UMASS University of Massachusetts Medical School

Transcript reconstruction

Summary I – Data types, file formats and utilities

- Annotation: Genomic regions
  - Genes
  - Peaks
  - *bedtools*
- Alignment: Map reads
  - BAM/SAM
  - *Samtools*
- Aggregation: Summary files
  - Wig (UCSC)
  - TDF (IGV)

Summary II – Data process

- Short read alignment (Bowtie, BWA)
  - Making the genome searchable: Hashing/BW
  - Seed an extend (hashing) vs suffix searches (BW)
  - New aligners are mix
- Spliced aligners (TopHat, STAR, GSNAP)
  - Map read fragments then strung them
  - Choosing the fragment size
  - Avoiding biases using information (junctions)
- Quantifying (RSEM/Cufflinks)
  - Read/Isoform assignment
  - Normalization procedures
- Differential expression (DESeq/EdgeR/Cufflinks)

Summary III – Using a graphical user interface

- Galaxy – for knowledgable users who are not comfortable with UNIX
- All tools available
- Not great for many samples

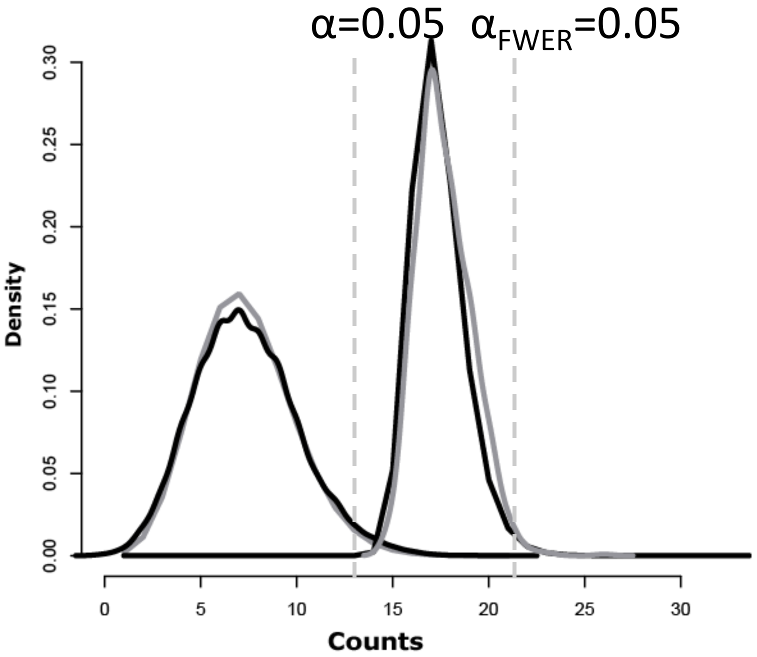
Todays topics

- Assemble transcripts from RNA-Seq (theory)
- Processing large number of samples using pre-designed pipelines.

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

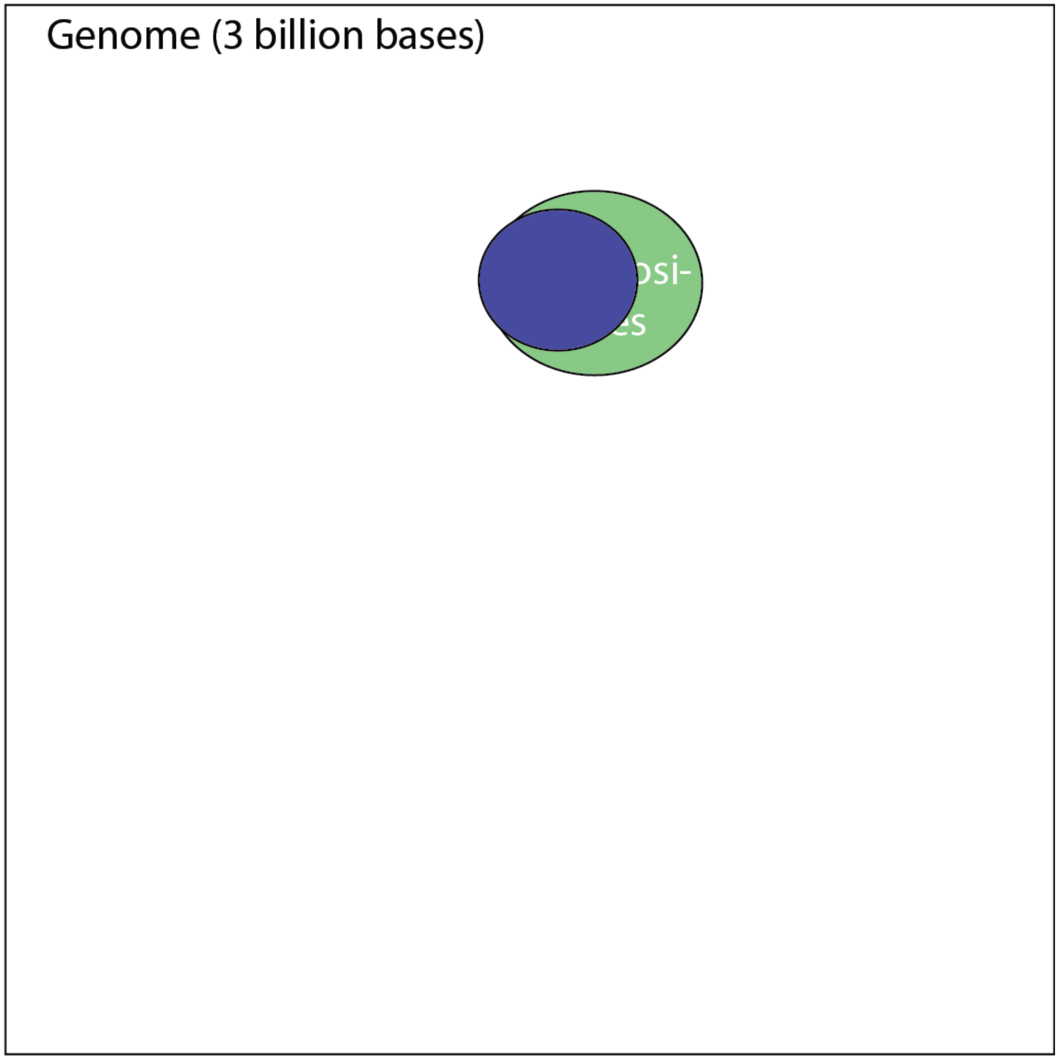
$P(k-1|\lambda w)$ is the Poisson probability of observing $k-1$ counts given an expected count of λw

