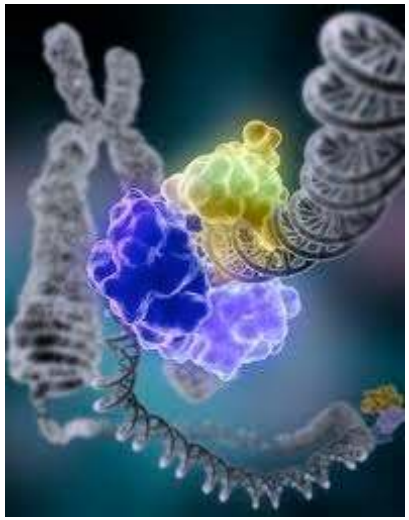
Sequence analysis with Scripture

Manuel Garber

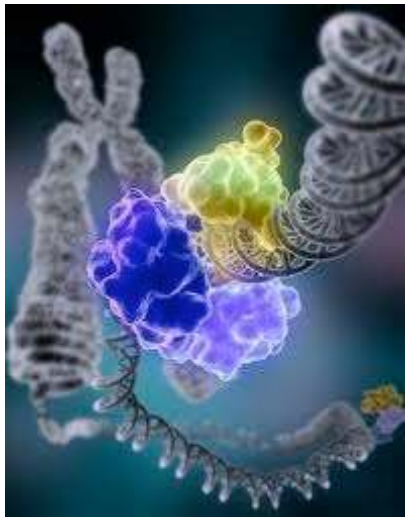
The dynamic genome

The dynamic genome

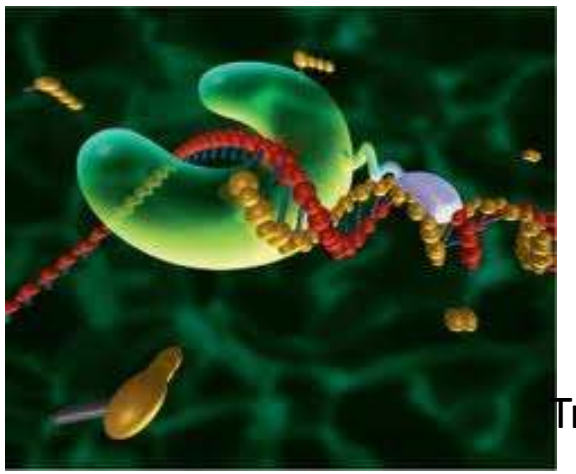


Copied and
Repaired

The dynamic genome

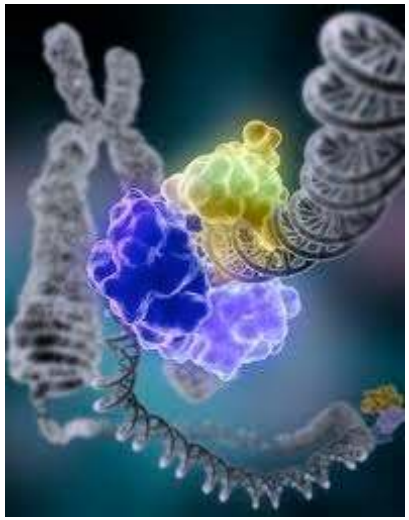


Copied and
Repaired

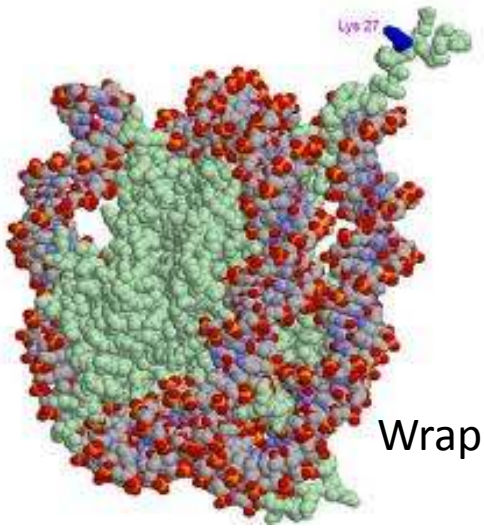


Transcribed

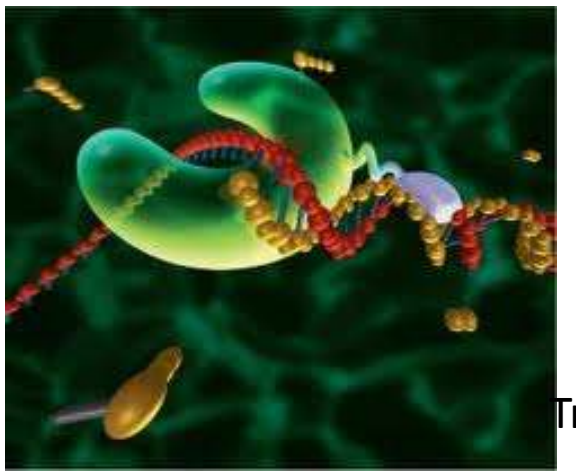
The dynamic genome



Copied and
Repaired

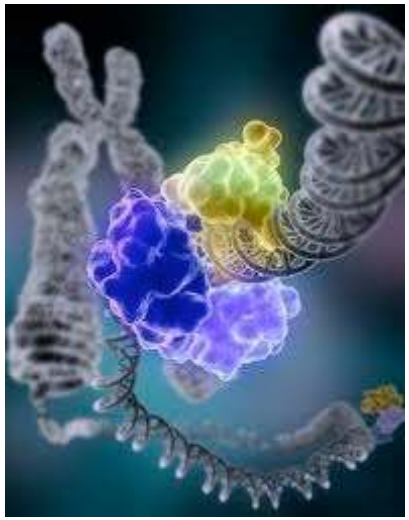


Wrapped in marked histones



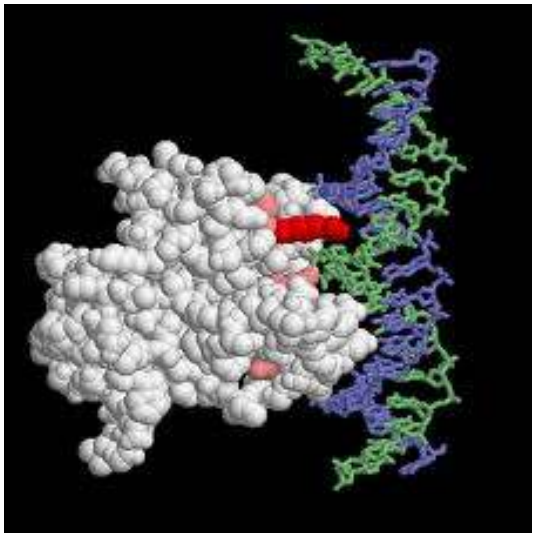
Transcribed

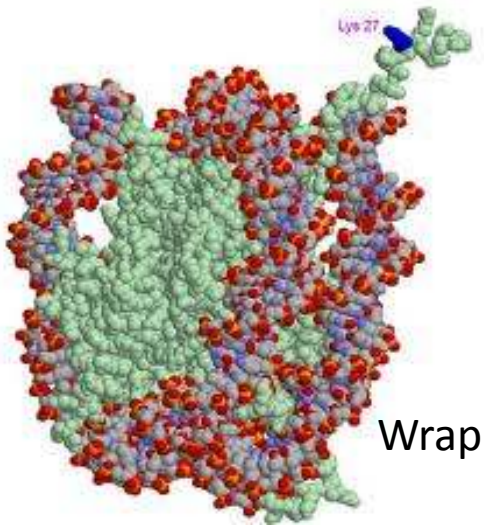
The dynamic genome



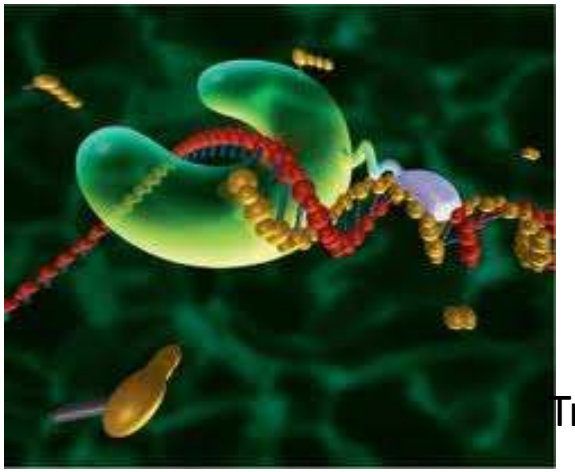
Copied and Repaired

Bound by Transcription Factors



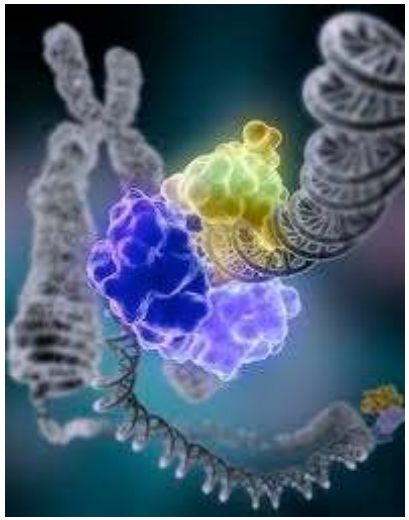


Wrapped in marked histones



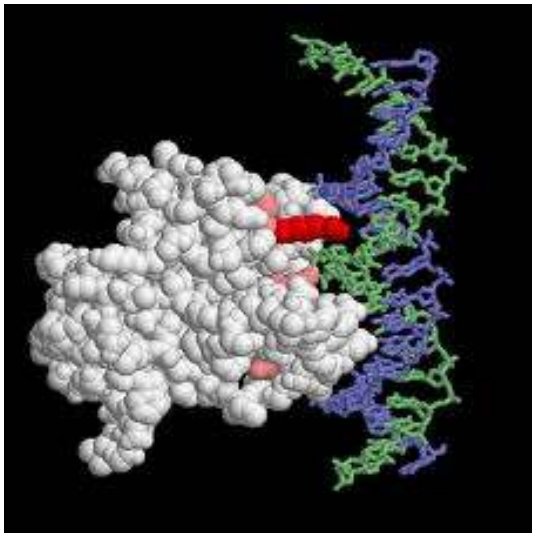
Transcribed

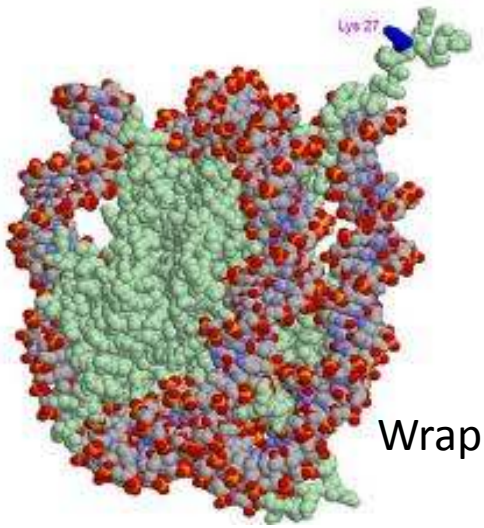
The dynamic genome



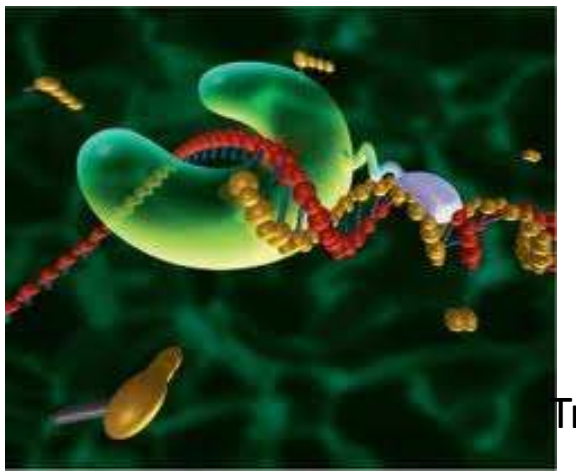
Copied and Repaired

Bound by Transcription Factors

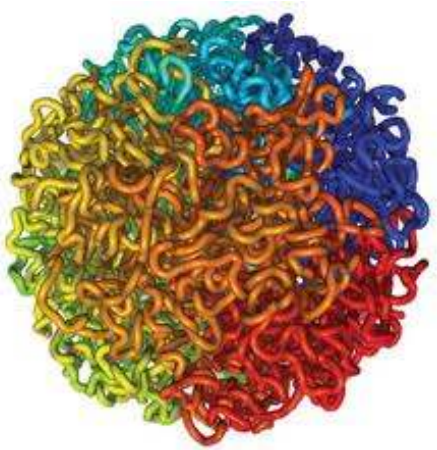




Wrapped in marked histones

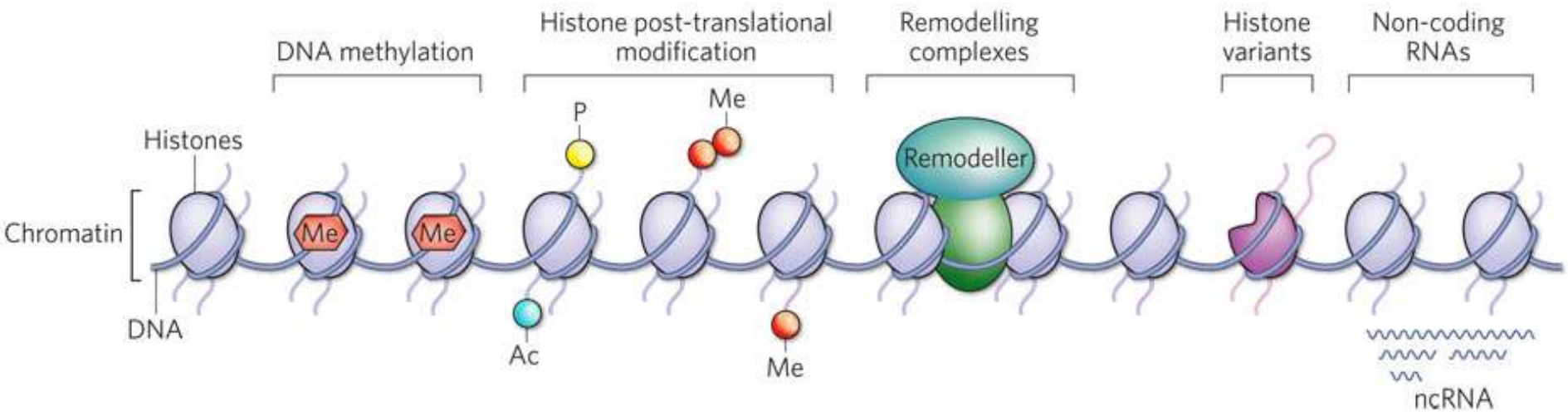


Folded into a fractal globule



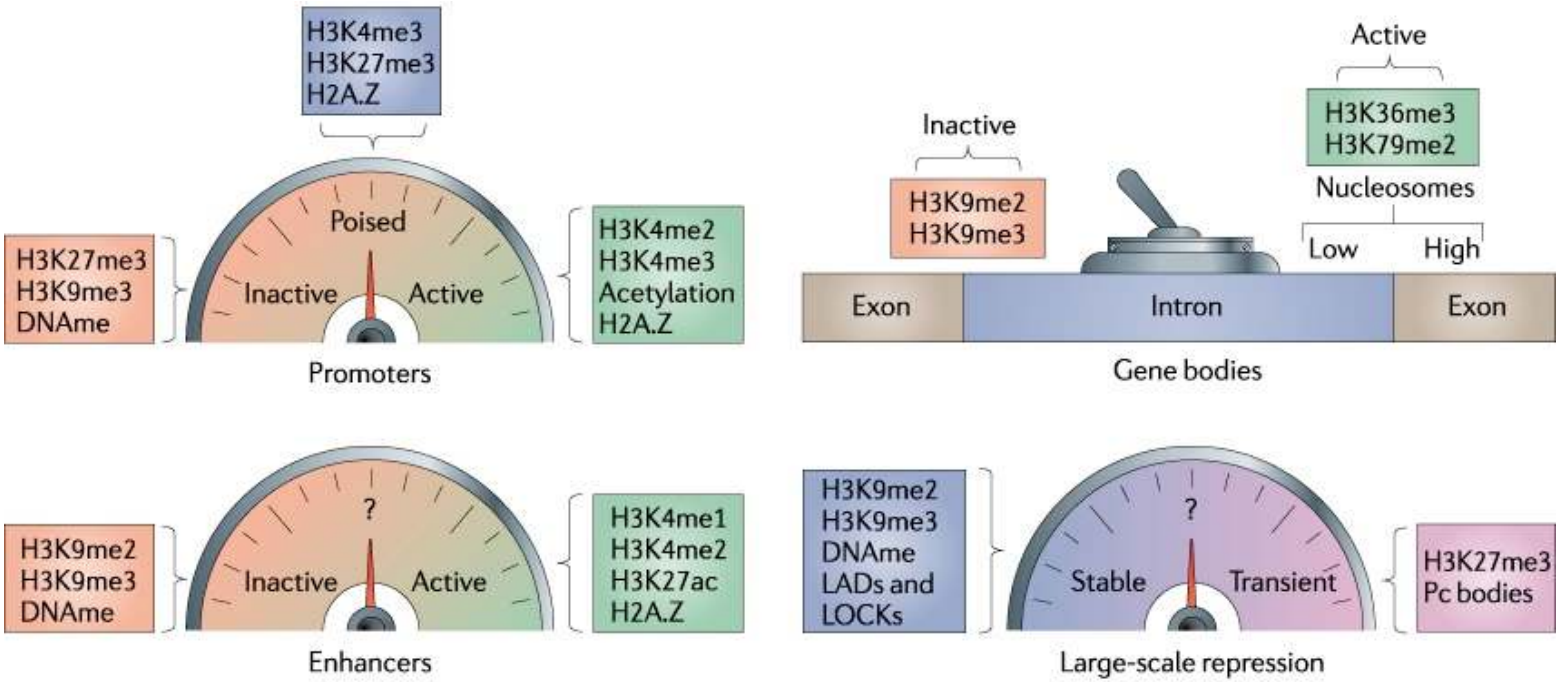
Transcribed

With a very dynamic epigenetic state



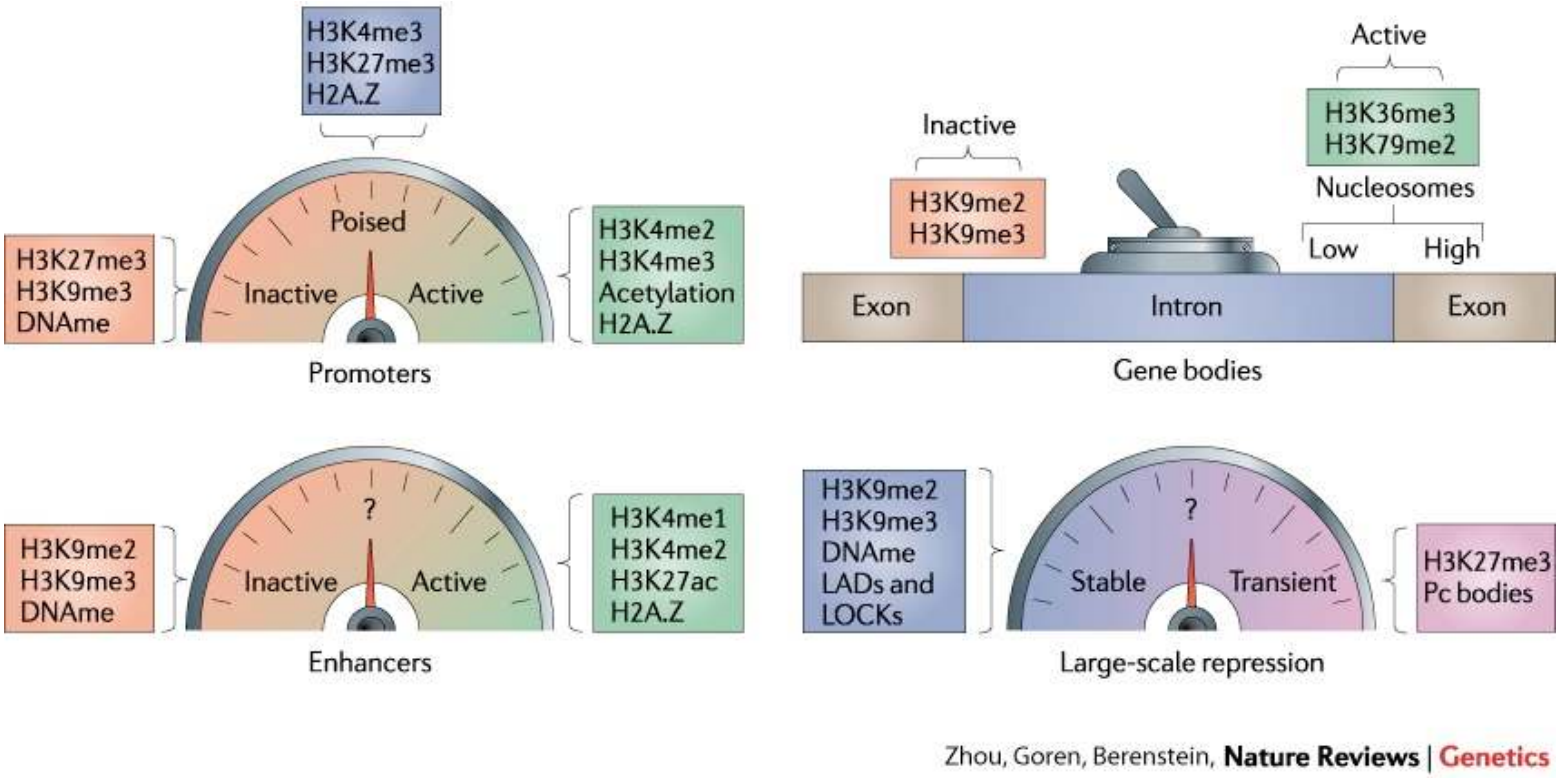
Catherine Dulac, Nature 2010

Which define the state of its functional elements



Zhou, Goren, Berenstein, **Nature Reviews** | Genetics

Which define the state of its functional elements

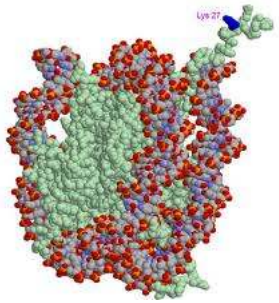


Zhou, Goren, Berenstein, **Nature Reviews** | Genetics

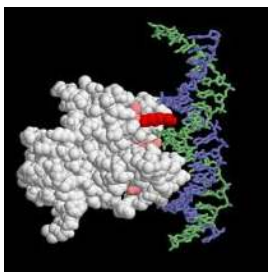
Motivation: find the genome state using sequencing data

We can use sequencing to find the genome state (Protein-DNA)

Histone Marks

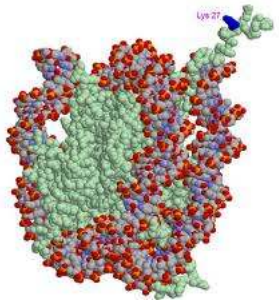


Transcription Factors

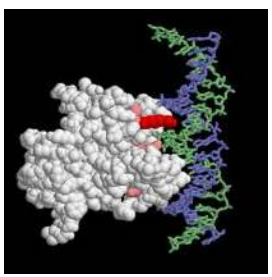


We can use sequencing to find the genome state (Protein-DNA)

