $P_1 = P_2$ or H_1: P_1 != P_2 are true.

To compare distributions Cufflinks utilizes an information content based metric of how different two distributions are called the Jensen-Shannon divergence:

$$JS(p^1, \ldots, p^m) = H\left(\frac{p^1 + \cdots + p^m}{m}\right) - \frac{\sum_{j=1}^{m} H(p^j)}{m}.$$

$$H(p) = -\sum_{i=1}^{n} p_i log p_i.$$

The square root of the JS distributes normal.

RNA-Seq differential expression software

	Underlying model	Notes
DegSeq	Normal. Mean and variance estimated from replicates	Works directly from reference transcriptome and read alignment
EdgeR	Negative Bionomial	Gene read counts table
DESeq	Negative Bionomial	Gene read counts table
Cufflinks	~~Poisson~~ Negative Bionomial	Works directly from the alignments
Myrna	Empirical	Sequence reads and reference transcriptome

The quest for inexpensive expression assays

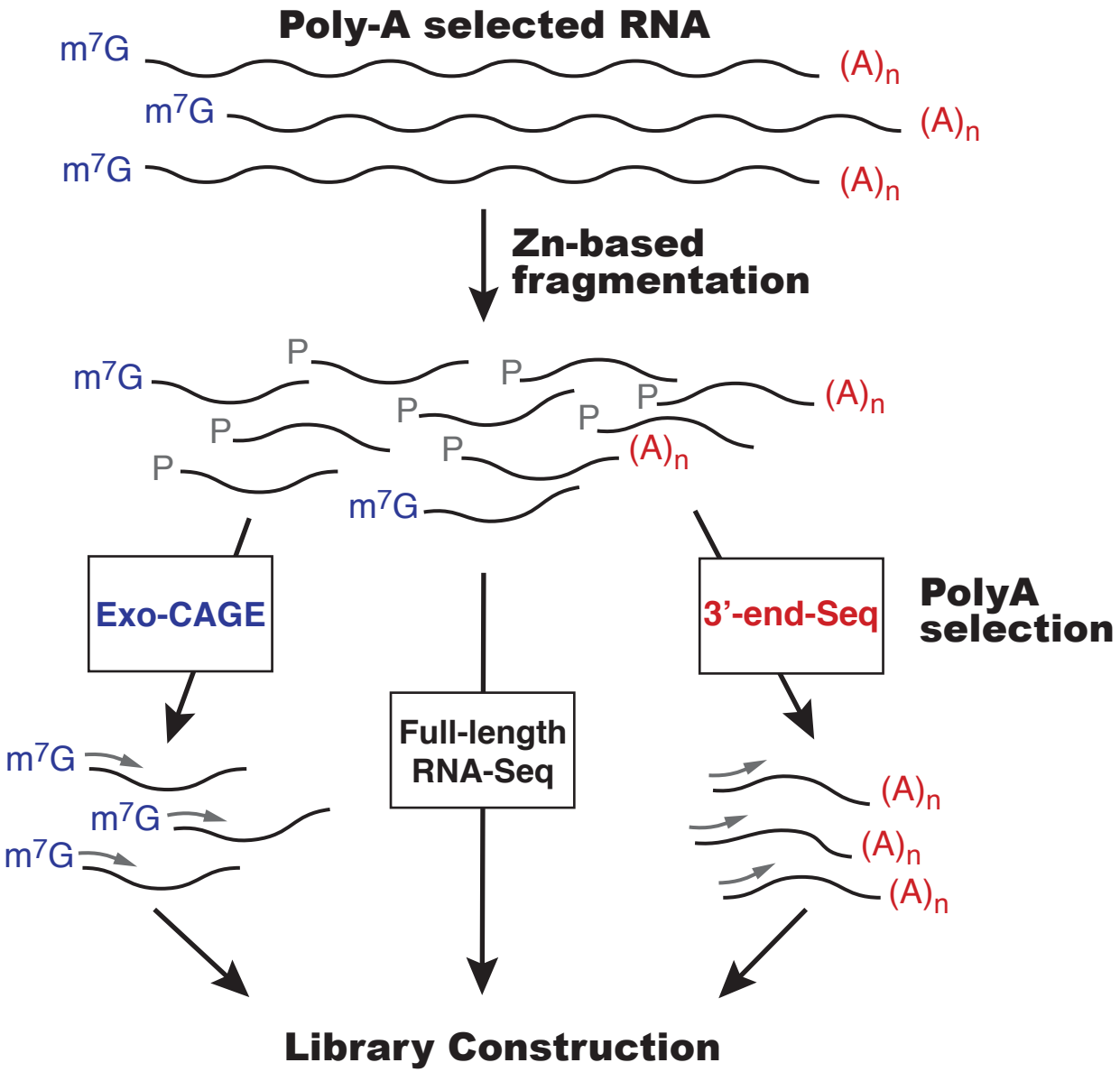
- Goal: Routinely profile hundreds of samples
- Why?
  - Human variability in health and disease
  - Perturbation studies
  - Clinical applications of expression profiling
- Current costs
  - Afffy ~$300-$400/sample
  - Illumina bead arrays $150/sample
  - RNA-Seq (20 mill reads) ~$400-$500/sample ($350 in sequencing)
- RNA-Seq disadvantages
  - Complex analysis
  - Length bias

Final considerations on quantification

- Using different libraries:
  - Targeting the 3' end
  - Targeting 5' end
- What depth do we really need?

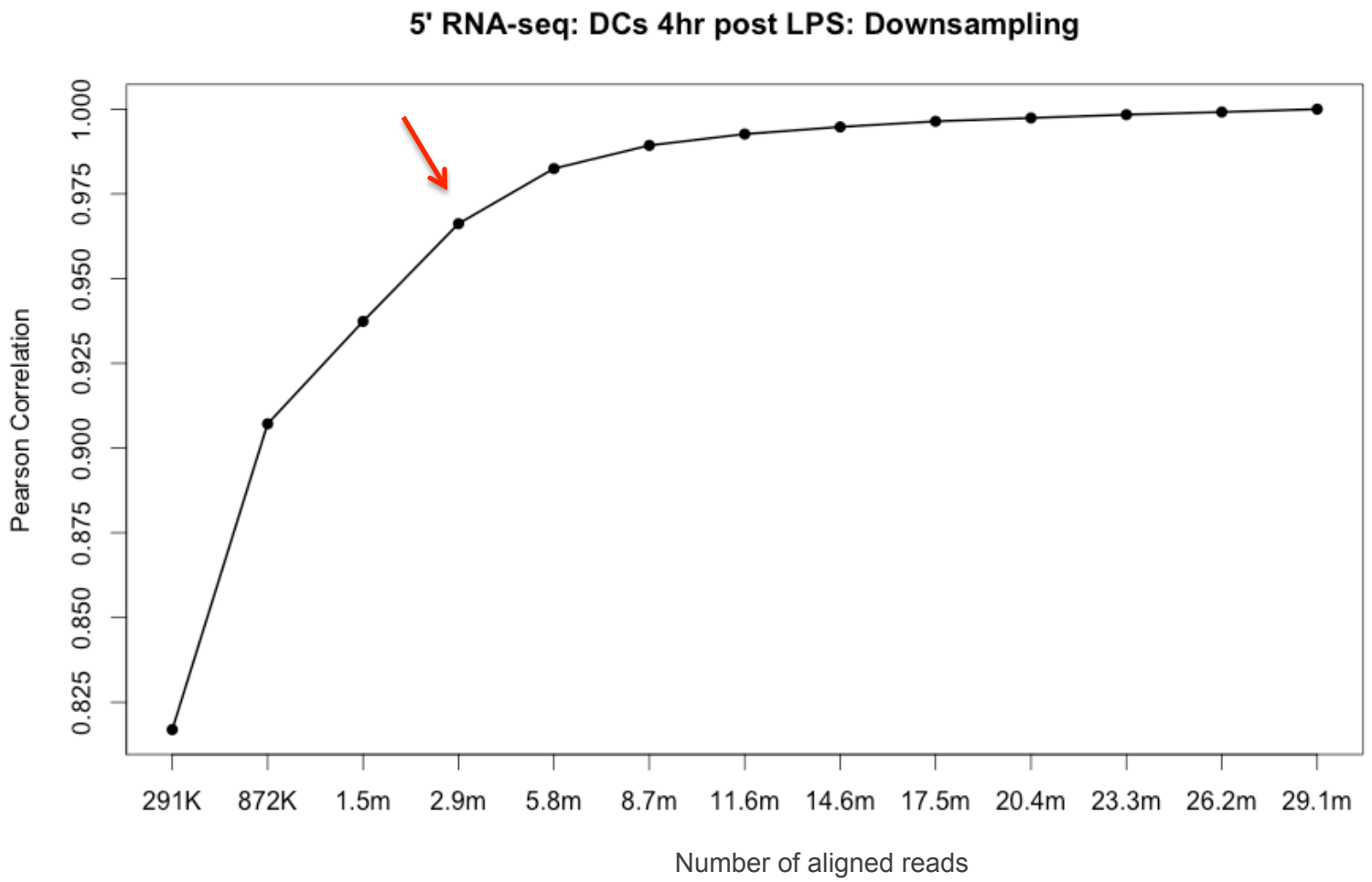
Alper Kucukural
Sabah Kadri
Maxim Artyomov

RNA-Seq libraries: Summary



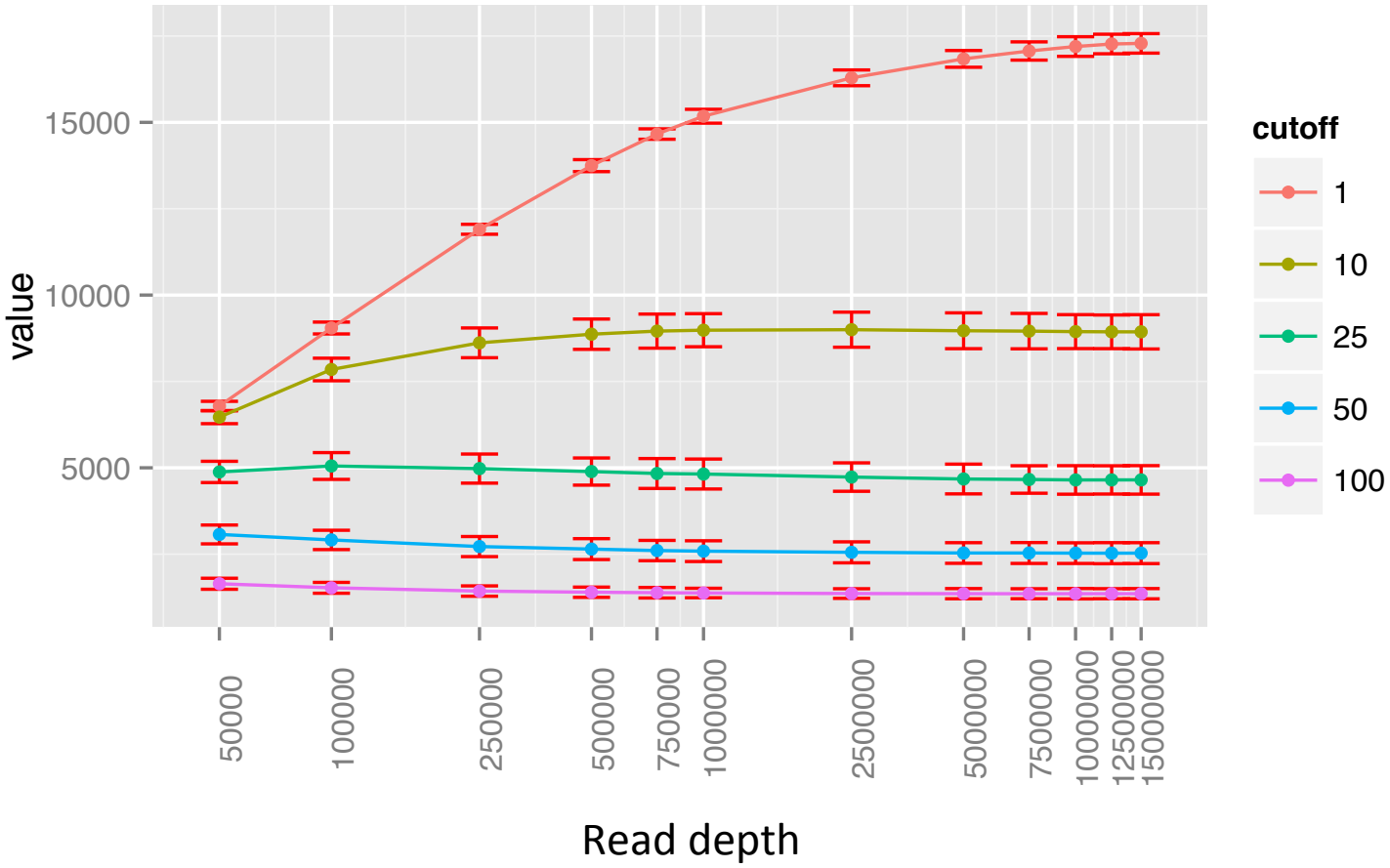
Sabah Kadri
Maxim Artyomov
Sebastian Kadener

Quantification works with shallow reads



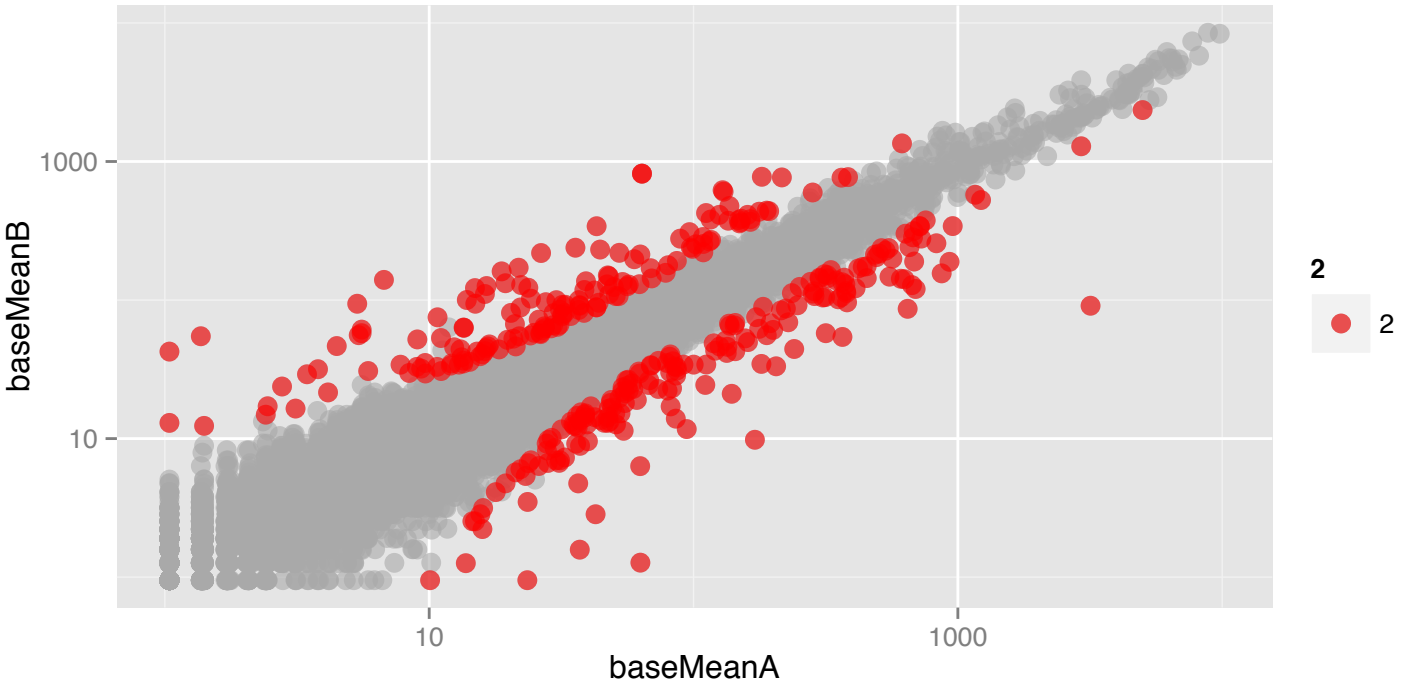
Requires just a fraction of the reads required for RNA-Seq quantification