and

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

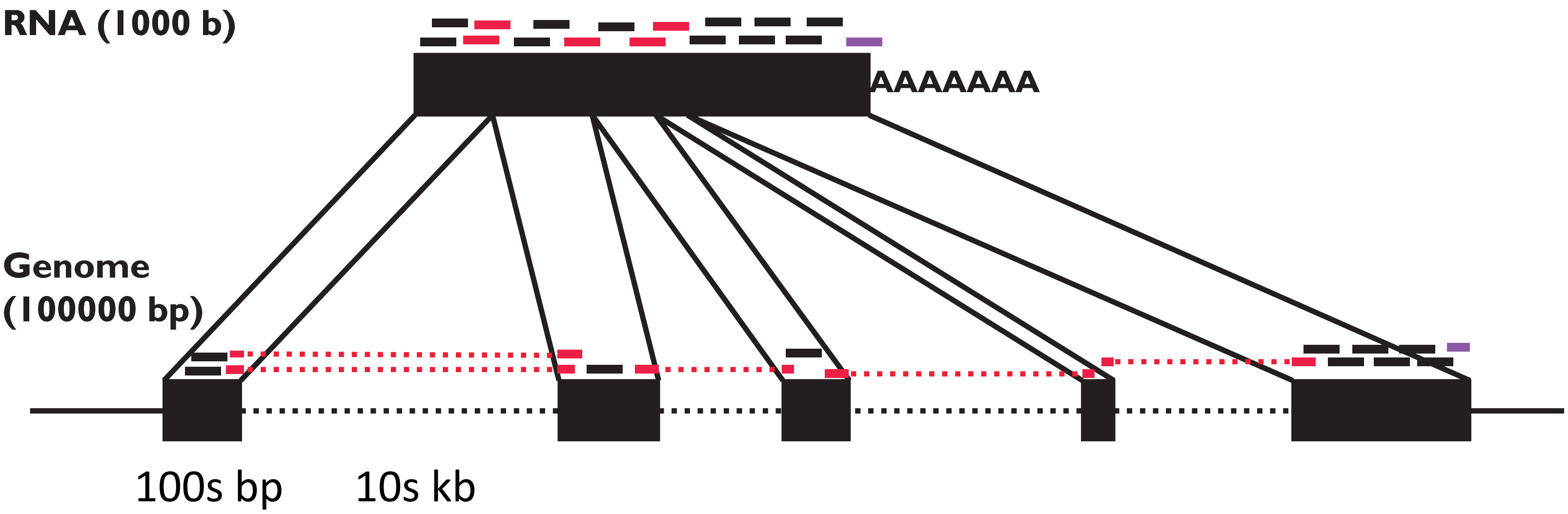
The scan distribution gives a computationally very efficient way to estimate the FWER

FWER-Scan Statistics



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

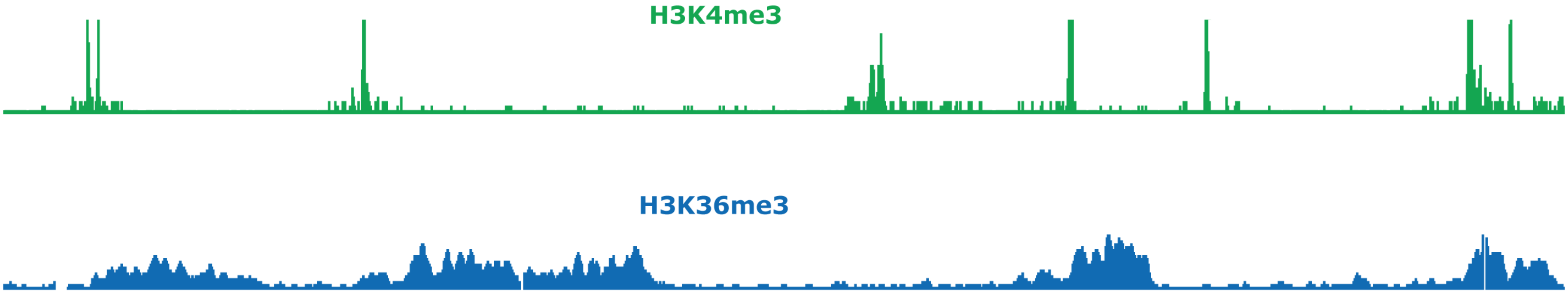
Transcript reconstruction problem as a segmentation problem



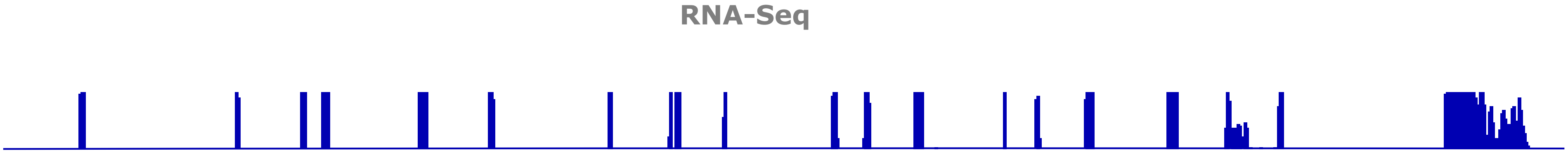
Challenges:

- 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

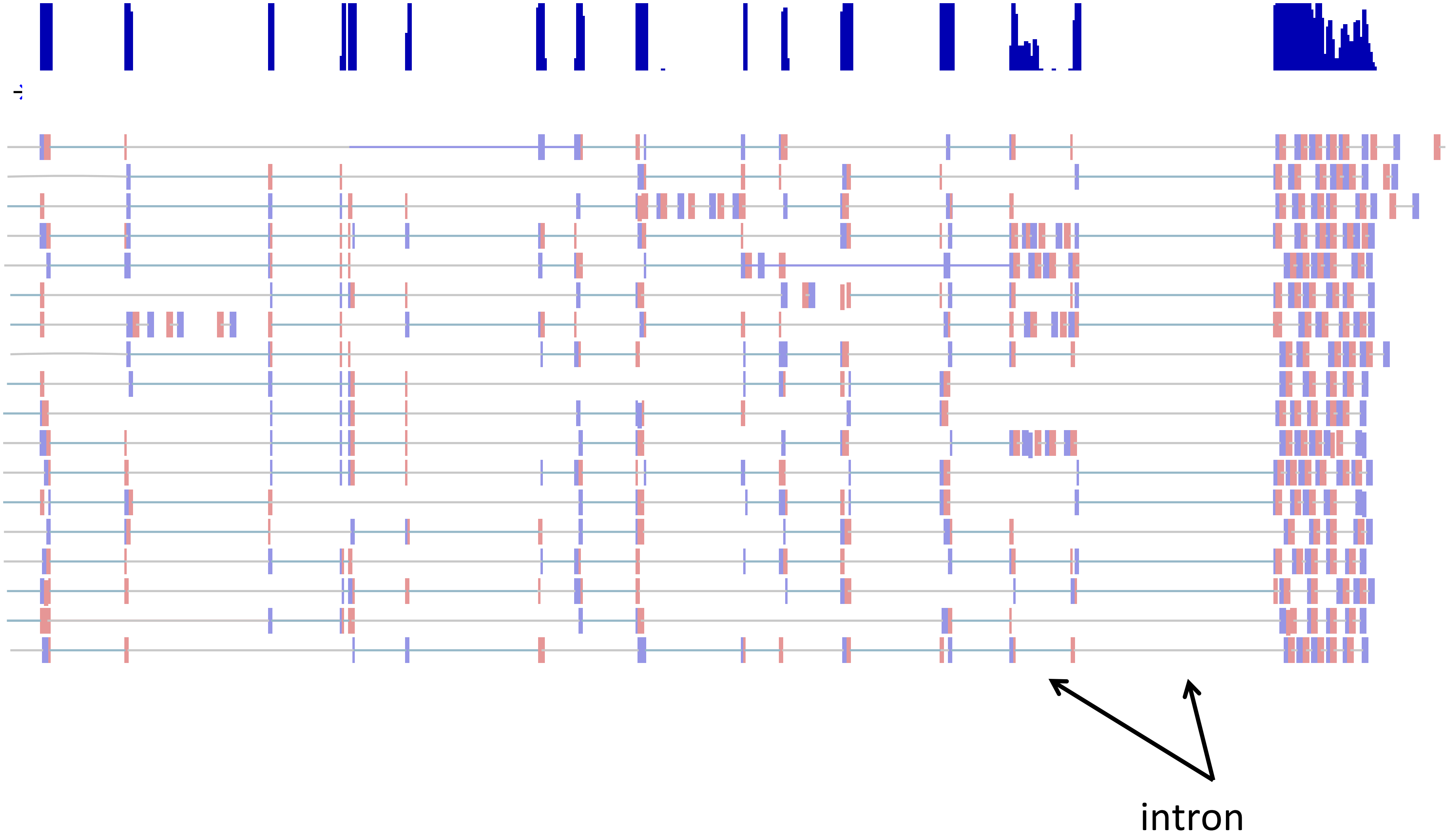


Statistical segmentation of chromatin modifications uses continuity of segments to increase power for interval detection



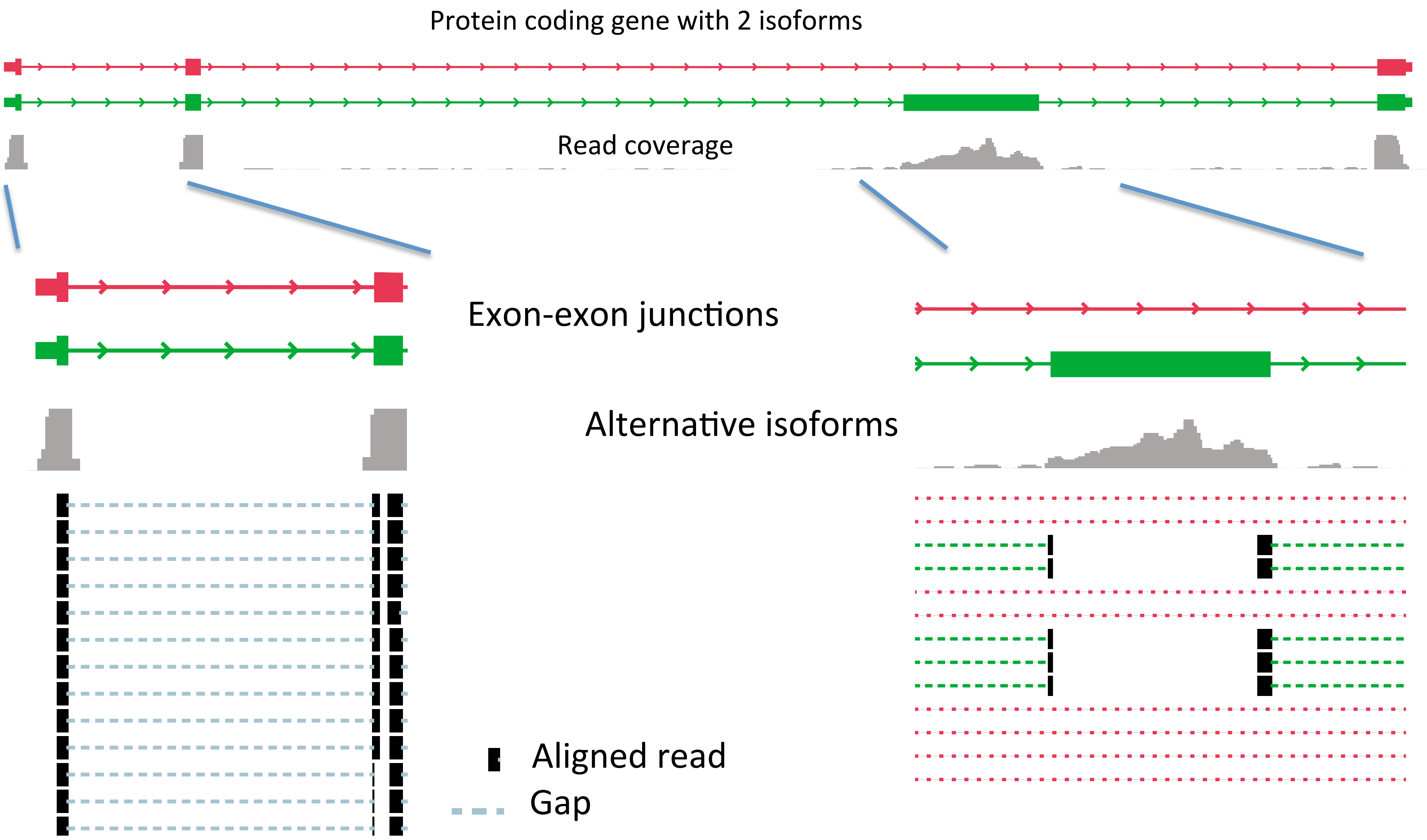
If we know the connectivity of fragments, we can increase our power to detect transcripts

Longer (76) reads increased number of junction reads



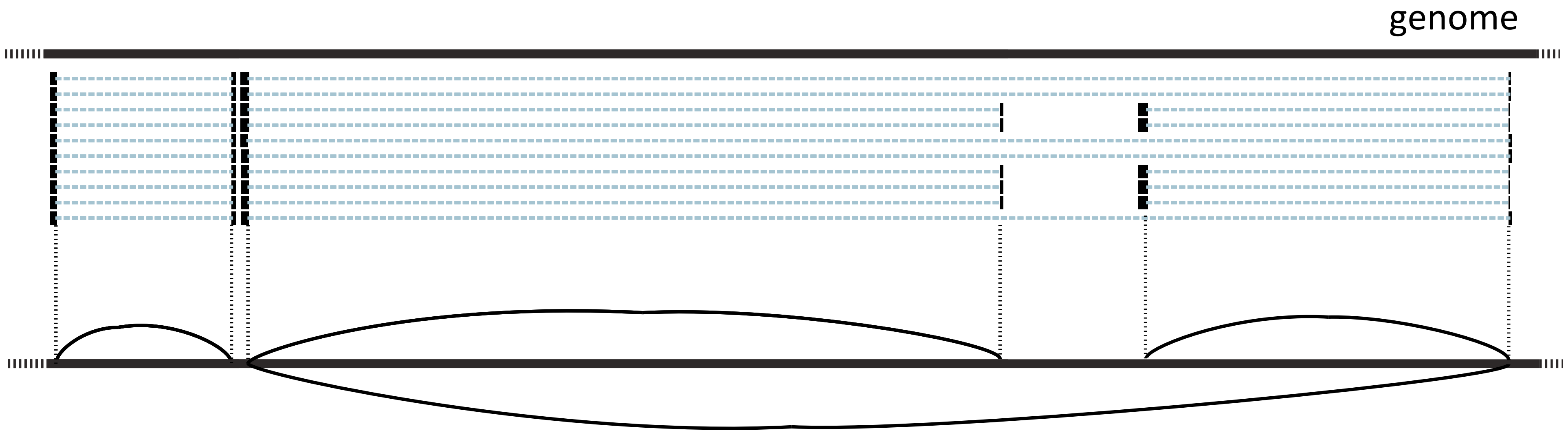
Exon junction spanning reads provide the connectivity information.