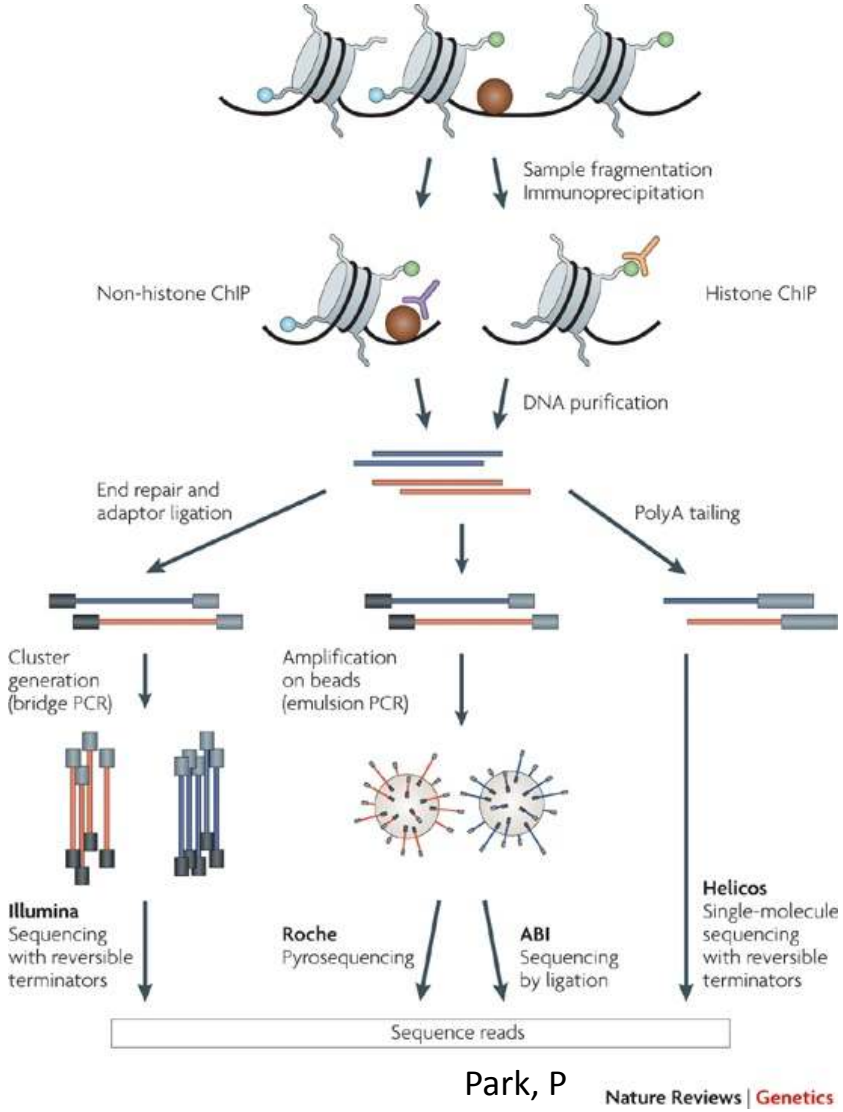
Histone Marks



Transcription Factors



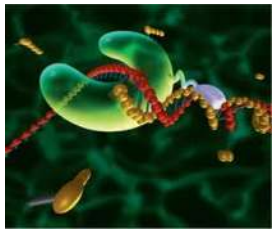
ChIP-Seq



Park, P

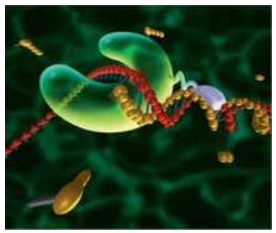
Nature Reviews | Genetics

We can use sequencing to find the genome state



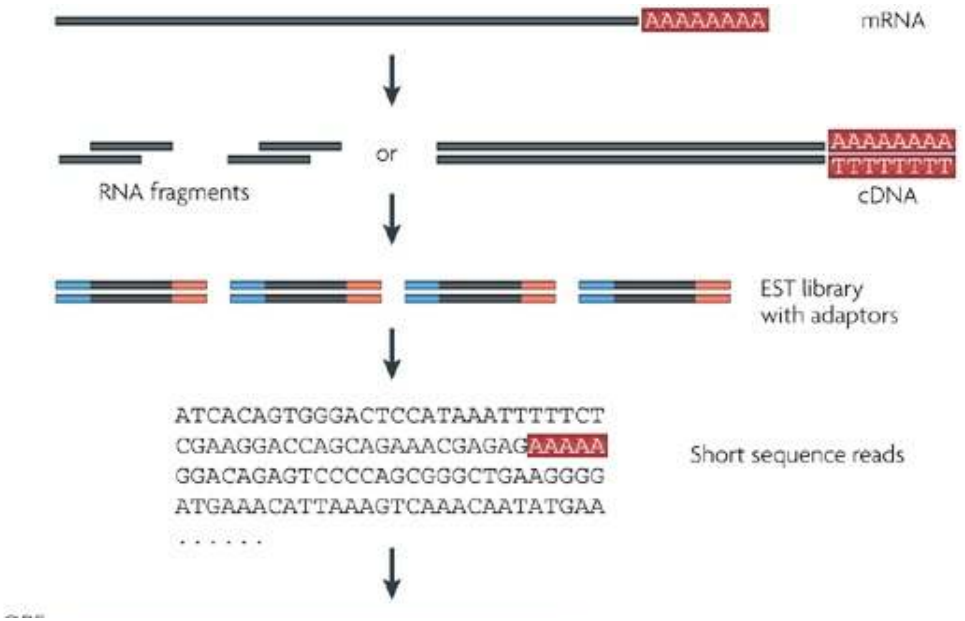
Transcription

We can use sequencing to find the genome state



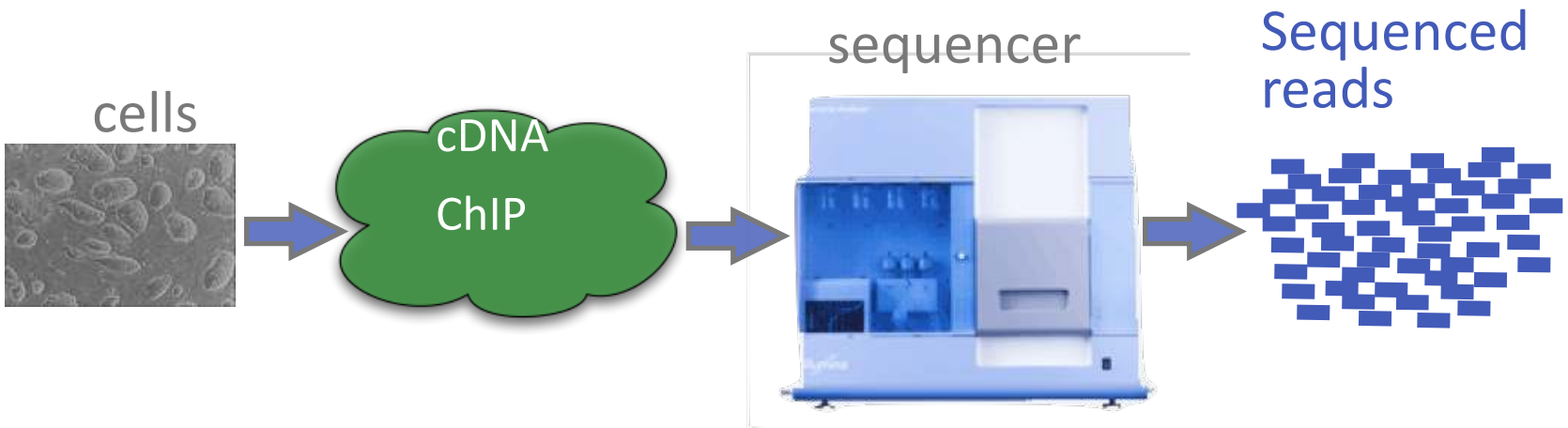
Transcription

RNA-Seq

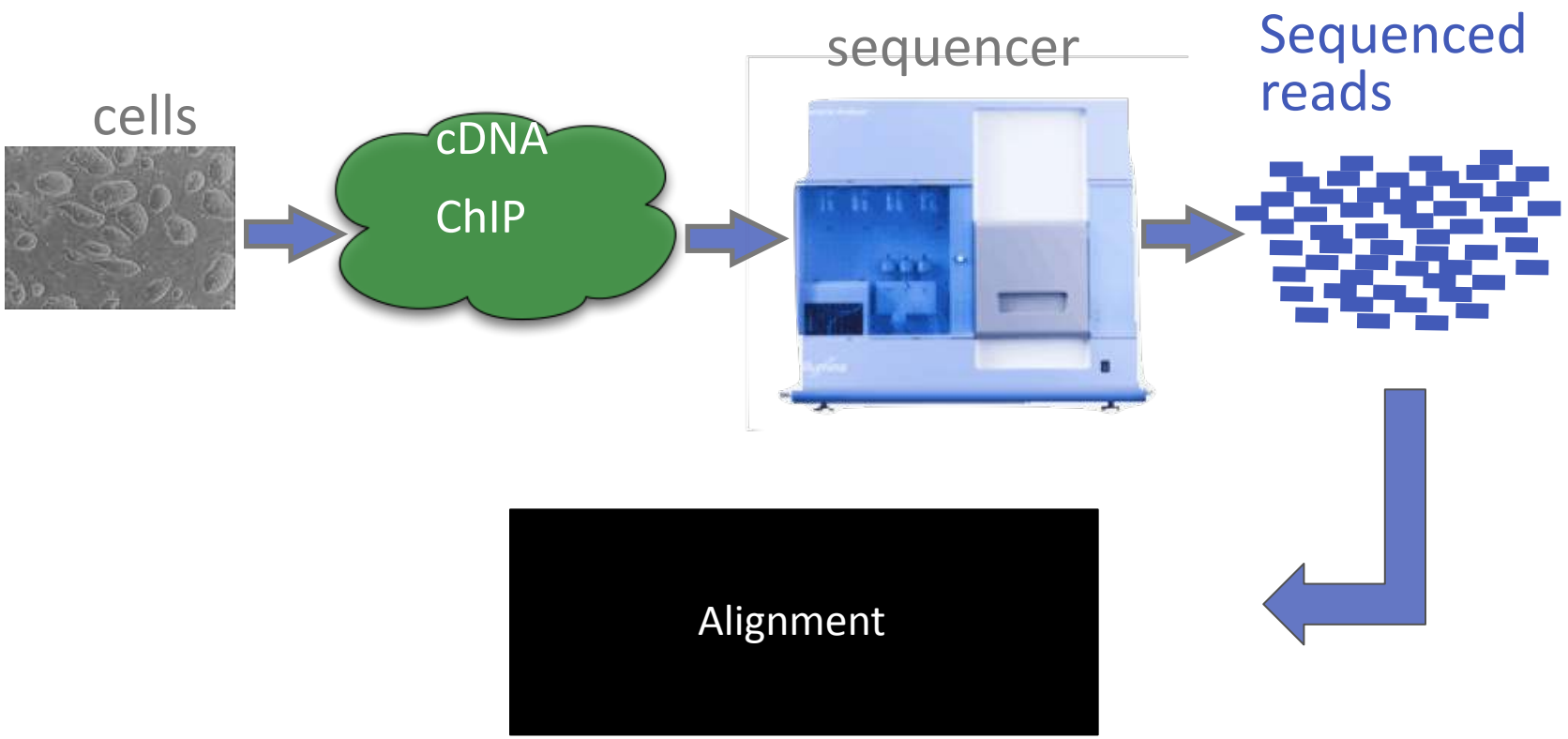


Wang, Z Nature Reviews Genetics 2009

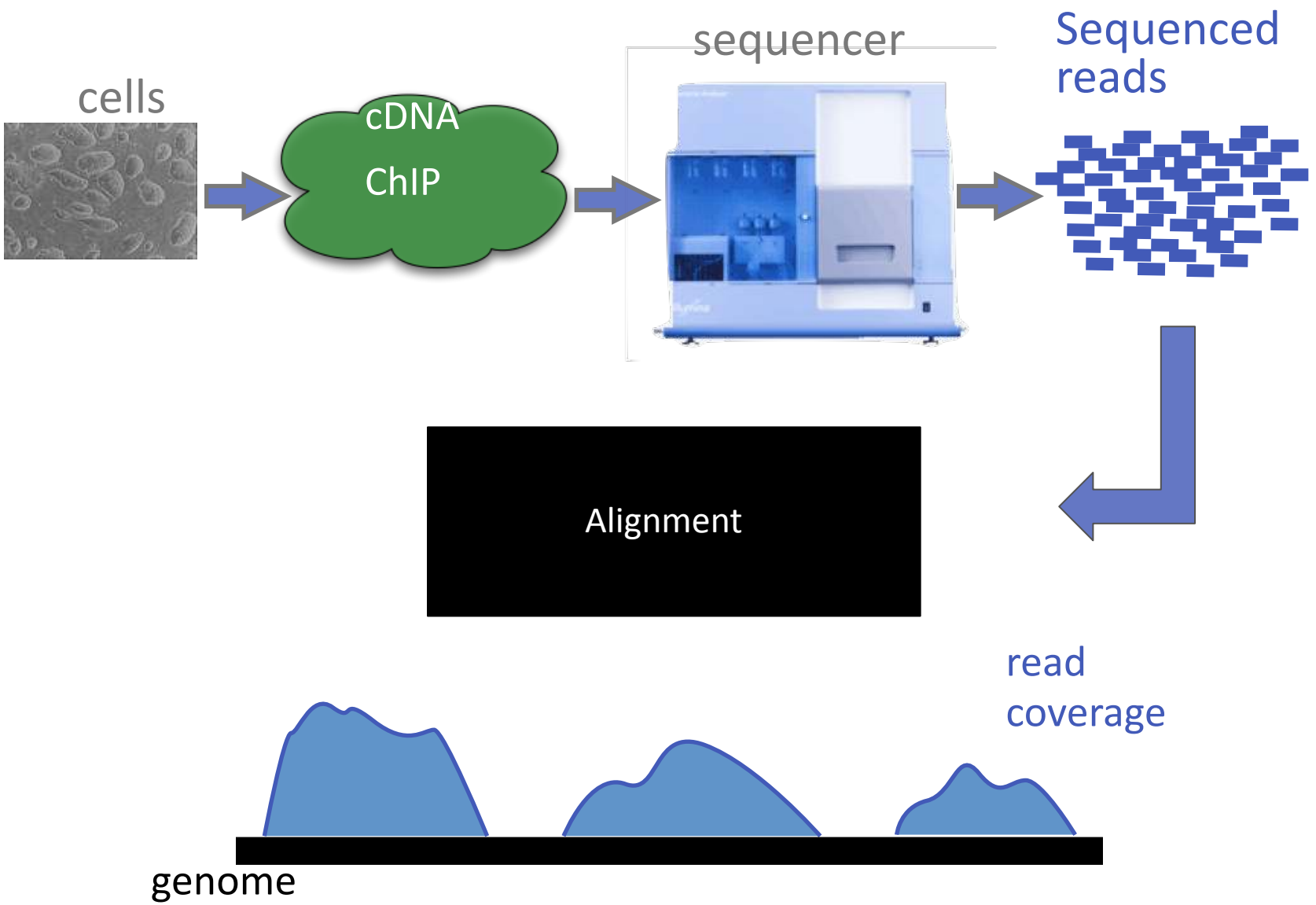
Once sequenced the problem becomes computational



Once sequenced the problem becomes computational



Once sequenced the problem becomes computational



Overview of the session

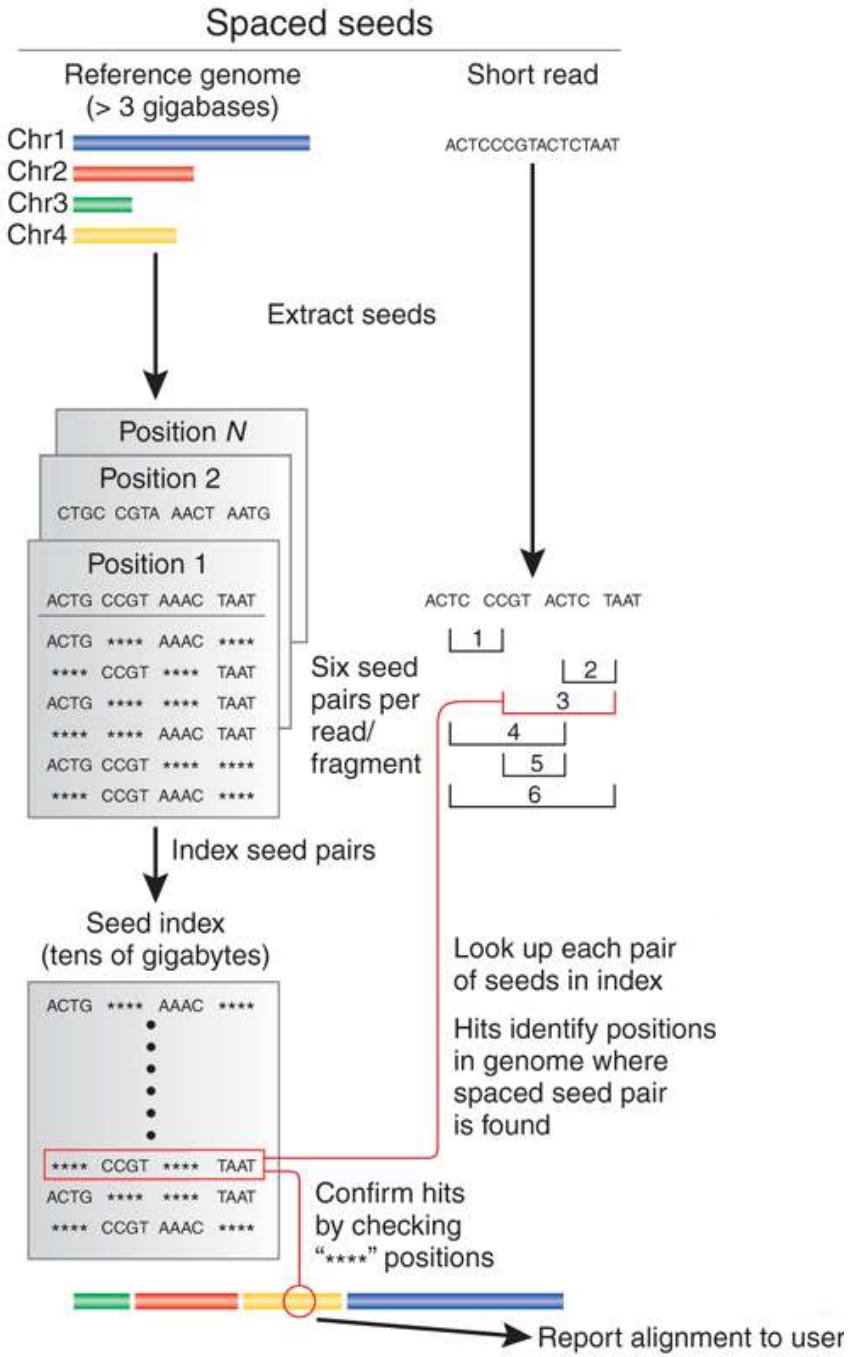
We'll cover the 3 main computational challenges of sequence analysis for *counting applications*:

- Read mapping (alignment): Placing short reads in the genome
- Reconstruction: Finding the regions that originated the reads
- Quantification:
    - Assigning scores to regions
    - Finding regions that are differentially represented between two or more samples.

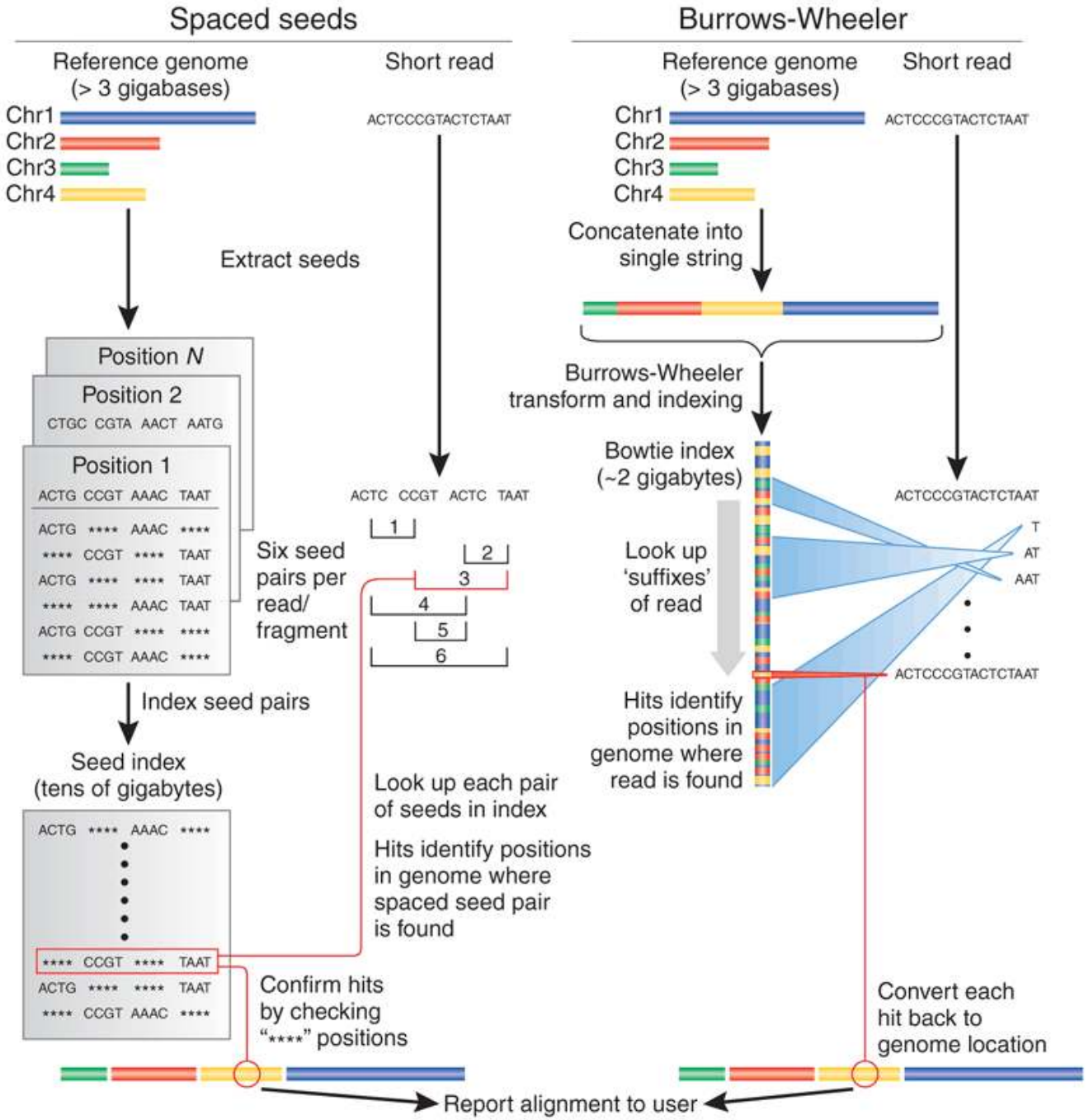
Overview of the session

We'll cover the 3 main computational challenges of sequence analysis for *counting applications*:

- Read mapping (alignment): Placing short reads in the genome
- Reconstruction: Finding the regions that originated the reads
- Quantification:
  - Assigning scores to regions
  - Finding regions that are differentially represented between two or more samples.



Trapnell, Salzberg, Nature Biotechnology 2009



Trapnell, Salzberg, Nature Biotechnology 2009

Short read mapping software for ChIP-Seq

Seed-extend

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

BWA

	Use Base qual
BWA	**YES**
Bowtie	NO
Soap2	NO
Stampy*	YES

Short read mapping software for ChIP-Seq

Seed-extend

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

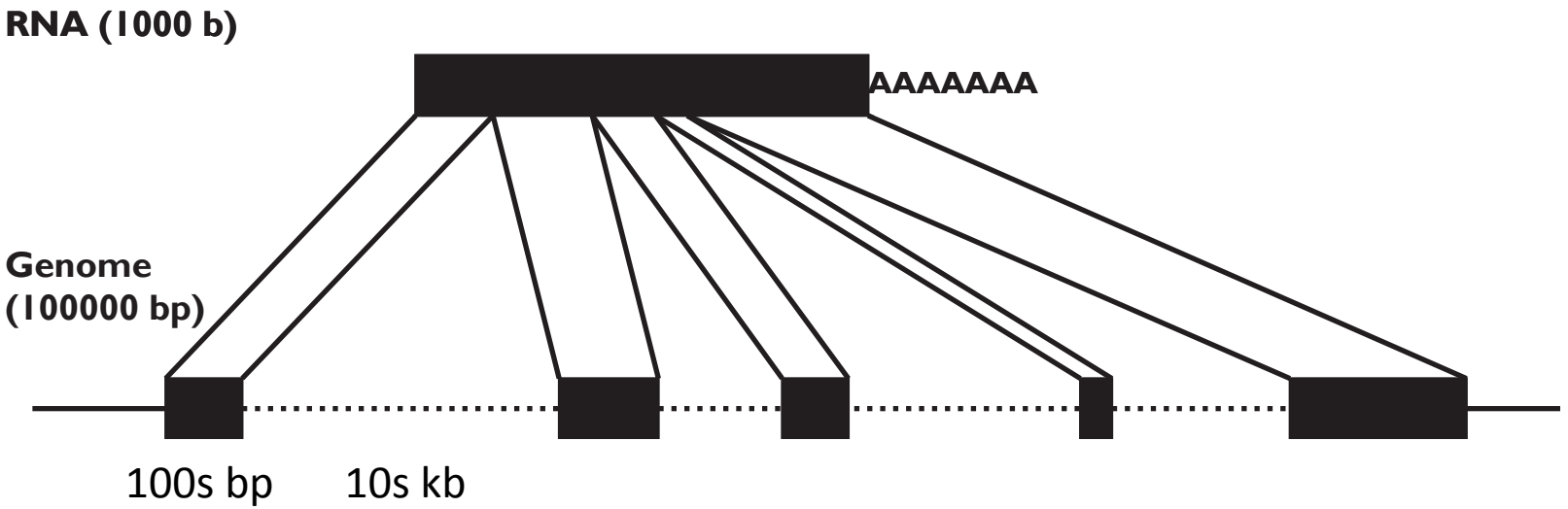
BWA

	Use Base qual
BWA	**YES**
Bowtie	NO
Soap2	NO
Stampy*	YES

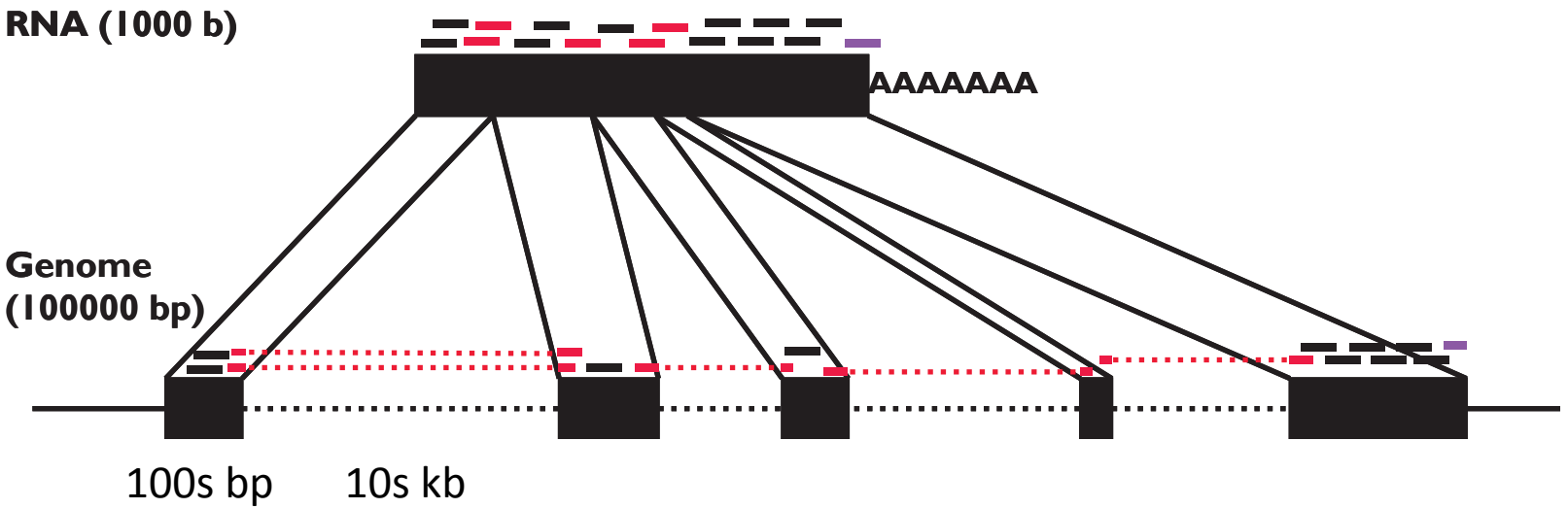
Stampy is a hybrid approach which first uses BWA to map reads then uses seed-extend only to reads not mapped by BWA

RNA-Seq read mapping is more complex

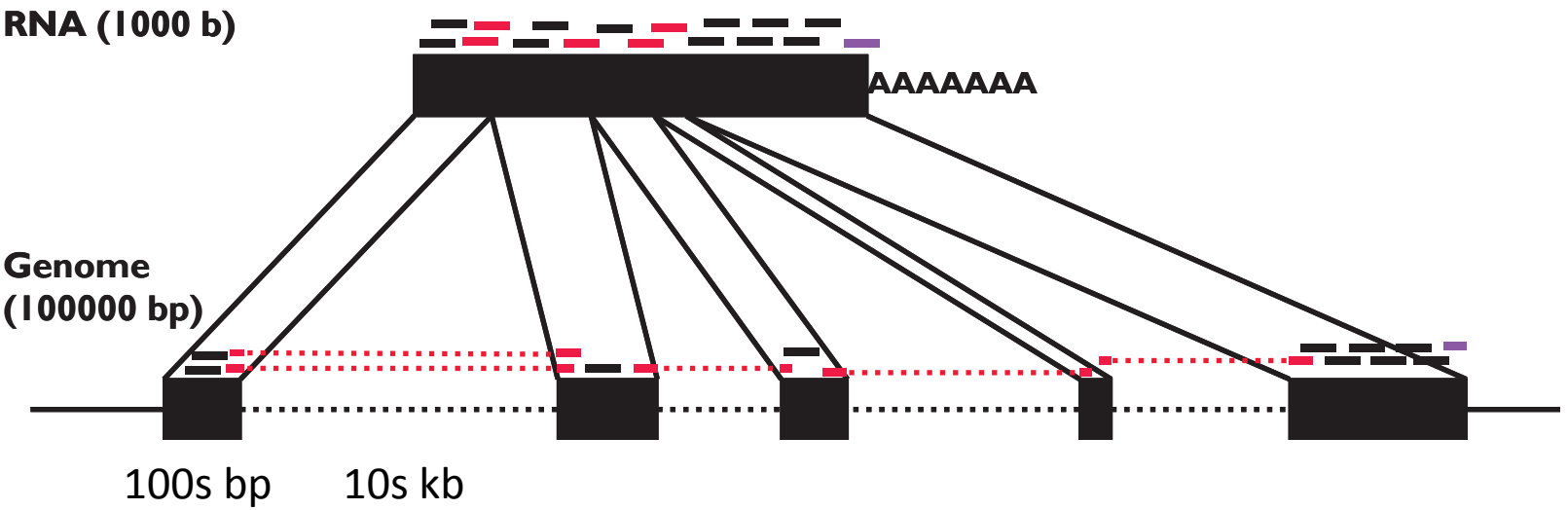
RNA-Seq read mapping is more complex



RNA-Seq read mapping is more complex



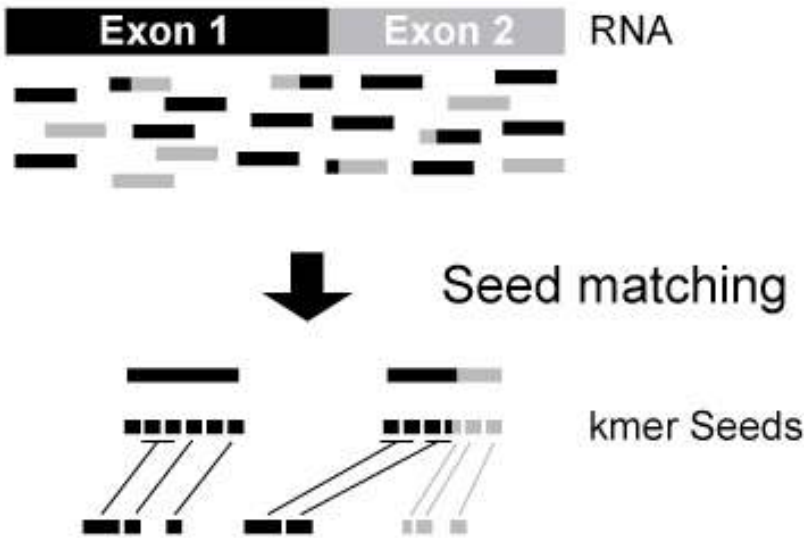
RNA-Seq read mapping is more complex



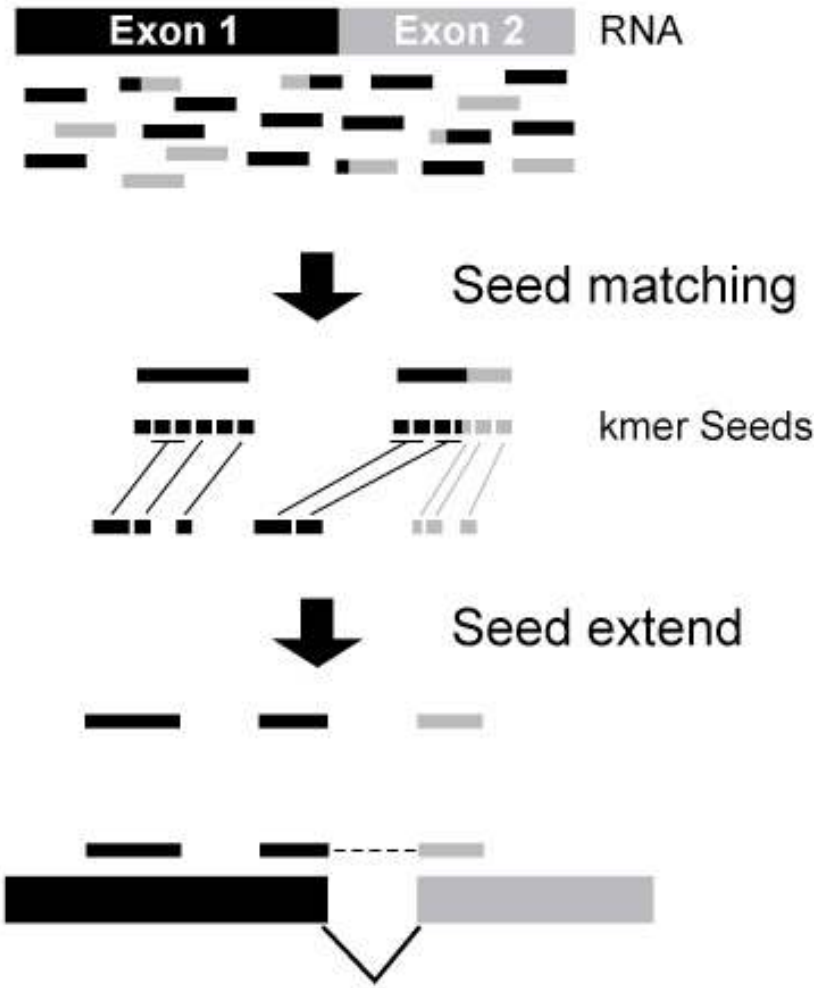
RNA-Seq reads can be spliced, and spliced reads are most informative