Robustness to low depth: Transcripts detected

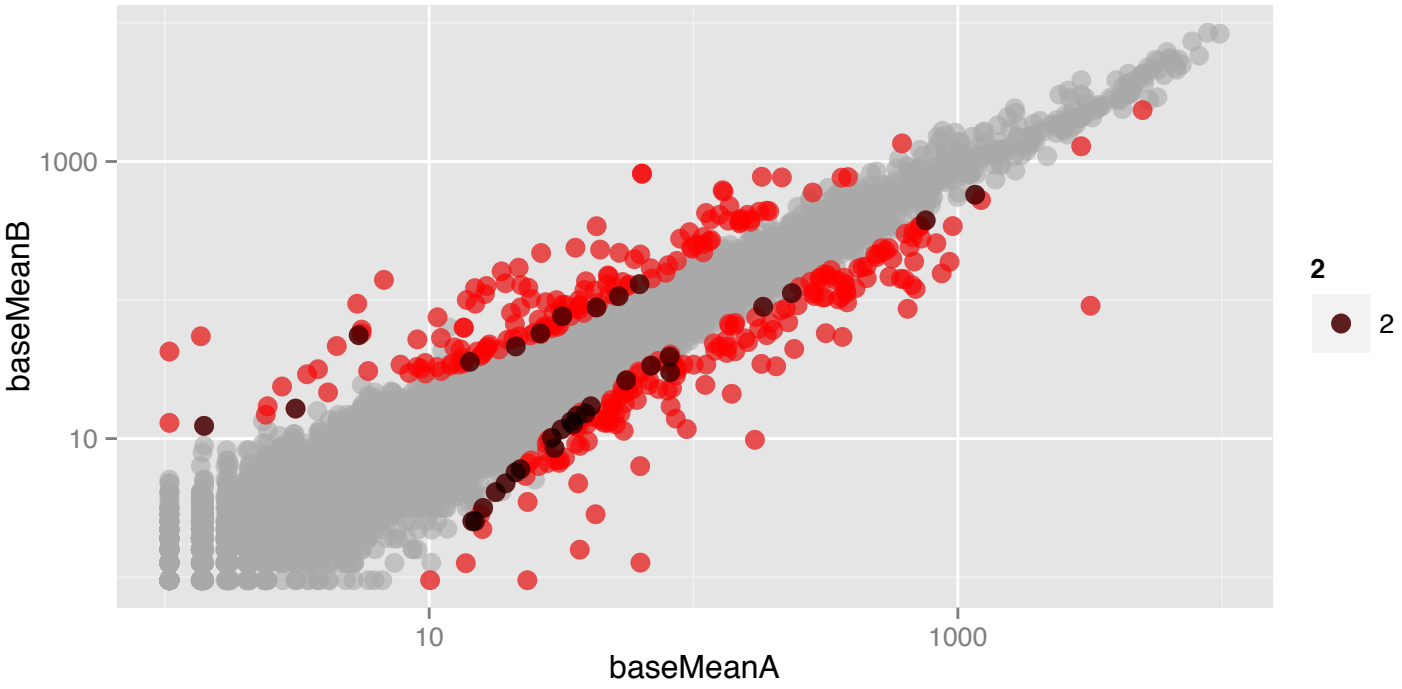


Alper Kucukural

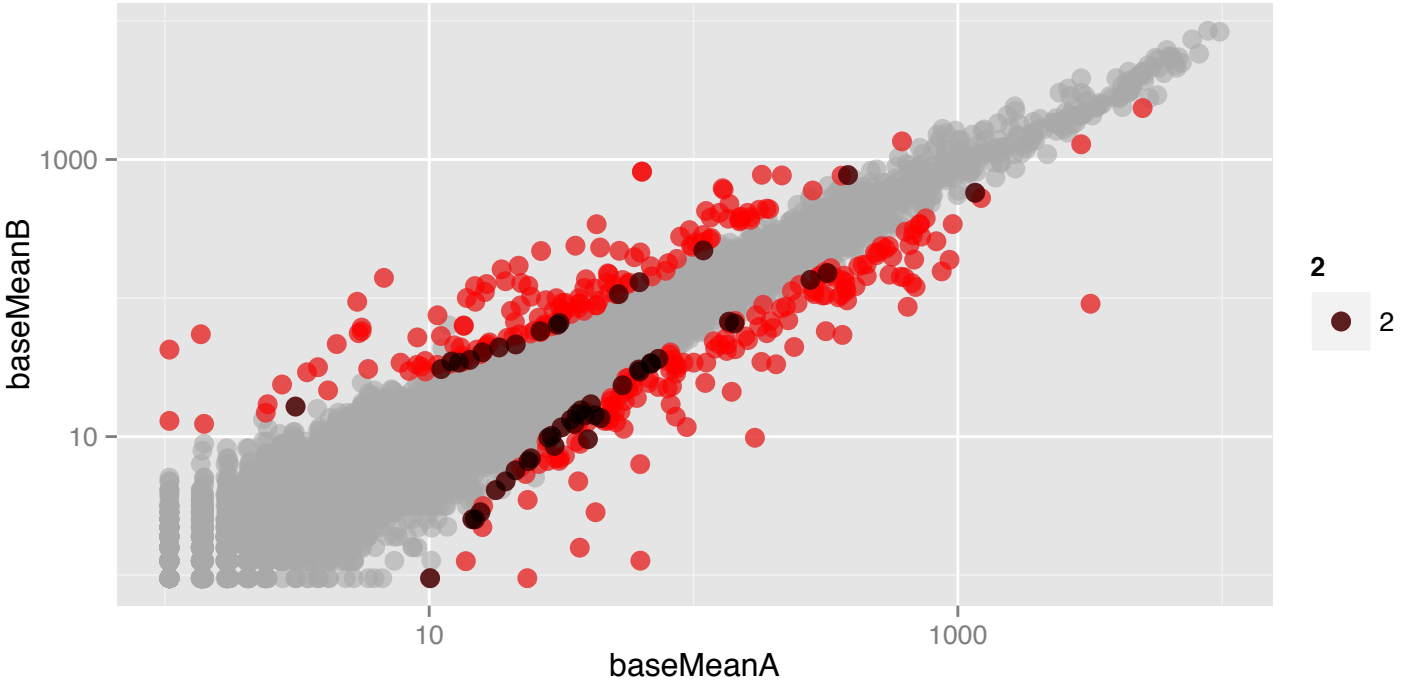
RSEM/DESeq: 15 mill reads in worm



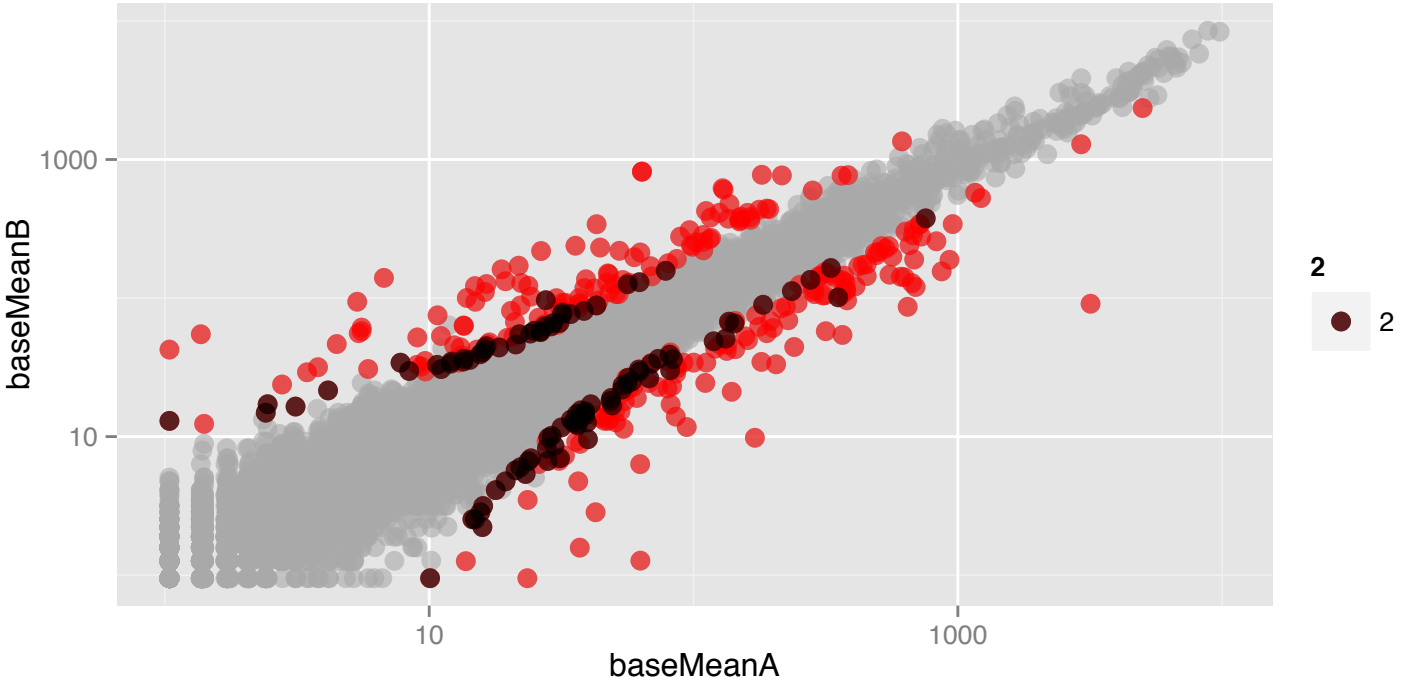
RSEM/DESeq: 10 mill reads in worm



RSEM/DESeq: 7.5 mill reads in worm

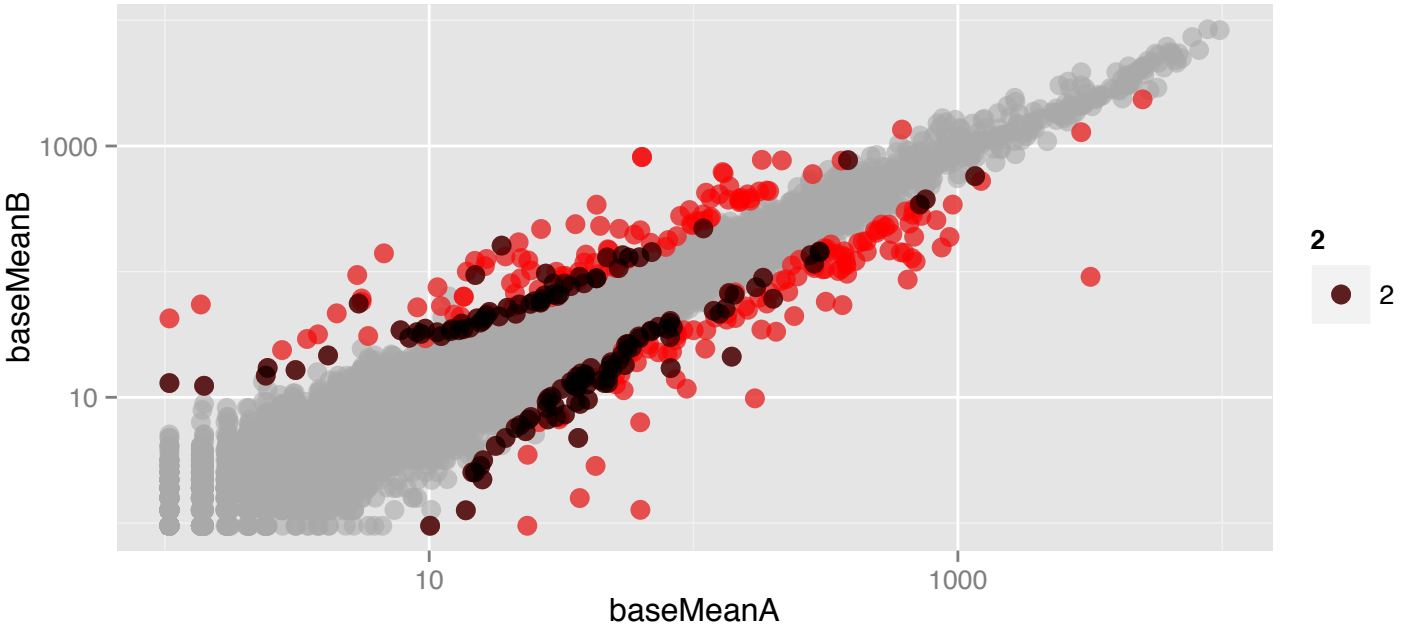


RSEM/DESeq: 5 mill reads in worm

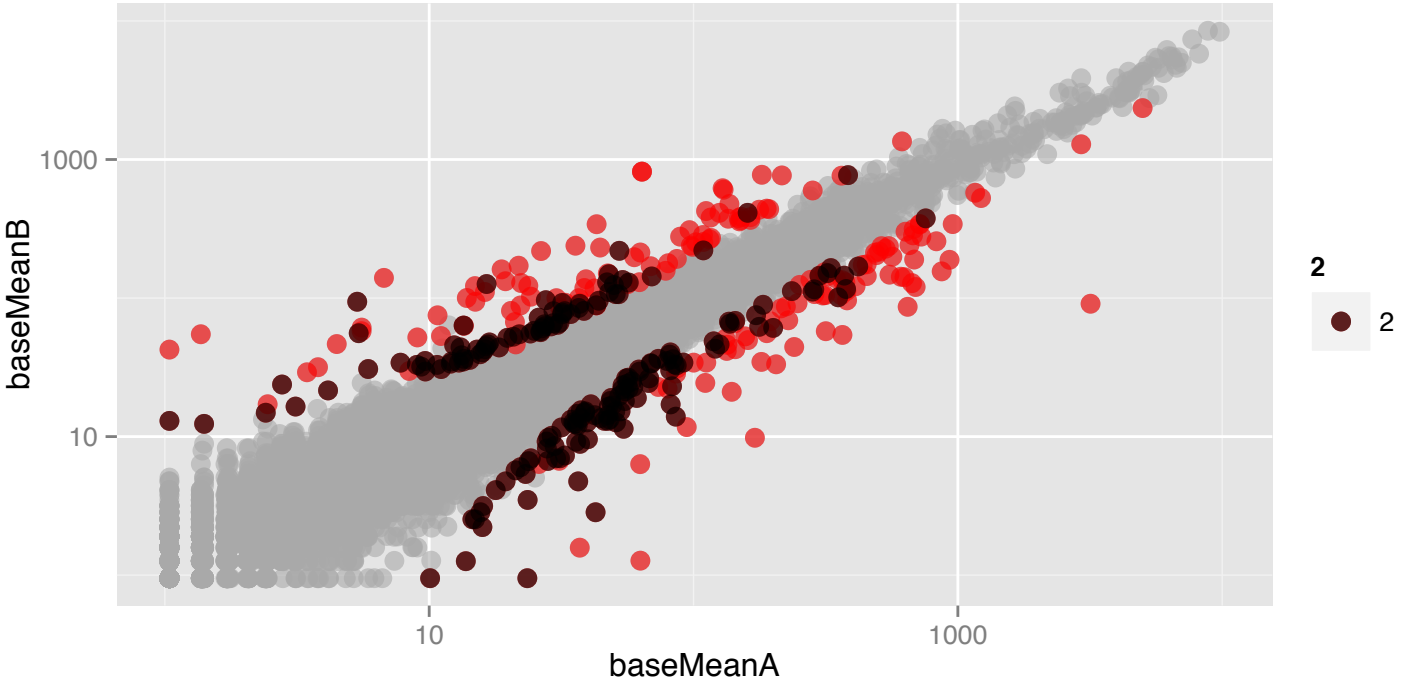


Alper Kucukural

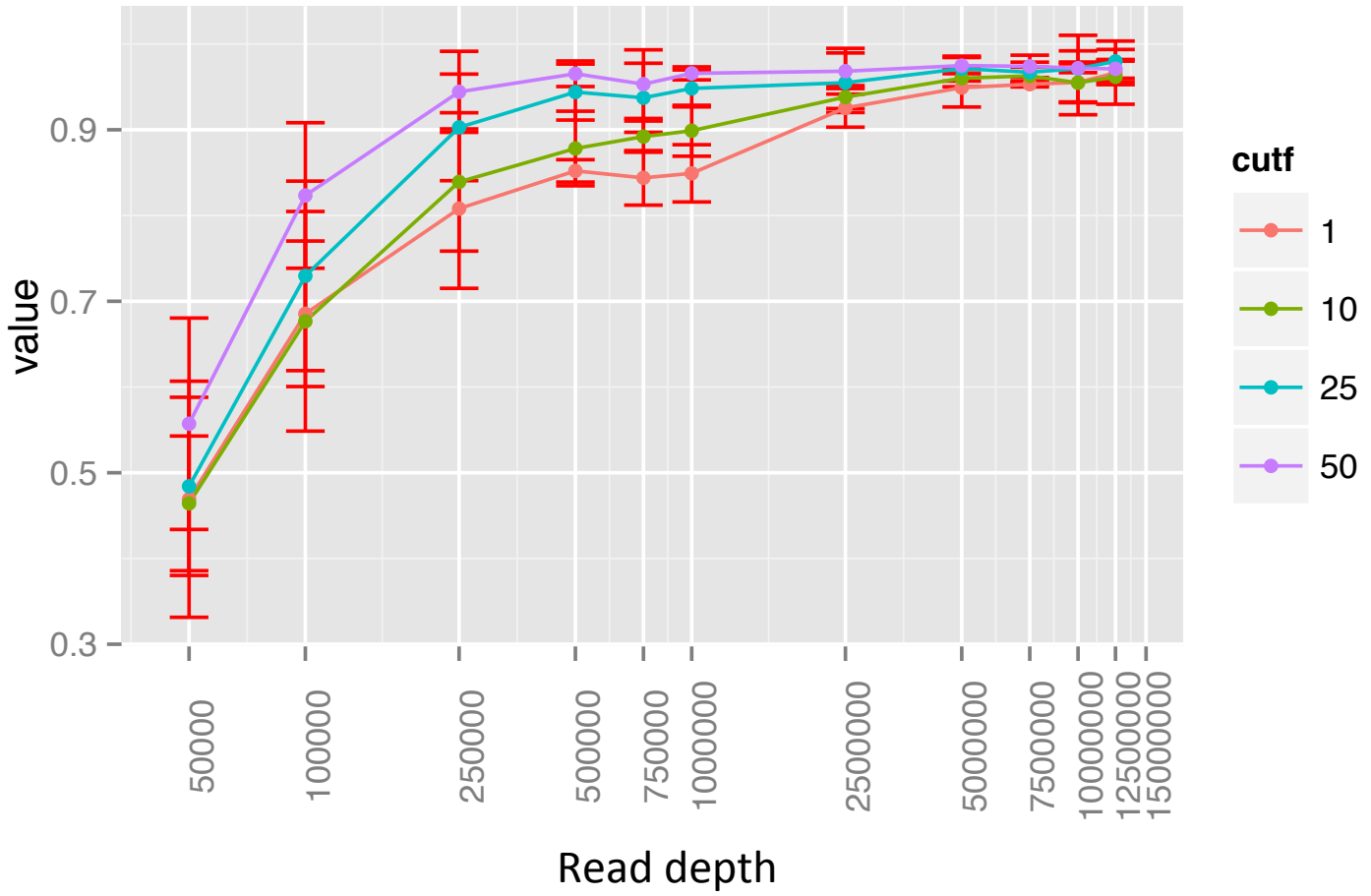
RSEM/DESeq: 2.5 mill reads in worm



RSEM/DESeq: 1 mill reads in worm



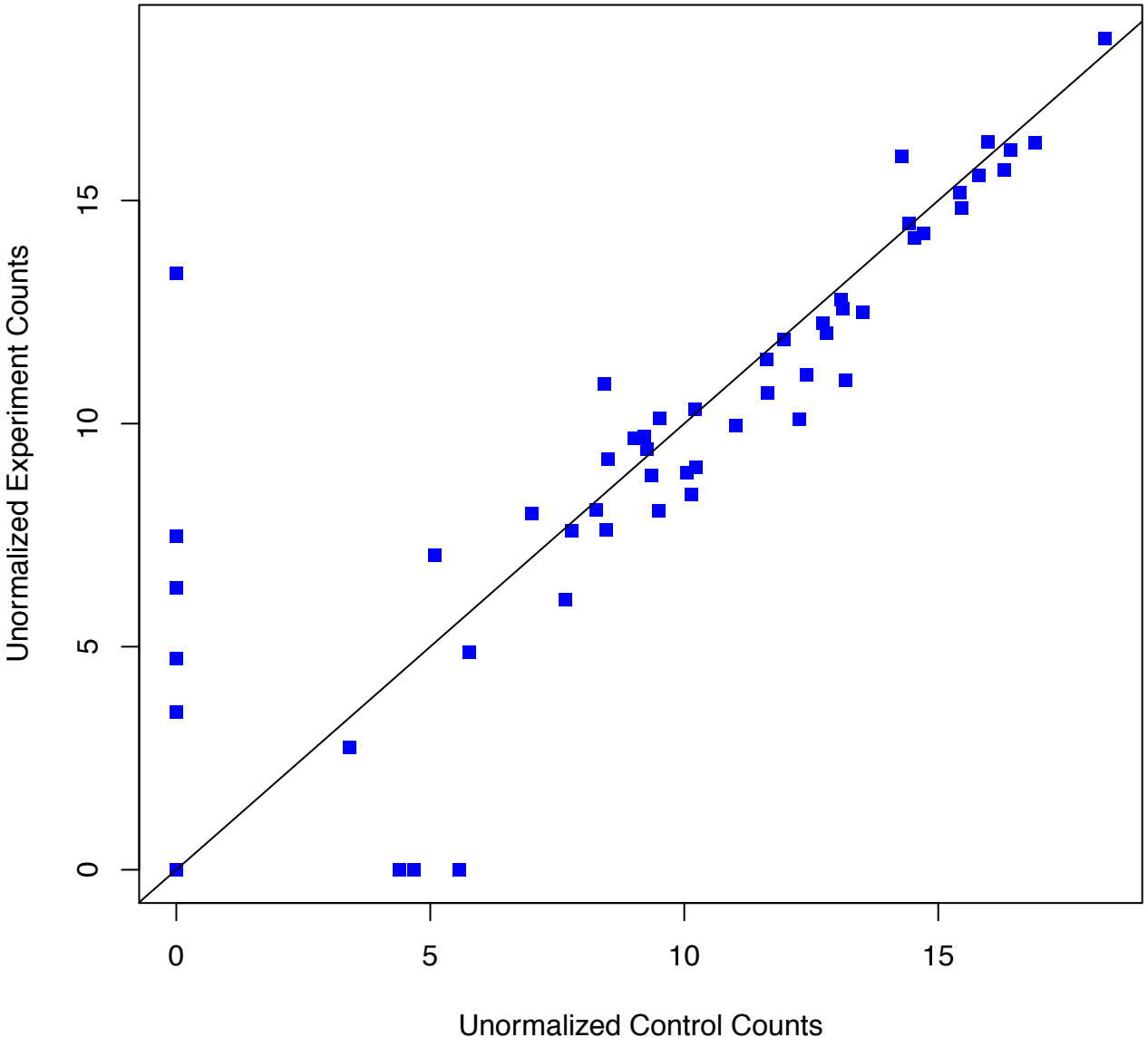
Robustness of DGE to low depth



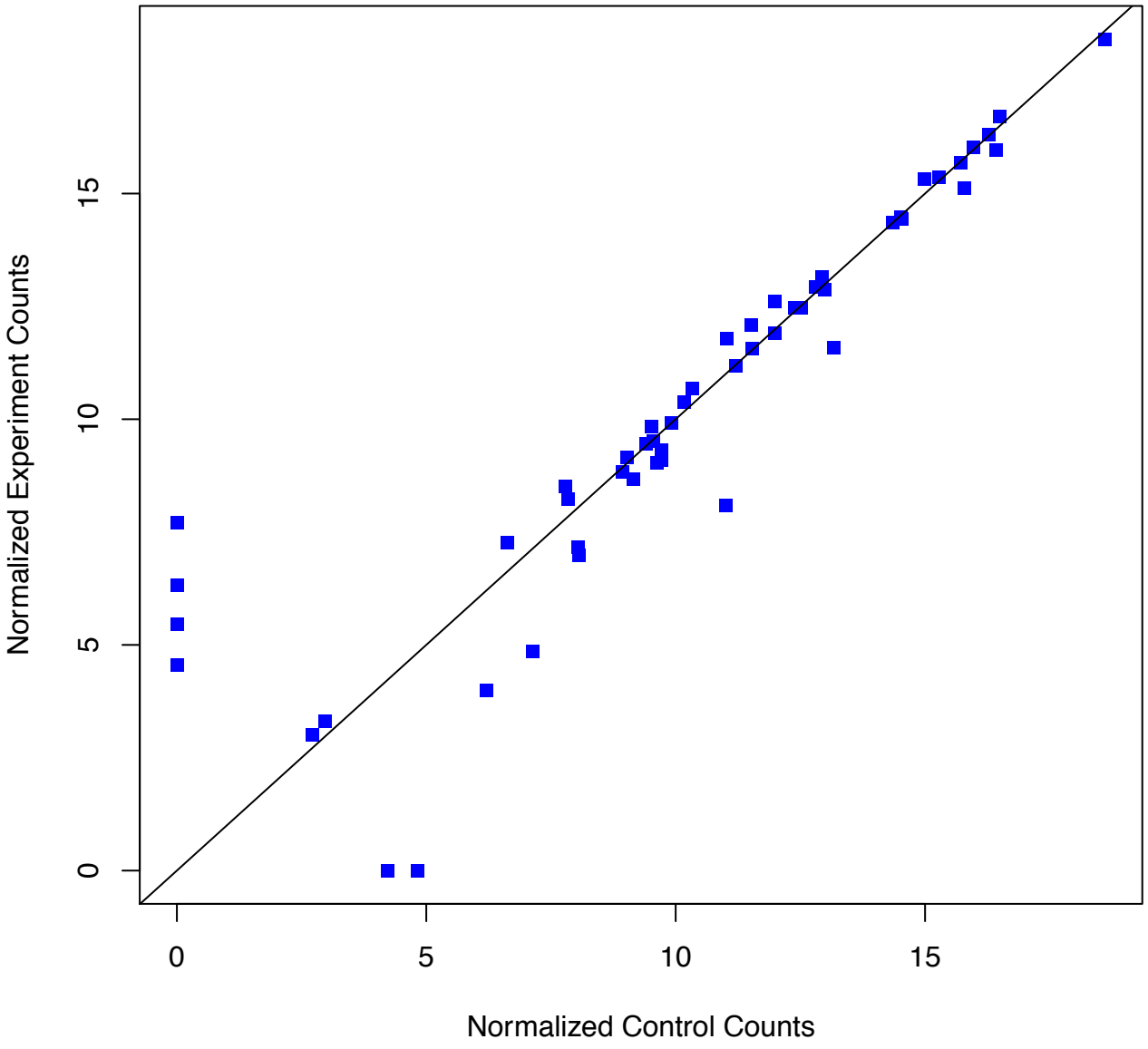
Final considerations: The steps of Sequencing analysis

- Filter reads (fastq file) by removing adapter, splitting barcodes.
  - Evaluate overall quality, look for drop in quality at ends. Trim reads if ends are of low quality
- Alignment to the genome
  - Use transcriptome if available
  - Filter out likely PCR duplicates (reads that align to the same place in the genome