The power of spliced alignments



Statistical reconstruction of the transcriptome

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

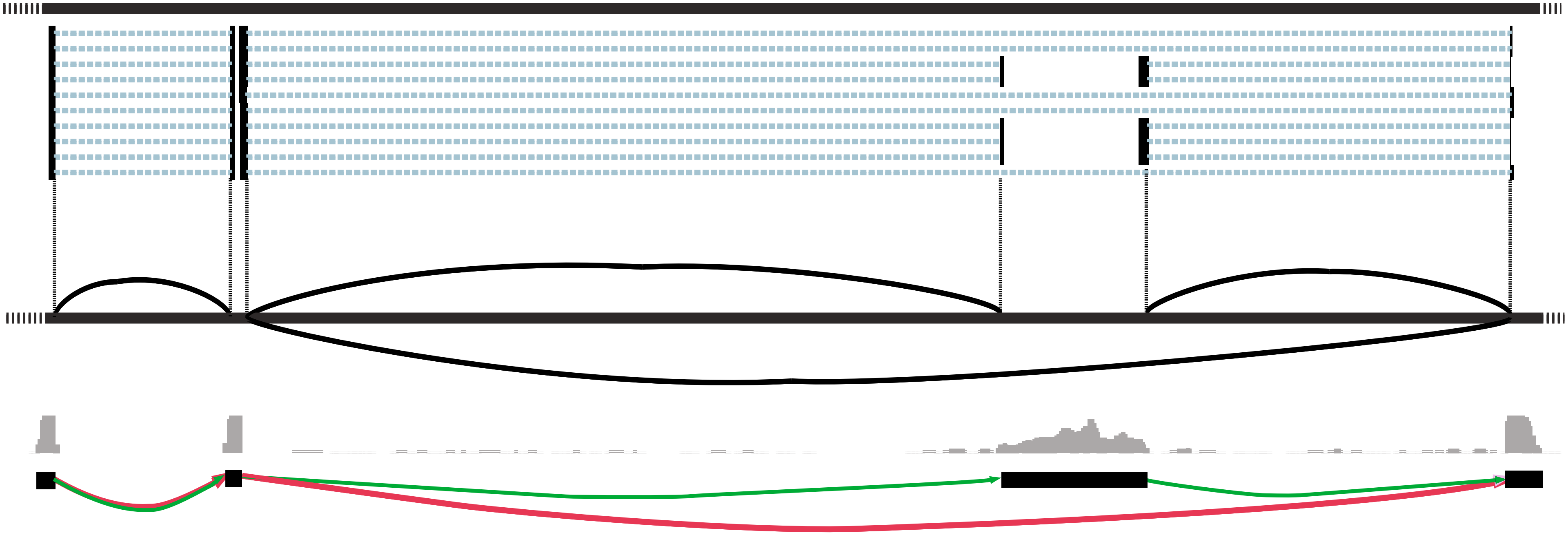


Step 2: Build an oriented connectivity graph using every spliced alignment and orienting edges using the flanking splicing motifs

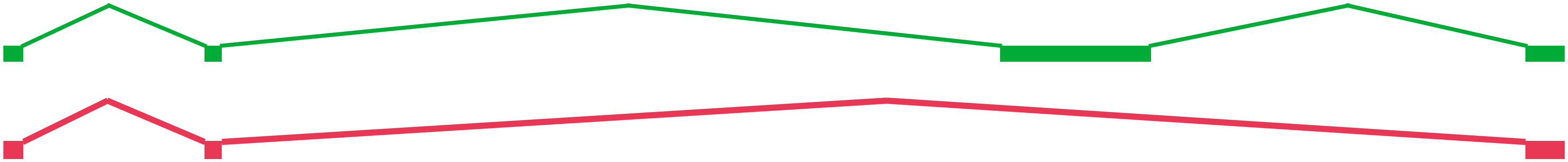
The "connectivity graph" connects all bases that are directly connected within the transcriptome

Statistical reconstruction of the transcriptome

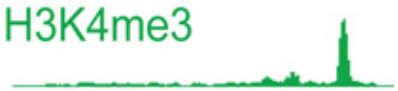
Step 3: Identify "segments" across the graph



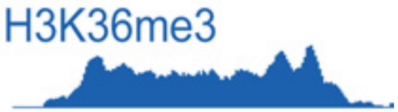
Step 4: Find significant segments



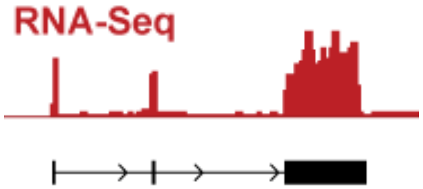
Can we identify enriched regions across different data types?