Method 1: Seed-extend spliced alignment

Method 1: Seed-extend spliced alignment

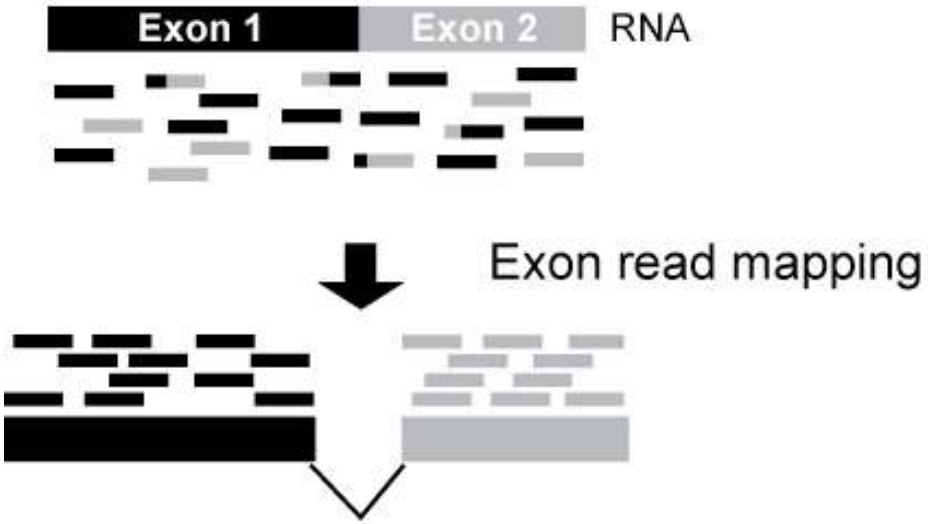


Method 1: Seed-extend spliced alignment

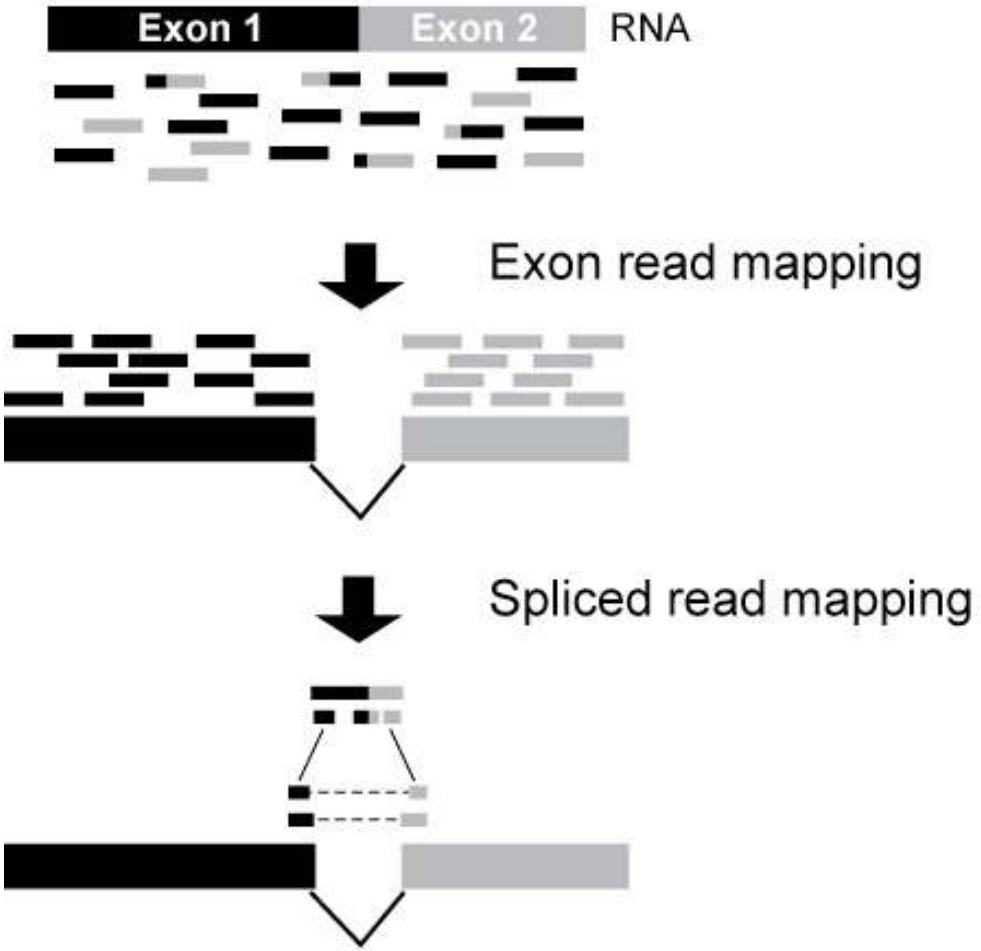


Method 11: Exon-first spliced alignment

Method 11: Exon-first spliced alignment



Method 11: Exon-first spliced alignment



Short read mapping software for RNA-Seq

Seed-extend

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

Exon-first

	Use base qual
MapSplice	NO
SpliceMap	NO
TopHat	NO

Short read mapping software for RNA-Seq

Seed-extend

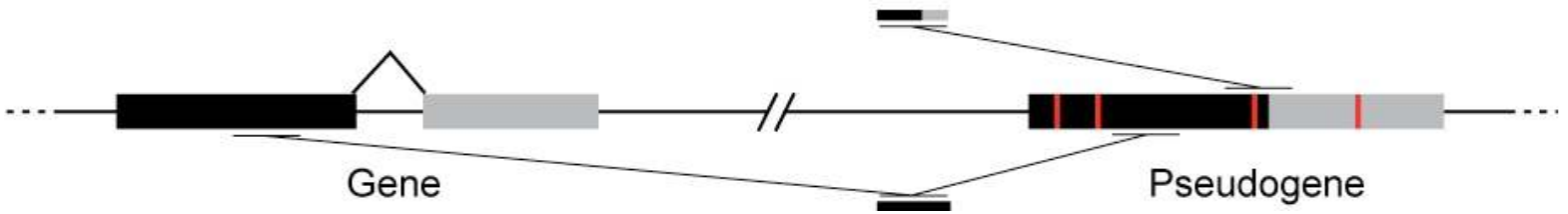
	Short indels	Use base qual
GSNAP	**No**	NO
QPALMA	Yes	NO
BLAT	Yes	NO

Exon-first

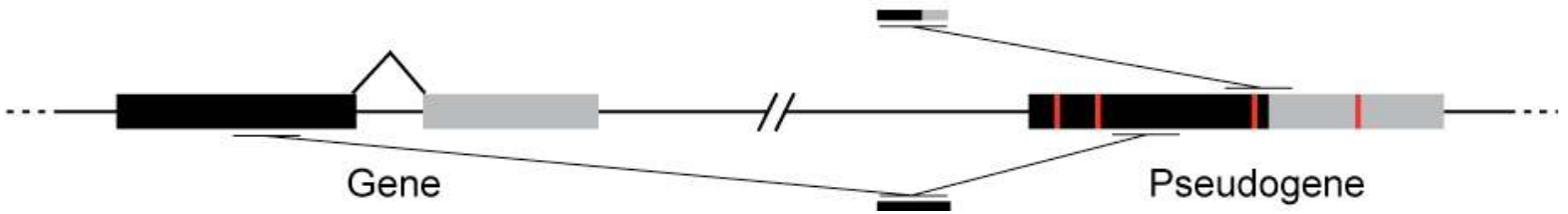
	Use base qual
MapSplice	NO
SpliceMap	NO
TopHat	NO

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

Exon-first aligners are faster but at cost



Exon-first aligners are faster but at cost



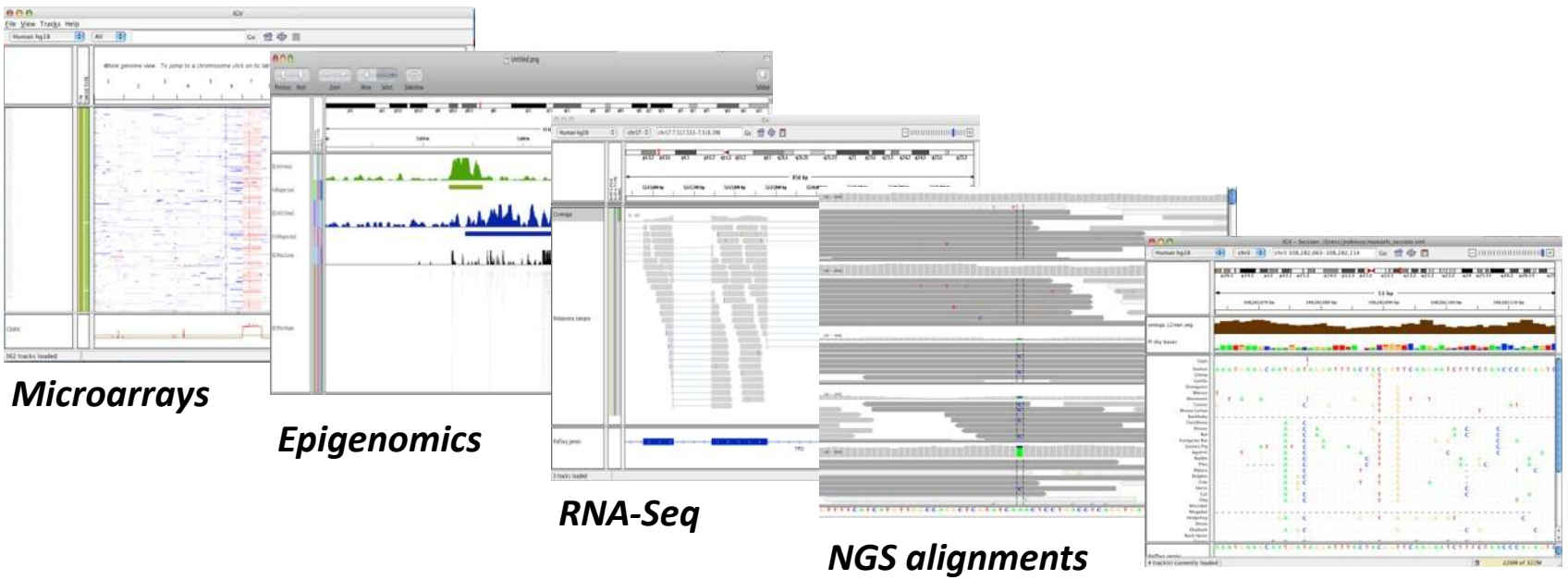
How do we visualize the results of these programs

IGV: Integrative Genomics Viewer

A desktop application

for the visualization and interactive exploration

of genomic data



Microarrays

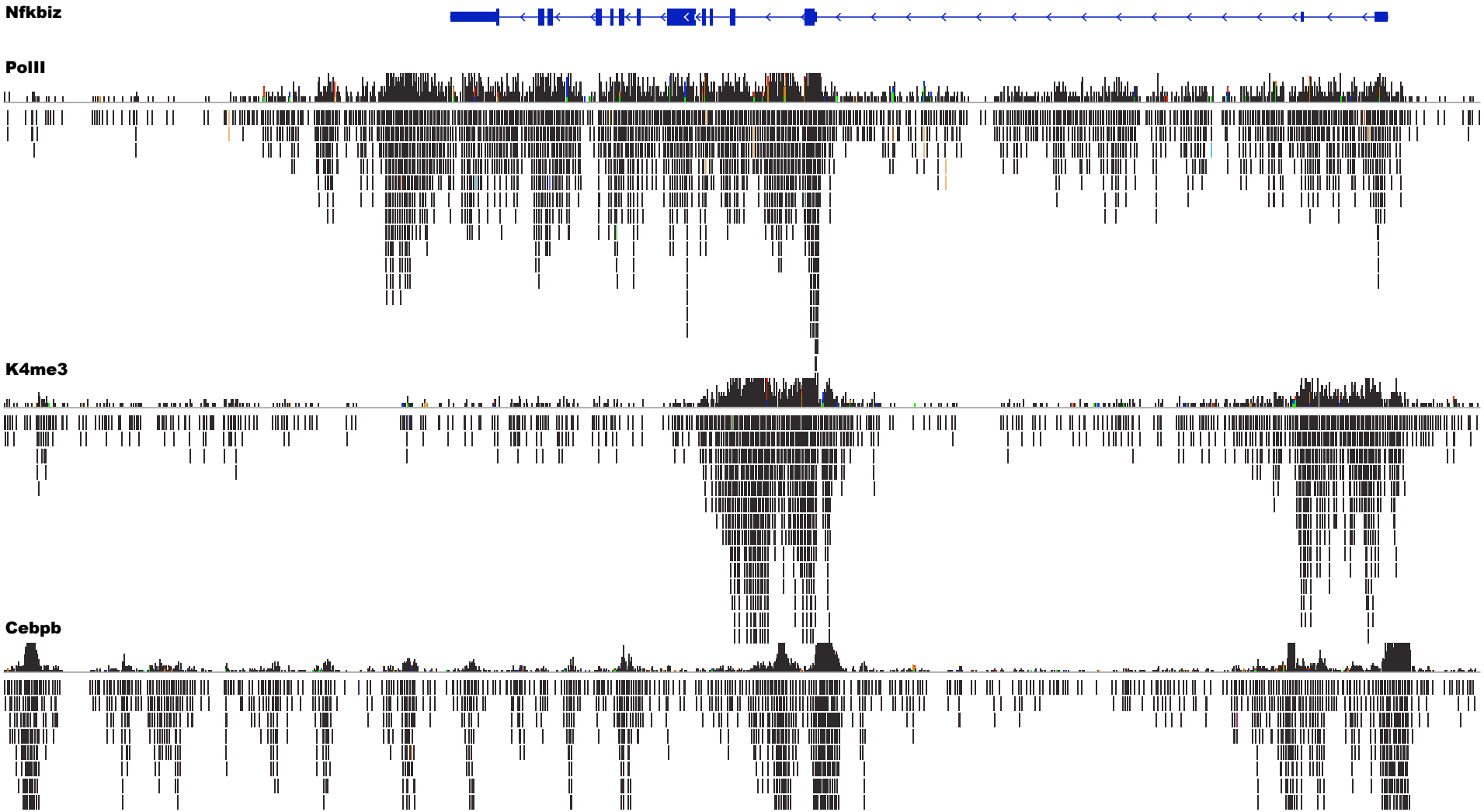
Epigenomics

RNA-Seq

NGS alignments

Comparative genomics

Visualizing read alignments with IGV

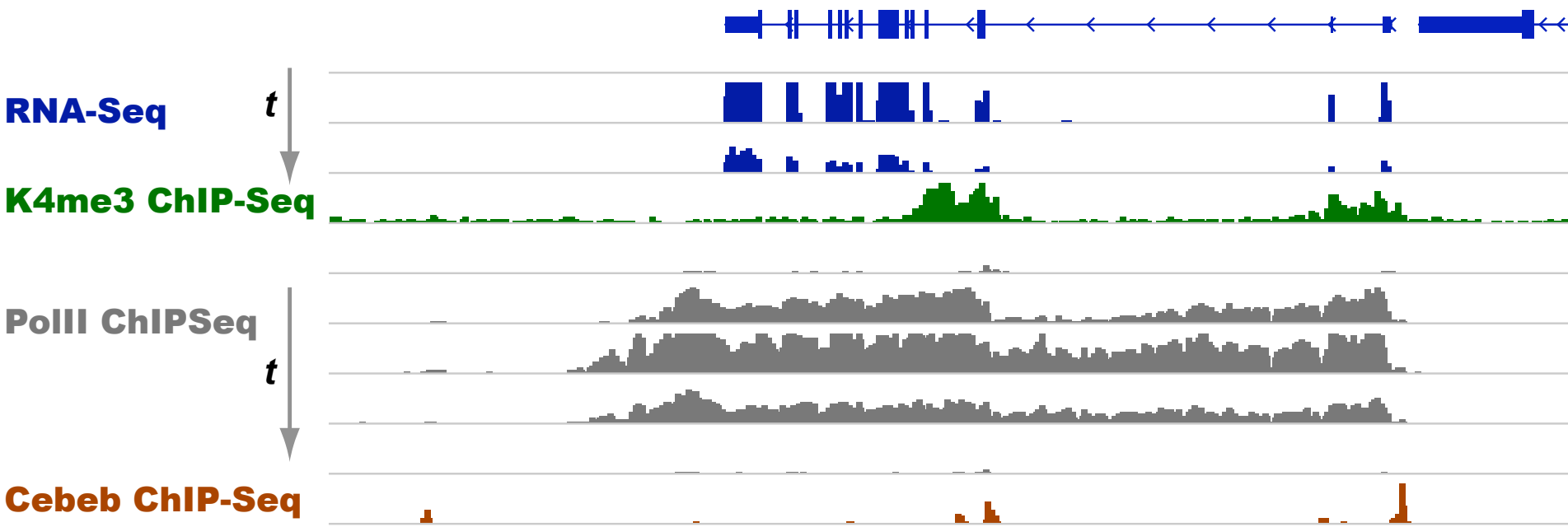


Visualizing read alignments with IGV — RNASeq

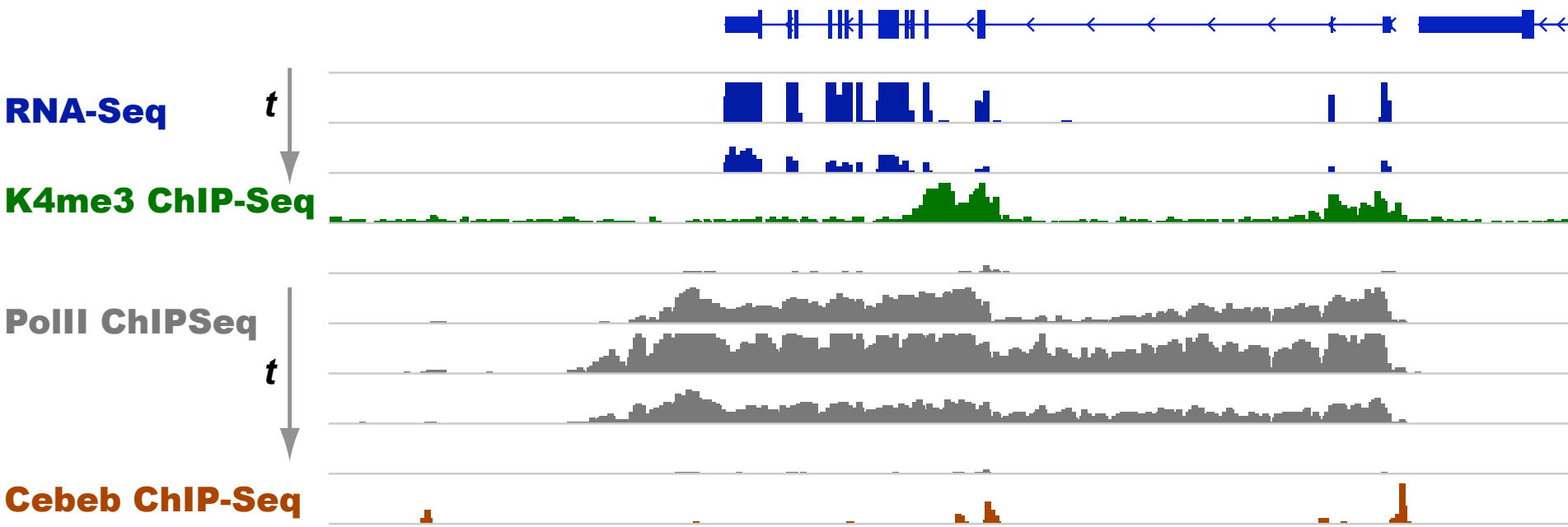
Visualizing read alignments with IGV — RNASeq

Visualizing read alignments with IGV — RNASeq close-up

Visualizing read alignments with IGV — zooming out



Visualizing read alignments with IGV — zooming out



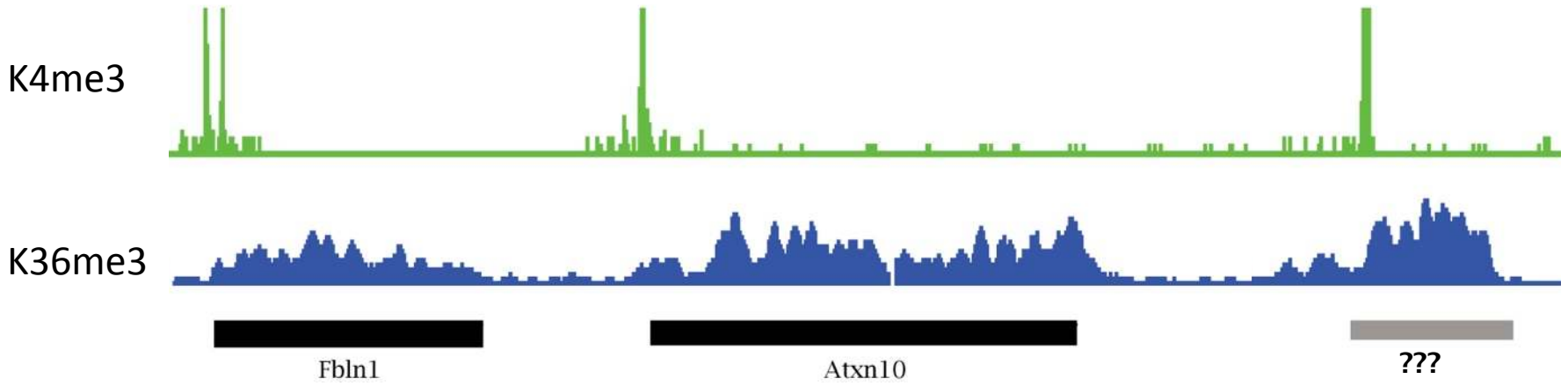
How can we identify regions enriched in sequencing reads?

Overview of the session

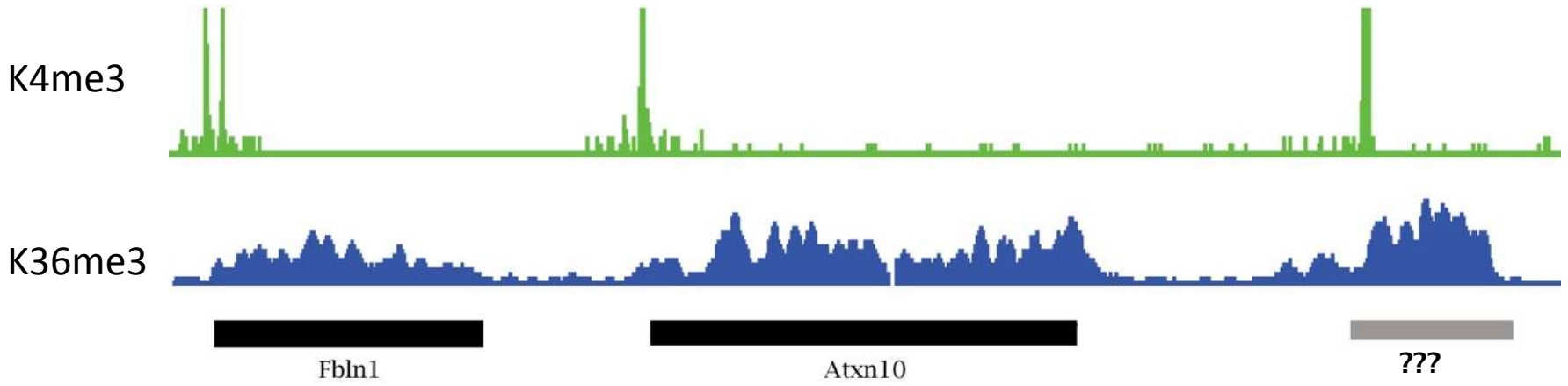
The 3 main computational challenges of sequence analysis for *counting applications*:

- Read mapping: Placing short reads in the genome
- Reconstruction: Finding the regions that originate the reads
- Quantification:
  - Assigning scores to regions
  - Finding regions that are differentially represented between two or more samples.

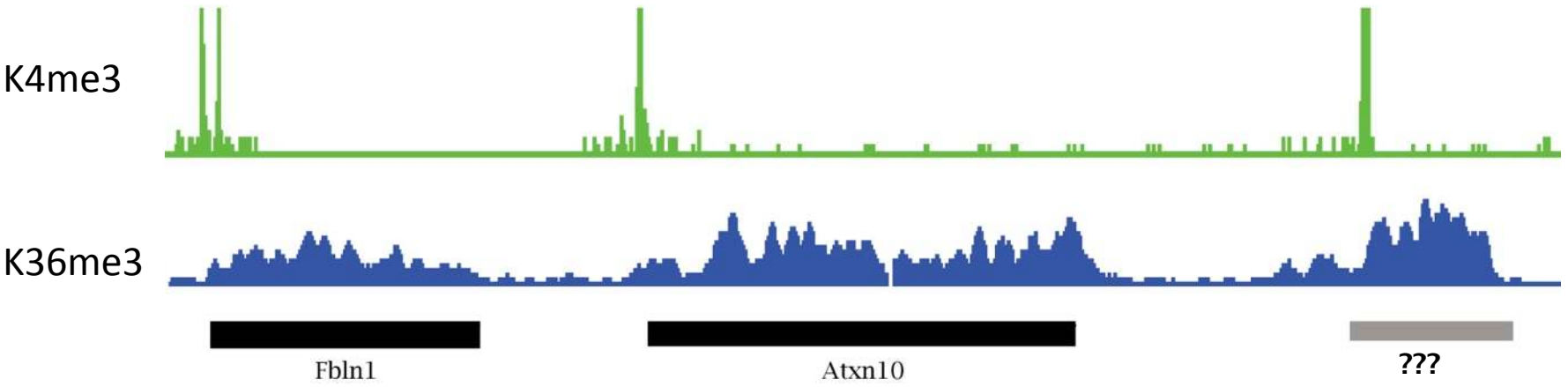
Chromatin domains demarcate interesting surprises in the transcriptome



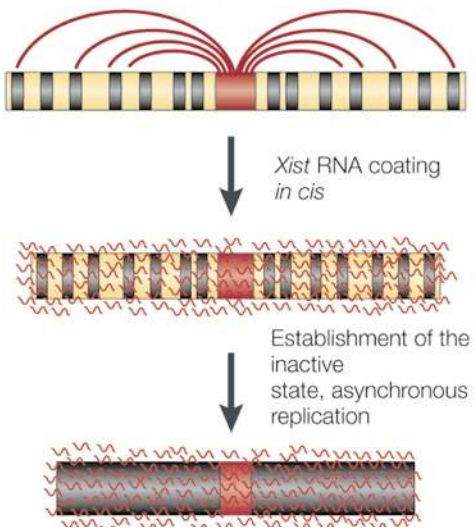
Chromatin domains demarcate interesting surprises in the transcriptome



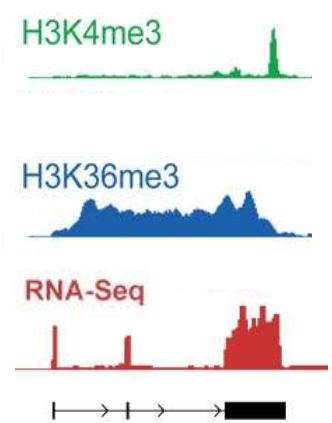
Chromatin domains demarcate interesting surprises in the transcriptome



XIST



How can we identify these chromatin marks and the genes within?

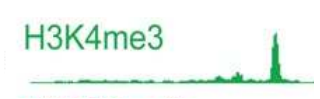


Short modification

Long modification

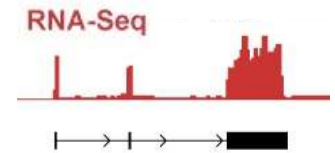
Discontinuous data

How can we identify these chromatin marks and the genes within?