  - Evaluate ribosomal contamination
  - What percent of reads aligned
- Reconstruct(?)
- Quantify
  - Normalize according to application

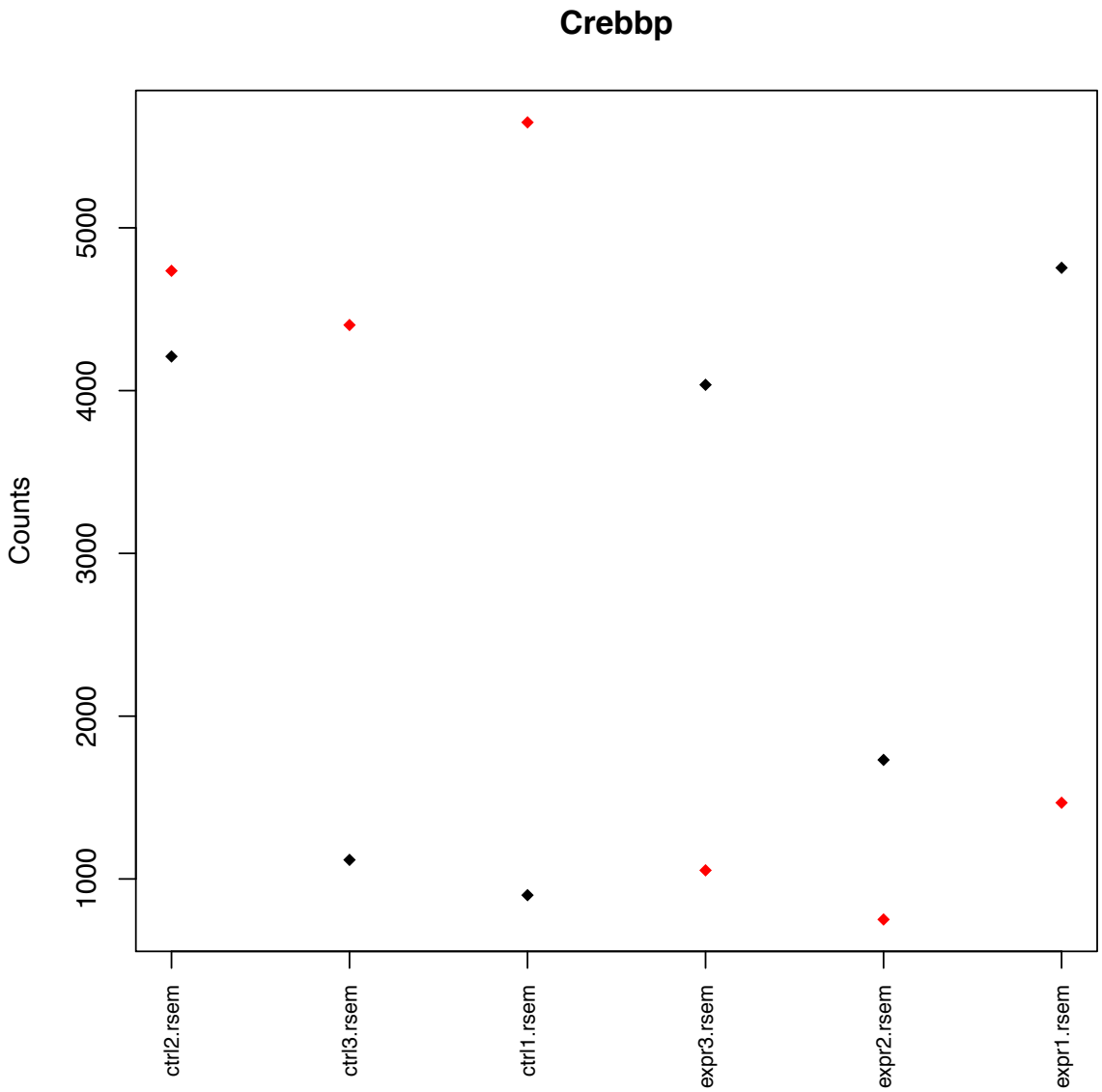
Revisiting Exercise 3



Transcript Reconstruction

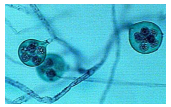


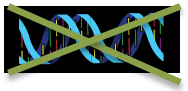
Comparing normalization

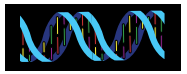


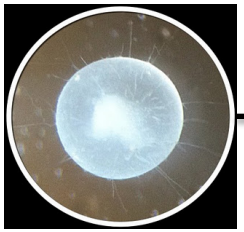
Transcript Reconstruction

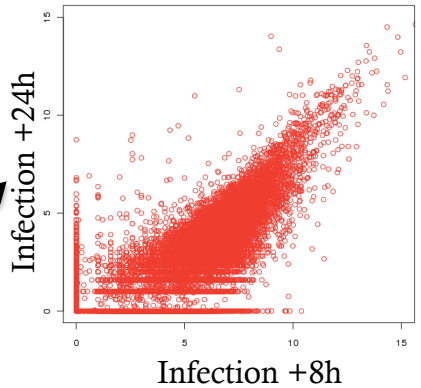
Advantages of RSEM, DESeq



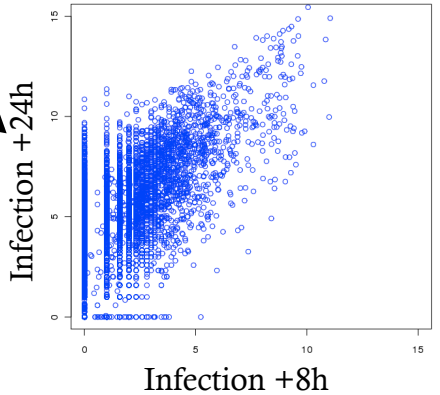








Trout



Saprolegnia

Infected fish cells

Brian Haas, Rays Jiang, Carsten Russ

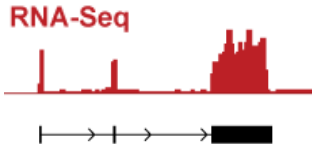
What does significance means?

- RNA-Seq: The gene is expressed
- ChIP-Seq: Factor binds the region
- CLIP-Seq: Protein binds RNA region
- Ribosomal footprinting:
  - Transcript is translated
  - Ribosomes stalling at region

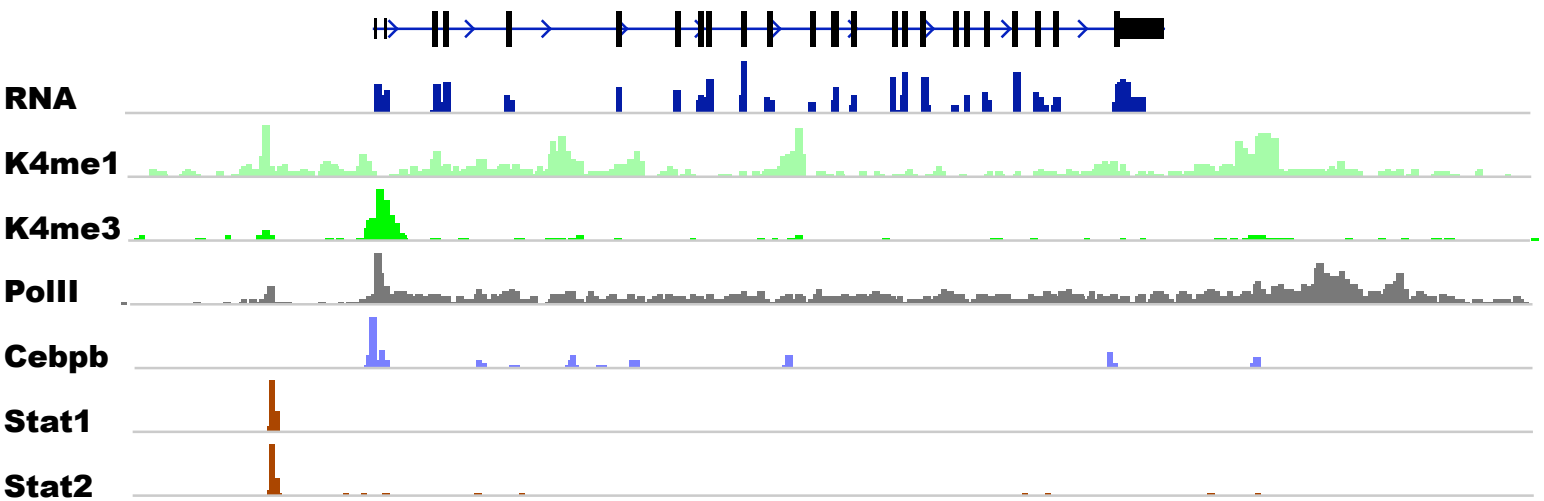
How do we find peaks?