Short modification ✓



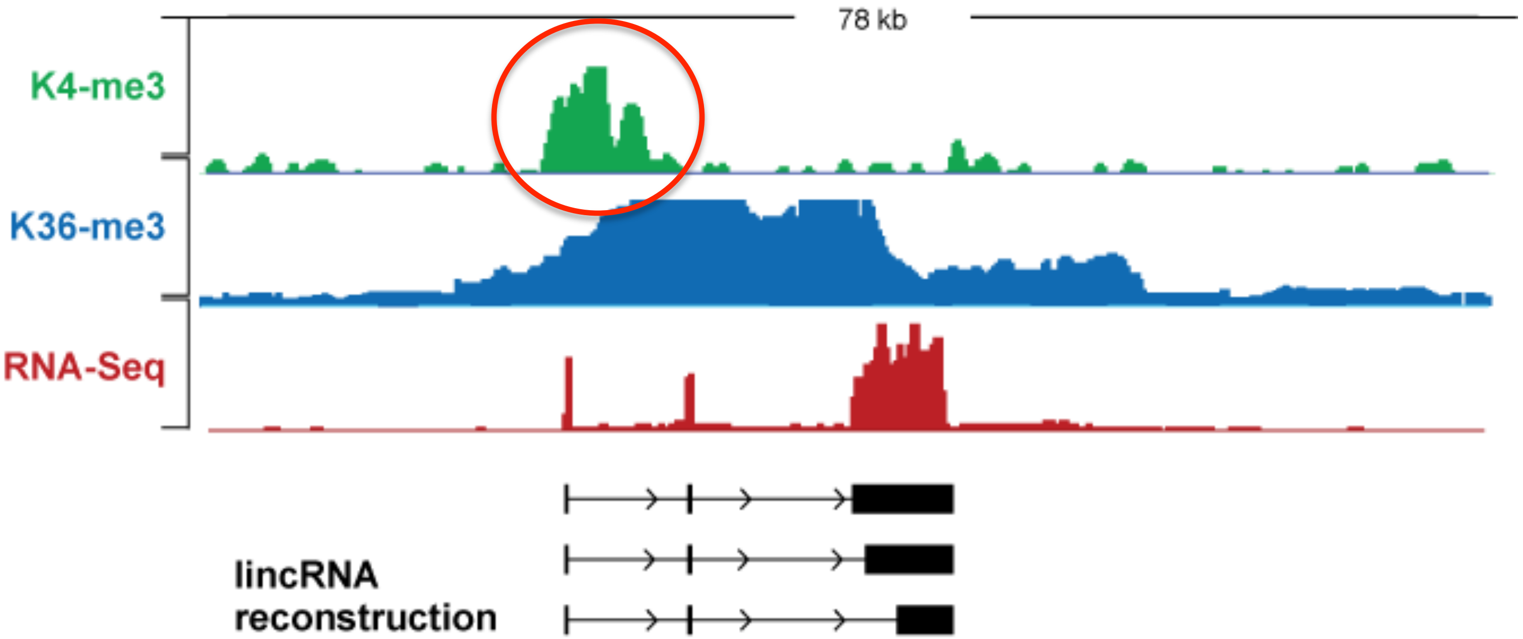
Long modification ✓



Discontinuous data ✓

Are we really sure reconstructions are complete?

RNA-Seq data is incomplete for comprehensive annotation

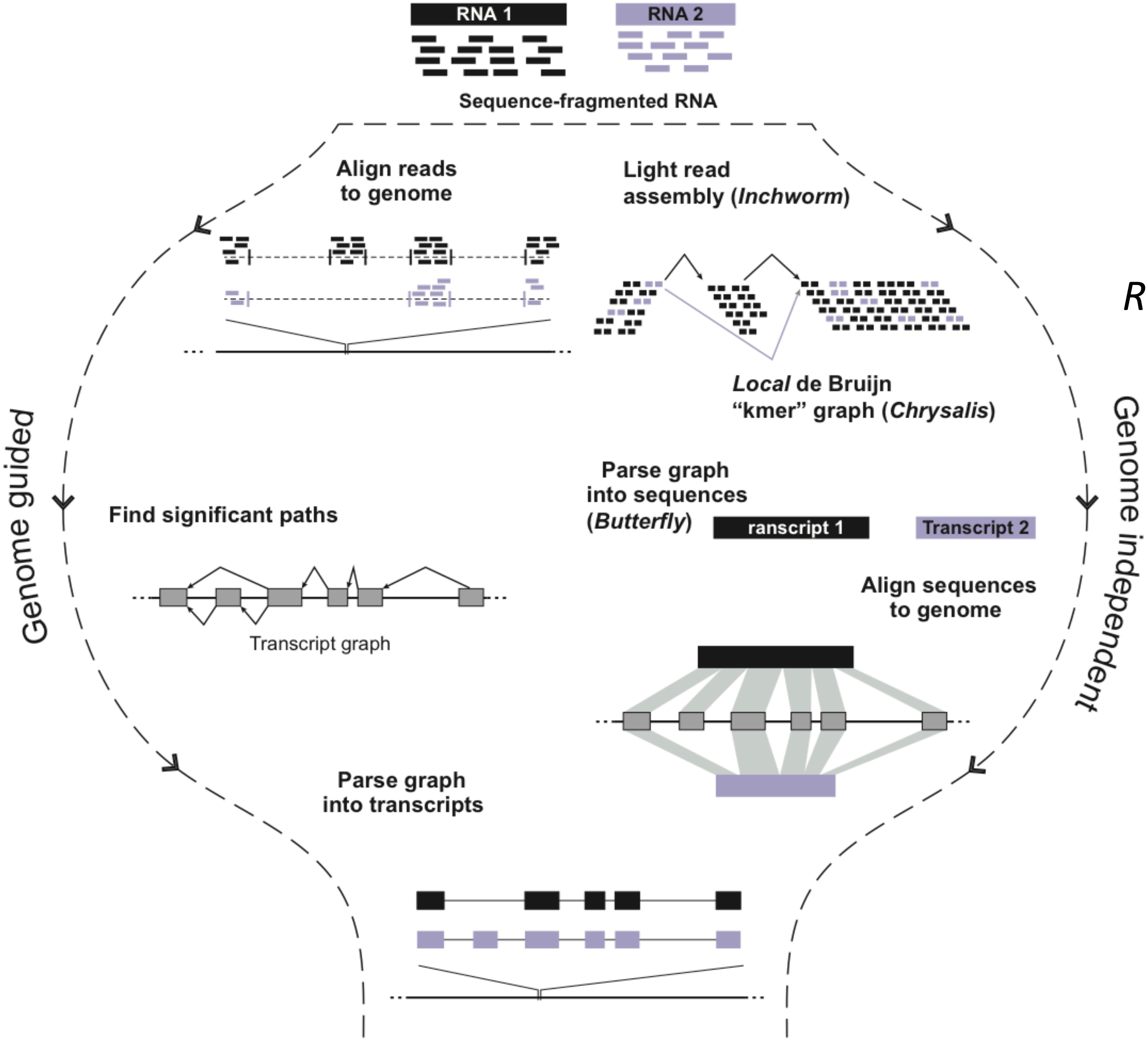


Library construction can help provide more information. More on this later

If there is no reference genome!
Genome independent methods

Rayan Chikhi lecture

Cufflinks
Sripture



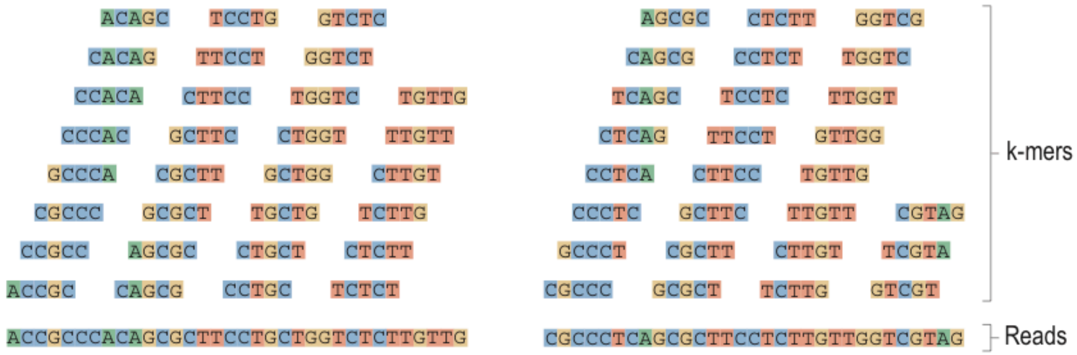
RNA 1
RNA 2
Sequence-fragmented RNA
Align reads to genome
Light read assembly (*Inchworm*)
Local de Bruijn "kmer" graph (*Chrysalis*)
Genome guided
Genome independent
Find significant paths
Transcript graph
Parse graph into sequences (*Butterfly*)
ranscript 1
Transcript 2
Align sequences to genome
Parse graph into transcripts