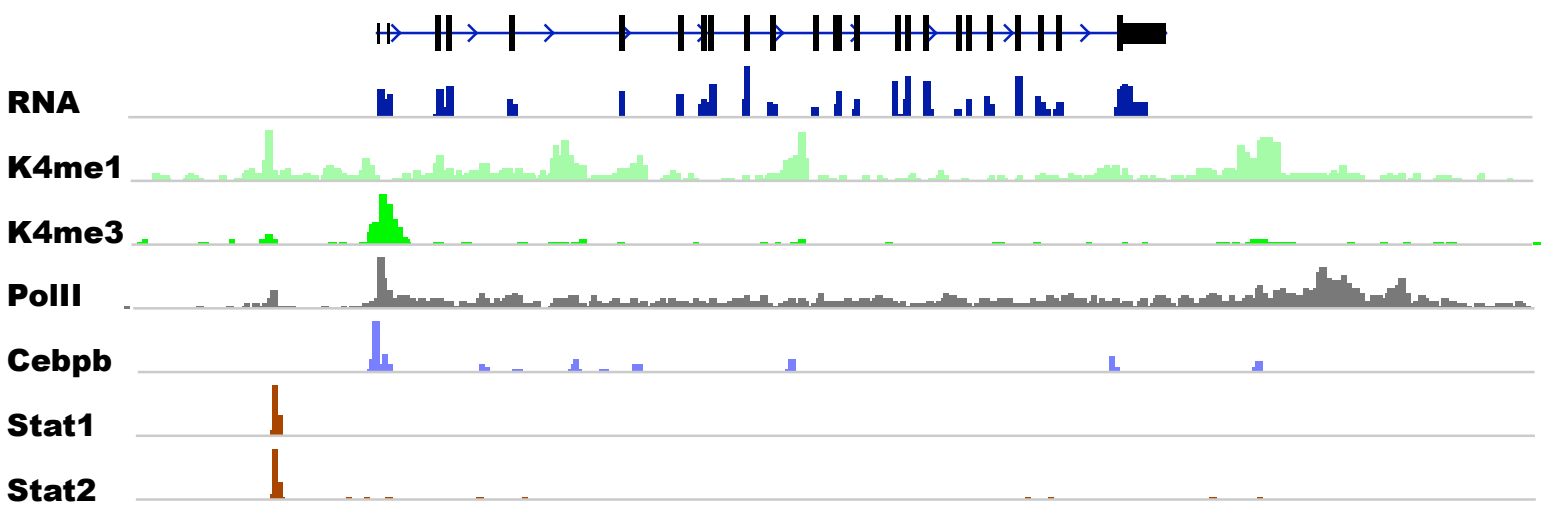


Short modification

Long modification



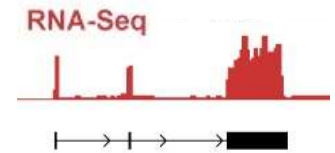
Discontinuous data



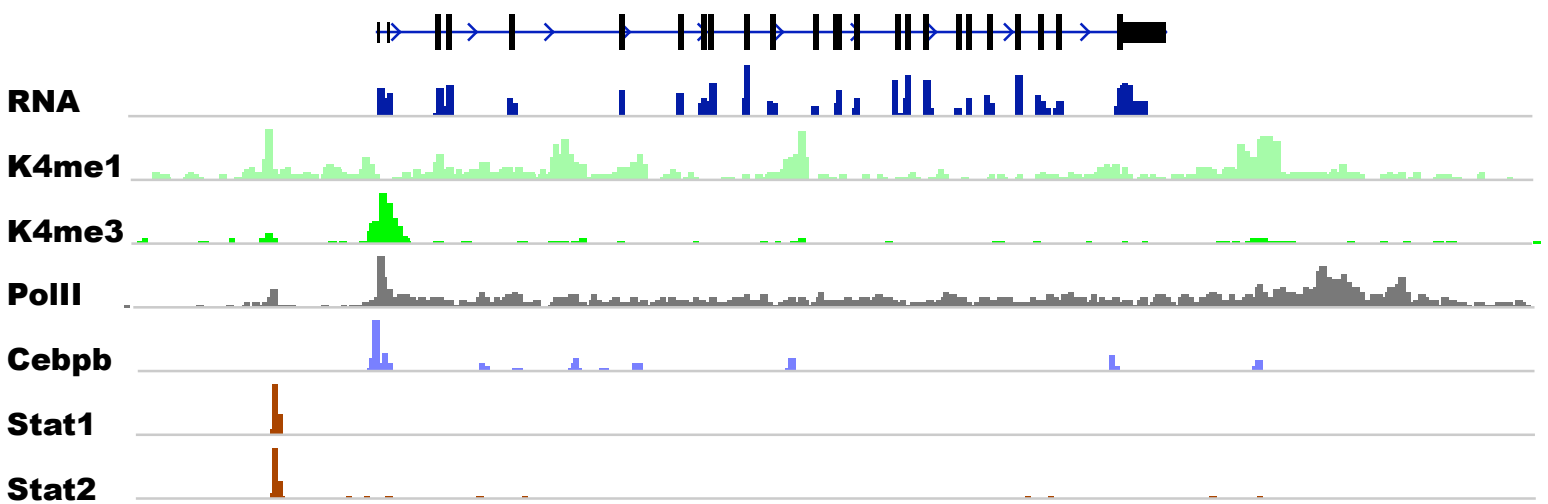
How can we identify these chromatin marks and the genes within?

Short modification

Long modification



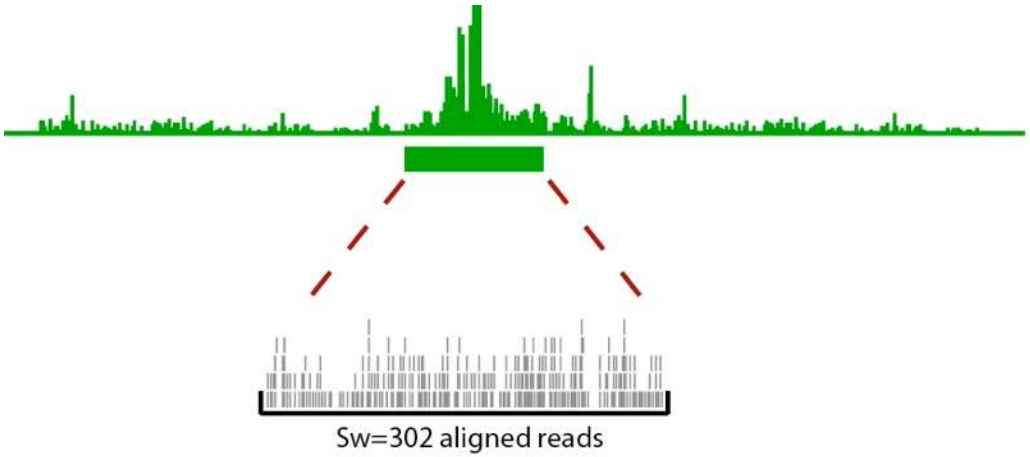
Discontinuous data



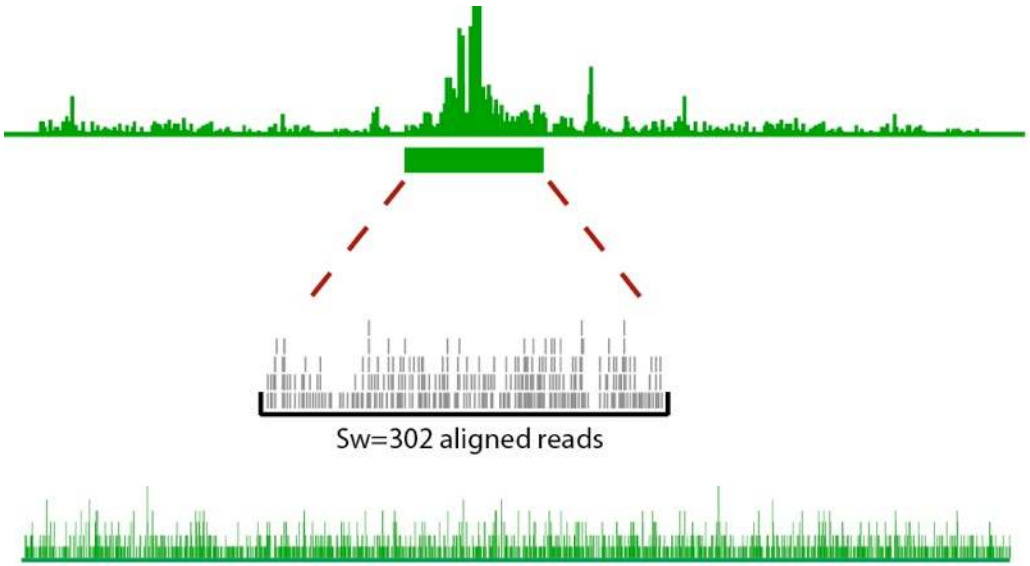
Scripture is a method to solve this general question

Our approach

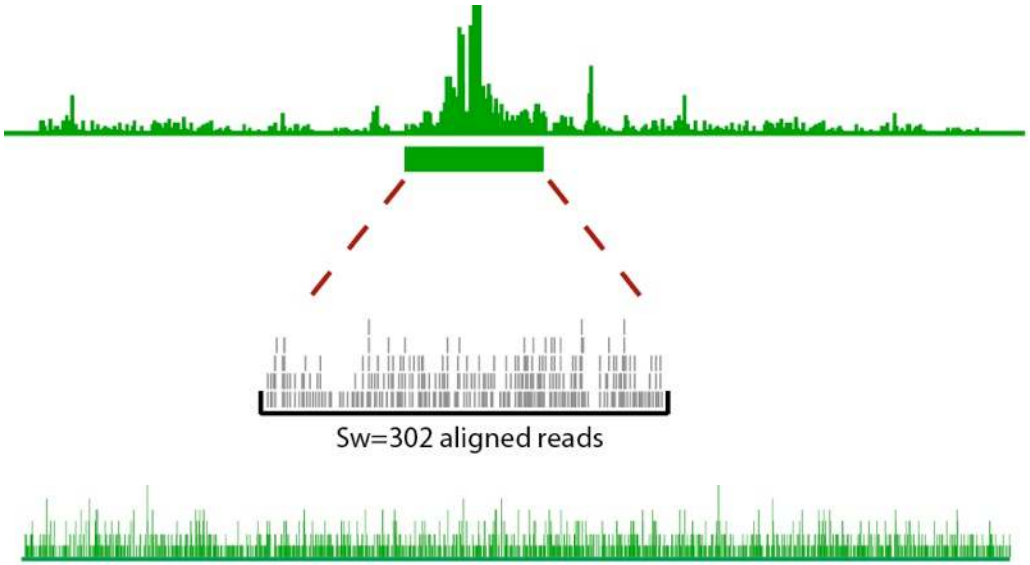
Our approach



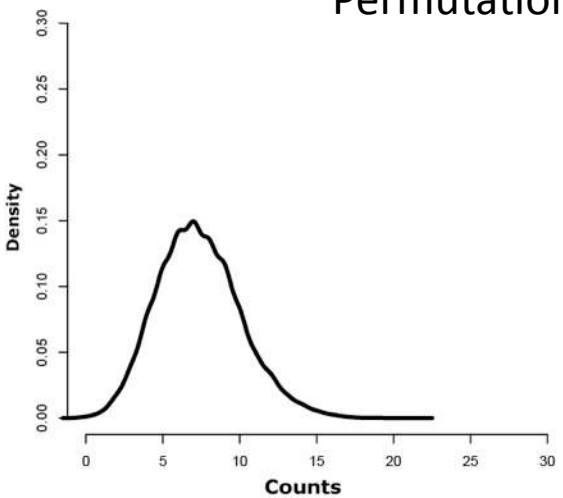
Our approach



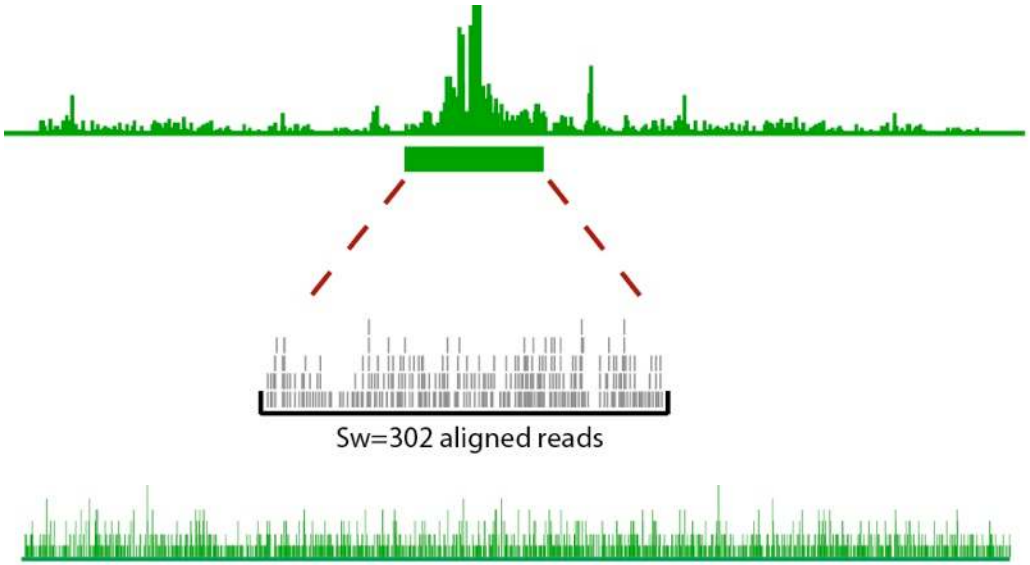
Our approach



Permutation

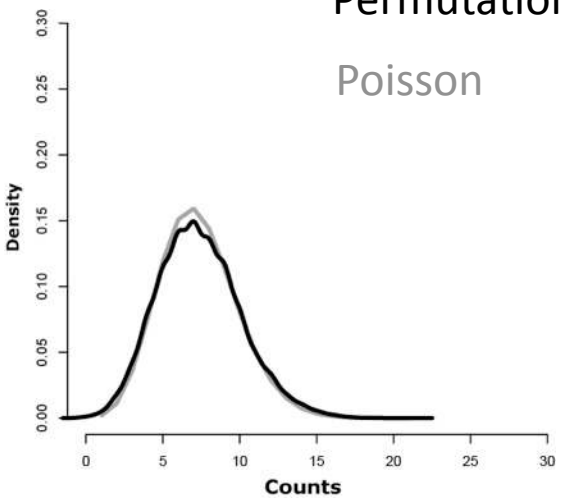


Our approach

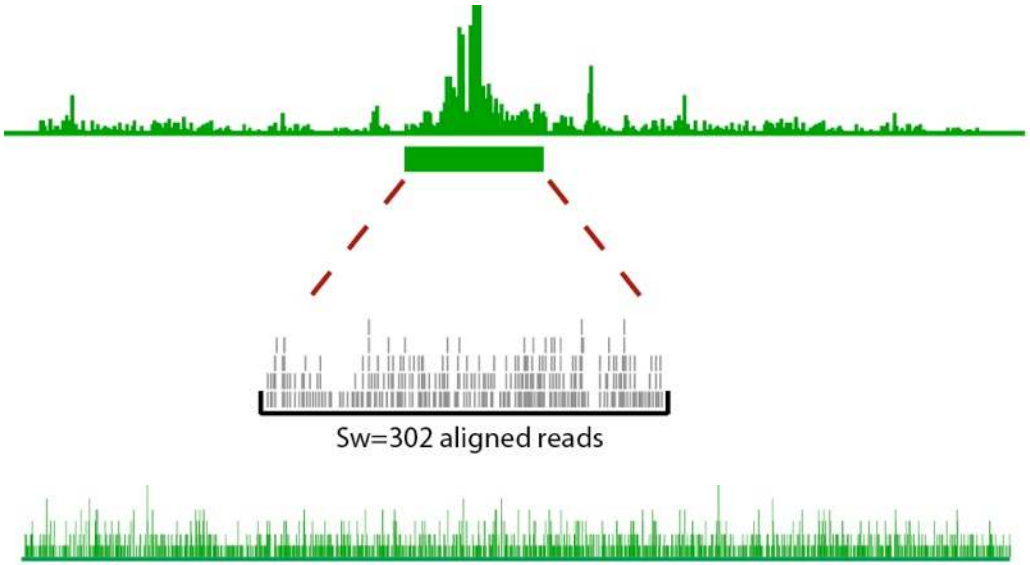


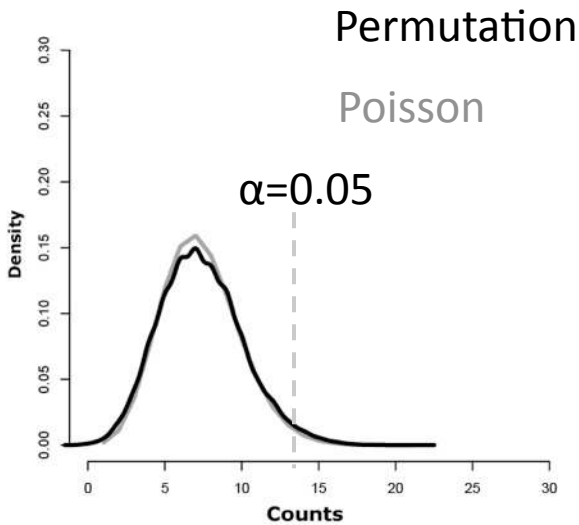
Permutation

Poisson

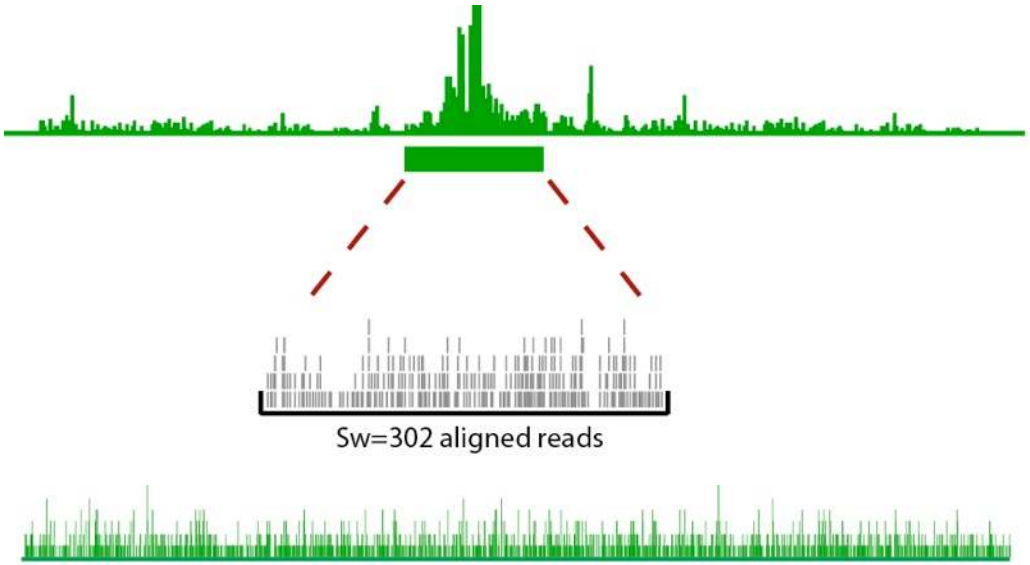


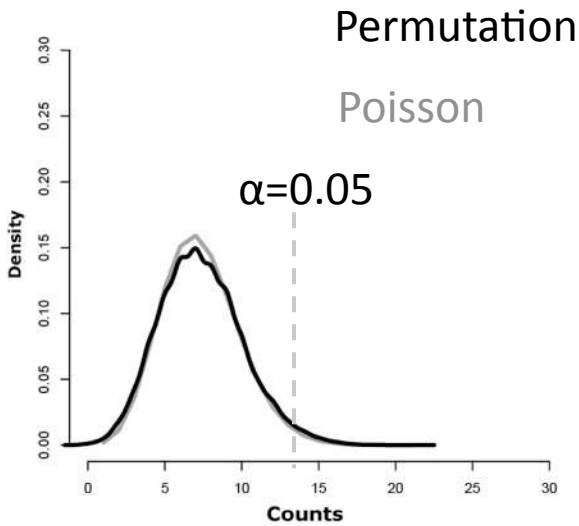
Our approach





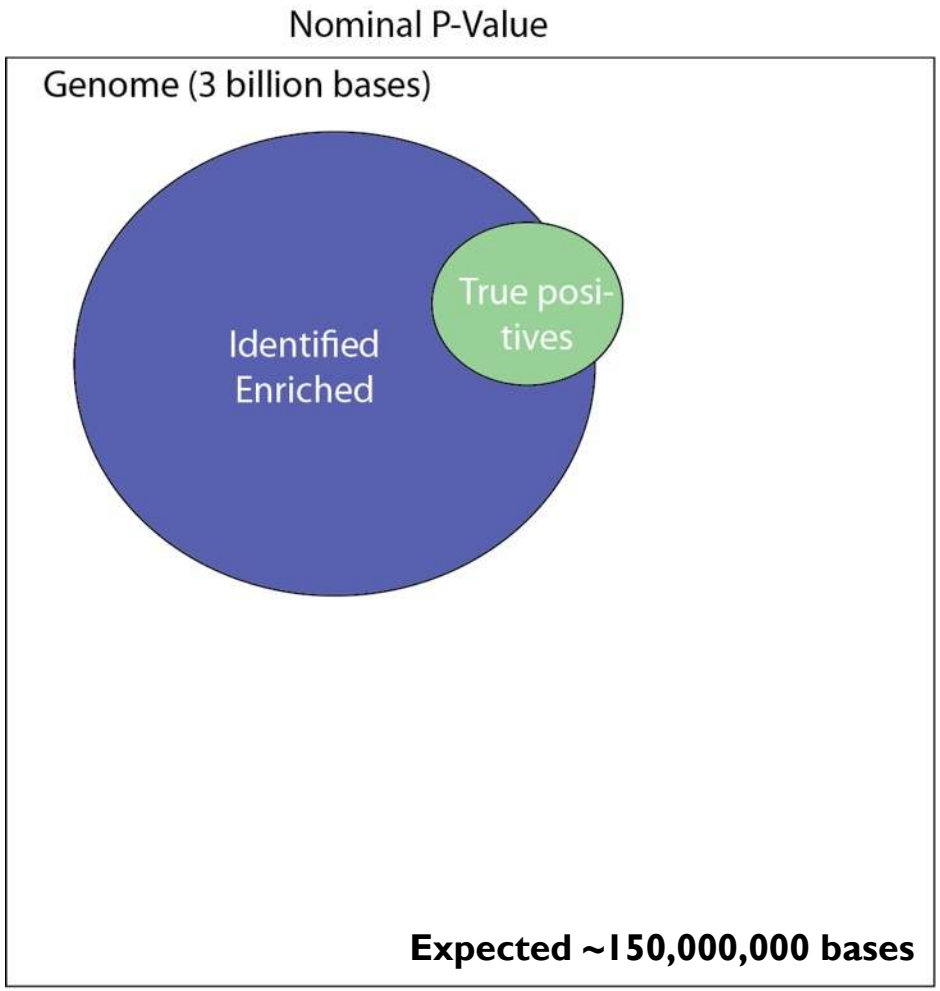
Our approach



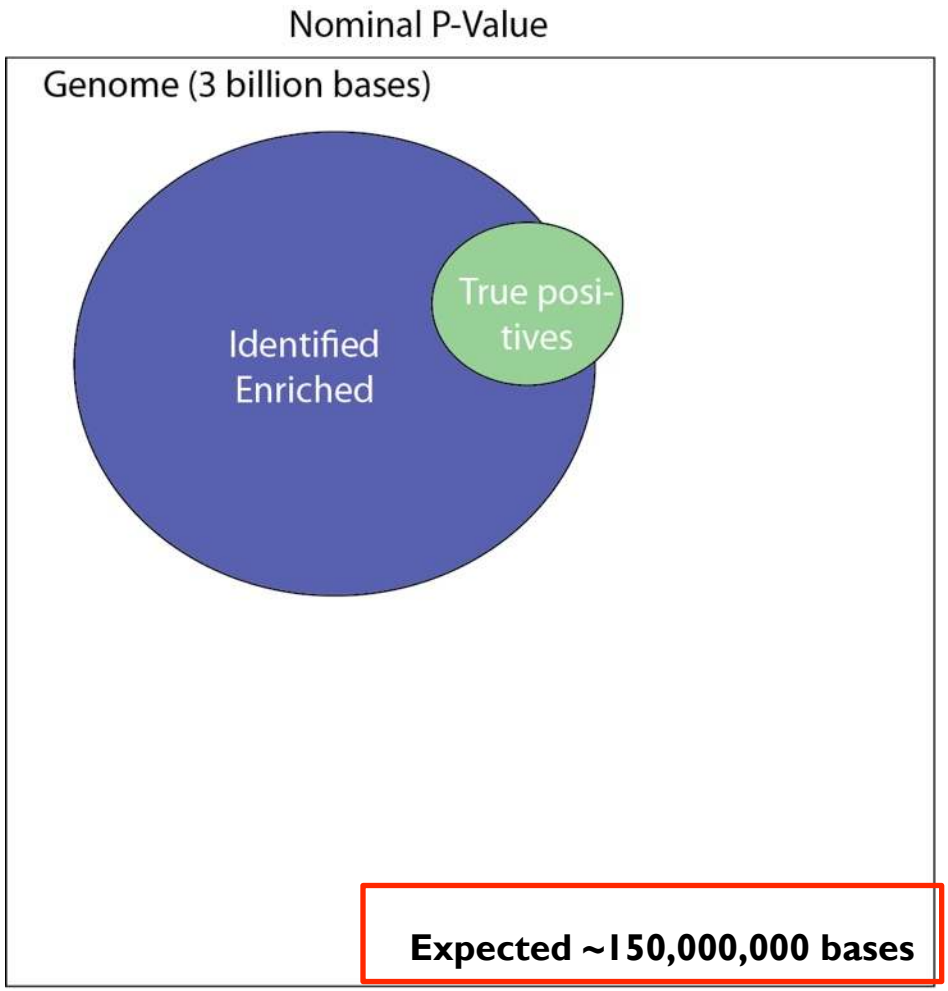


We have an efficient way to compute read count p-values …

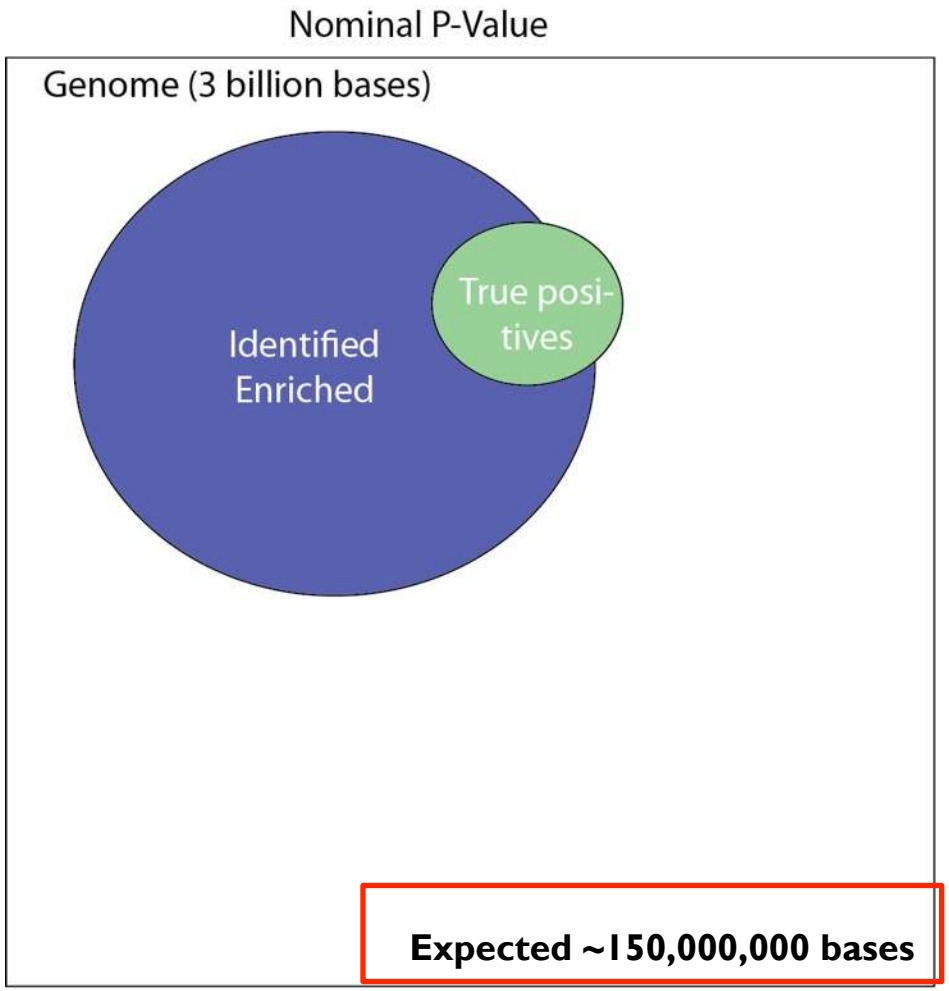
The genome is big: A lot happens by chance