Short modification

Long modification

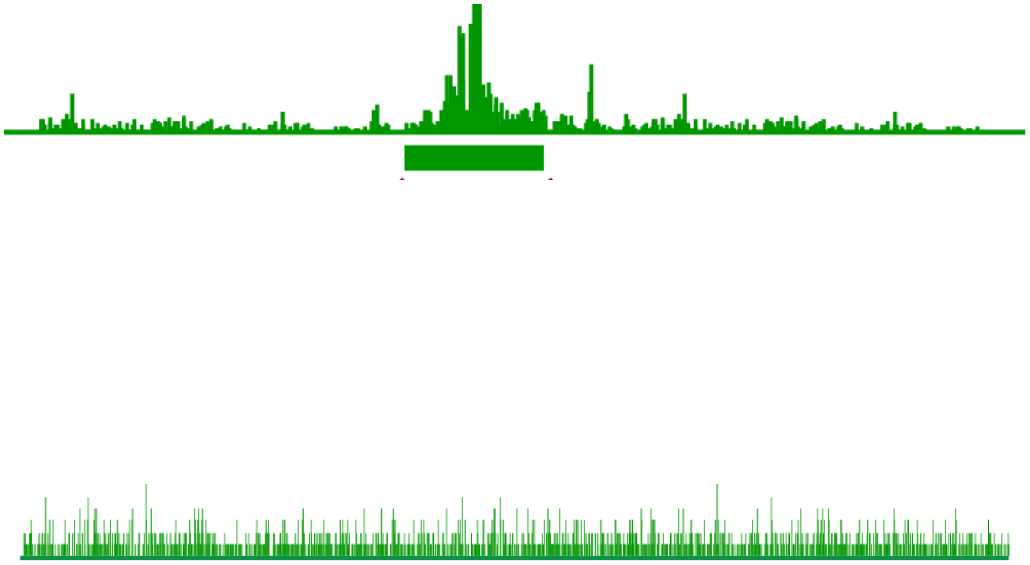


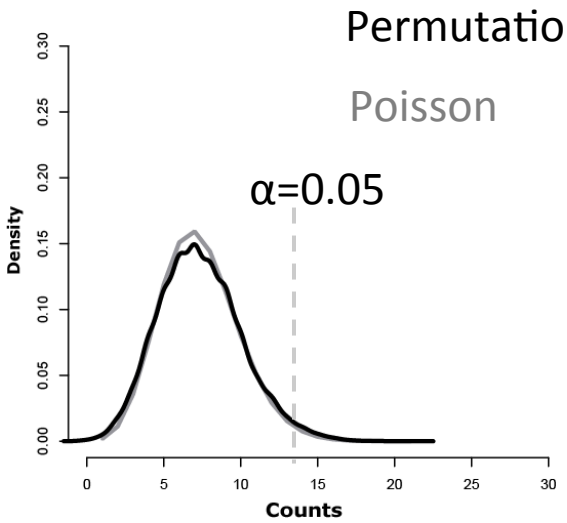
Discontinuous data



Scripture is a method to solve this general question

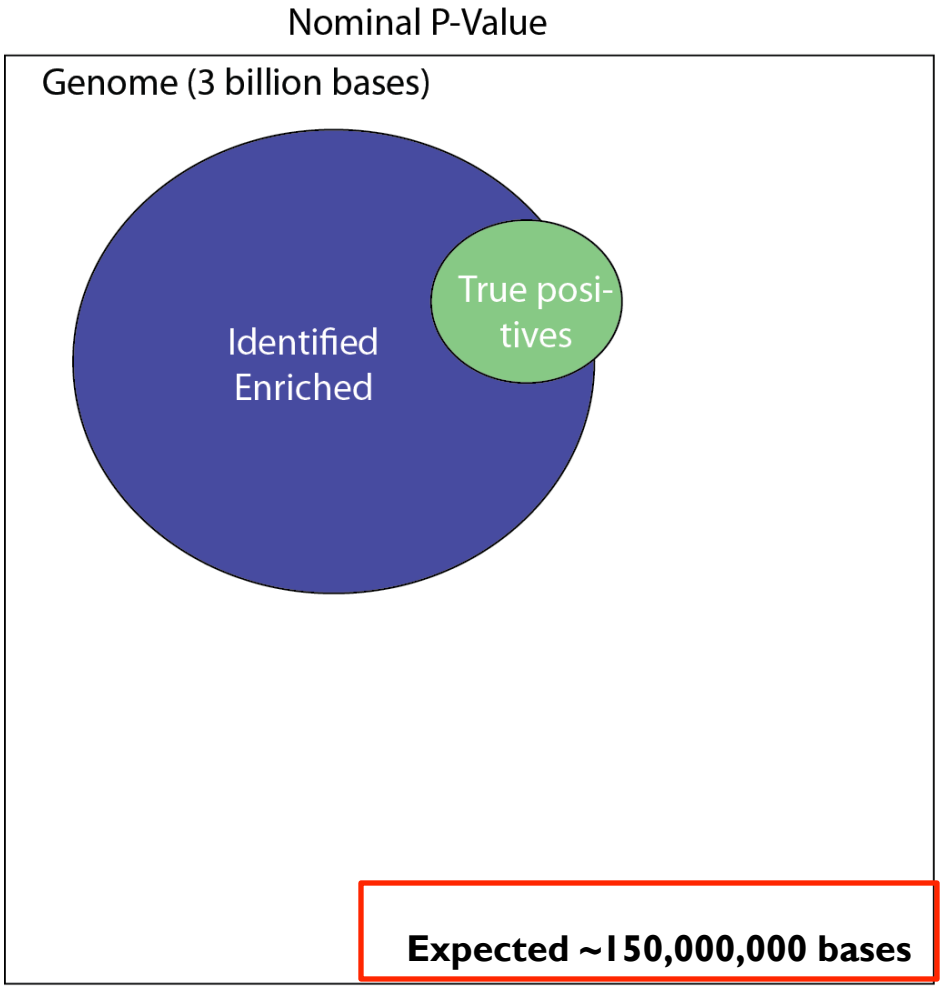
Our approach





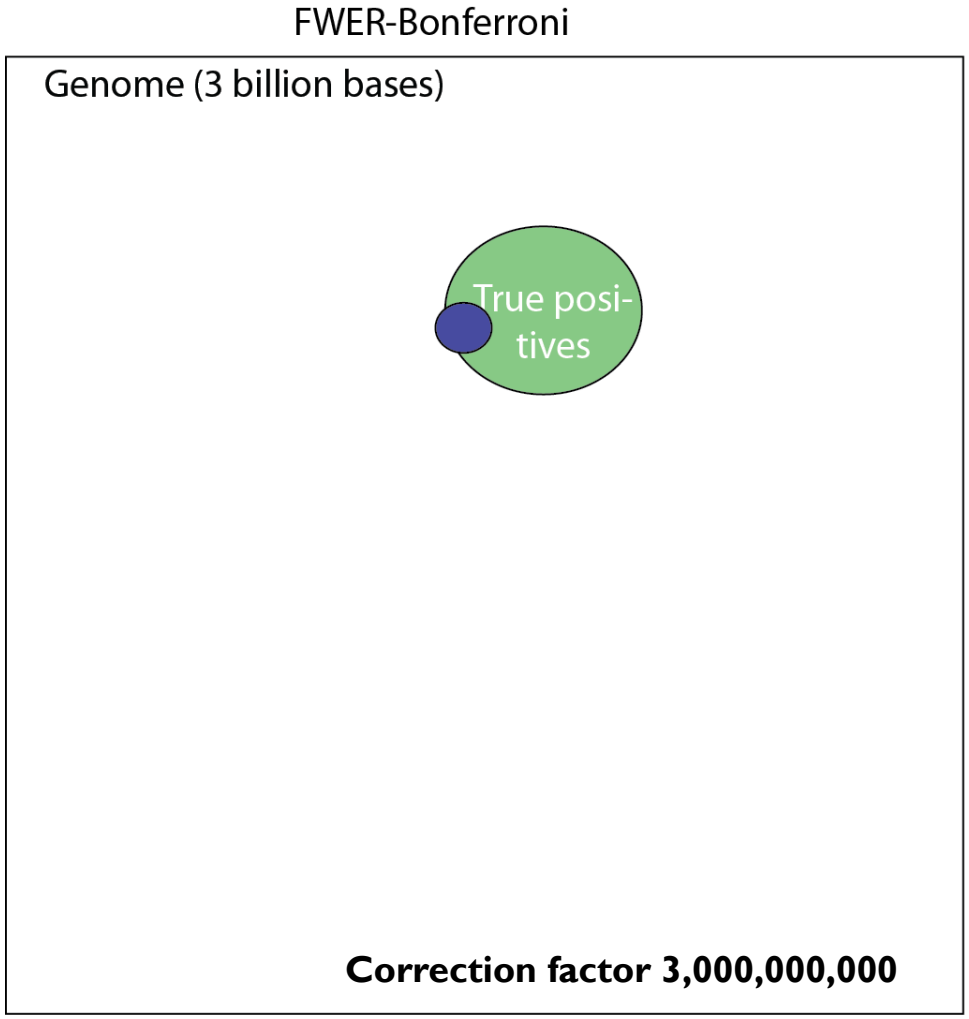
We have an efficient way to compute read count p-values …

The genome is large, many things happen by chance



We need to correct for multiple hypothesis testing

Bonferroni correction is way to conservative

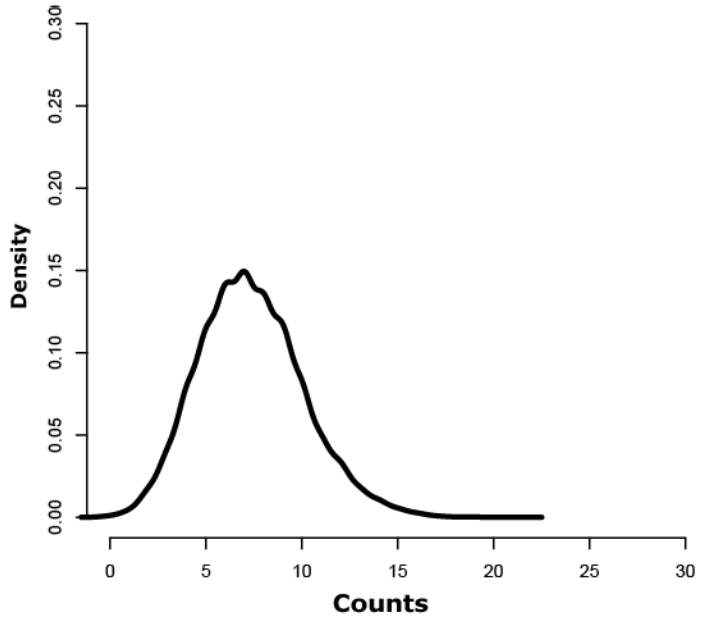


Bonferroni corrects the number of hits but misses many true hits because its too conservative – How do we get more power?

Controlling FWER

Max Count distribution

α=0.05 α_{FWER}=0.05



Count distribution (Poisson)

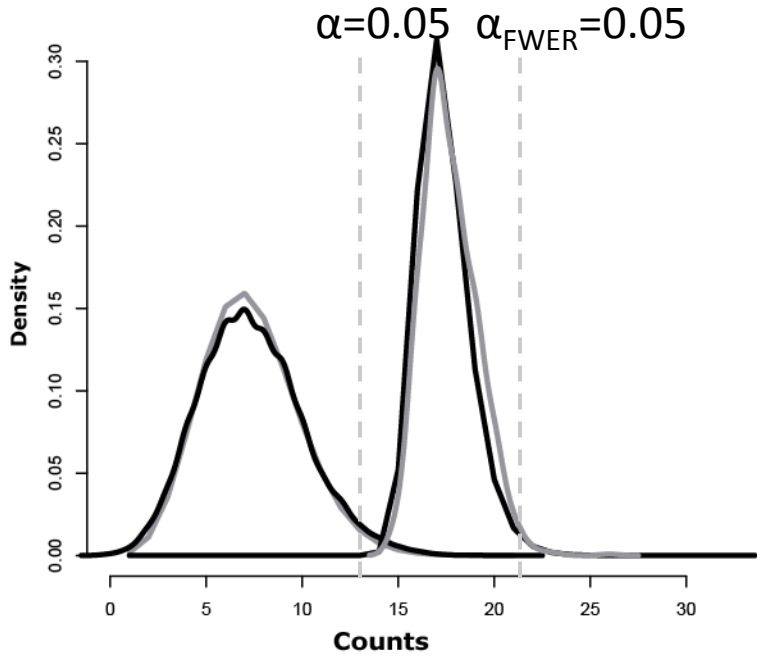
Given a region of size w and an observed read count n. What is the probability that one or more of the $3x10^9$ regions of size w has read count >= n under the null distribution?

We could go back to our permutations and compute an FWER: **max of the genome-wide distributions of same sized region**)→
but really really really slow!!!

Scan distribution, an old problem

- Is the observed number of read counts over our region of interest high?
- Given a set of Geiger counts across a region find clusters of high radioactivity
- Are there time intervals where assembly line errors are high?

Scan distribution



Poisson distribution

Thankfully, the ***Scan Distribution*** computes a closed form for this distribution.

ACCOUNTS for dependency of overlapping windows thus more powerful!

Scan distribution for a Poisson process

The probability of observing k reads on a window of size w in a genome of size L given a total of N reads can be approximated by (Alm 1983):

$$P(k|\lambda w, N, L) \approx 1 - F_p(k-1|\lambda w)e^{-\frac{k-w\lambda}{k}\lambda(T-w)P(k-1|\lambda w)}$$

where

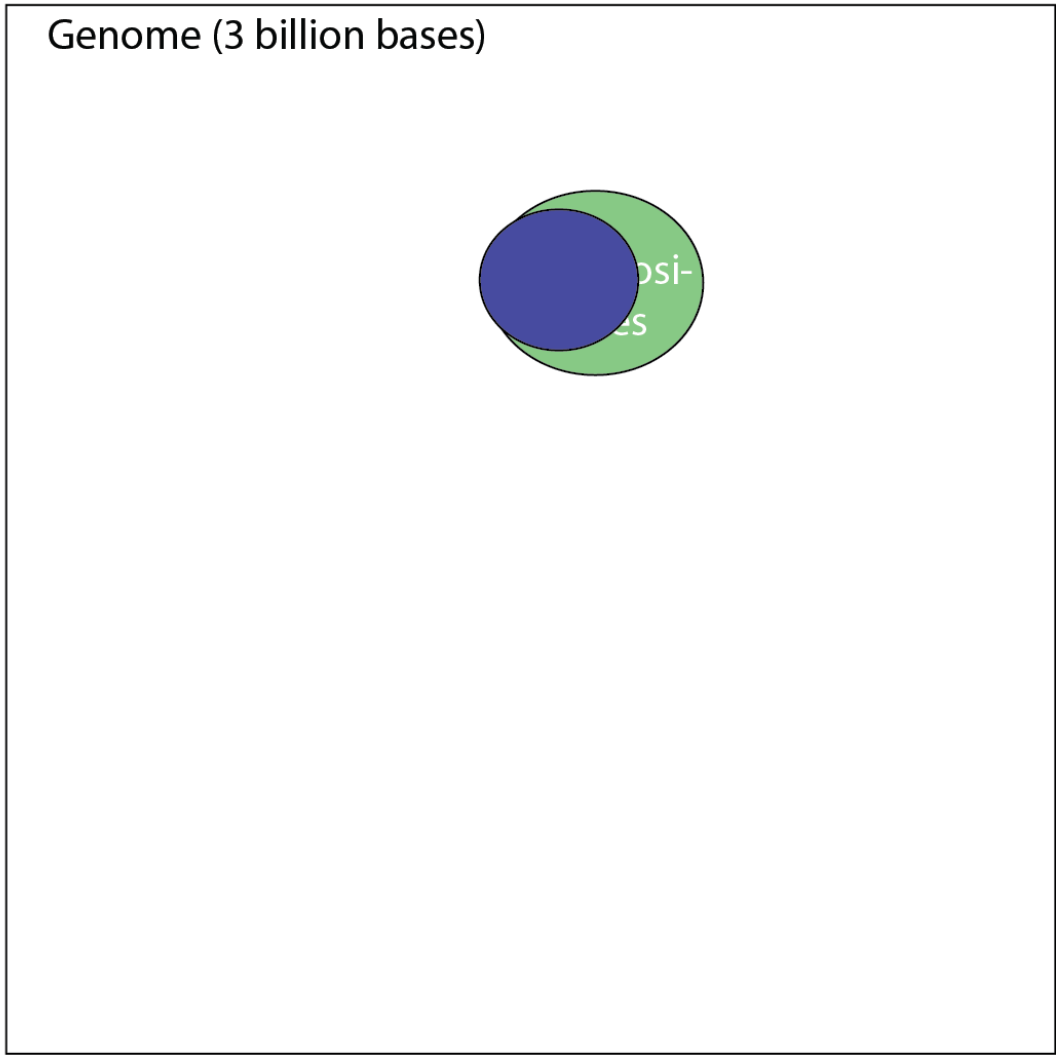
$P(k-1|\lambda w)$ is the Poisson probability of observing $k-1$ counts given an expected count of λw

and

$F_p(k-1|\lambda w)$ is the Poisson probability of observing $k-1$ or fewer counts given an expectation of λw reads