Abyss
Trinity
Velvet

Garber et al, Nature Methods 2011

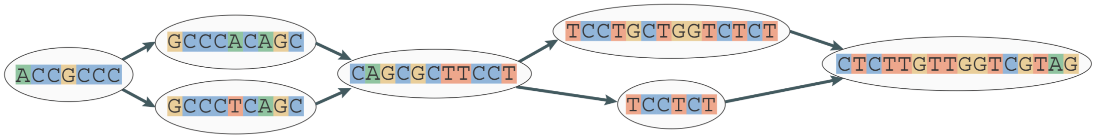
Assembly approach

1) Extract all substring of length k from reads



Assembly approach

3) Collapse graph



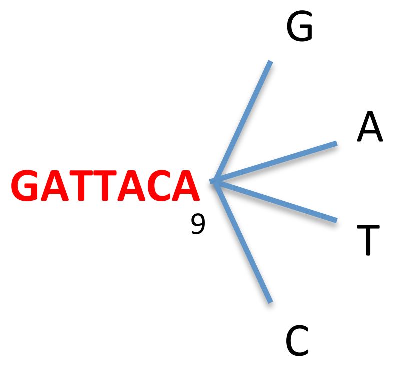
But this challenging already with DNA and RNA has many different challenges

The Trinity approach: Localize

Decompose all reads into overlapping Kmers (25-mers)

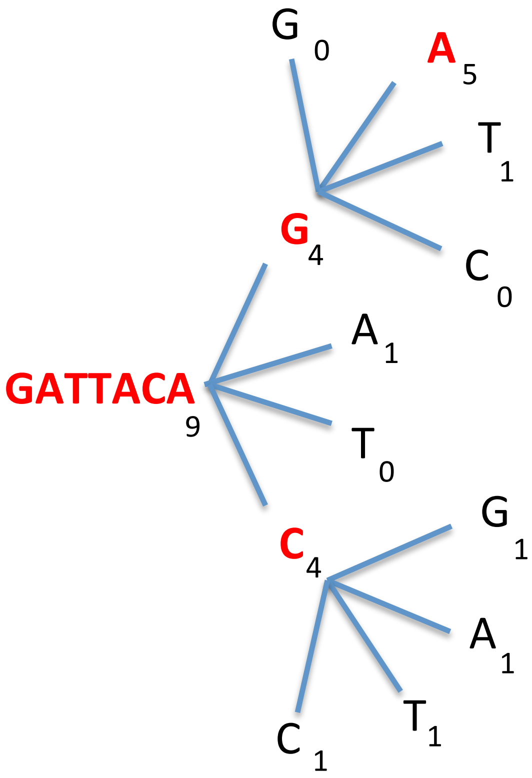
Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.

Extend kmer at 3’ end, guided by coverage.



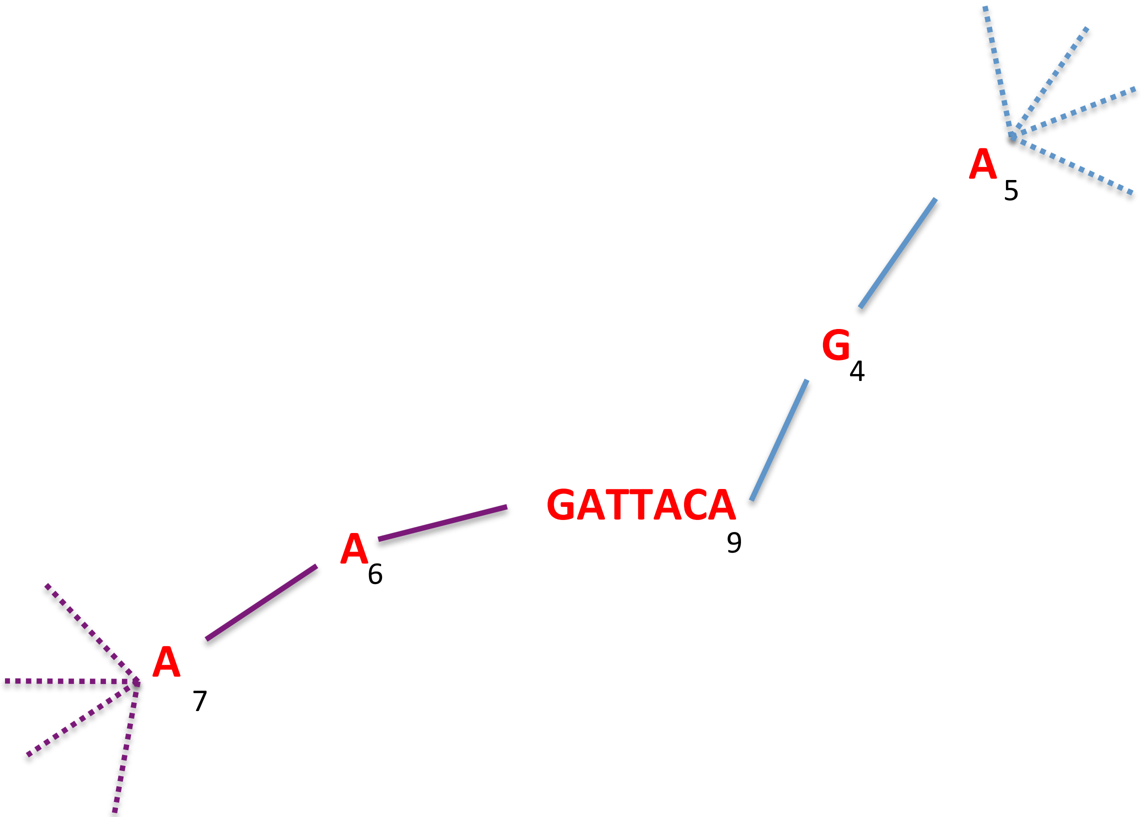
Briah Haas

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.