The genome is big: A lot happens by chance

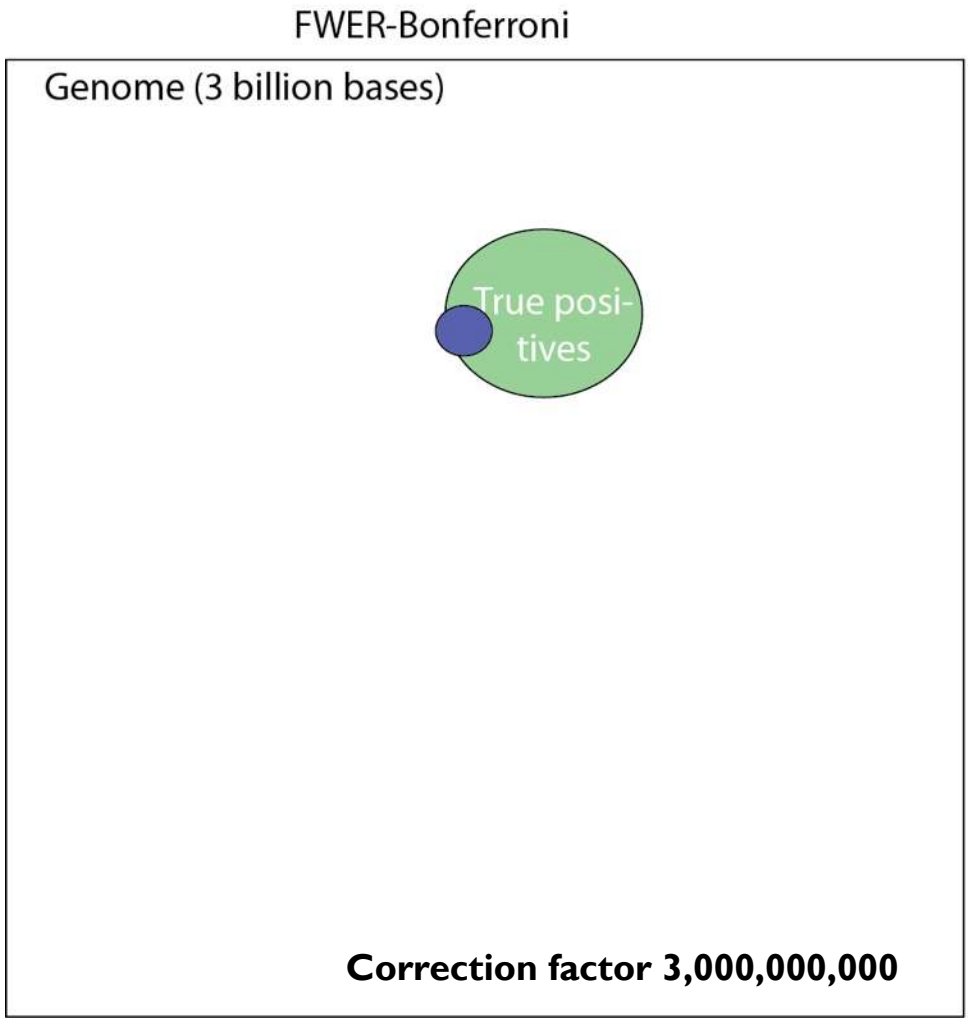


The genome is big: A lot happens by chance



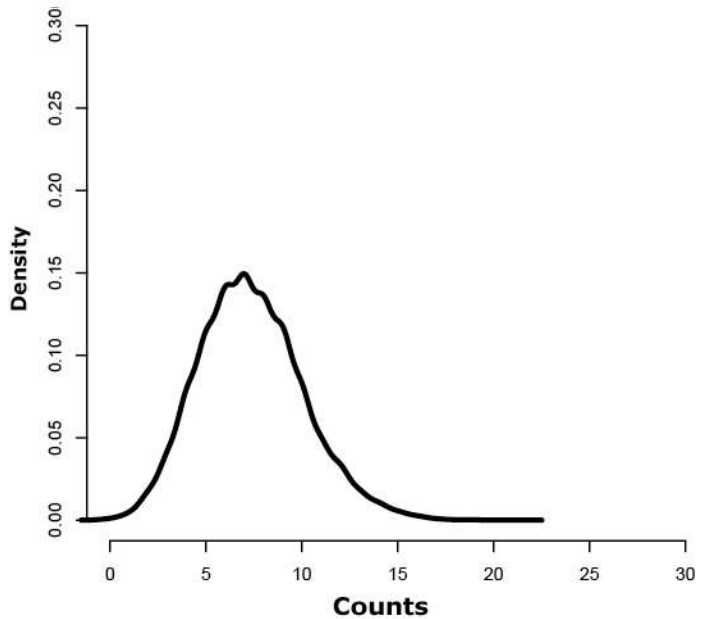
We need to correct for multiple hypothesis testing

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



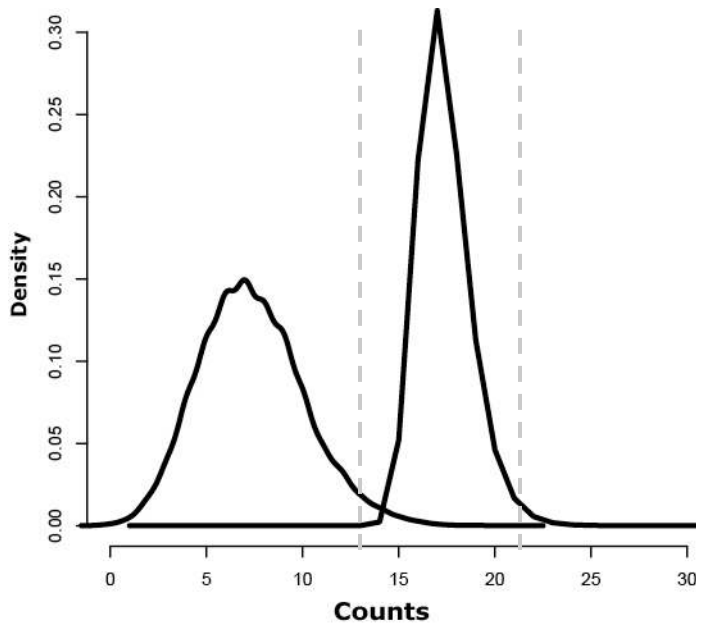
Count distribution (Poisson)

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?

Controlling FWER

Max Count distribution

α=0.05 α_{FWER}=0.05



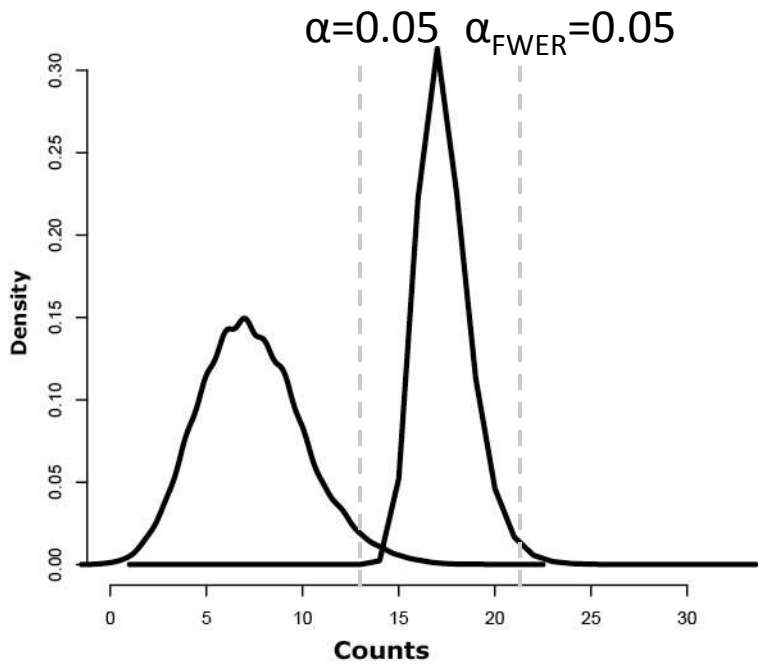
Count distribution (Poisson)

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!!!

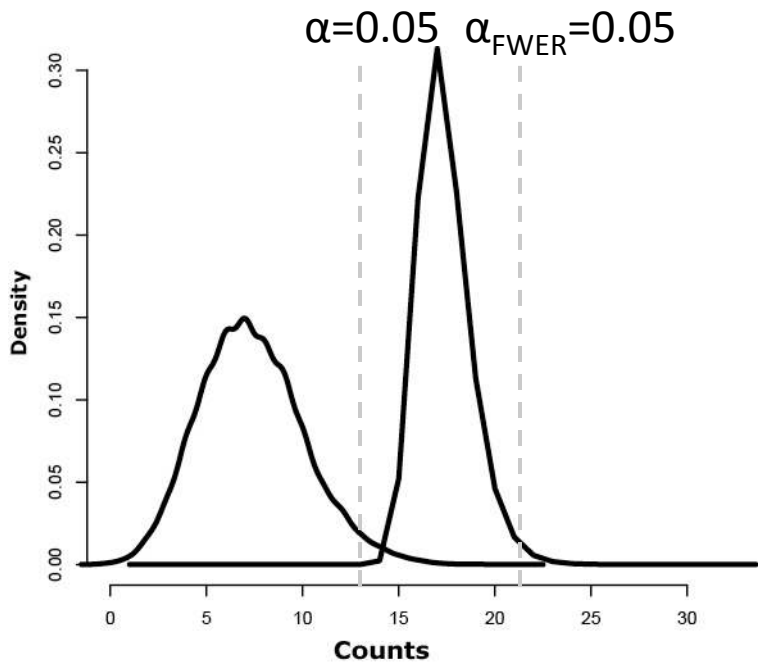
Scan distribution, an old problem

- 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, 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?



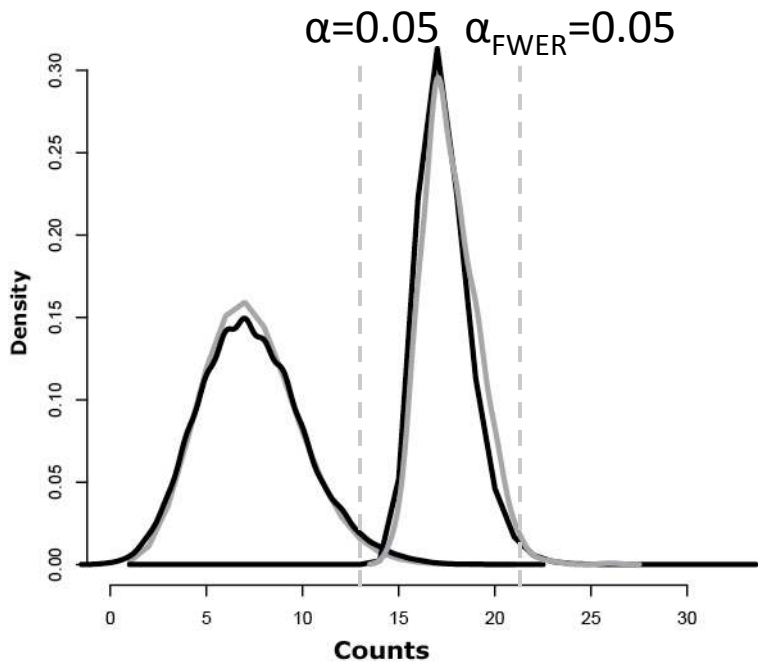
Thankfully, there is a distribution called the Scan Distribution which computes a closed form for this distribution.

ACCOUNTS for dependency of overlapping windows thus more powerful!

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, there is a distribution called the Scan Distribution which 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):

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

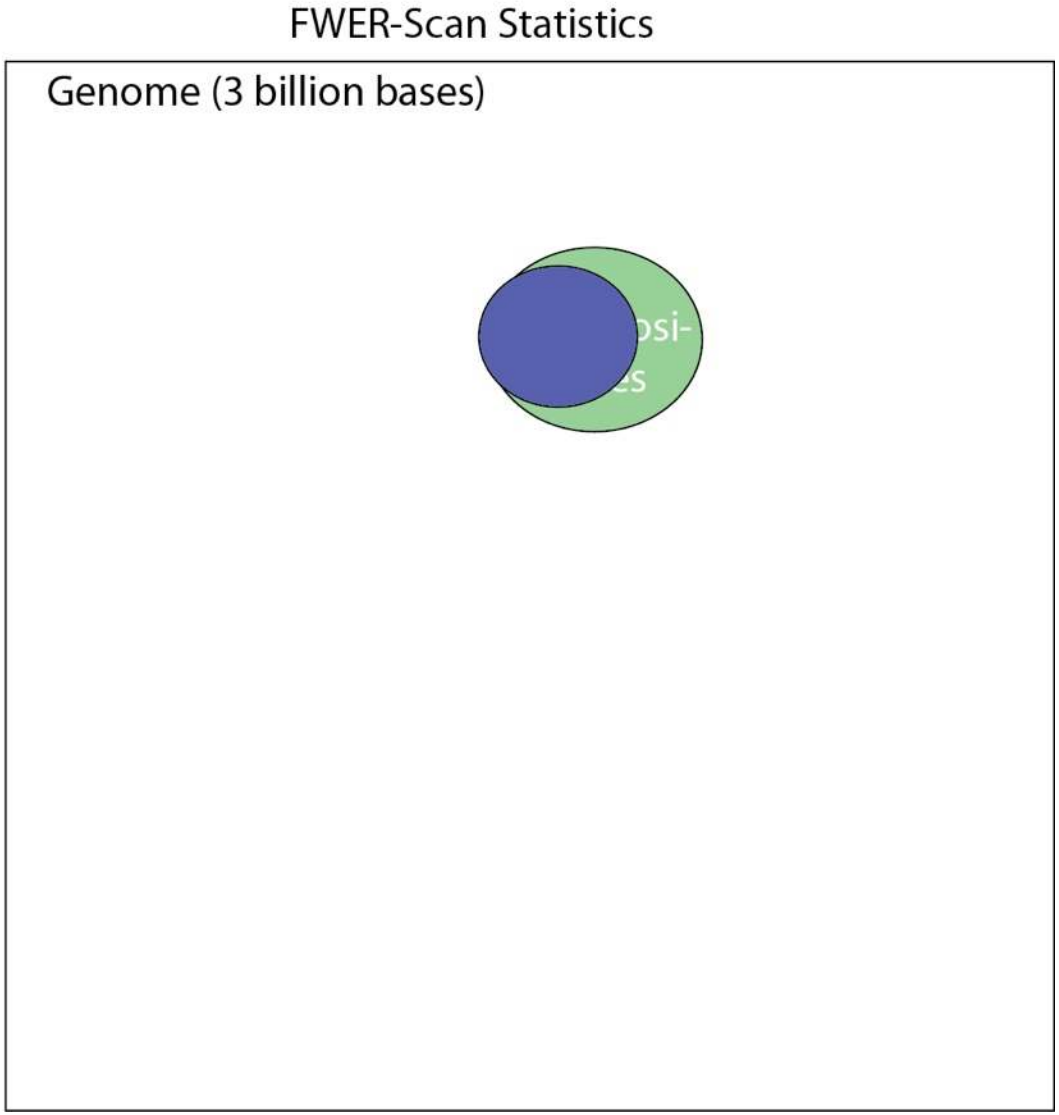
where

$P(k-1|\lambda w)$ 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

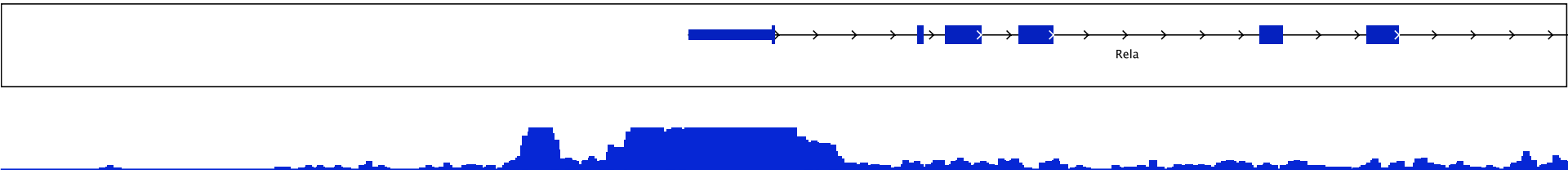
The scan distribution gives a computationally very efficient way to estimate the FWER



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

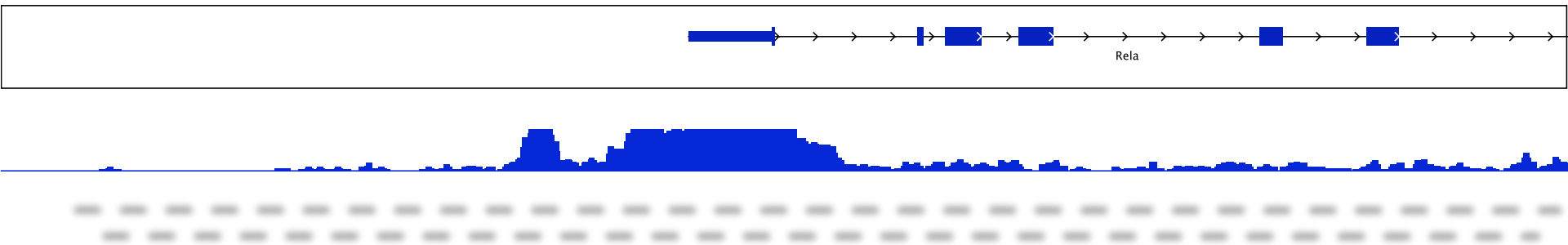
Segmentation method for contiguous regions

Example : PolII ChIP



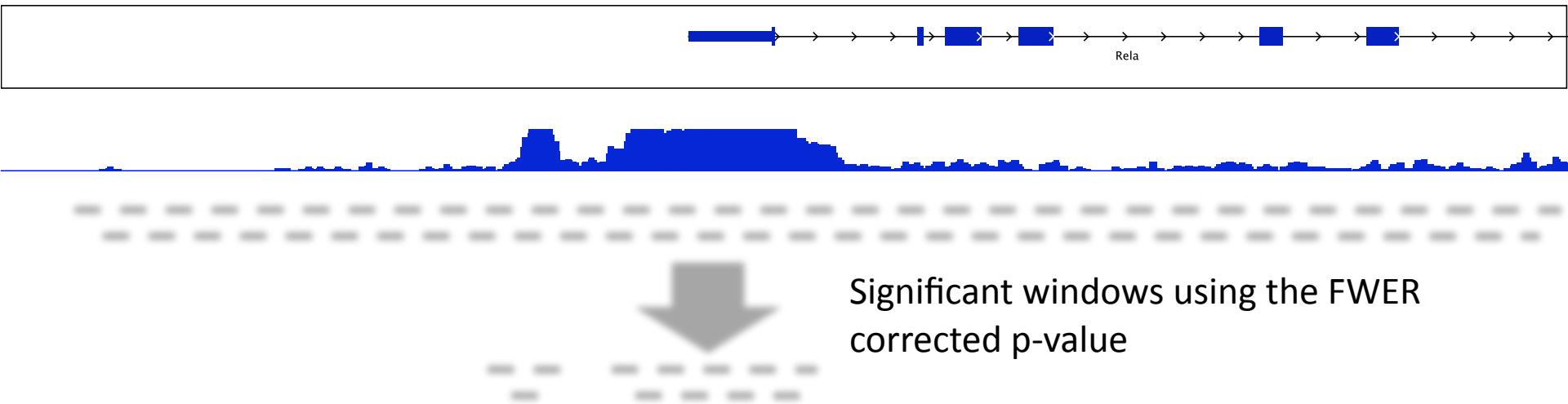
Segmentation method for contiguous regions

Example : PolII ChIP



Segmentation method for contiguous regions

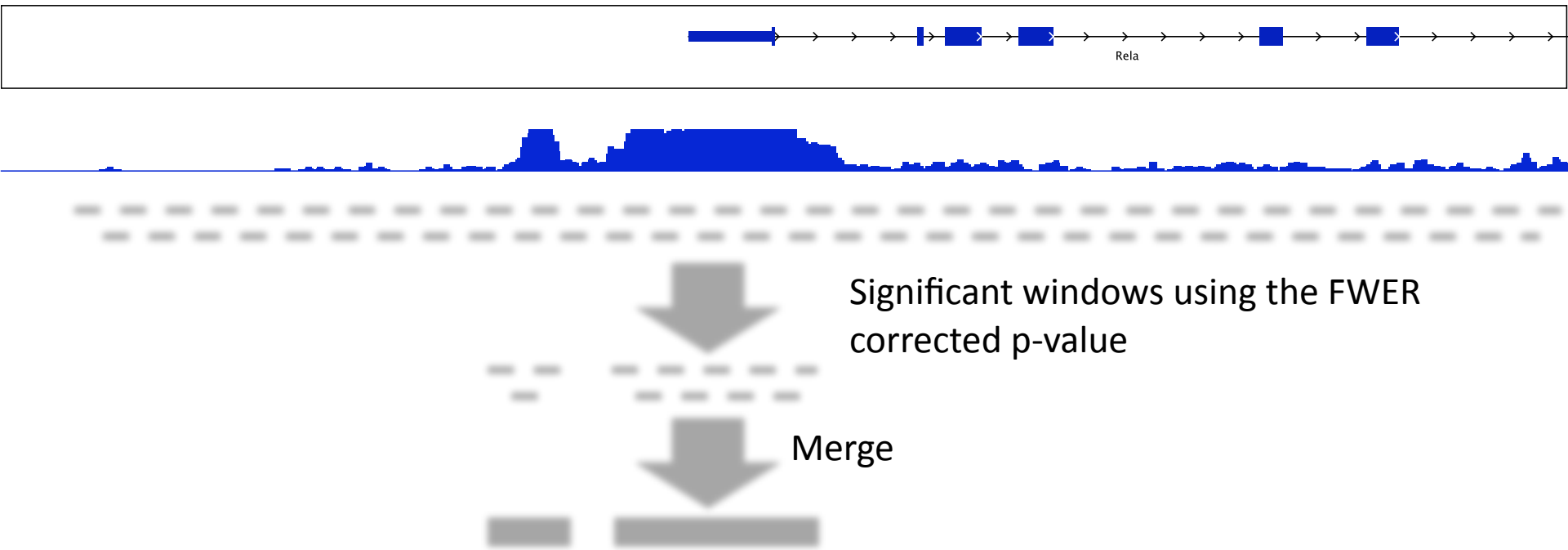
Example : PolII ChIP



Significant windows using the FWER corrected p-value

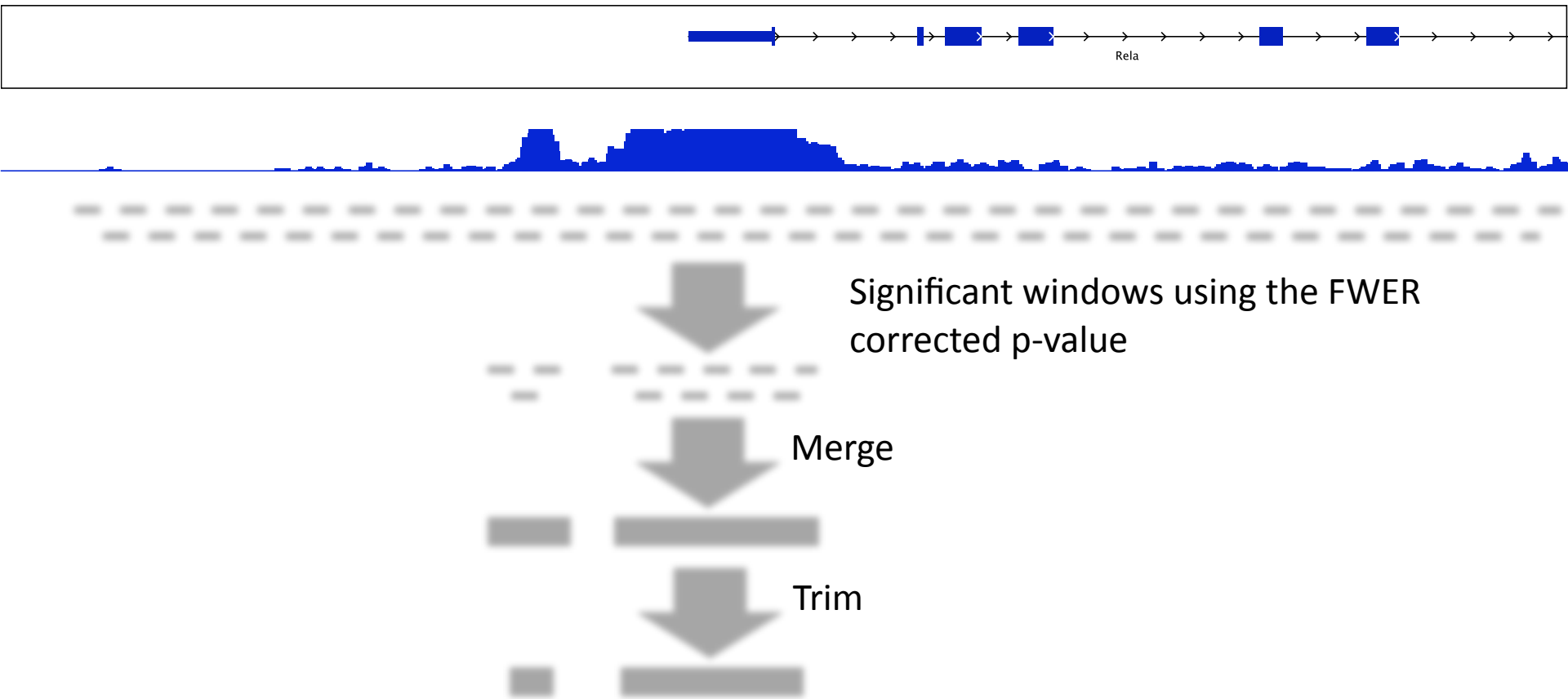
Segmentation method for contiguous regions

Example : PolII ChIP



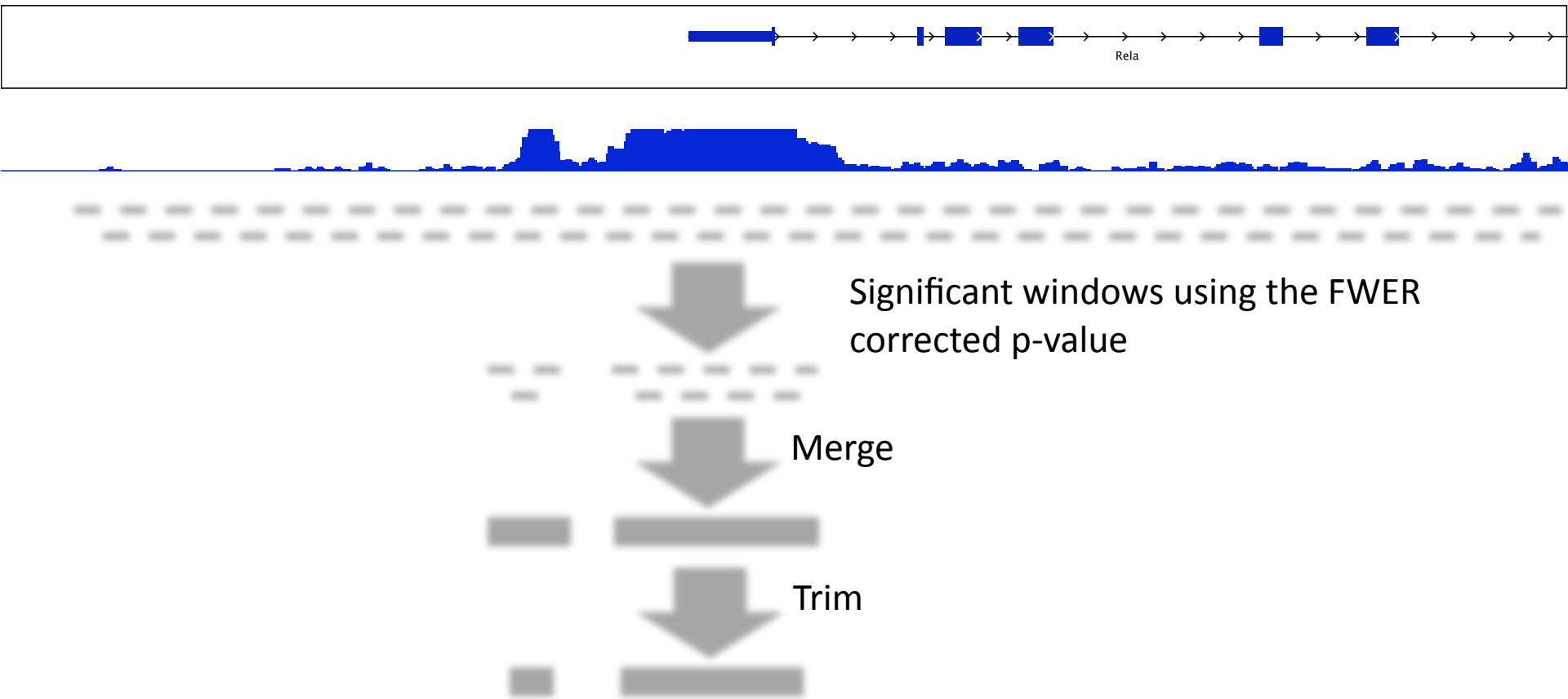
Segmentation method for contiguous regions

Example : PolII ChIP



Segmentation method for contiguous regions

Example : PolII ChIP

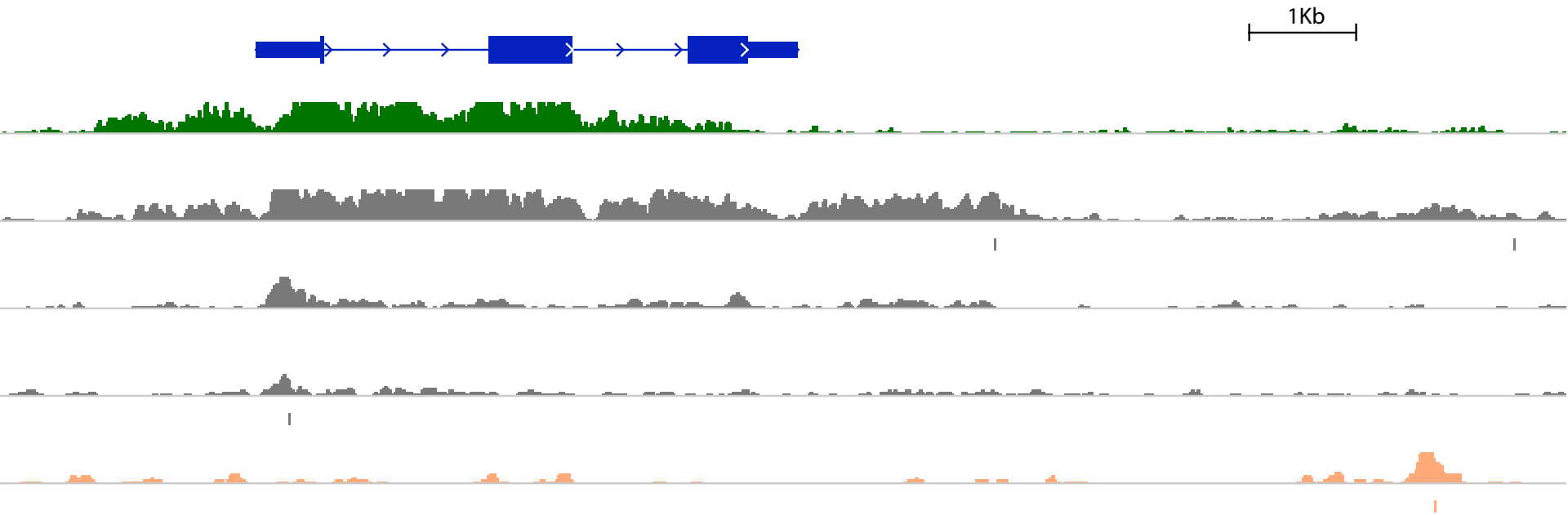


But, which window?