The scan distribution gives a computationally very efficient way to estimate the FWER

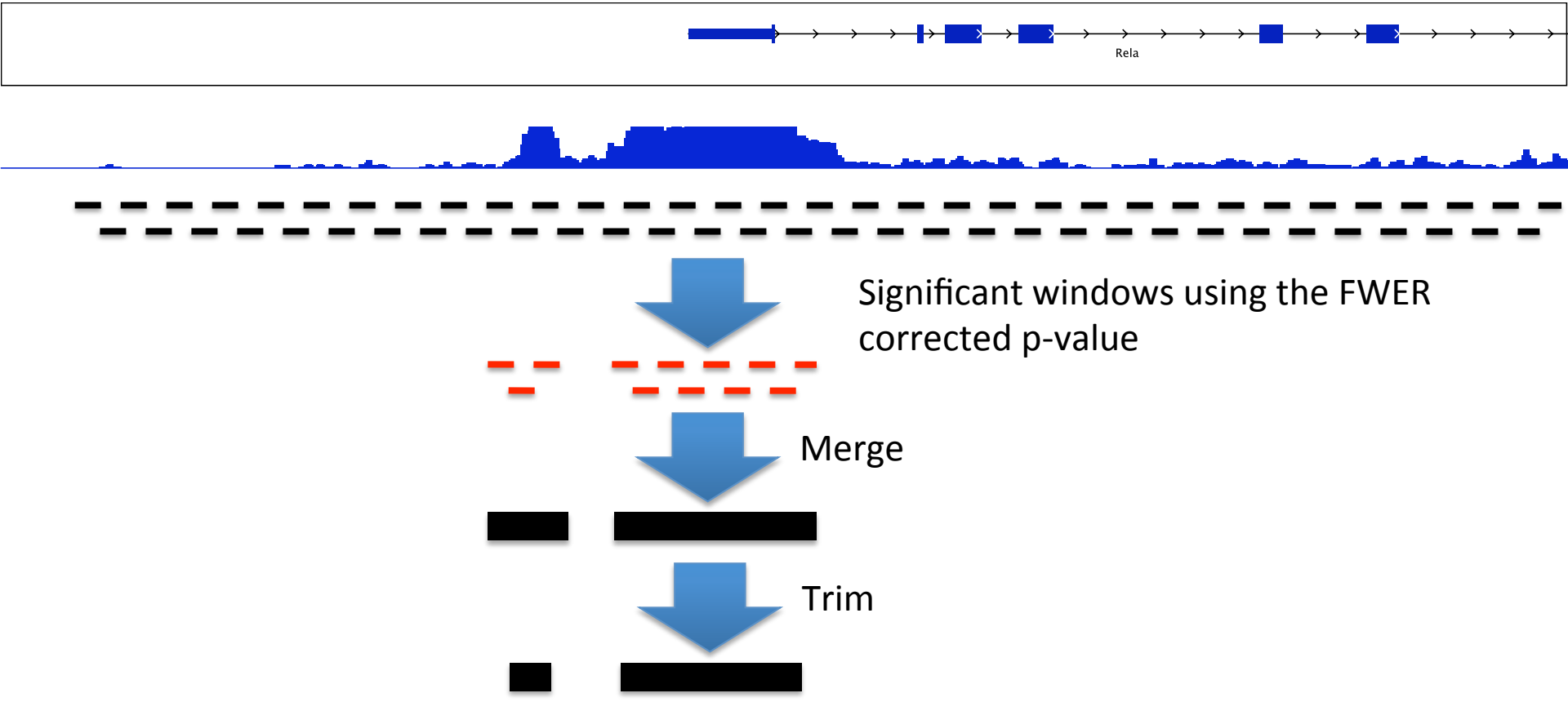
FWER-Scan Statistics



By utilizing the dependency of overlapping windows we have greater power, while still controlling the same genome-wide false positive rate.

Segmentation method for contiguous regions

Example : PolII ChIP

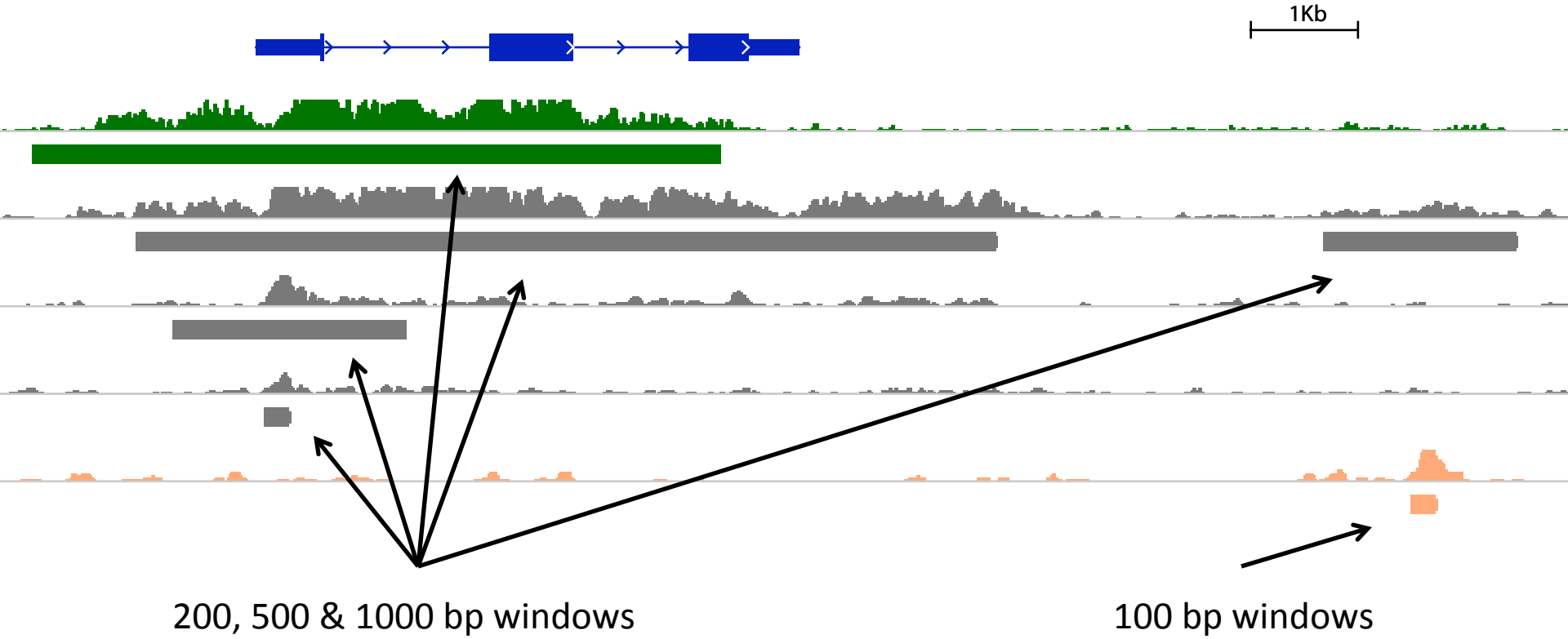


But, which window?

We use multiple windows

- Small windows detect small punctuate regions.
- Longer windows can detect regions of moderate enrichment over long spans.
- In practice we scan different windows, finding significant ones in each scan.
- In practice, it helps to use some prior information in picking the windows although globally it might be ok.

Applying Scripture to a variety of ChIP-Seq data

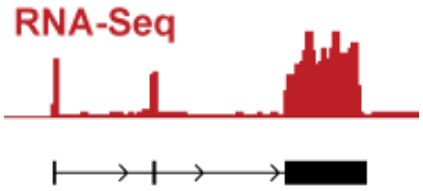


Can we identify enriched regions across different libraries?

Short modification ✓

Long modification ✓

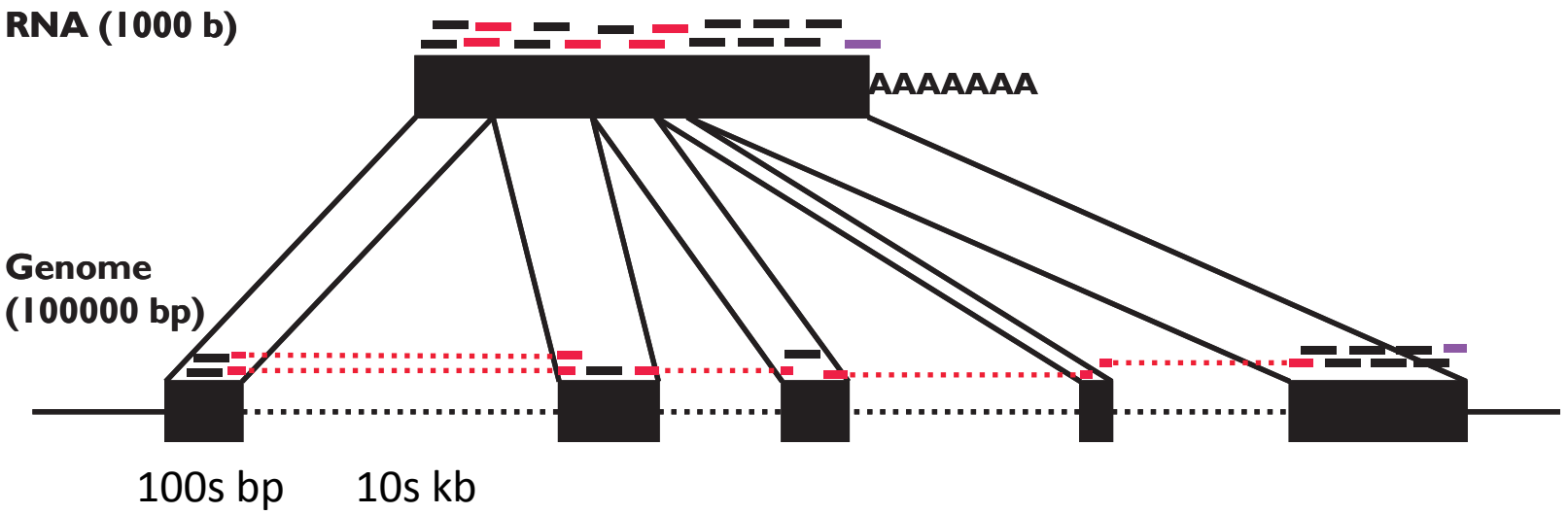
Using chromatin signatures we discovered hundreds of putative genes.
What is their structure?



Discontinuous data: RNA-Seq to find gene structures for this gene-like regions

Scripture for RNA-Seq:
Extending segmentation to discontiguous regions

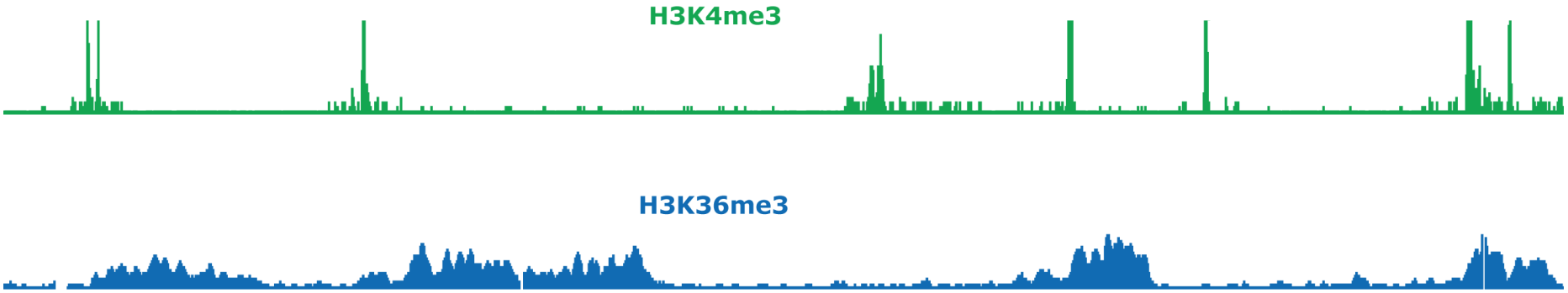
Transcript reconstruction problem as a segmentation problem



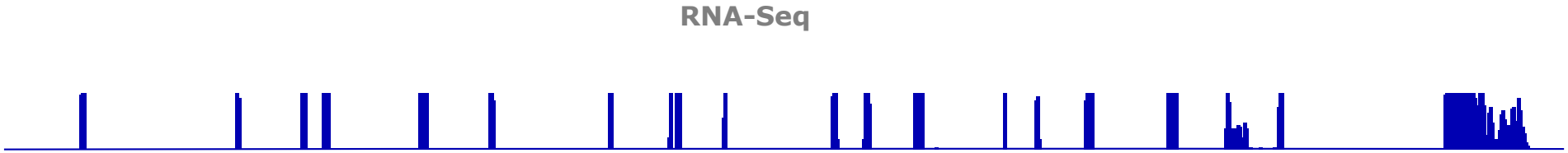
Challenges:

- Genes exist at many different expression levels, spanning several orders of magnitude.
- Reads originate from both mature mRNA (exons) and immature mRNA (introns) and it can be problematic to distinguish between them.
- Reads are short and genes can have many isoforms making it challenging to determine which isoform produced each read.

Scripture: Genome-guided transcriptome reconstruction

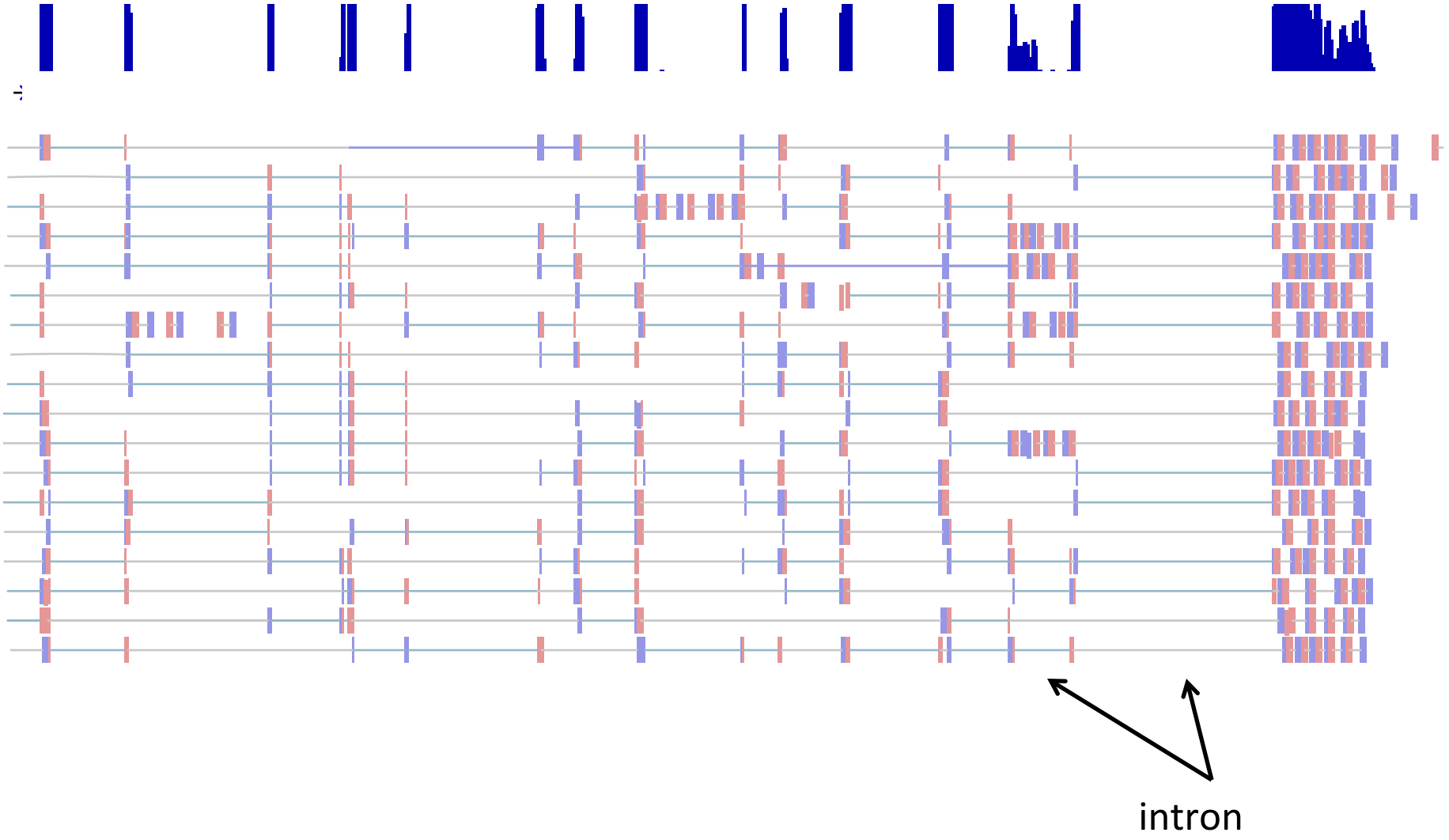


Statistical segmentation of chromatin modifications uses continuity of segments to increase power for interval detection