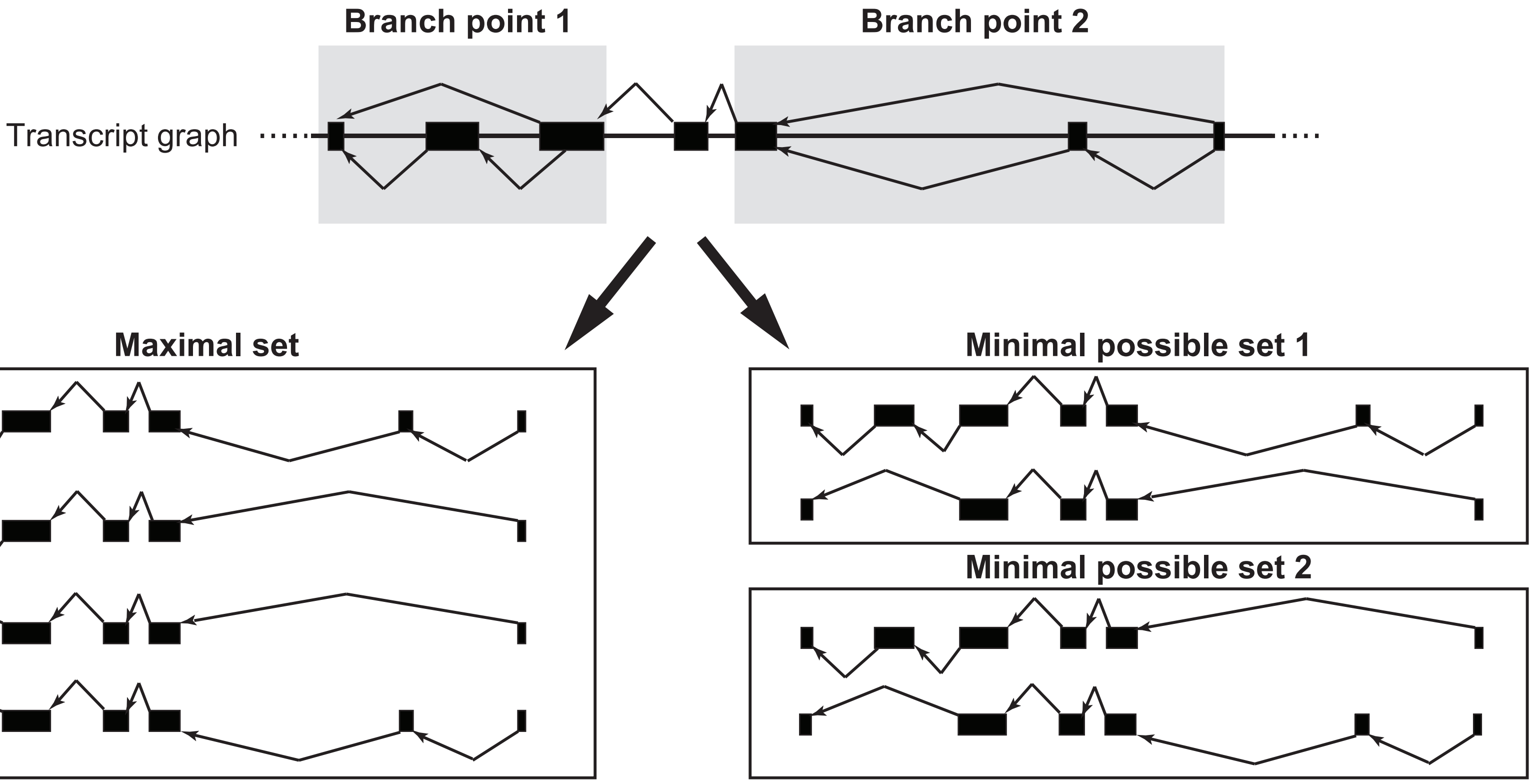
RNA-Seq transcript reconstruction software

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

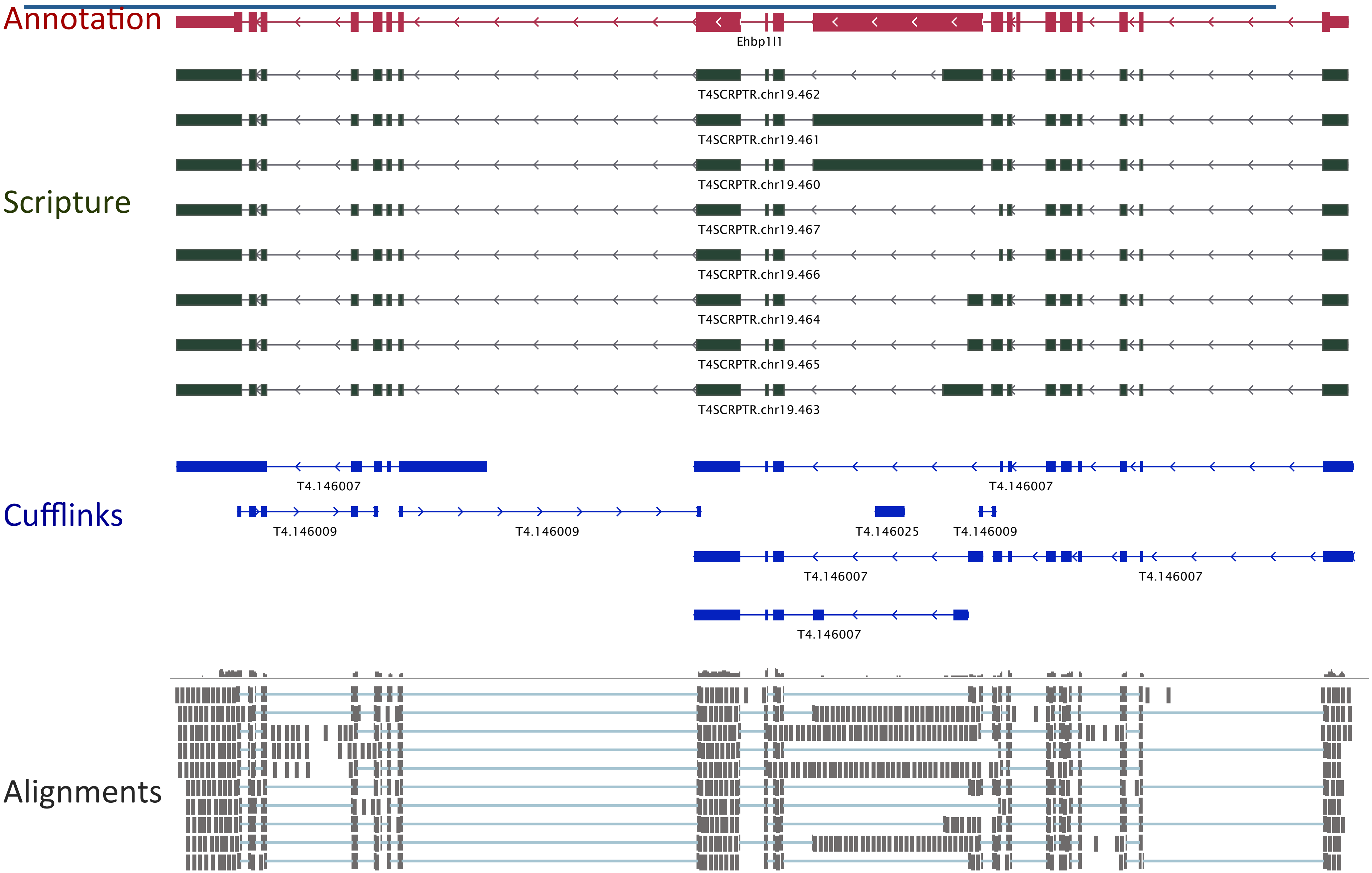
Differences between Cufflinks and Scripture