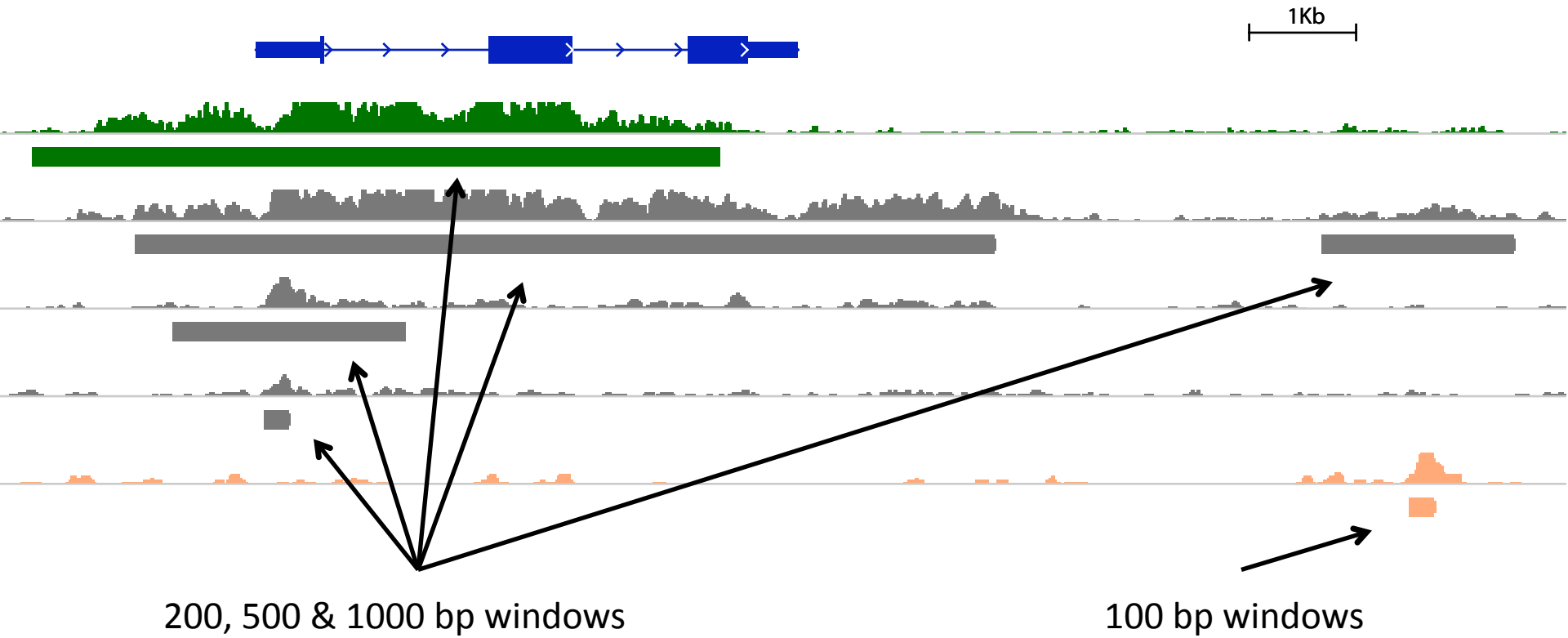
We use multiple windows

- 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.

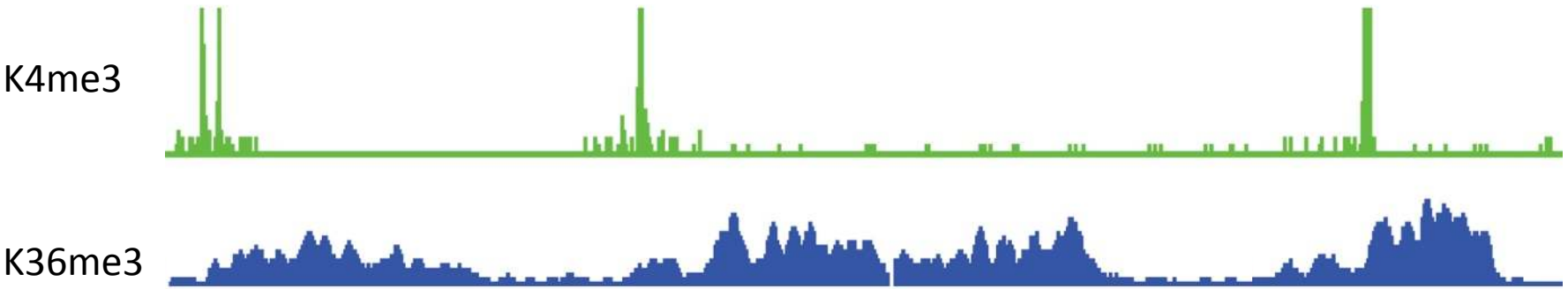
Applying Scripture to a variety of ChIP-Seq data



Applying Scripture to a variety of ChIP-Seq data

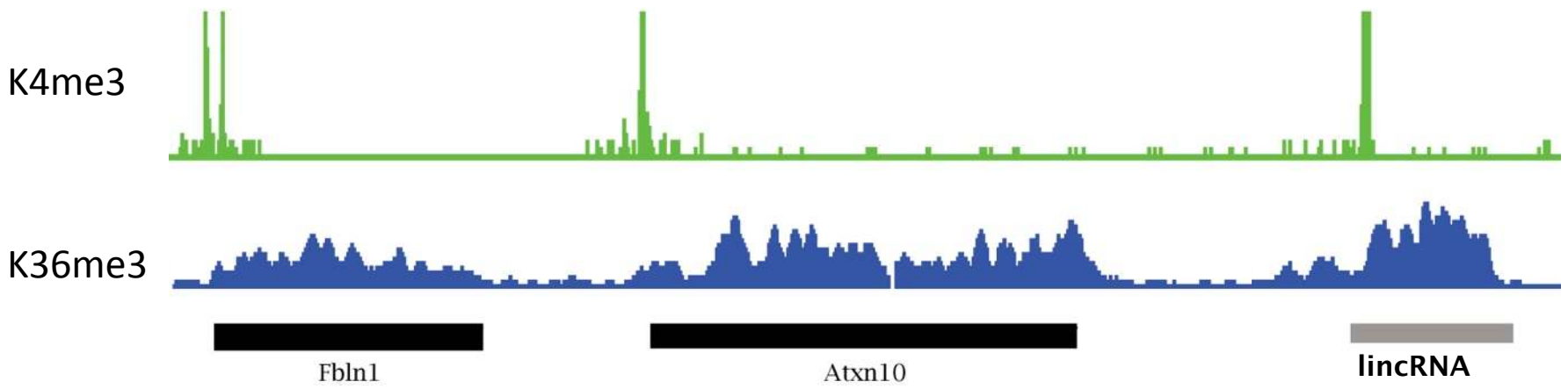


Application of scripture to mouse chromatin state maps



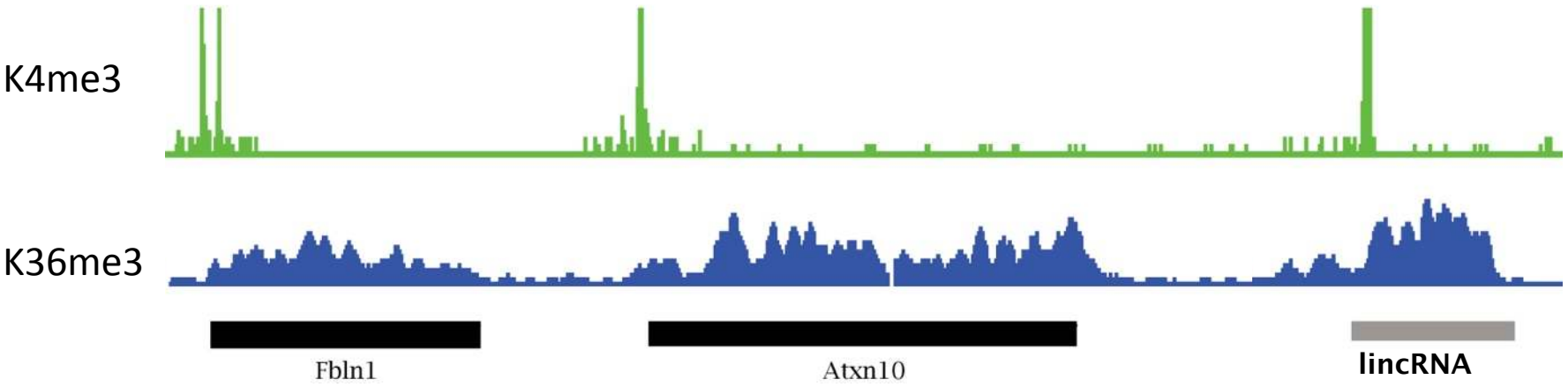
Mitch Guttman

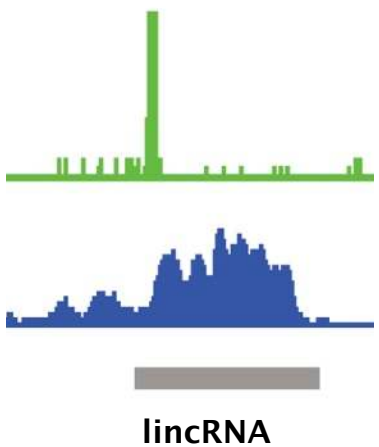
Application of scripture to mouse chromatin state maps



Mitch Guttman

Application of scripture to mouse chromatin state maps





Identifed

- ~1500 lincRNAs
  - Conserved
  - Noncoding
  - Robustly expressed

Mitch Guttman

Can we identify enriched regions across different data types?

Short modification ✓

Long modification ✓

Can we identify enriched regions across different data types?

Short modification ✓

Long modification ✓

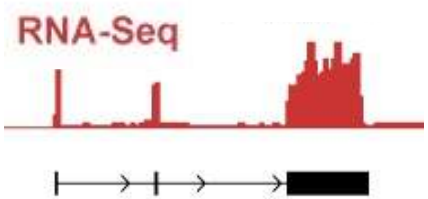
Using chromatin signatures we discovered hundreds of putative genes.
What is their structure?

Can we identify enriched regions across different data types?

Short modification ✓

Long modification ✓

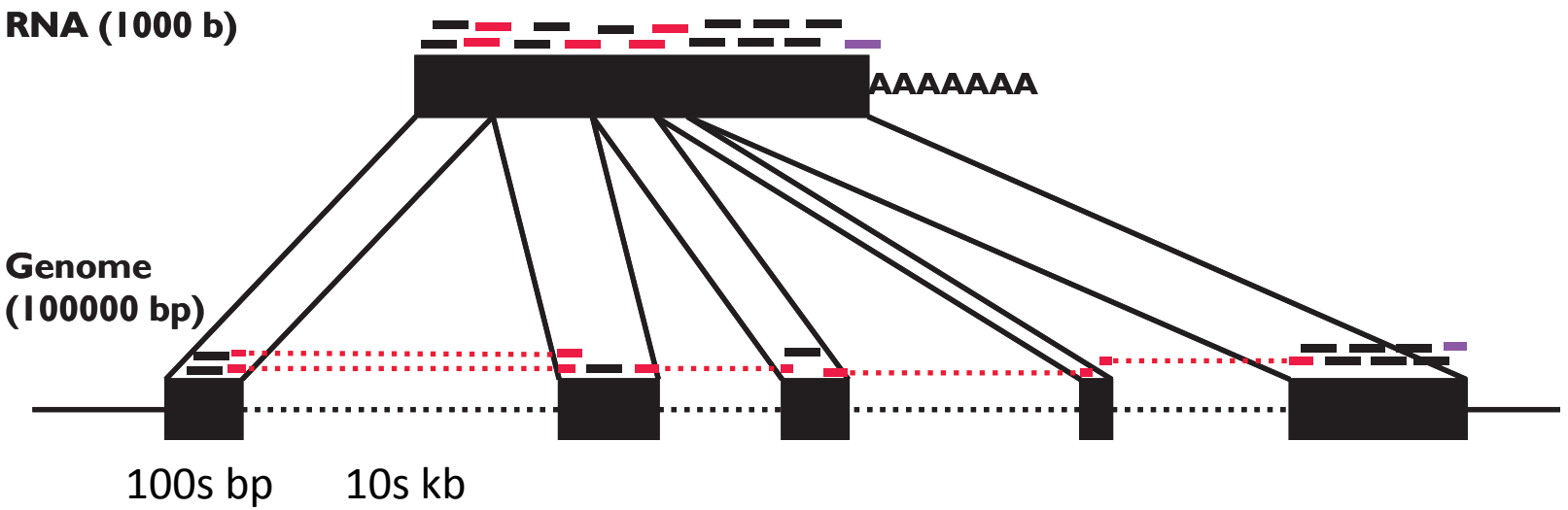
Using chromatin signatures we discovered hundreds of putative genes.
What is their structure?



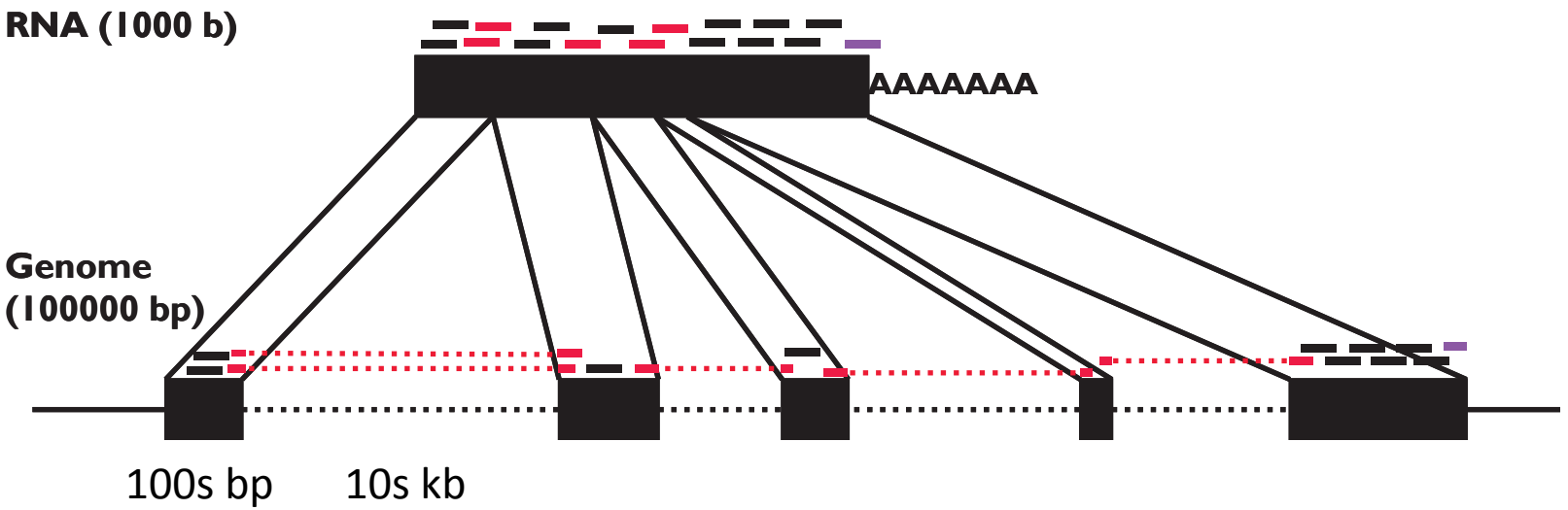
Discontinuous data: RNA-Seq to find gene structures for this gene-like regions

Scripture for RNA-Seq:
Extending segmentation to discontiguous regions

The transcript reconstruction problem as a segmentation problem



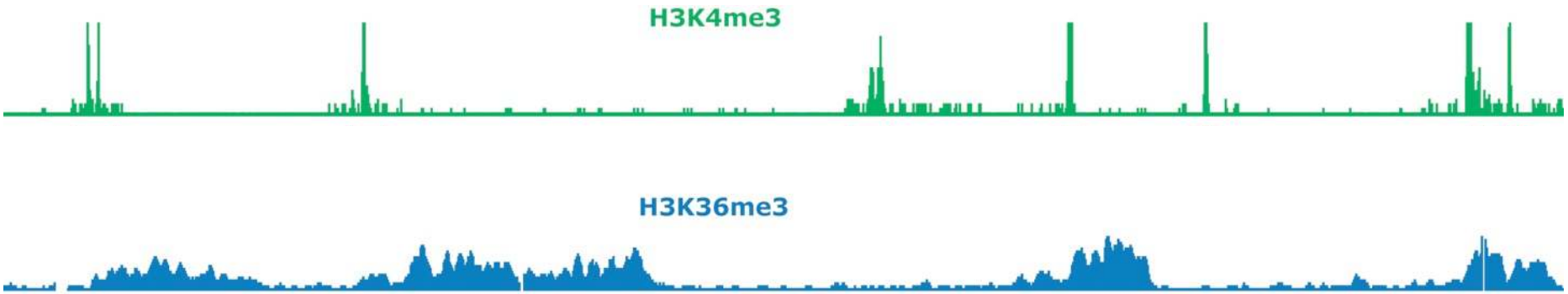
The transcript reconstruction problem as a segmentation problem



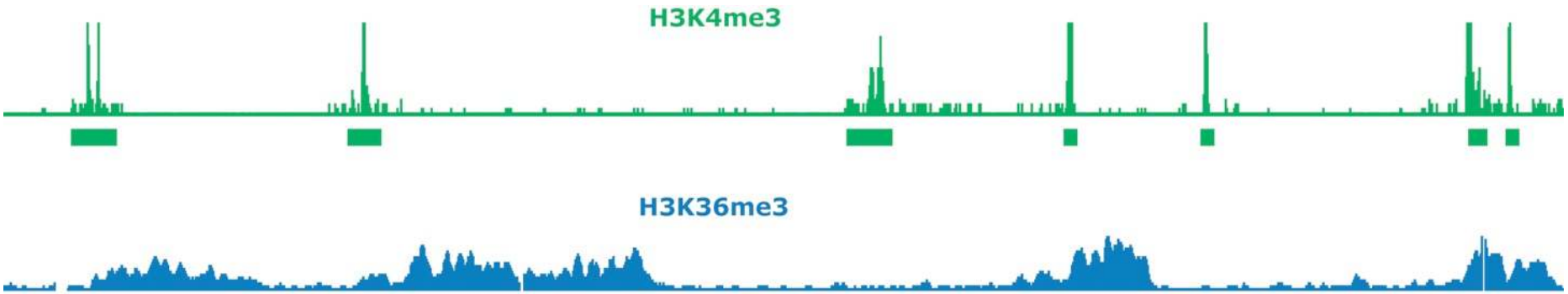
Challenges:

- 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.

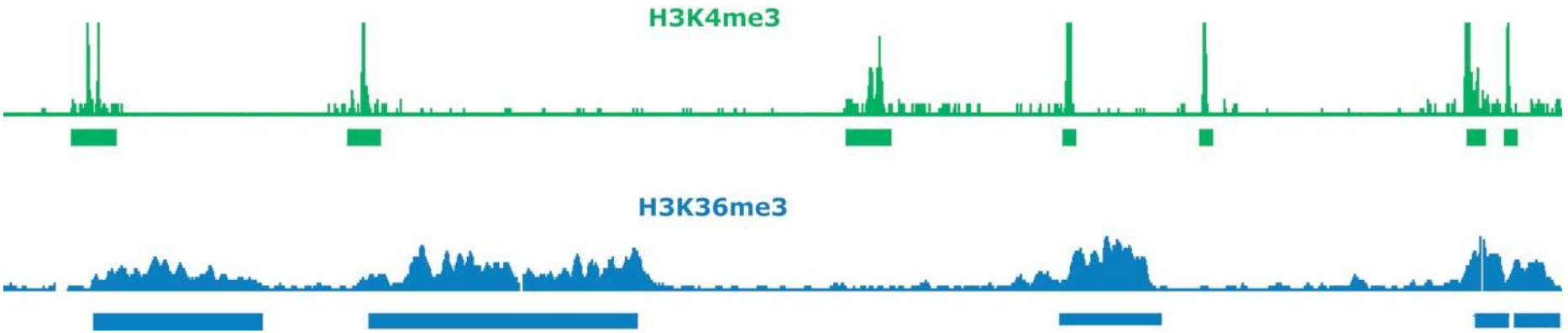
Scripture: A statistical genome-guided transcriptome reconstruction



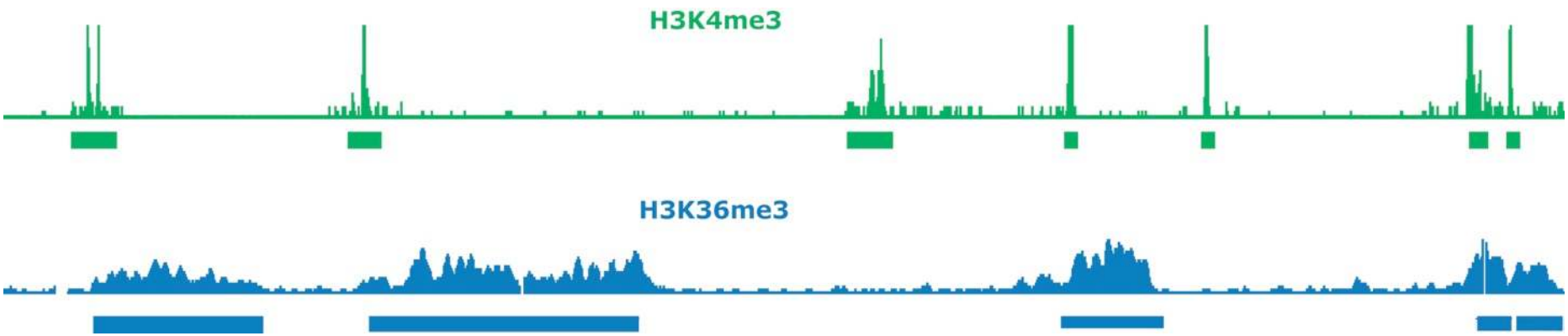
Scripture: A statistical genome-guided transcriptome reconstruction



Scripture: A statistical genome-guided transcriptome reconstruction

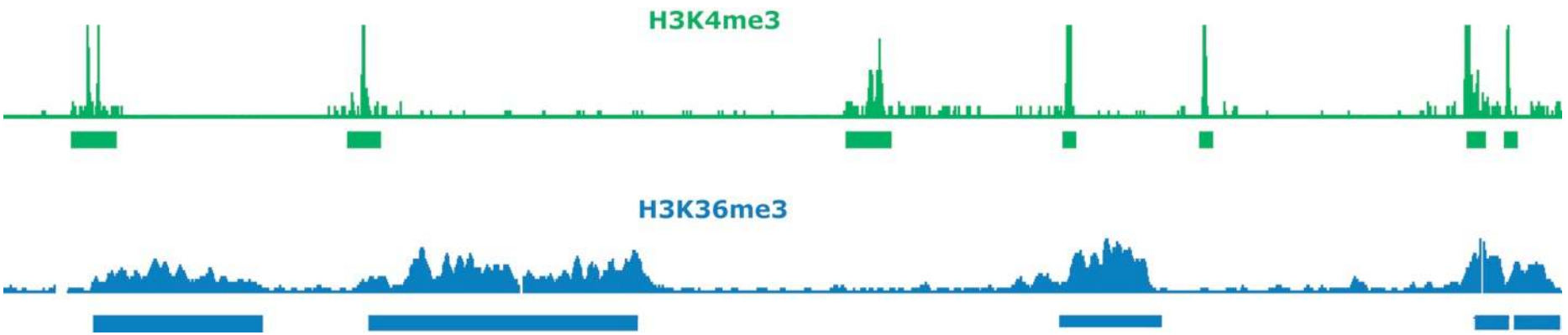


Scripture: A statistical genome-guided transcriptome reconstruction

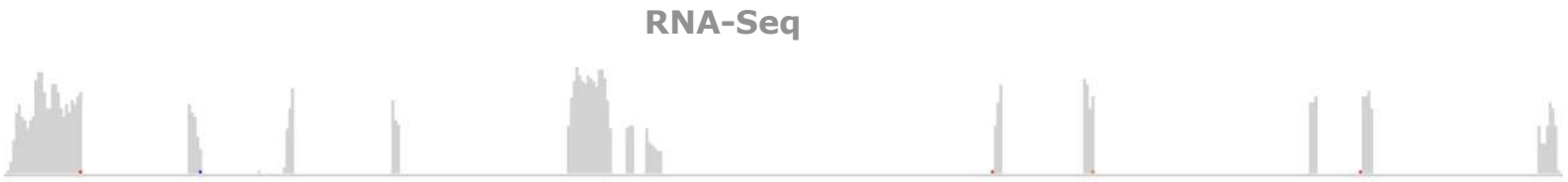


Statistical segmentation of chromatin modifications uses continuity of segments to increase power for interval detection

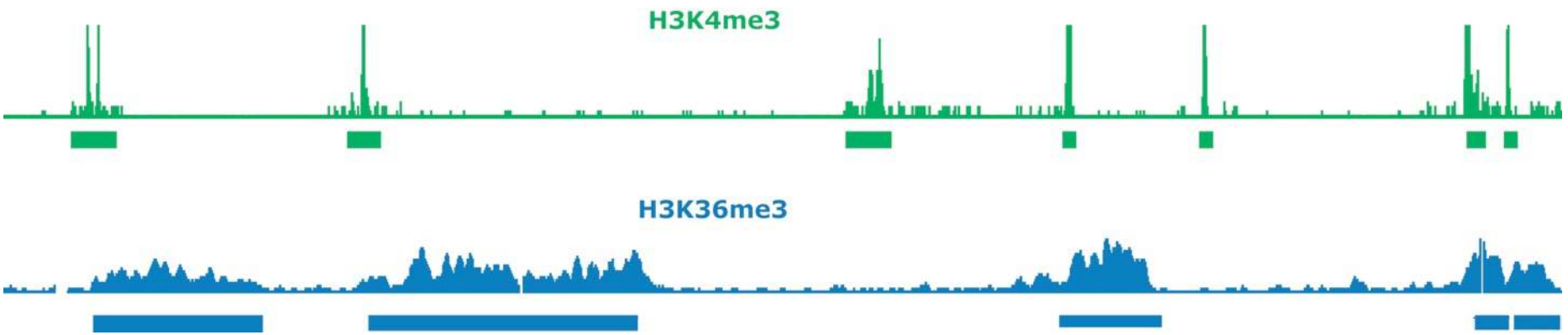
Scripture: A statistical genome-guided transcriptome reconstruction



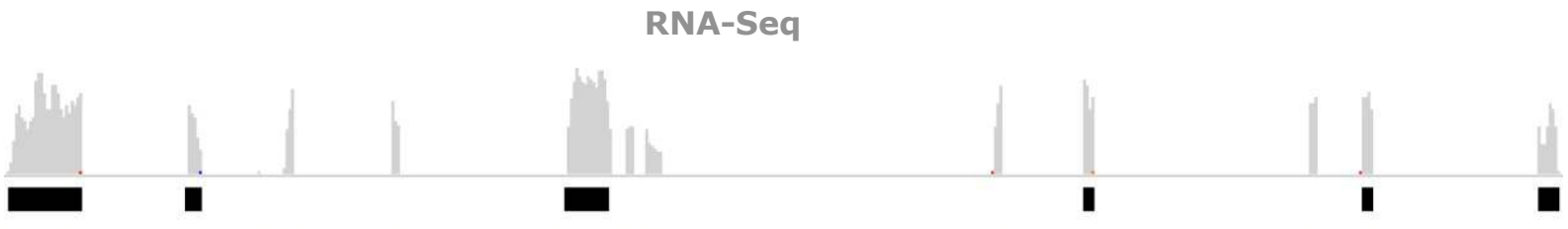
Statistical segmentation of chromatin modifications uses continuity of segments to increase power for interval detection