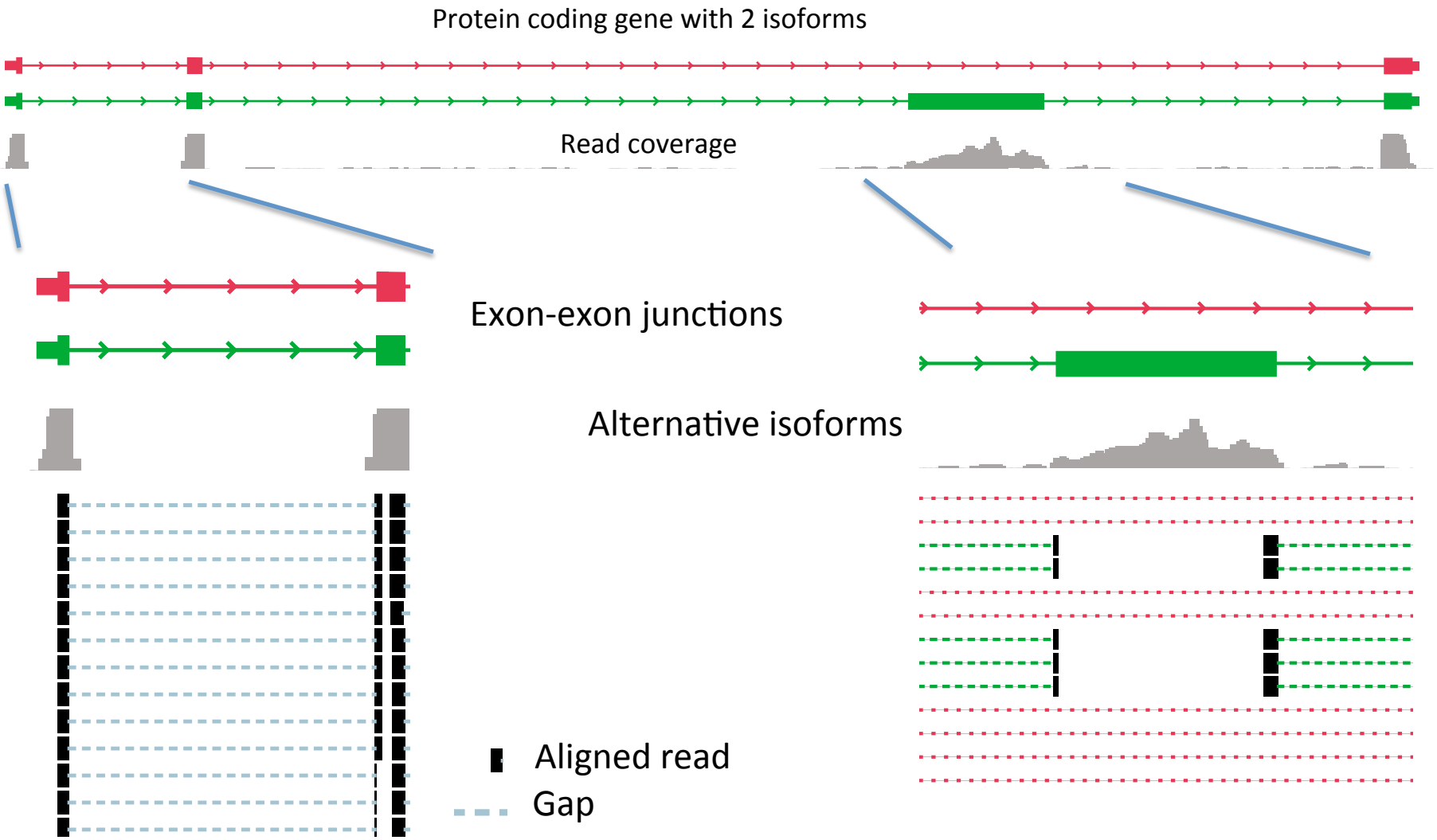
If we know the connectivity of fragments, we can increase our power to detect transcripts

Longer (76) reads increased number of junction reads



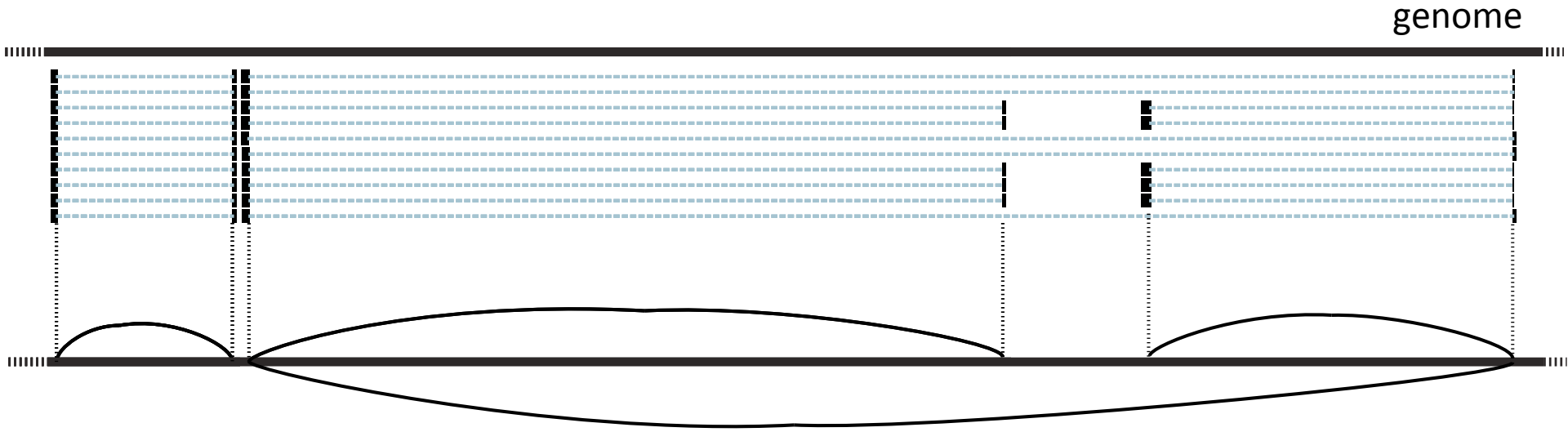
Exon junction spanning reads provide the connectivity information.

The power of spliced alignments



Statistical reconstruction of the transcriptome

Step 1: Align Reads to the genome allowing gaps flanked by splice sites

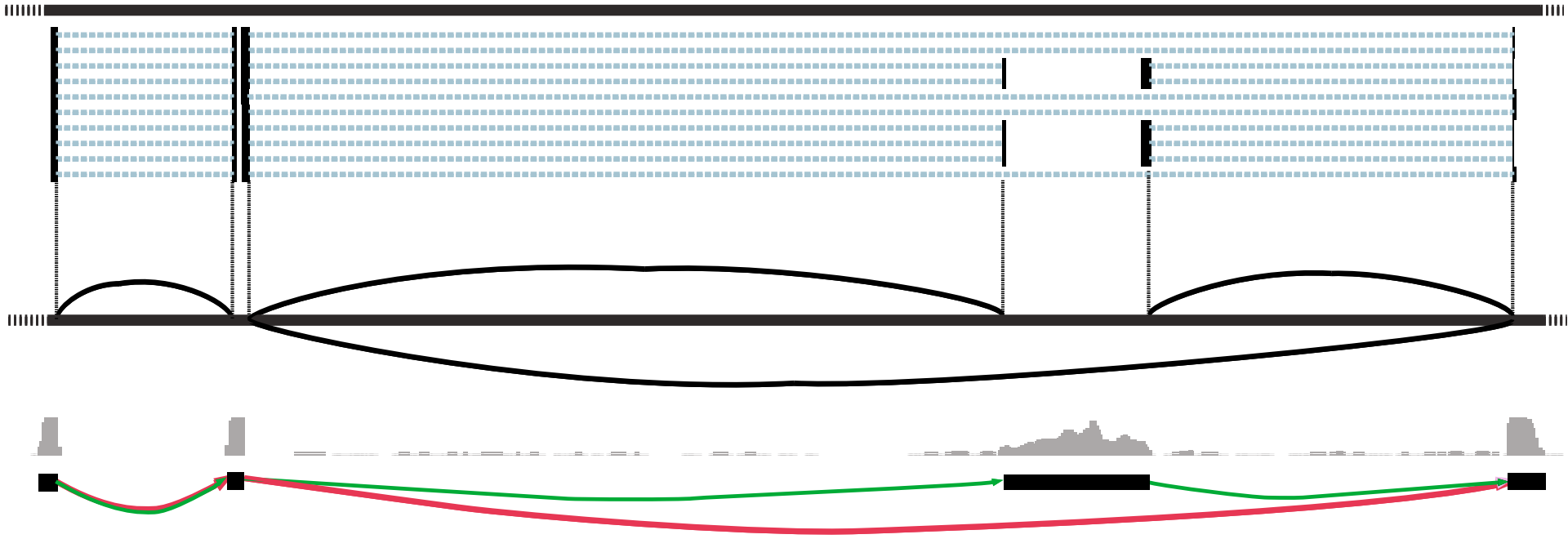


Step 2: Build an oriented connectivity graph using every spliced alignment and orienting edges using the flanking splicing motifs

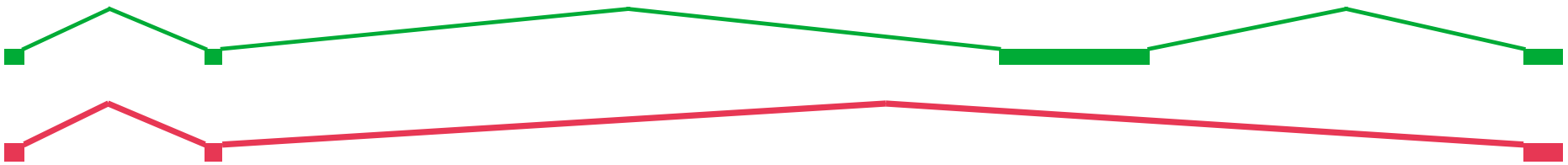
The "connectivity graph" connects all bases that are directly connected within the transcriptome

Statistical reconstruction of the transcriptome

Step 3: Identify “segments” across the graph



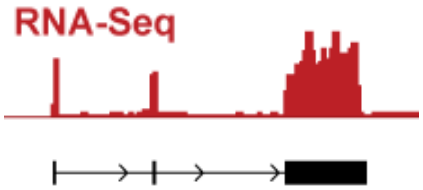
Step 4: Find significant segments



Can we identify enriched regions across different data types?

Short modification ✓

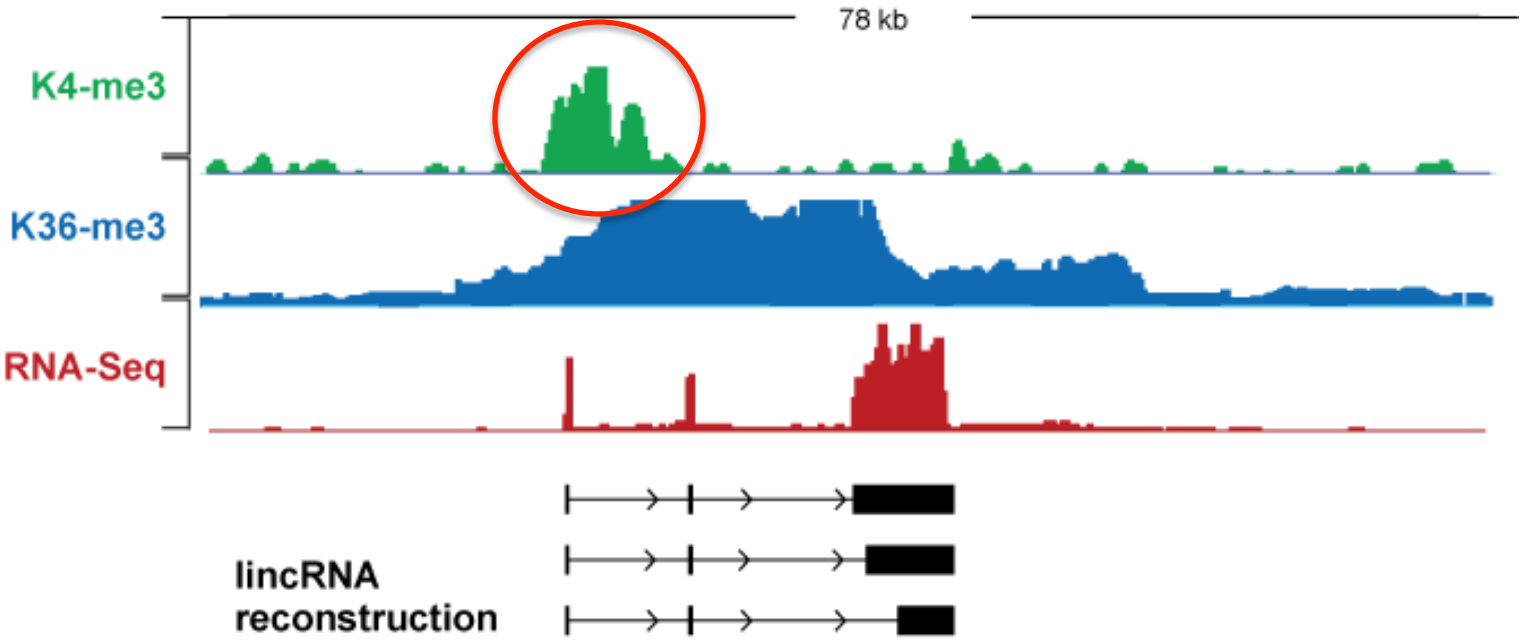
Long modification ✓



Discontinuous data ✓

Are we really sure reconstructions are complete?

RNA-Seq data is incomplete for comprehensive annotation

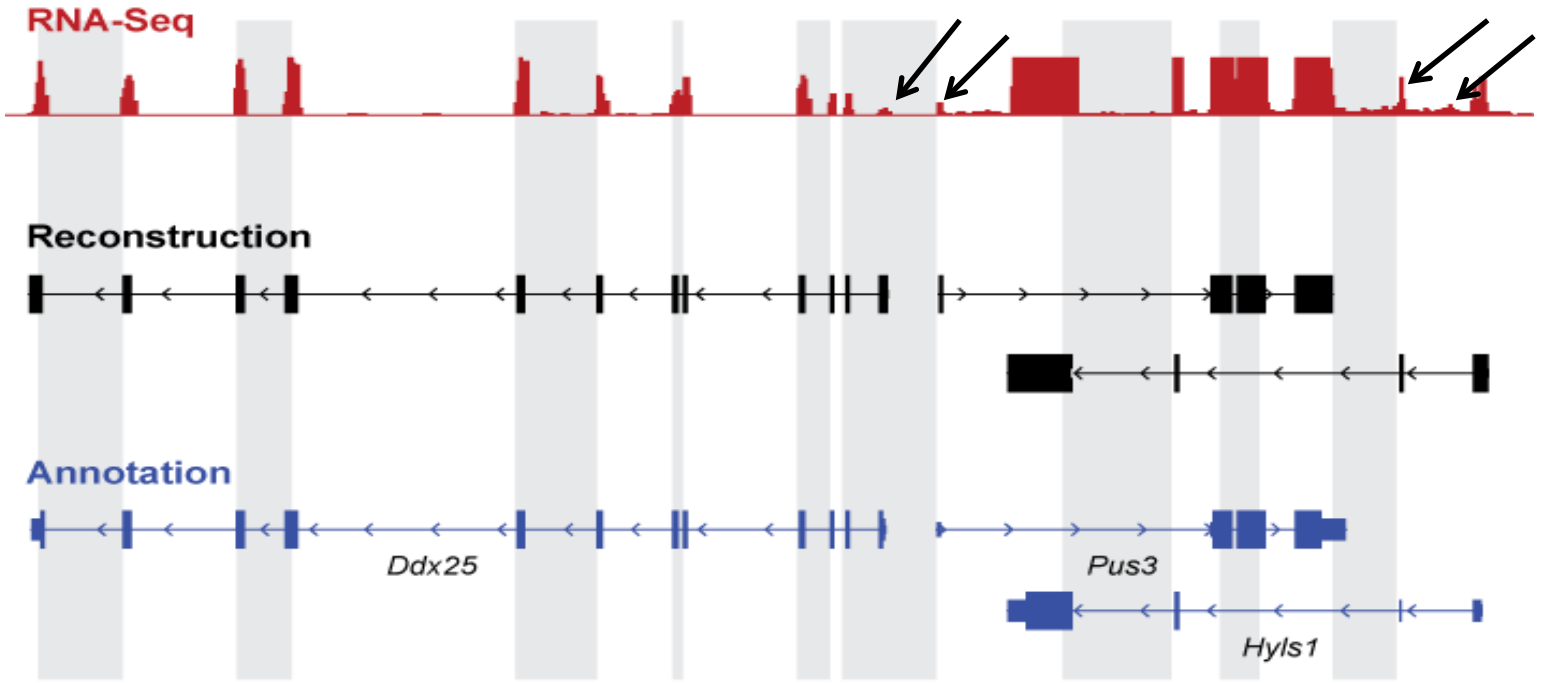


Library construction can help provide more information. More on this later

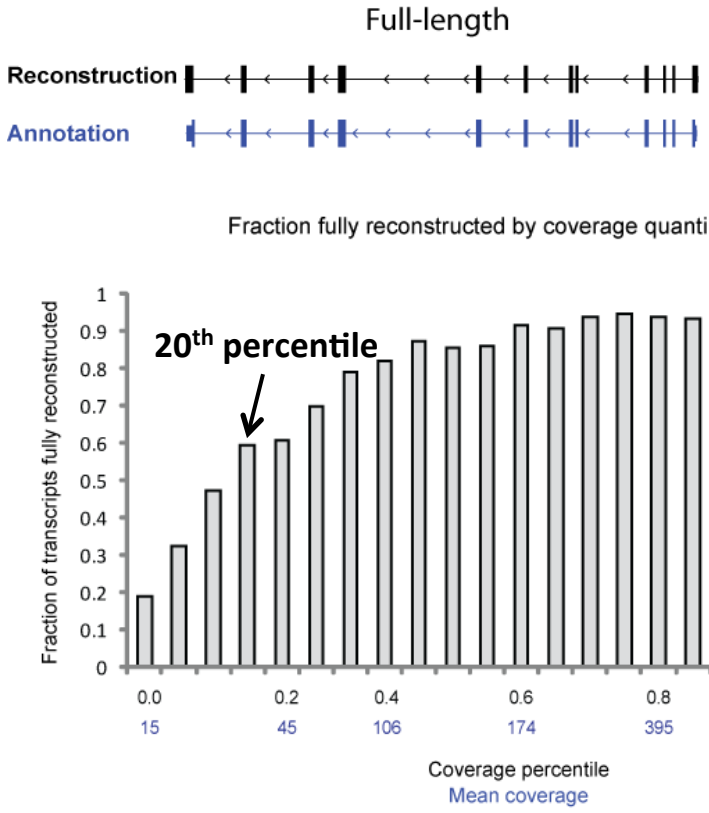
Applying scripture: Annotating the mouse transcriptome