- 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



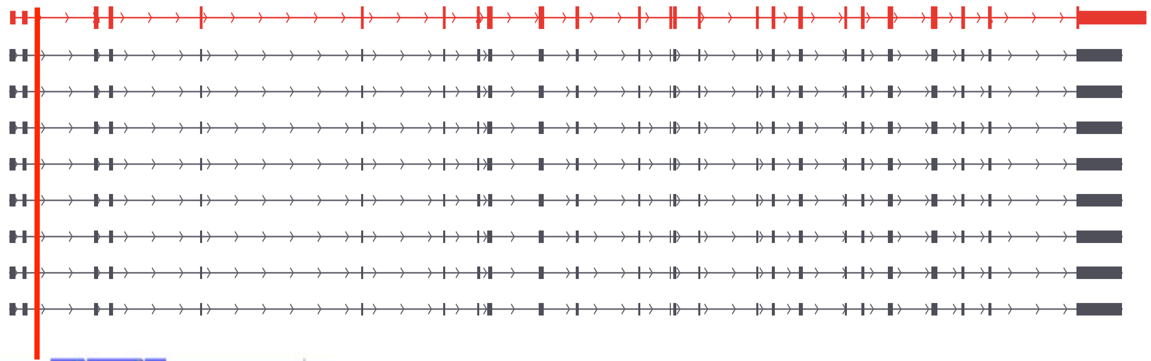
Comparing reconstructions

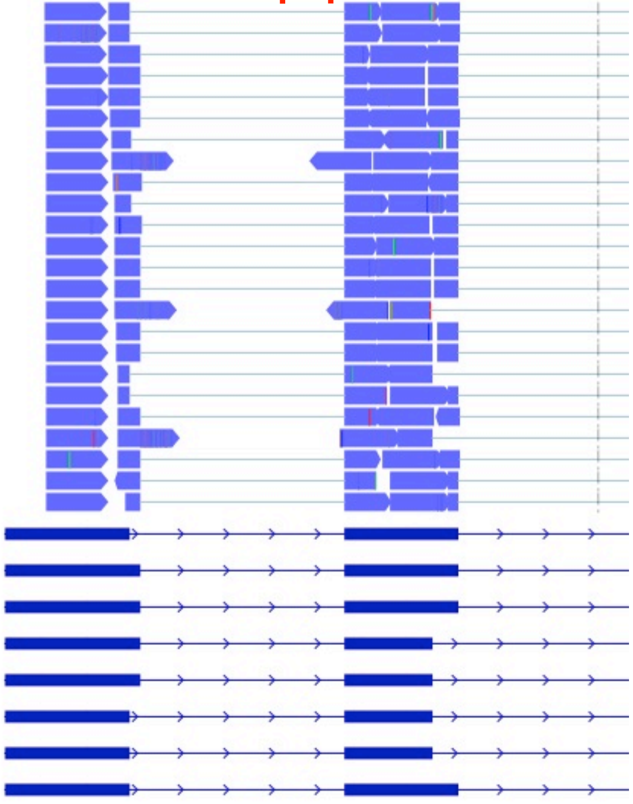
	CPU Hours	Total Memory	Genes fully reconstructed	Mean isoforms per reconstruction	Mean fragments per known annotation	Number of fragments predicted
Cufflinks	10	1.4 G	5,994	1.2	1.4	**159,856**
Scripture	16	3.5 G	6,221	**1.6**	1.3	61,922
Trans- Abyss	650	120 G[4]	3,330	4.7	2.6	3,117,238

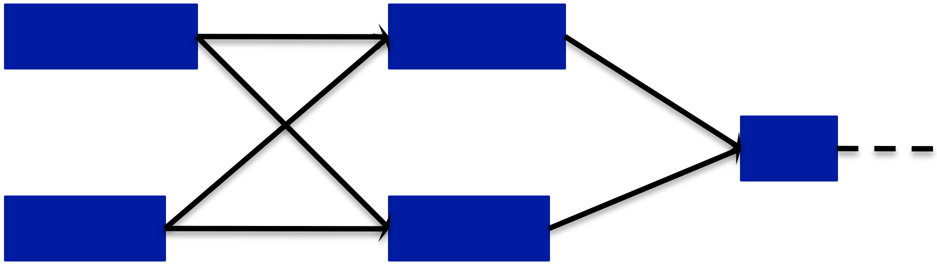
Many of the bogus locus and isoforms are due to alignment artifacts

Garber et al, Nature Methods 2011

Why so many isoforms



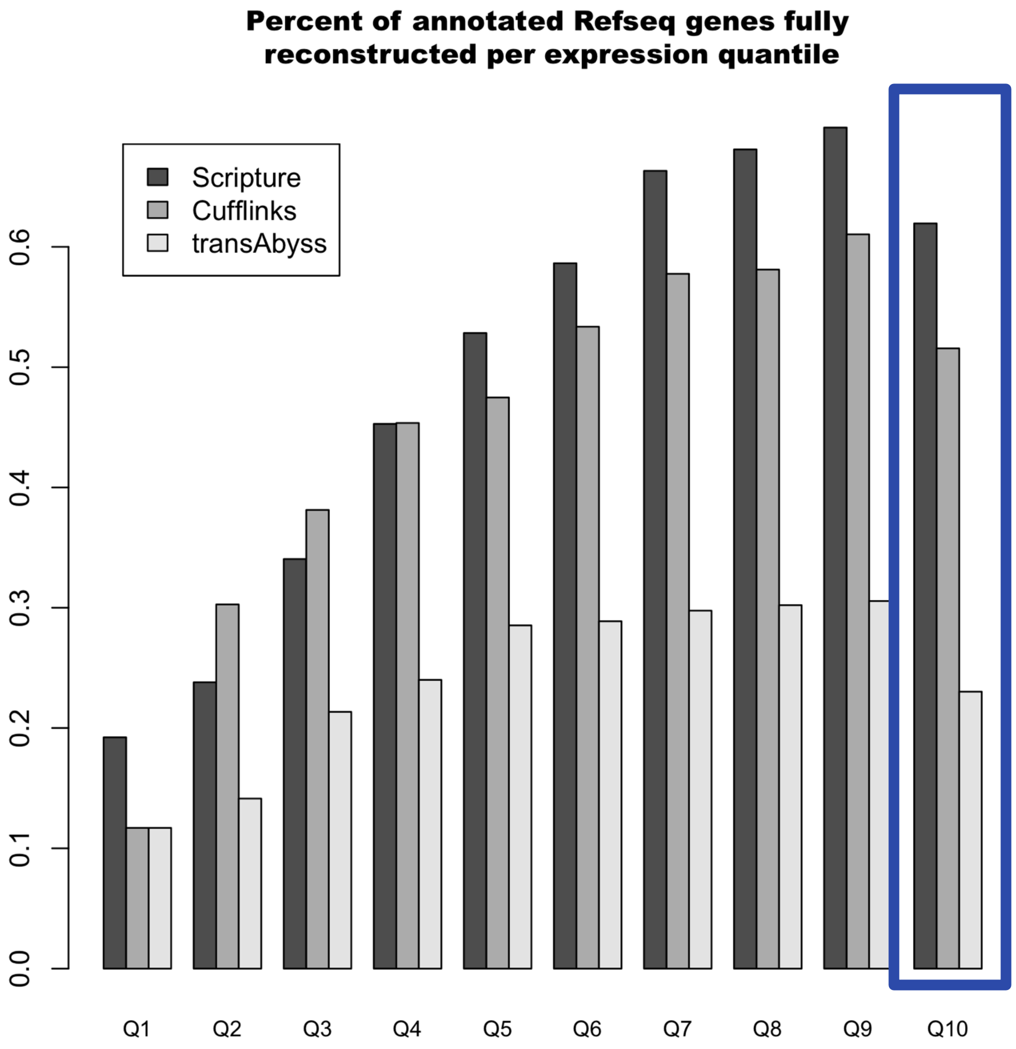




Every such splicing event or alignment artifact doubles the number of isoforms reported

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