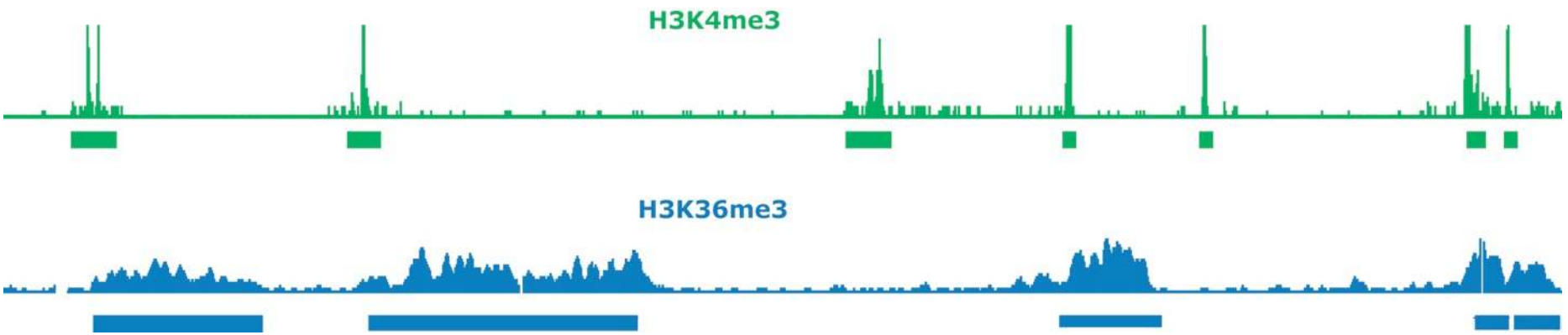
Scripture: A statistical genome-guided transcriptome reconstruction



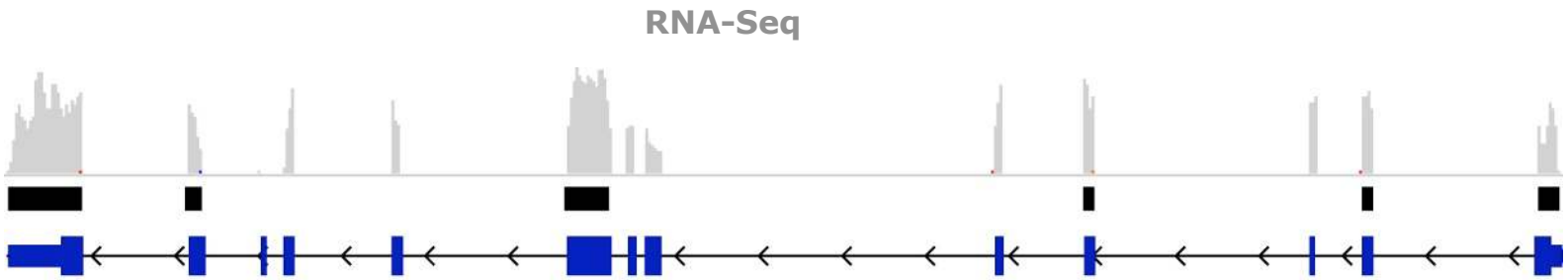
Statistical segmentation of chromatin modifications uses continuity of segments to increase power for interval detection



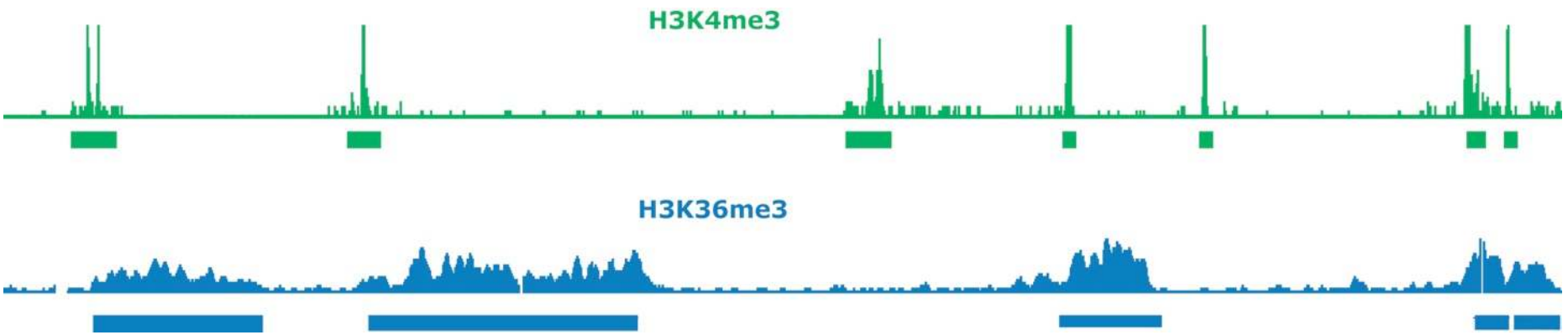
Scripture: A statistical genome-guided transcriptome reconstruction



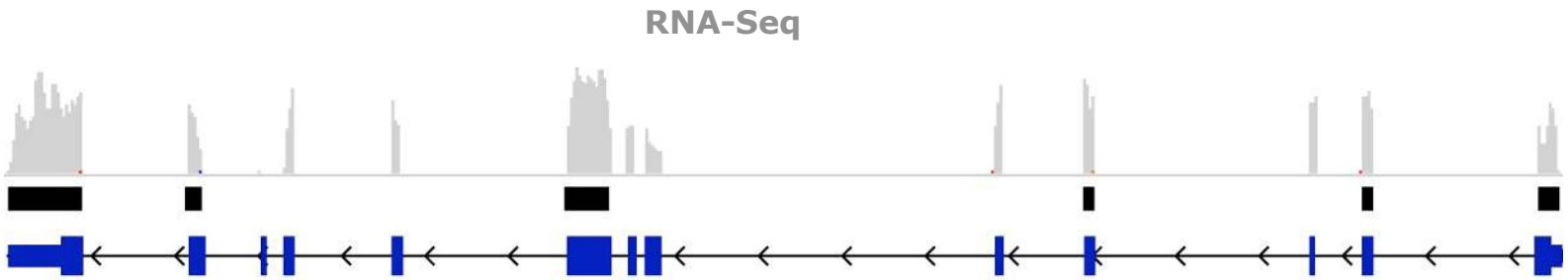
Statistical segmentation of chromatin modifications uses continuity of segments to increase power for interval detection



Scripture: A statistical genome-guided transcriptome reconstruction



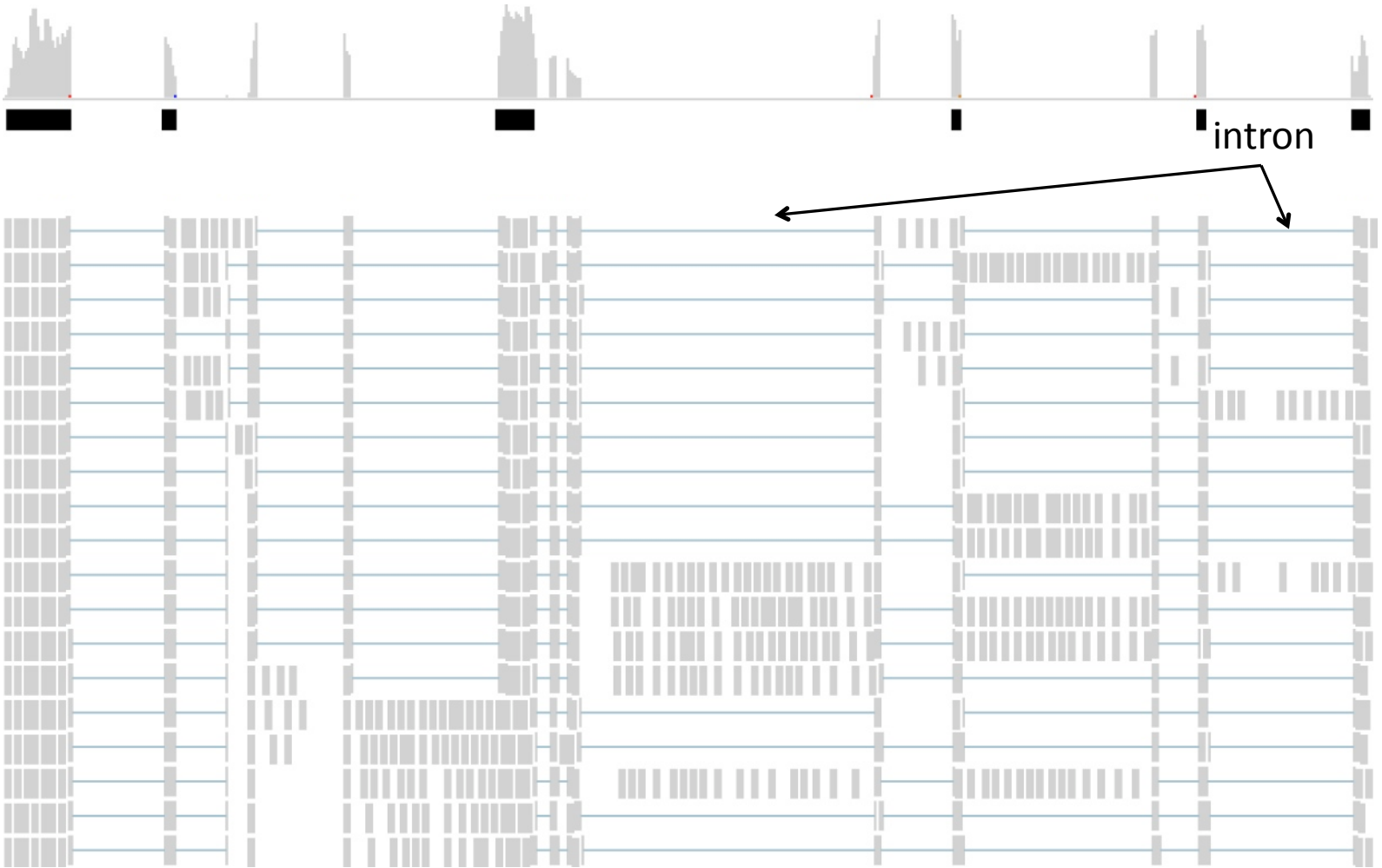
Statistical segmentation of chromatin modifications uses continuity of segments to increase power for interval detection



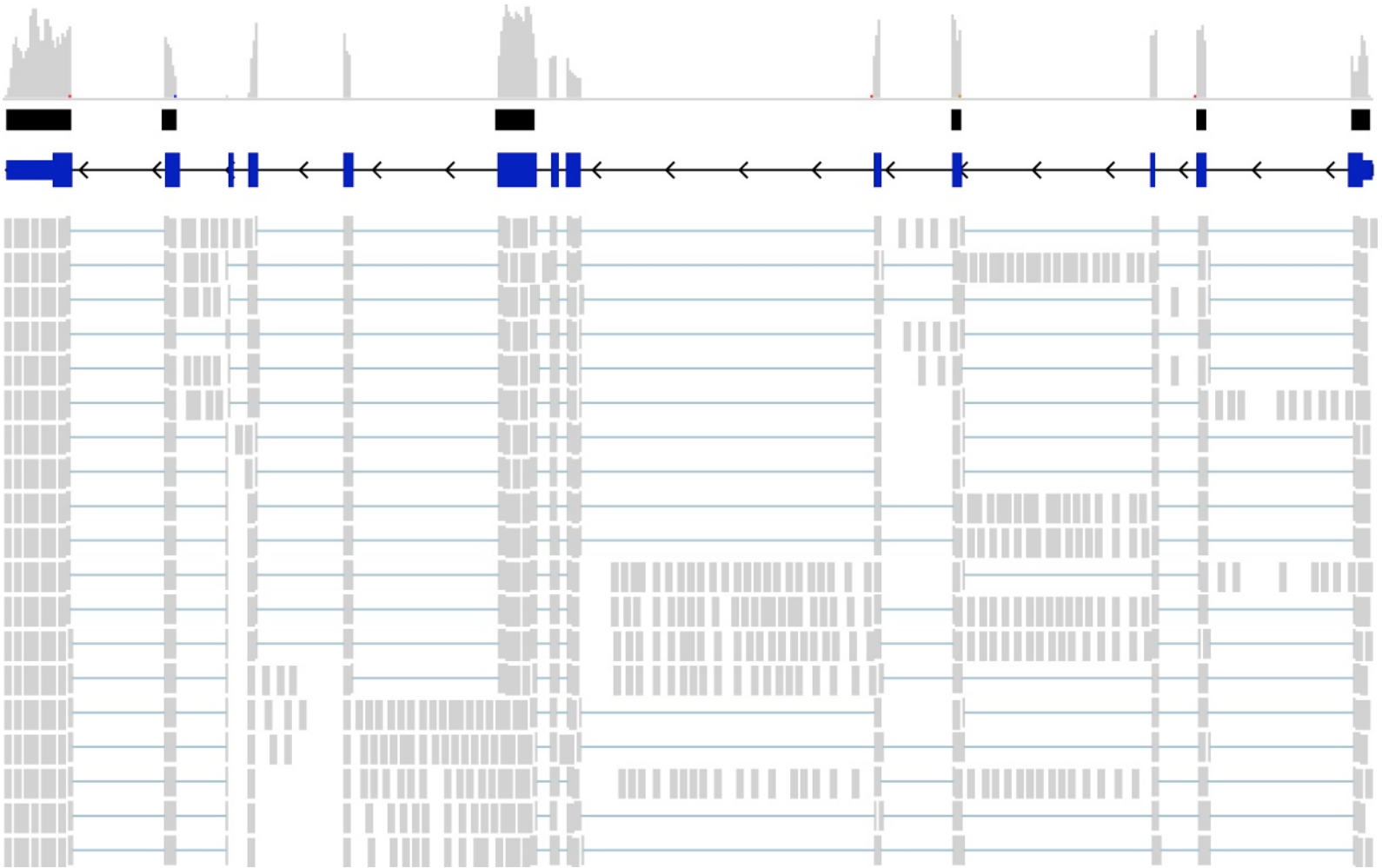
If we know the connectivity of fragments, we can increase our power to detect transcripts

Longer (76) reads provide increased number of junction reads

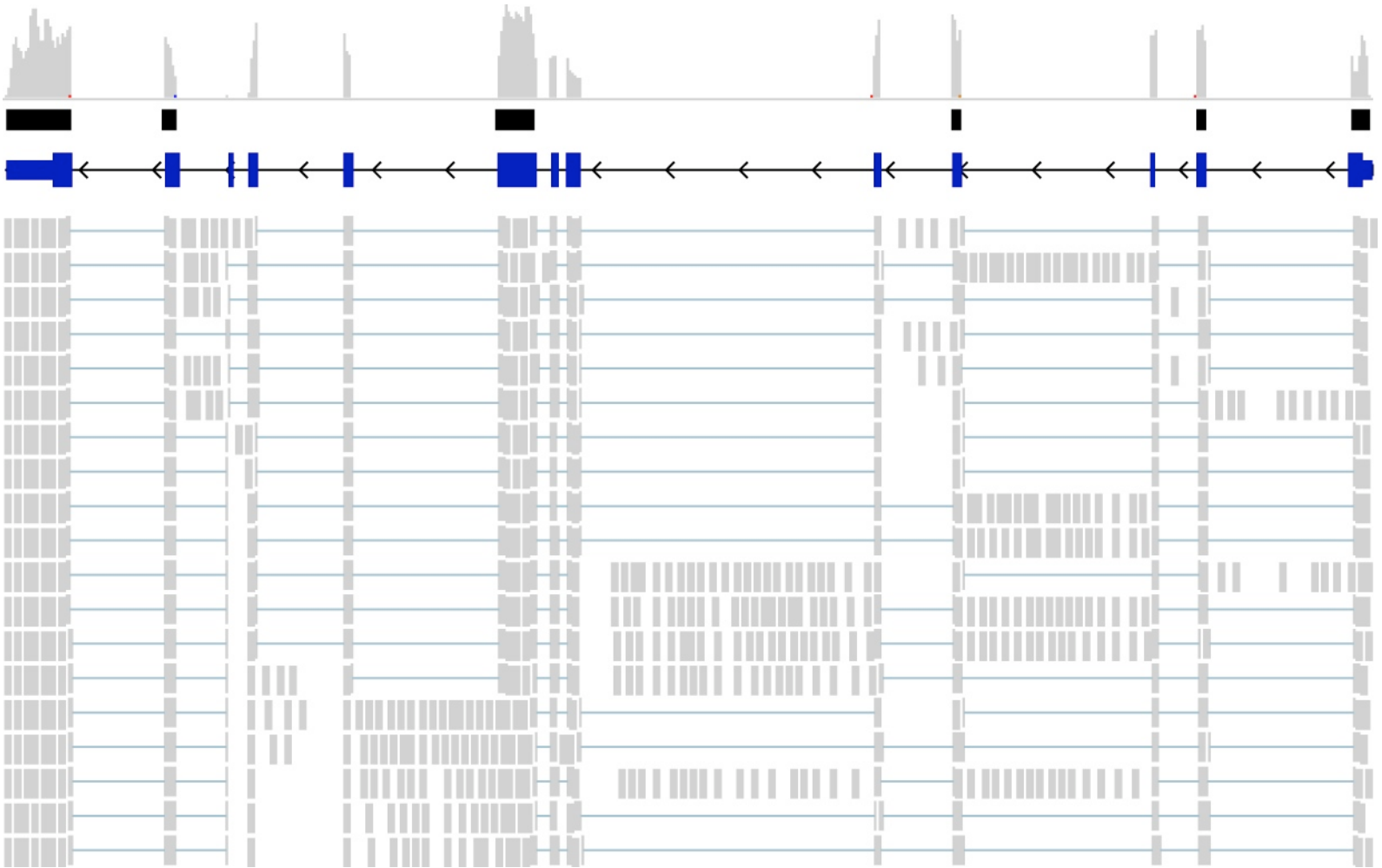
Longer (76) reads provide increased number of junction reads



Longer (76) reads provide increased number of junction reads

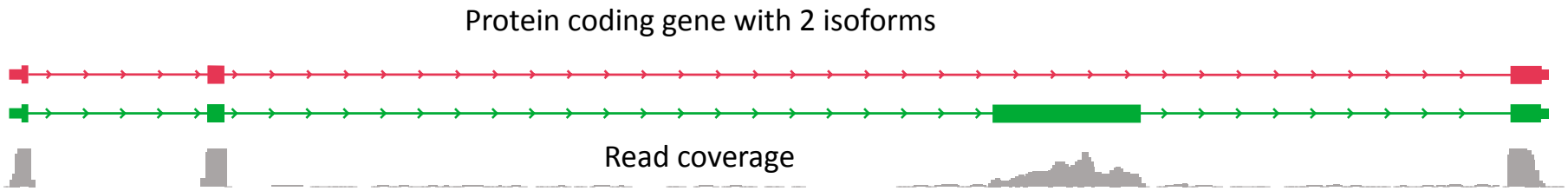


Longer (76) reads provide increased number of junction reads

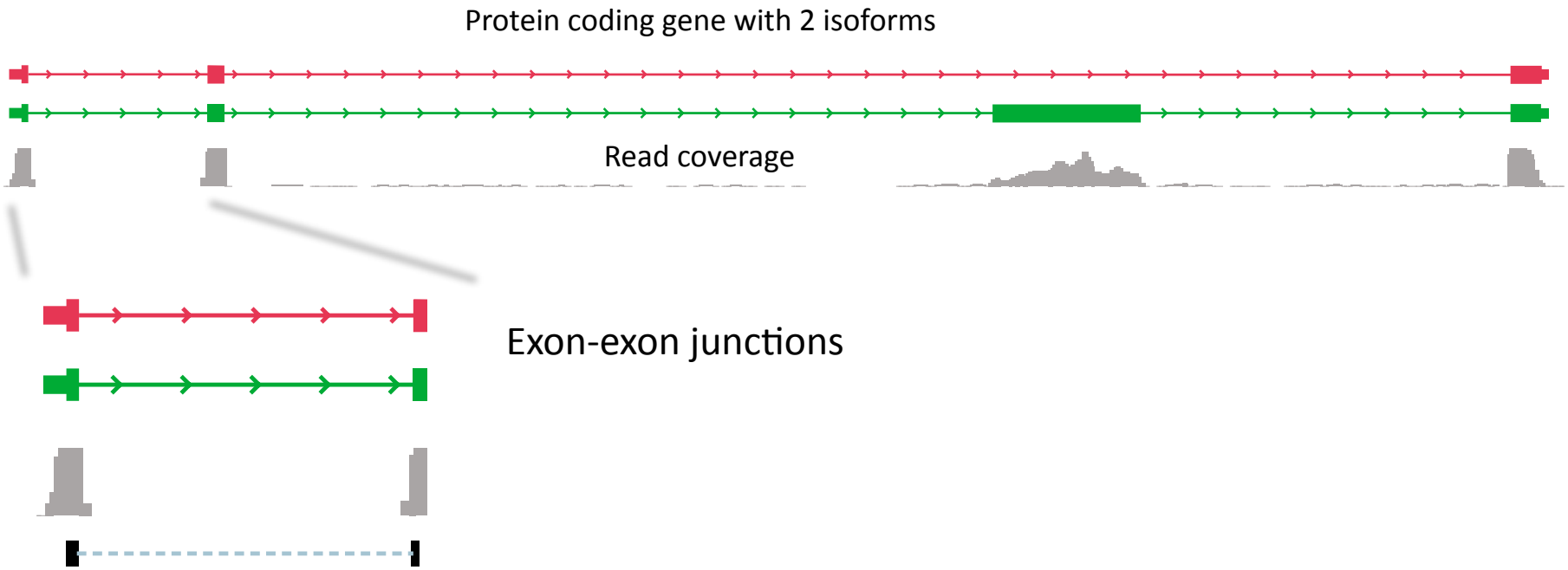


Exon junction spanning reads provide the connectivity information.

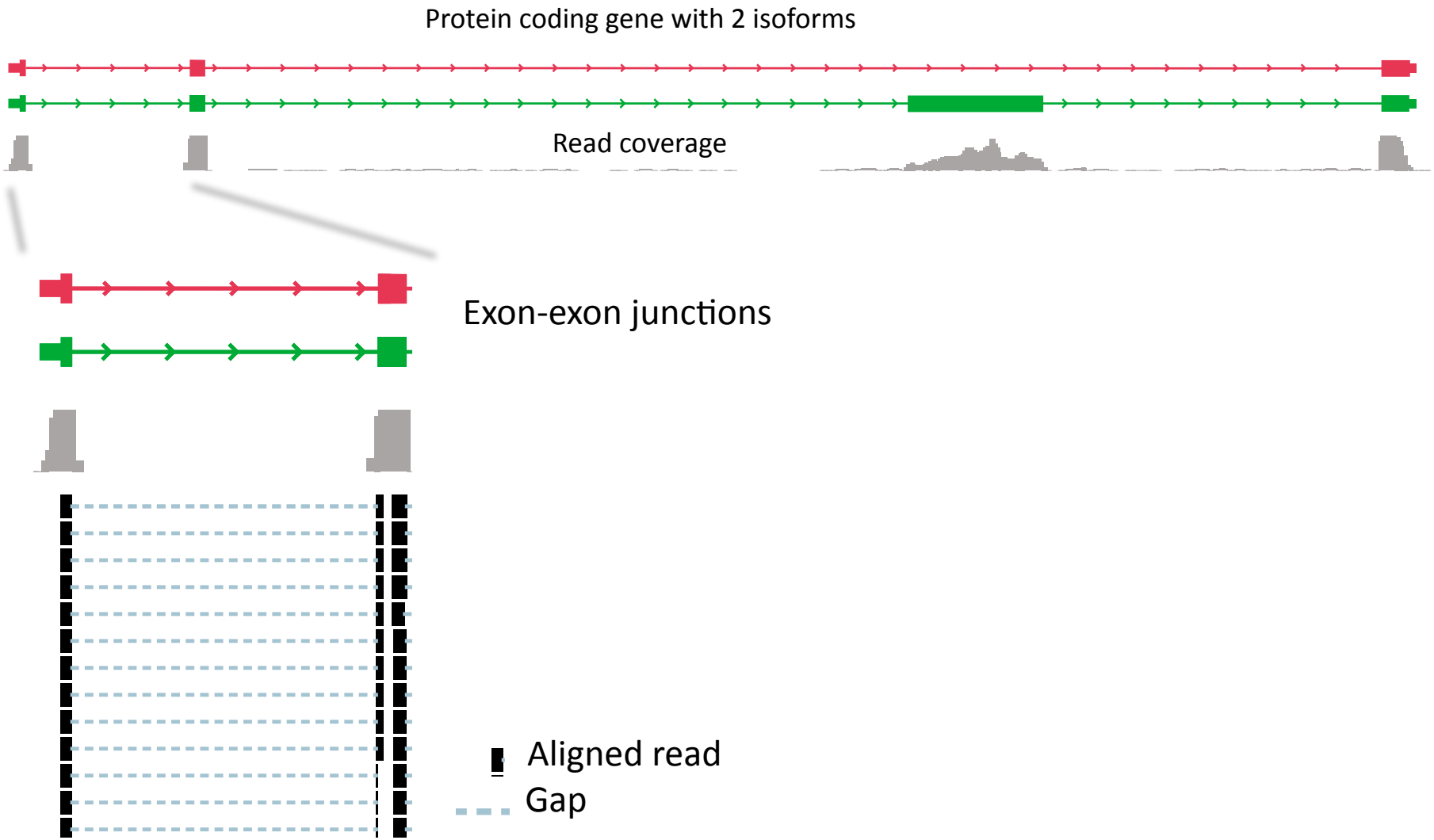
The power of spliced alignments



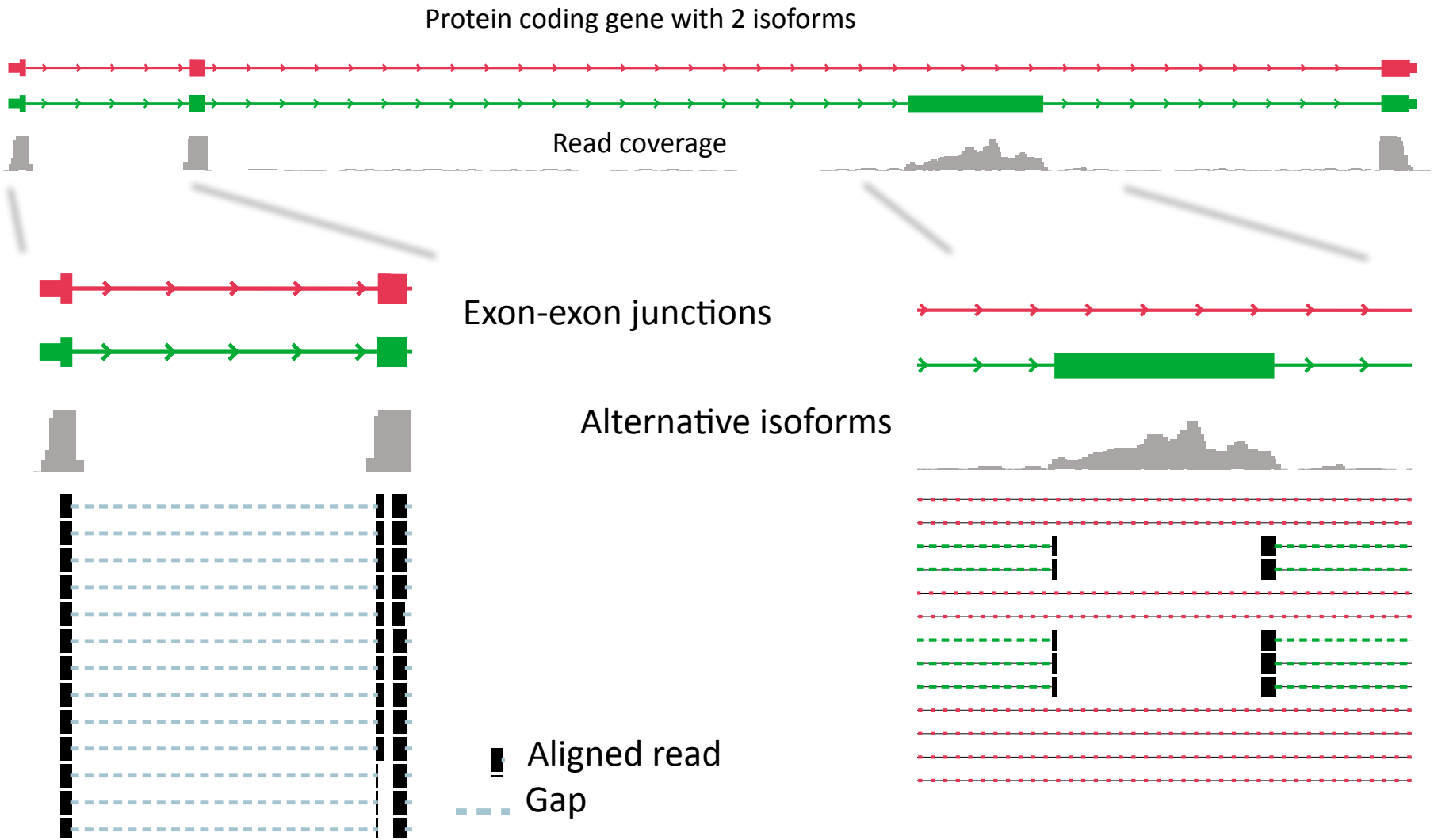
The power of spliced alignments



The power of spliced alignments

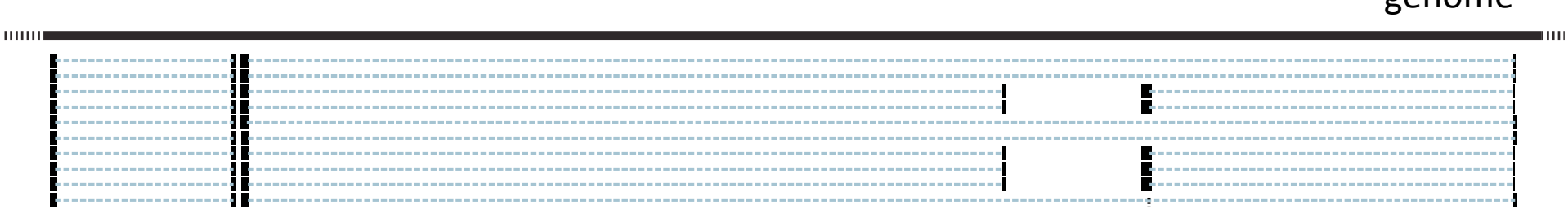


The power of spliced alignments



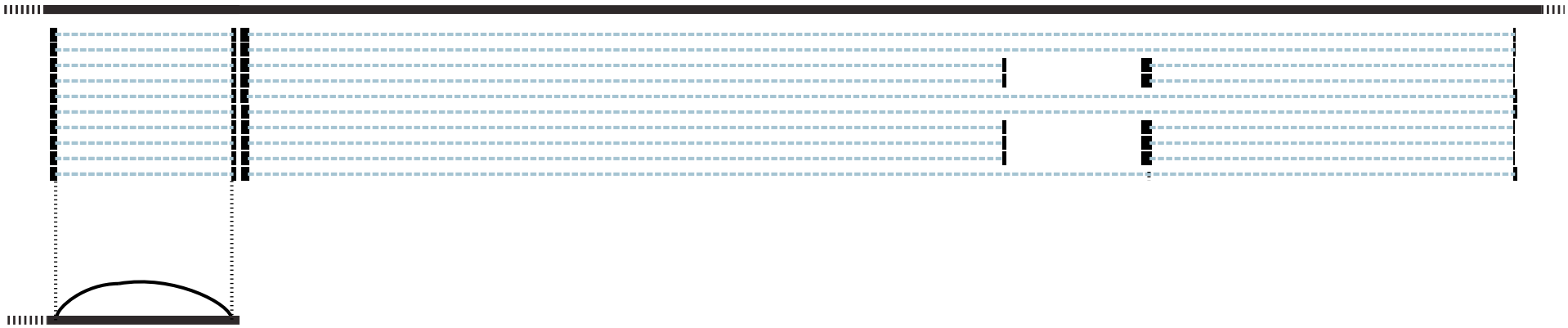
Statistical reconstruction of the transcriptome