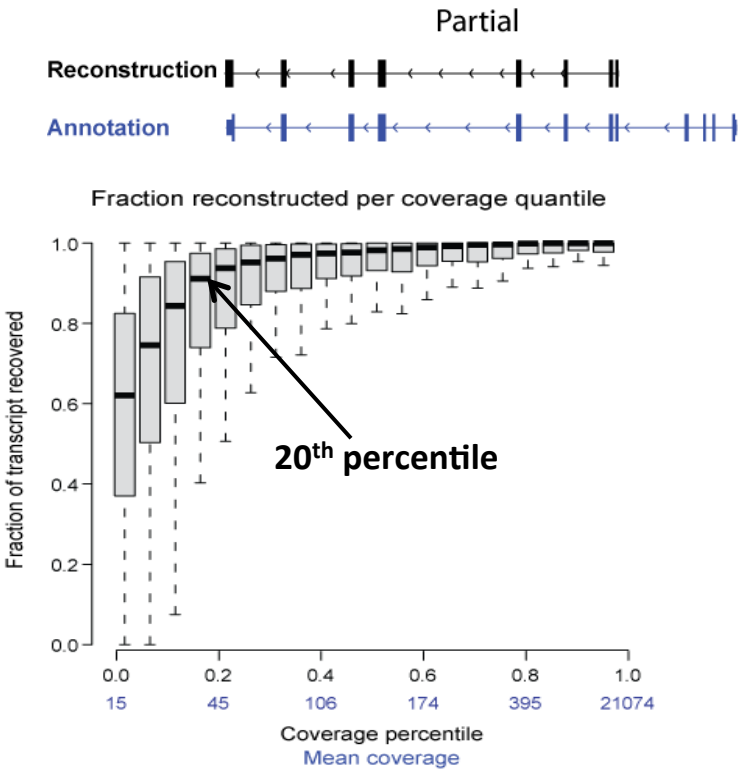
Reconstructing the mouse transcriptome (45M paired reads)

Mouse Cell Types → **Sequence** → **Reconstruct**



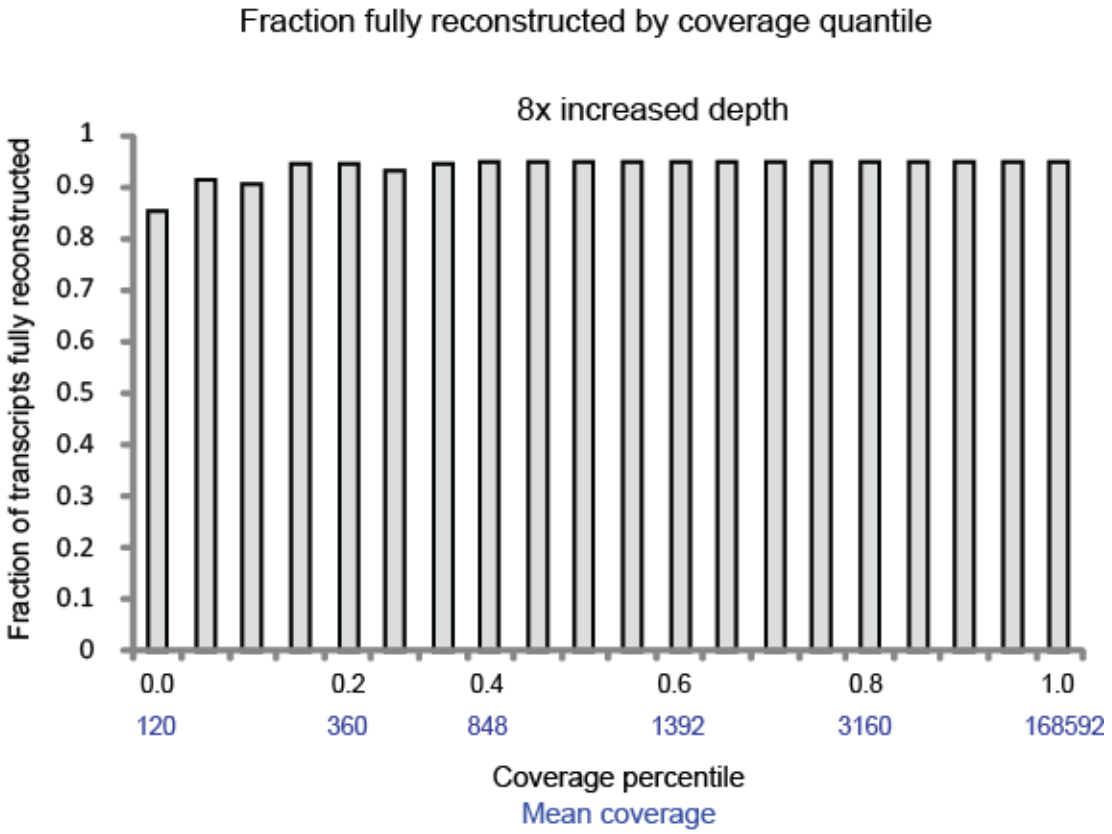
Sensitivity across expression levels





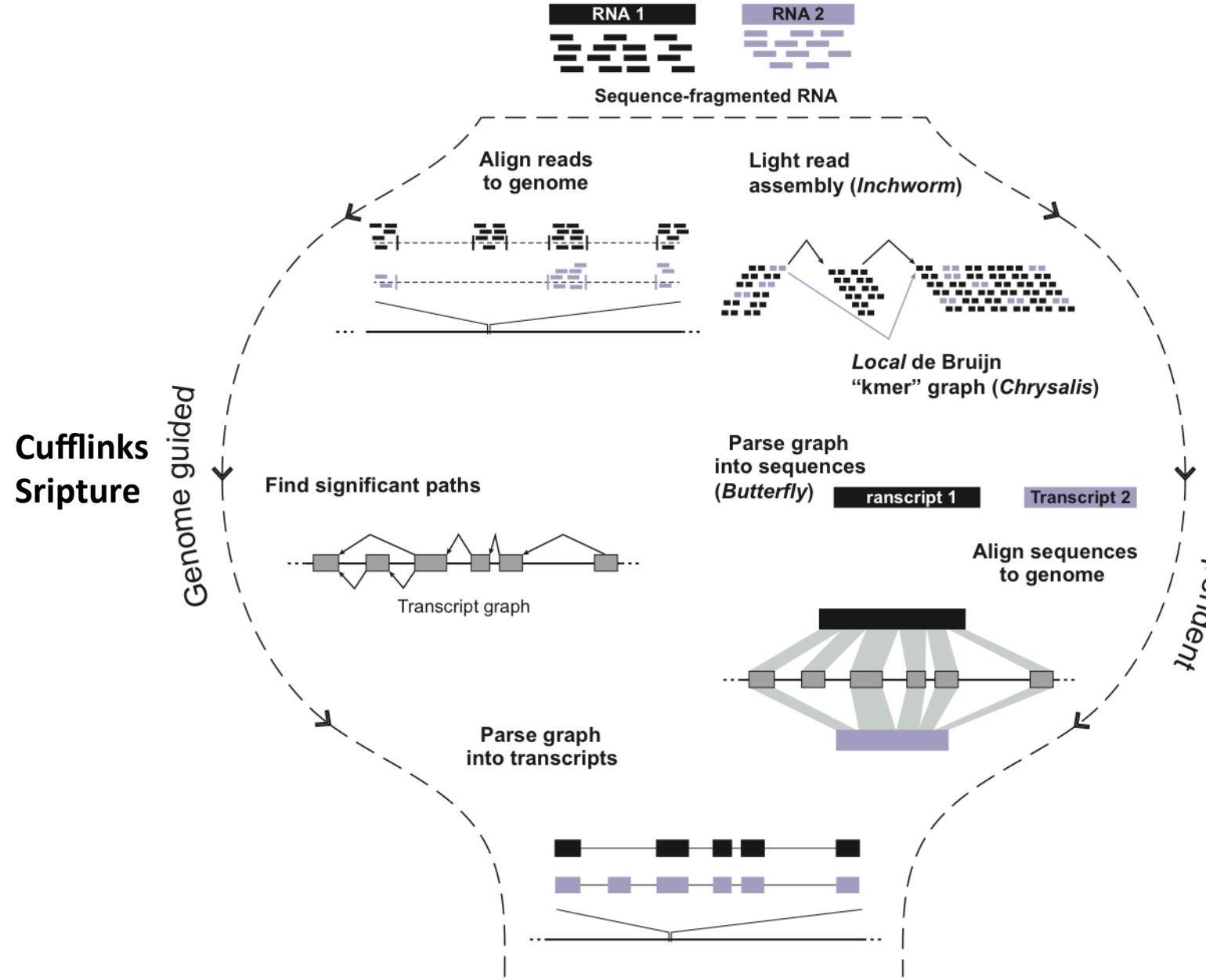
Even at low expression (20th percentile), we have:
average coverage of transcript is ~95% and 60% have full coverage

Sensitivity at low expression levels improves with depth



As coverage increases we are able to fully reconstruct a larger percentage of known protein-coding genes

If there is no reference genome!
Genome independent methods



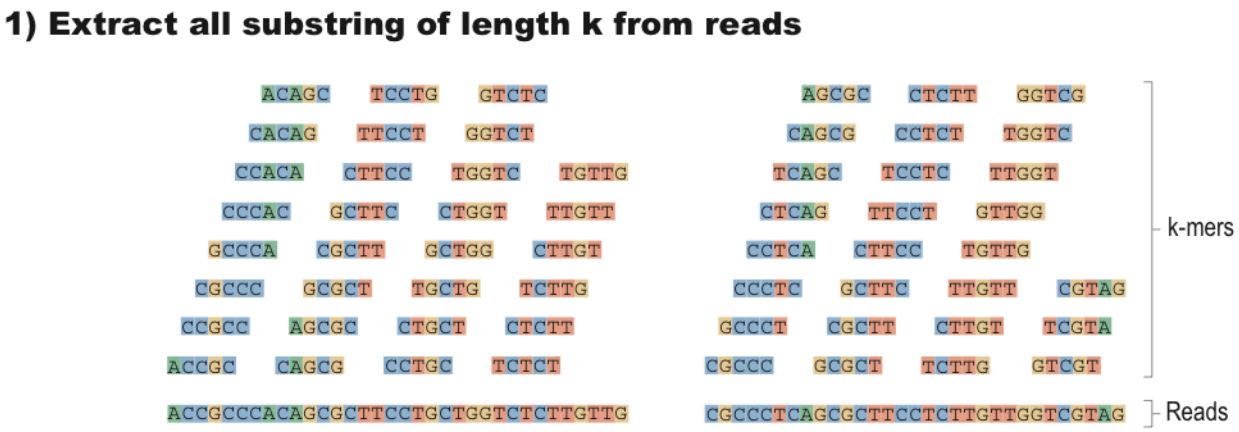
Rayan Chikhi lecture

Cufflinks
Sripture

Abyss
Trinity
Velvet

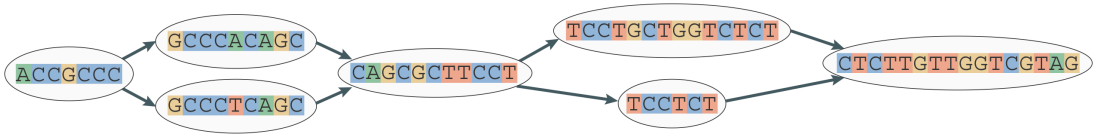
Garber et al, Nature Methods 2011

Assembly approach



Assembly approach

3) Collapse graph



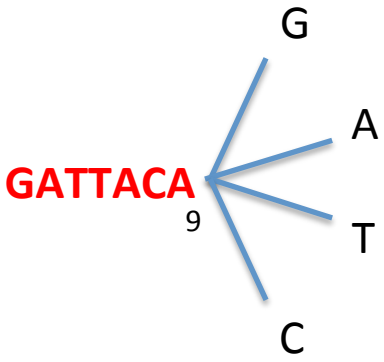
But this challenging already with DNA and RNA has many different challenges

The Trinity approach: Localize

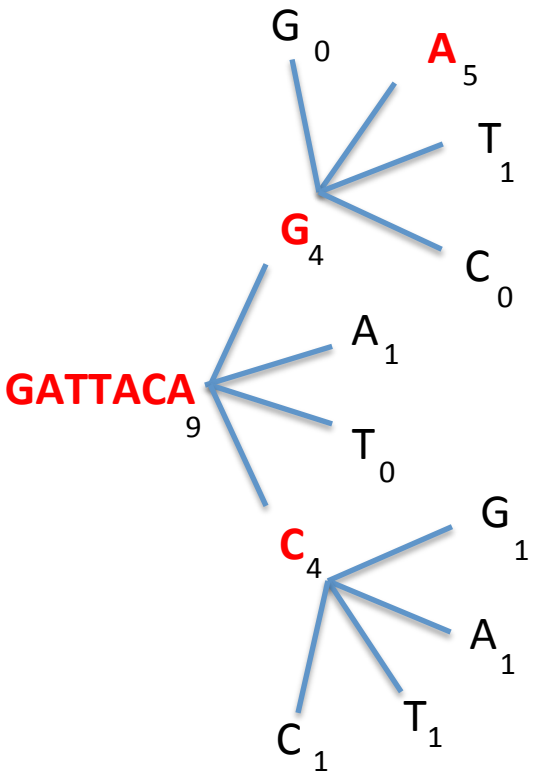
Decompose all reads into overlapping Kmers (25-mers)

Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.

Extend kmer at 3' end, guided by coverage.

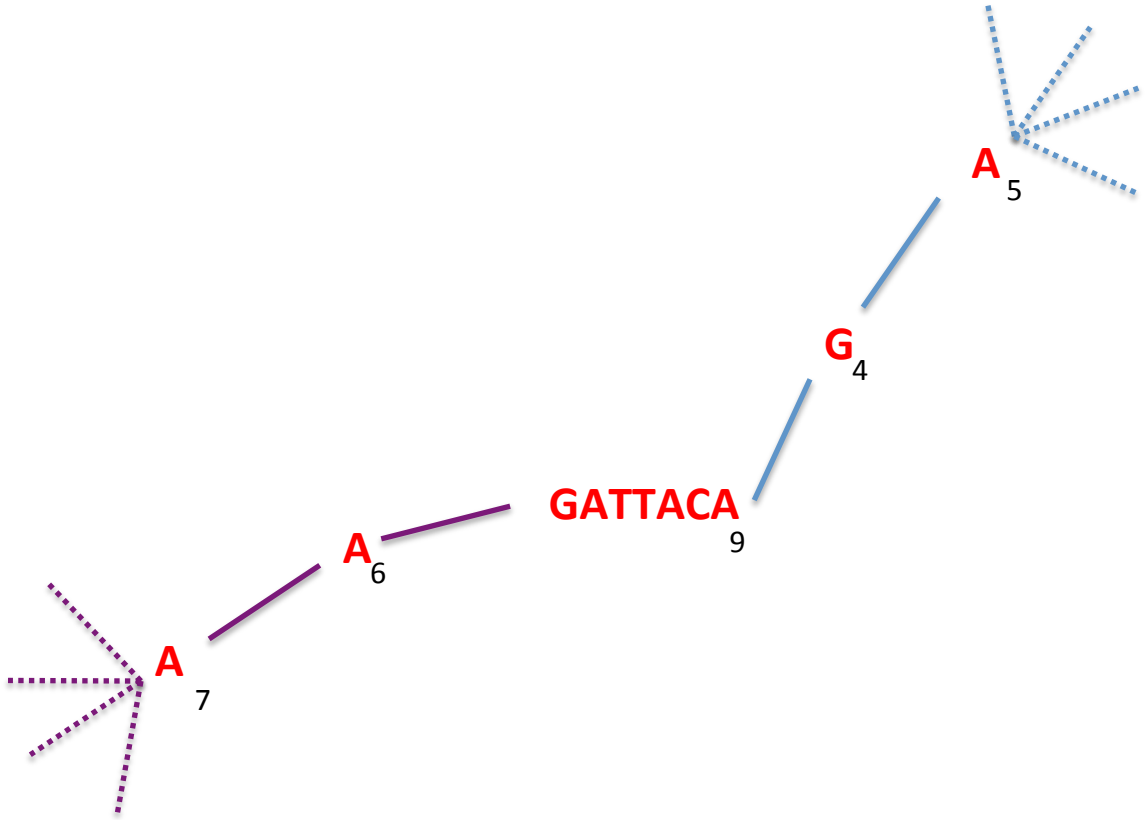


The Trinity approach: Localize



Briah Haas

The Trinity approach: Localize



Report contig: ....AAGATTACAGA....

Remove assembled kmers from catalog, then repeat the entire process.

Briah Haas

Trinity approach: Assemble

RNA-Seq reads

Group similar contigs

key: localize the assembly problem

Pros and cons of each approach

- Transcript assembly methods are the obvious choice for organisms without a reference sequence.
- Genome-guided approaches are ideal for annotating high-quality genomes and expanding the catalog of expressed transcripts and comparing transcriptomes of different cell types or conditions.
- Hybrid approaches for lesser quality or transcriptomes that underwent major rearrangements, such as in cancer cell.
- More than 1000 fold variability in expression leves makes assembly a harder problem for transcriptome assembly compared with regular genome assembly.
- Genome guided methods are very sensitive to alignment artifacts.

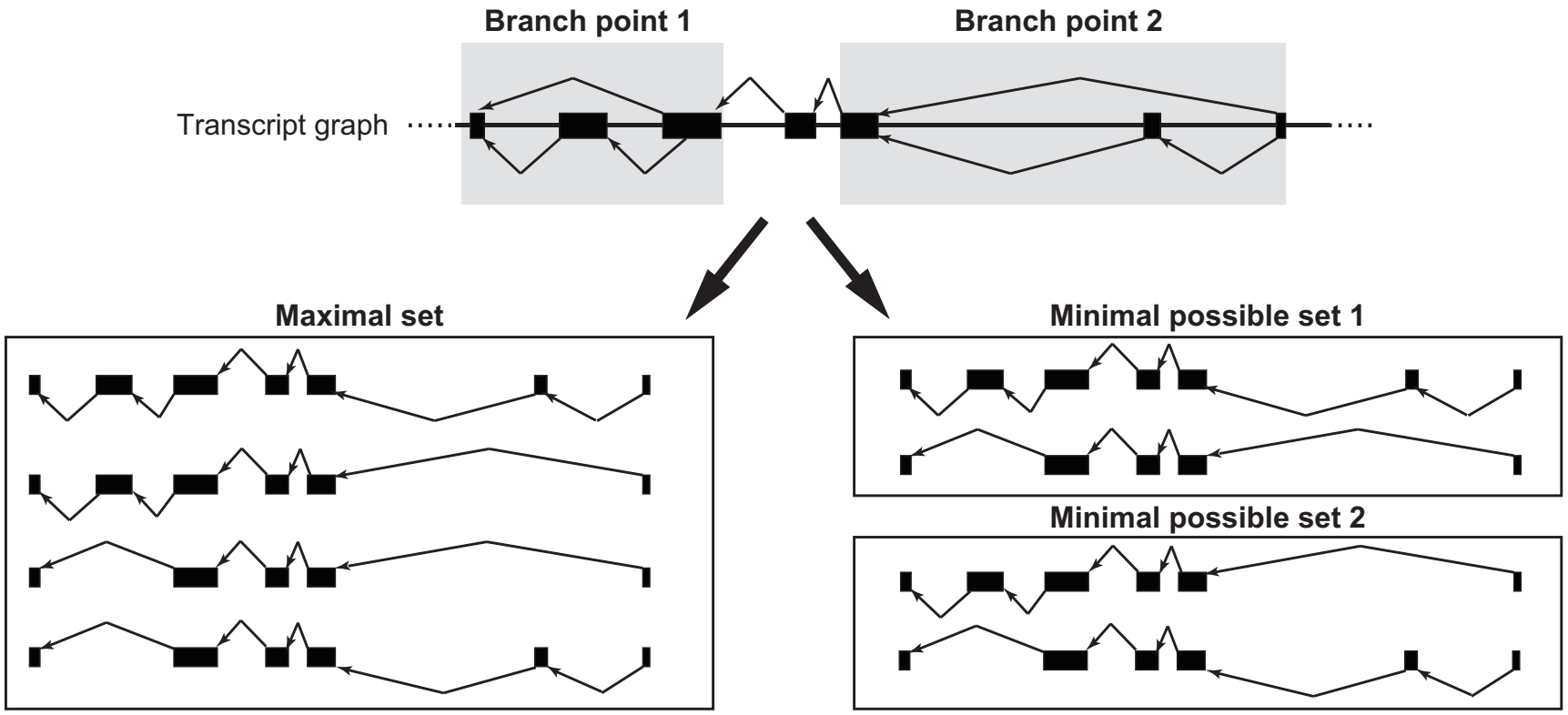
RNA-Seq transcript reconstruction software

Assembly	Genome Guided
Oasis (velvet)	Cufflinks
Trans-ABySS	Scripture
Trinity	

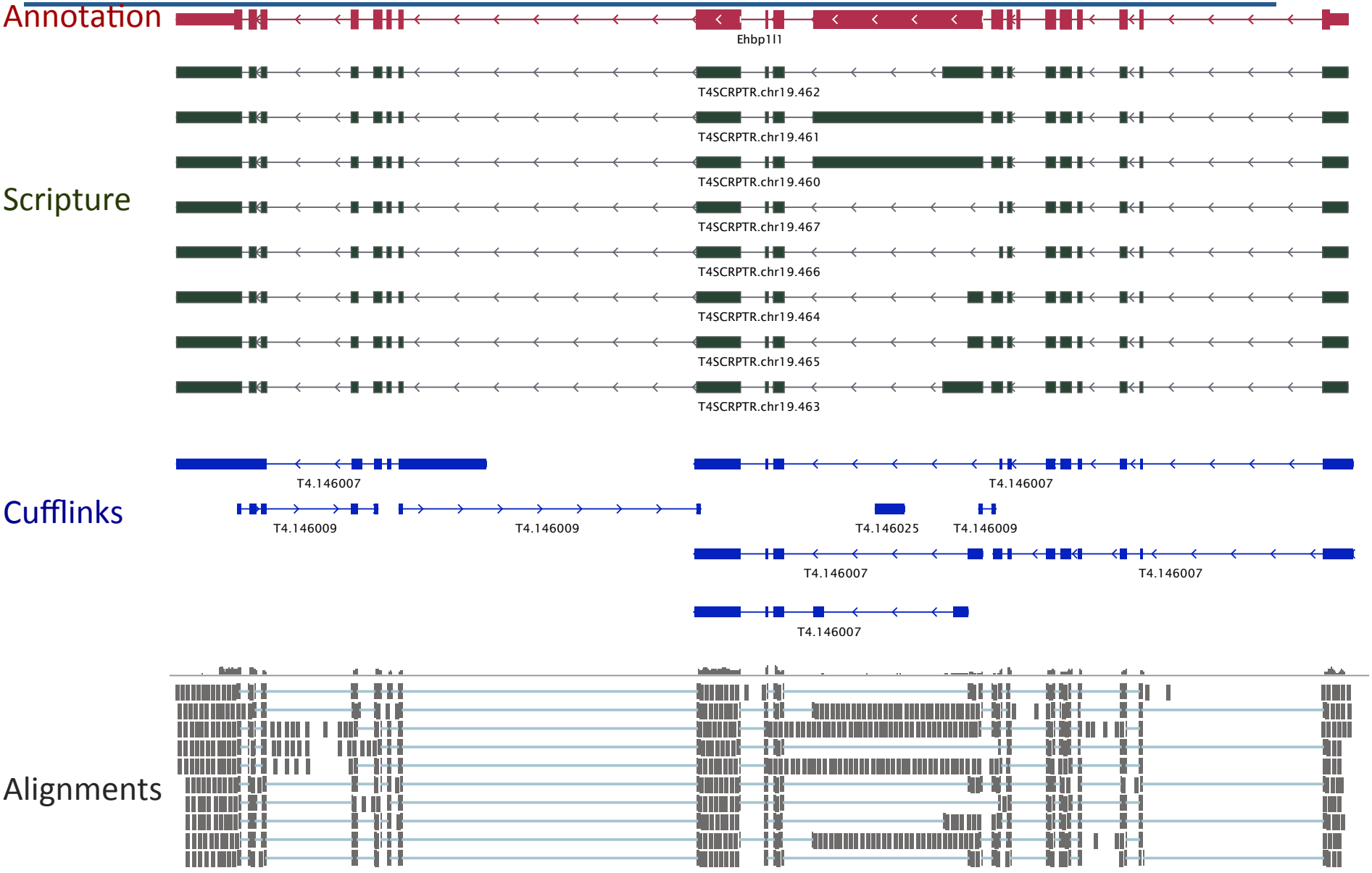
Differences between Cufflinks and Scripture

- Scripture was designed with annotation in mind. It reports all possible transcripts that are *significantly expressed* given the aligned data (*Maximum sensitivity*).
- Cuffllinks was designed with quantification in mind. It limits reported isoforms to the minimal number that explains the data (*Maximum precision*).

Maximum sensitivity vs. maximal precision



Differences between Cufflinks and Scripture - Example



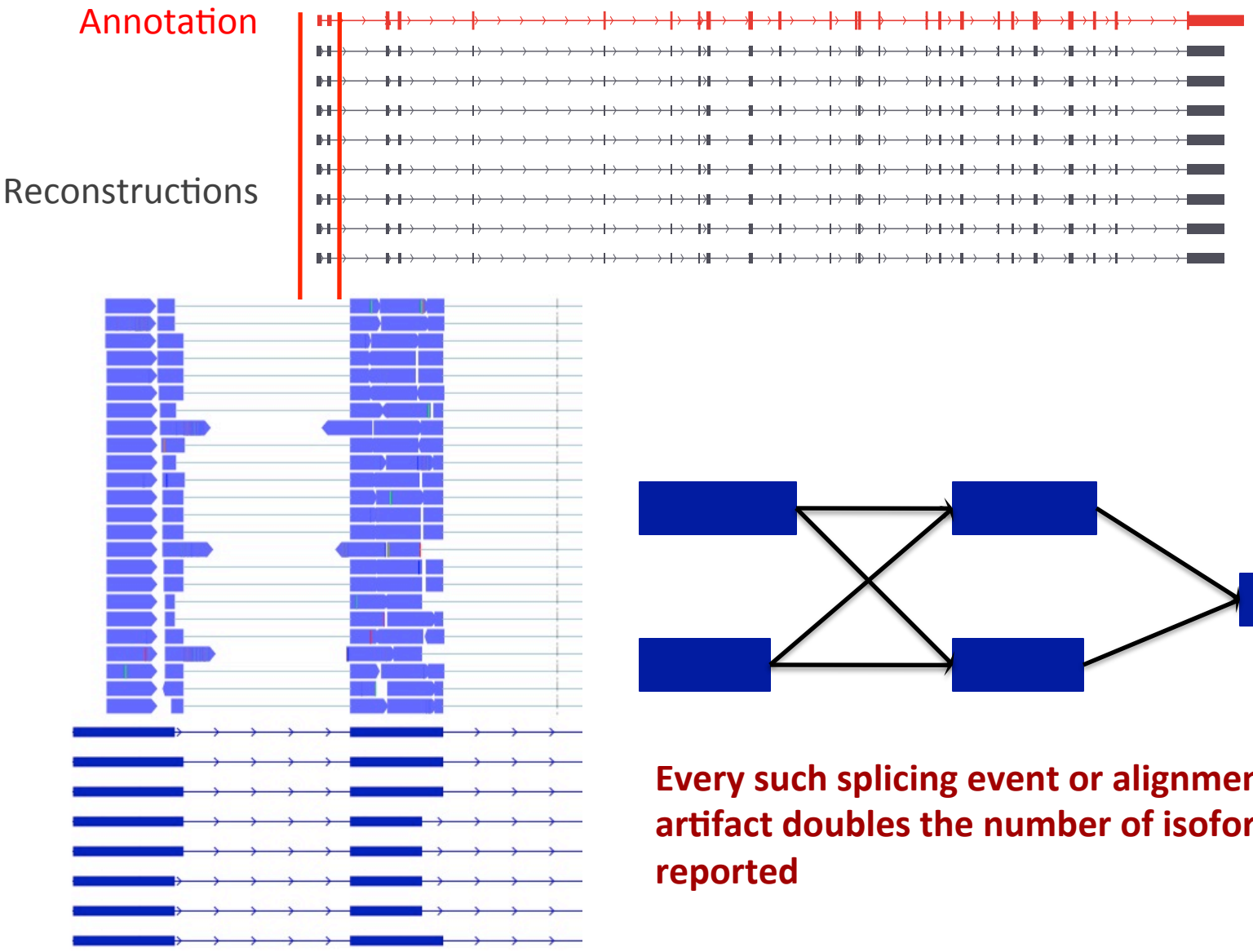
Comparing reconstructions

	CPU Hours	Total Memory	Genes fully reconstructed	Mean isoforms per reconstruction	Mean fragments per known annotation	Number of fragments predicted
Cufflinks	10	1.4 G	5,994	1.2	1.4	**159,856**
Scripture	16	3.5 G	6,221	**1.6**	1.3	61,922
Trans- Abyss	650	120 G[4]	3,330	4.7	2.6	3,117,238

Many of the bogus locus and isoforms are due to alignment artifacts

Garber et al, Nature Methods 2011

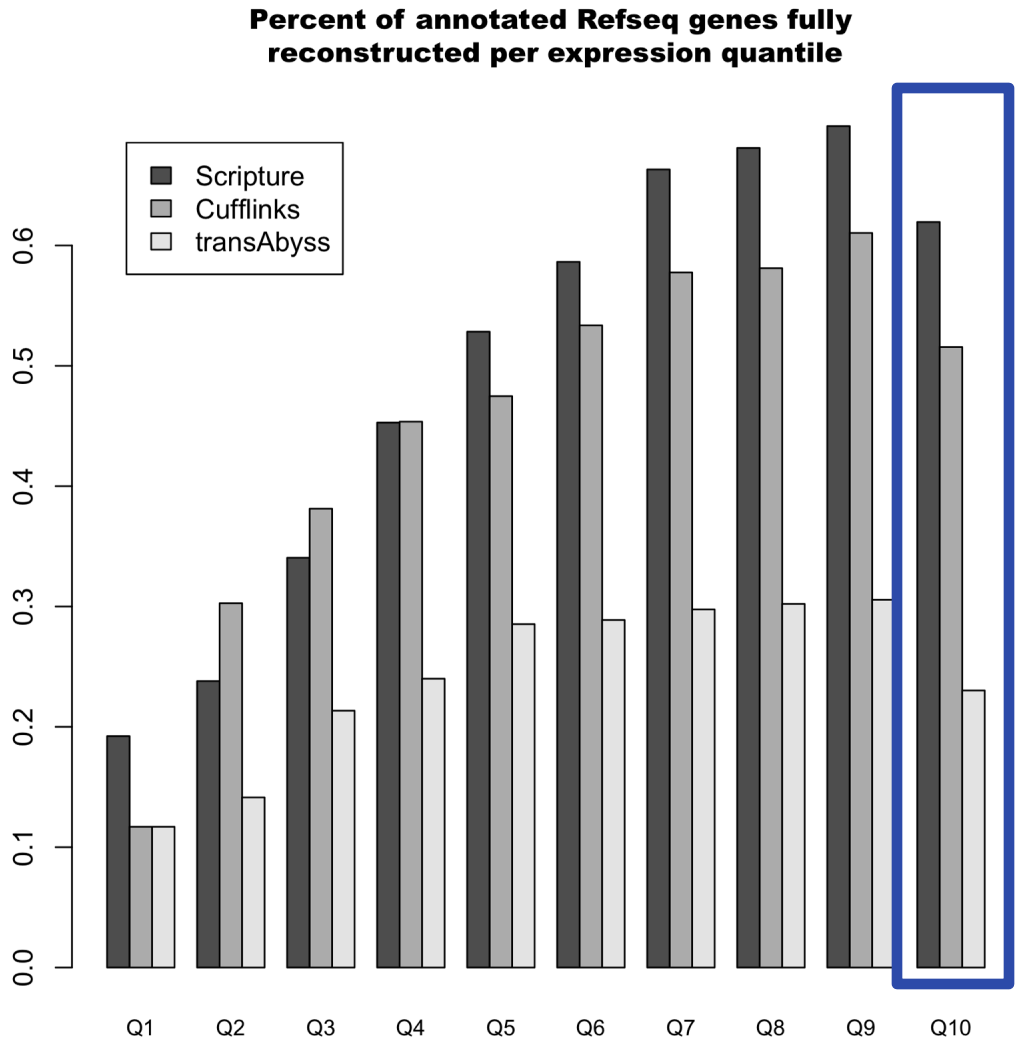
Why so many isoforms



Every such splicing event or alignment artifact doubles the number of isoforms reported

Longer reads (already possible) will reduce the uncertainty and possibilities

Reconstruction comparison



Too much of a good thing is not handled well by most reconstruction methods