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



Statistical reconstruction of the transcriptome

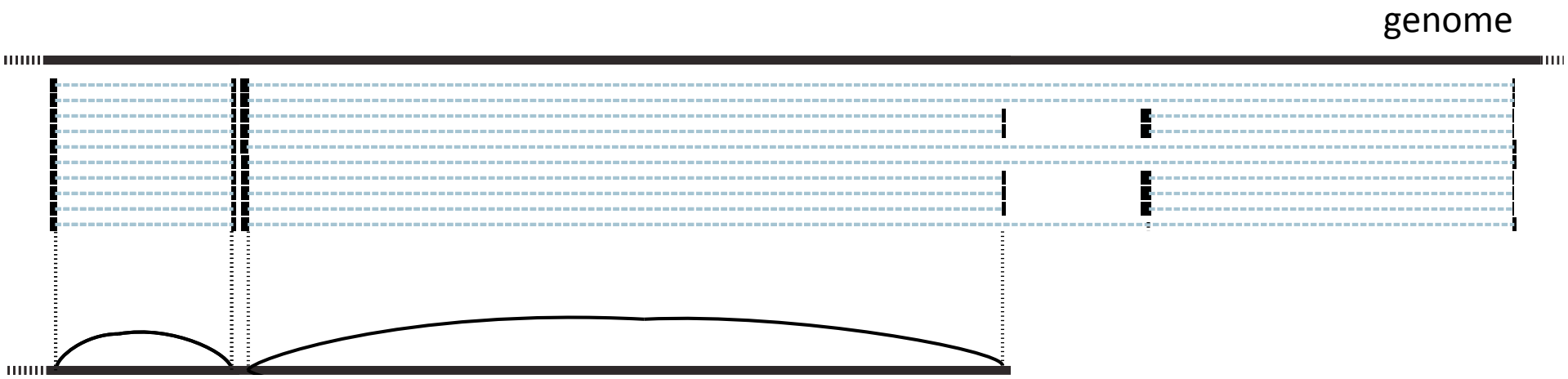
Step 1: Align Reads to the genome allowing gaps flanked by splice sites



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

Statistical reconstruction of the transcriptome

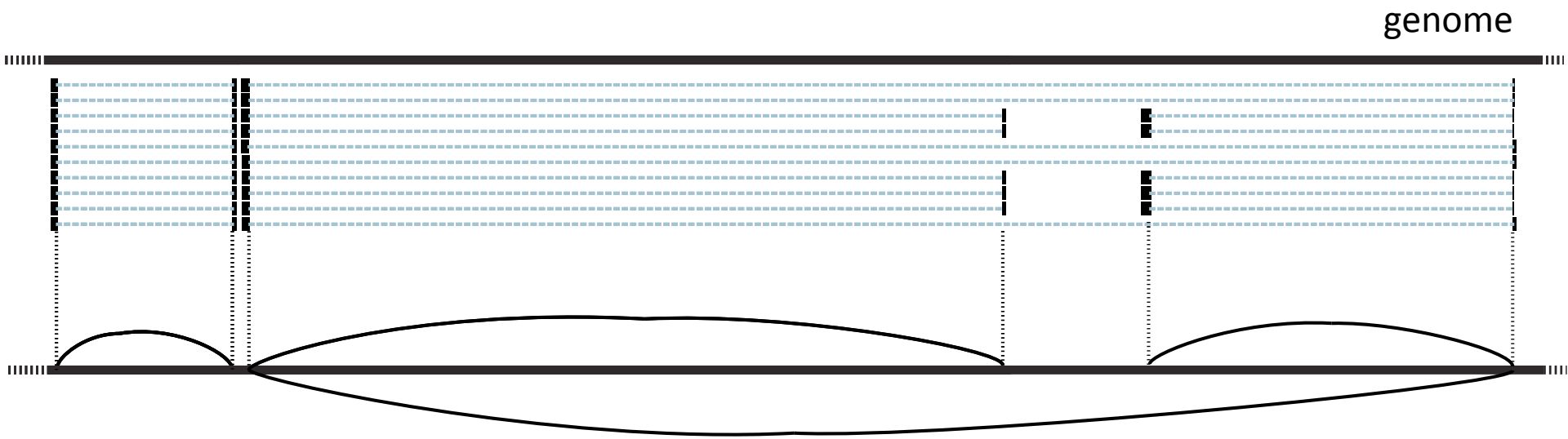
Step 1: Align Reads to the genome allowing gaps flanked by splice sites



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

Statistical reconstruction of the transcriptome

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

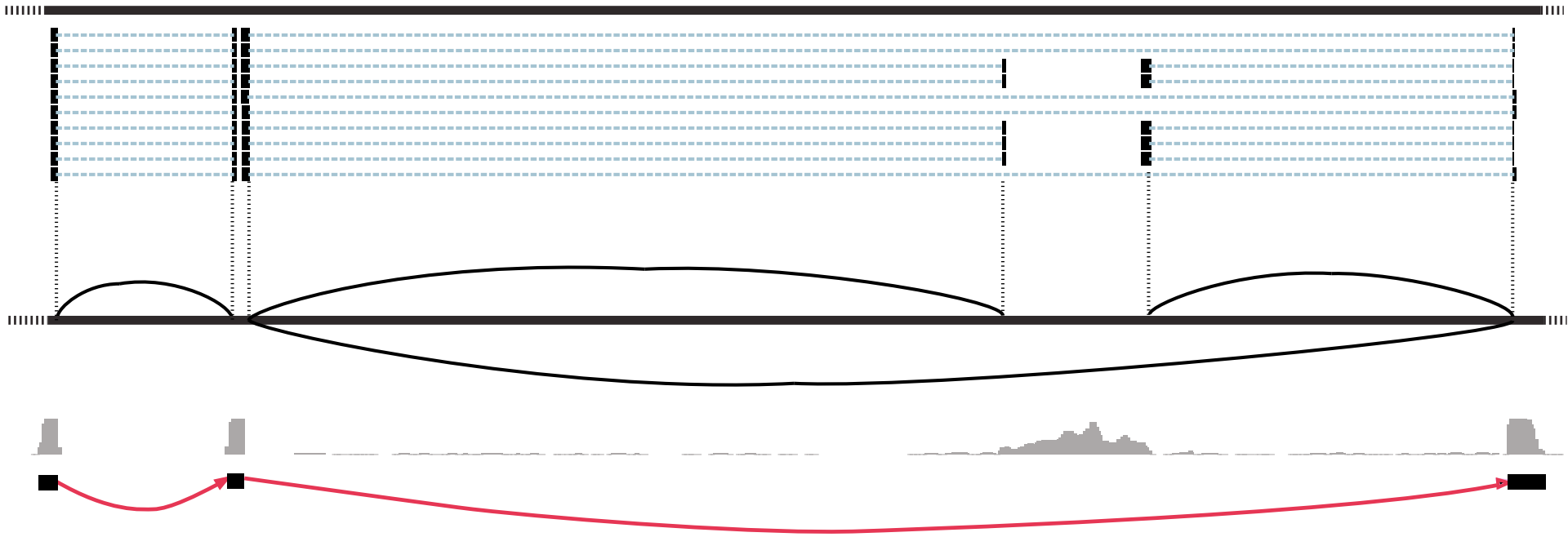


Step 2: Build an oriented connectivity graph using every spliced alignment and orienting edges using the flanking splicing motifs

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

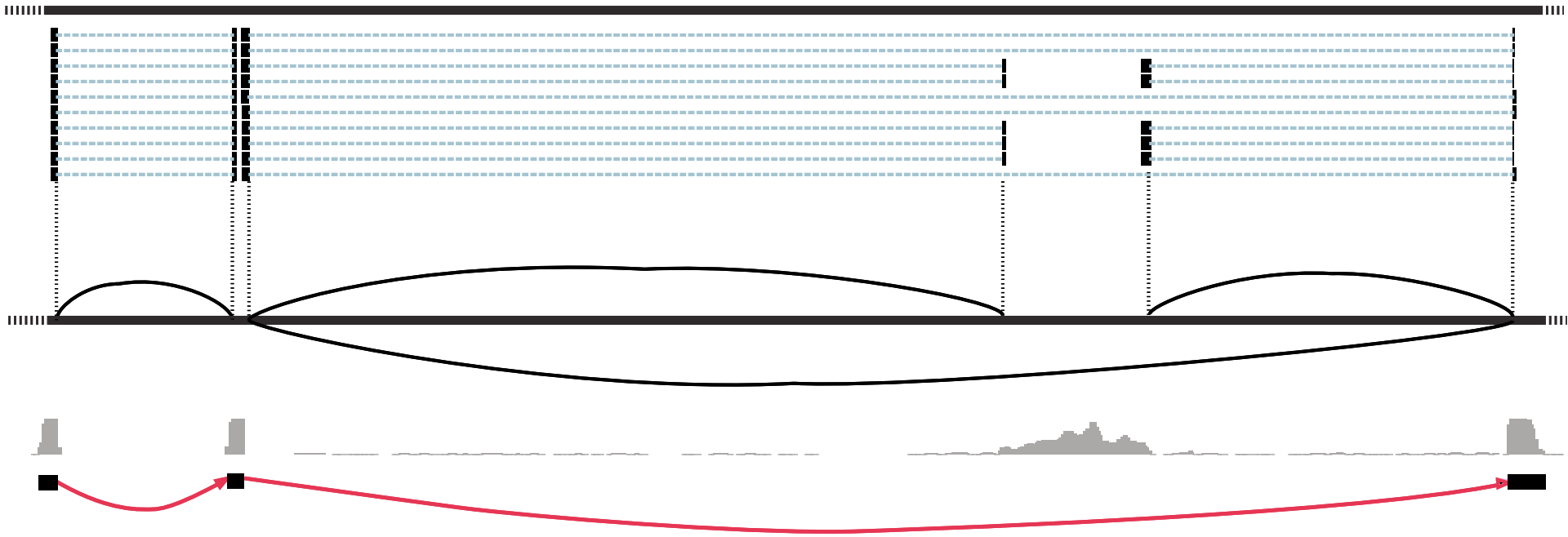
Statistical reconstruction of the transcriptome

Step 3: Identify “segments” across the graph



Statistical reconstruction of the transcriptome

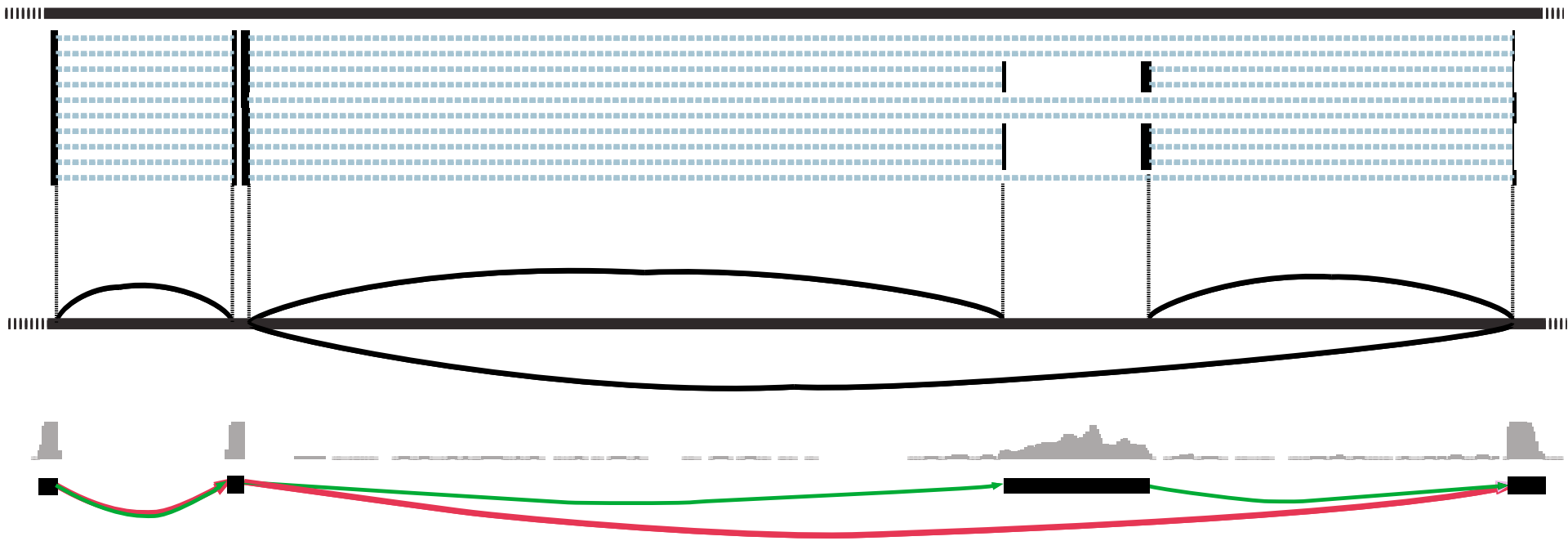
Step 3: Identify “segments” across the graph



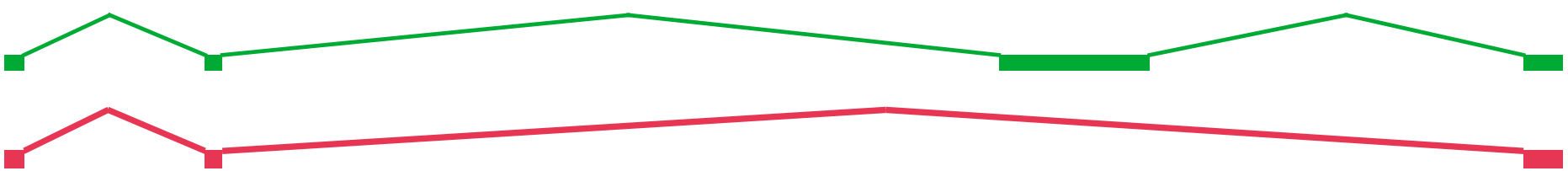
Step 4: Find significant segments

Statistical reconstruction of the transcriptome

Step 3: Identify “segments” across the graph



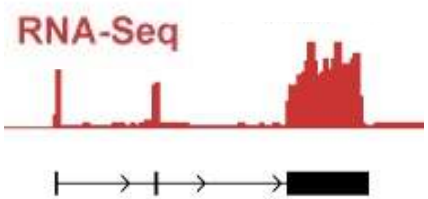
Step 4: Find significant segments



Can we identify enriched regions across different data types?

Short modification ✓

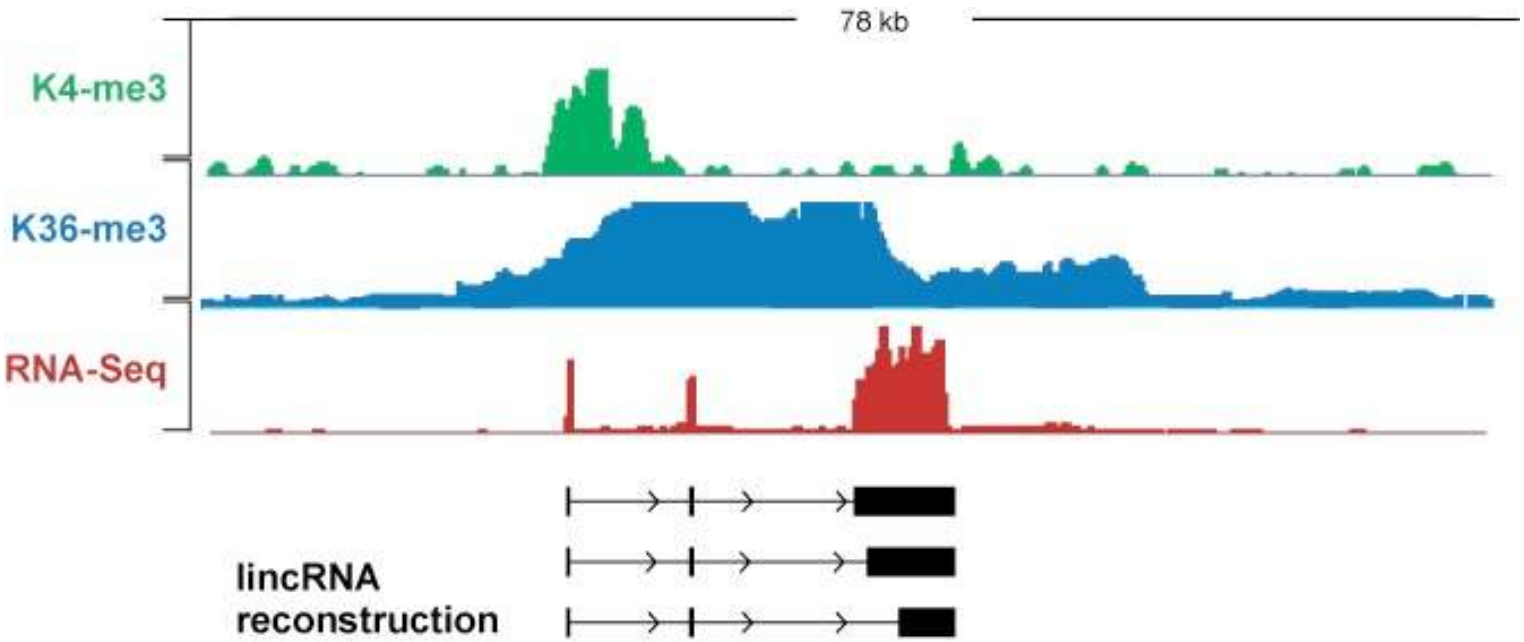
Long modification ✓



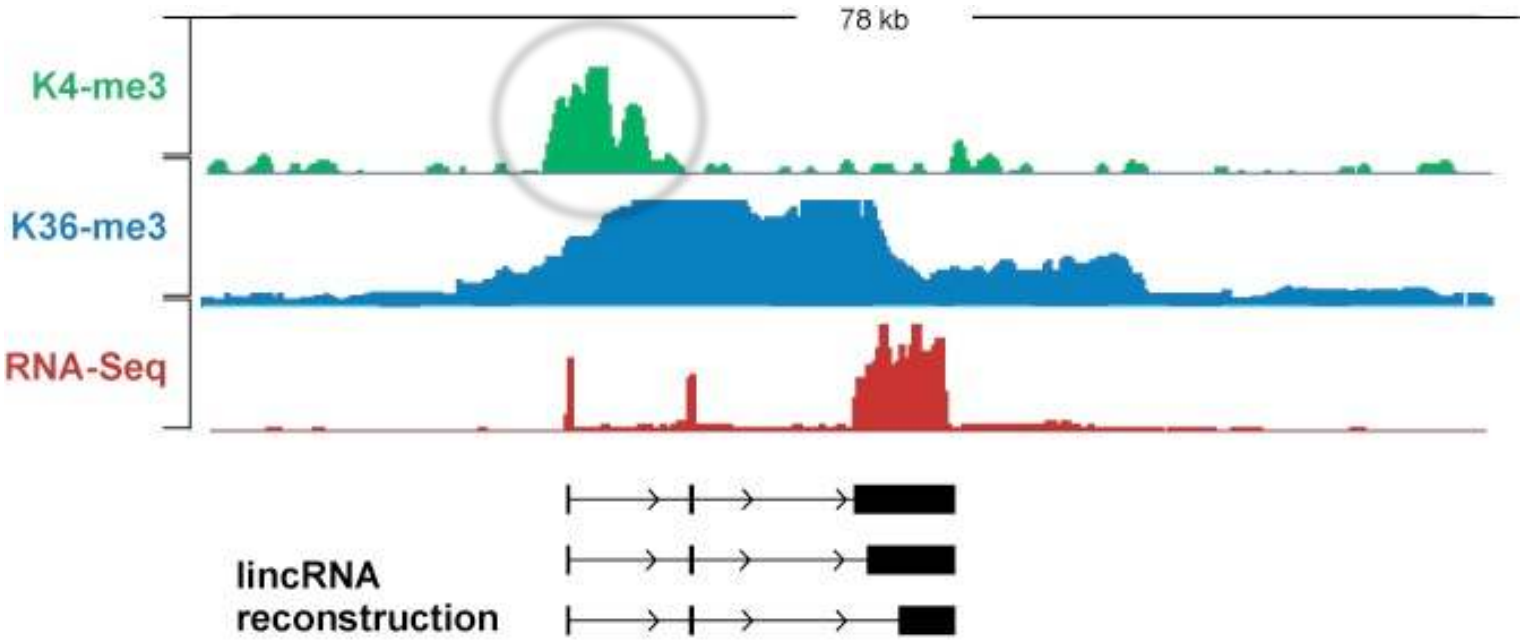
Discontinuous data ✓

Are we really sure reconstructions are complete?

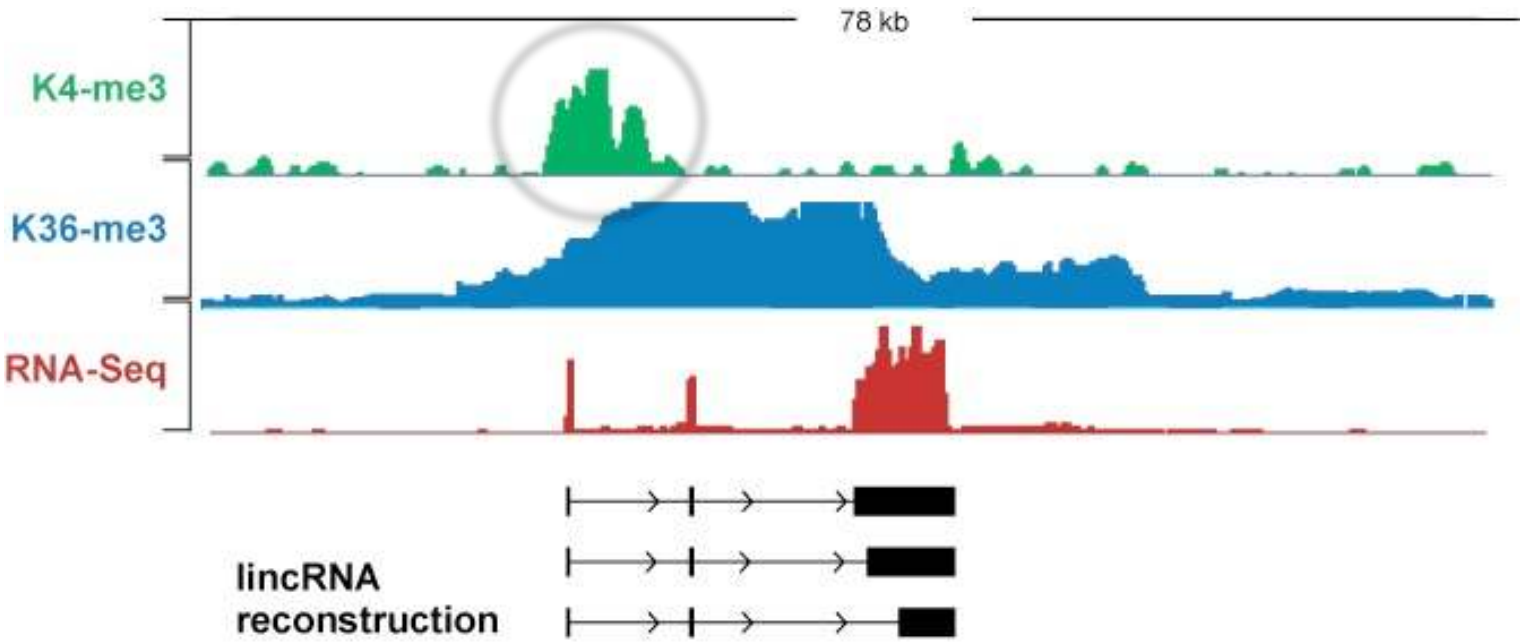
RNA-Seq data is incomplete for comprehensive annotation



RNA-Seq data is incomplete for comprehensive annotation



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 transcriptome of mouse cell types

Mouse Cell Types

Sequence

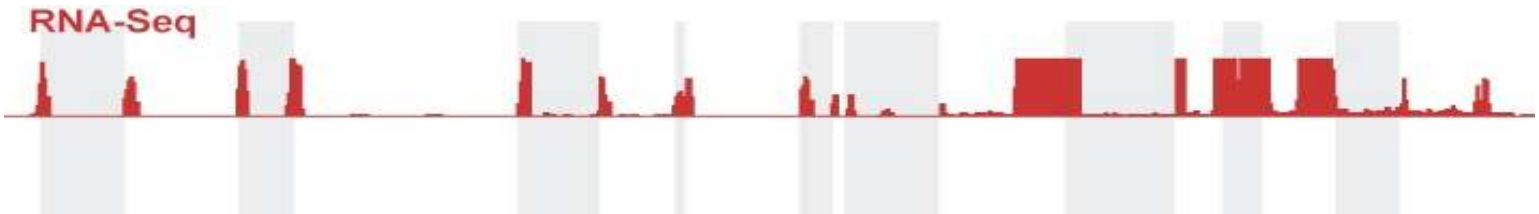
Reconstruct

Reconstructing the transcriptome of mouse cell types

Mouse Cell Types

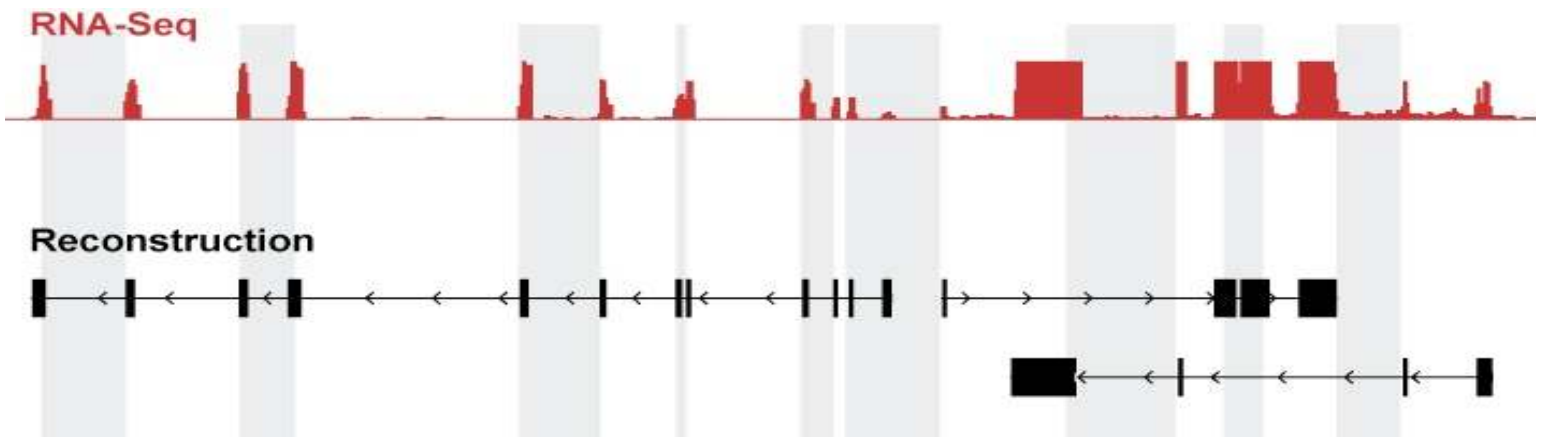
Sequence

Reconstruct



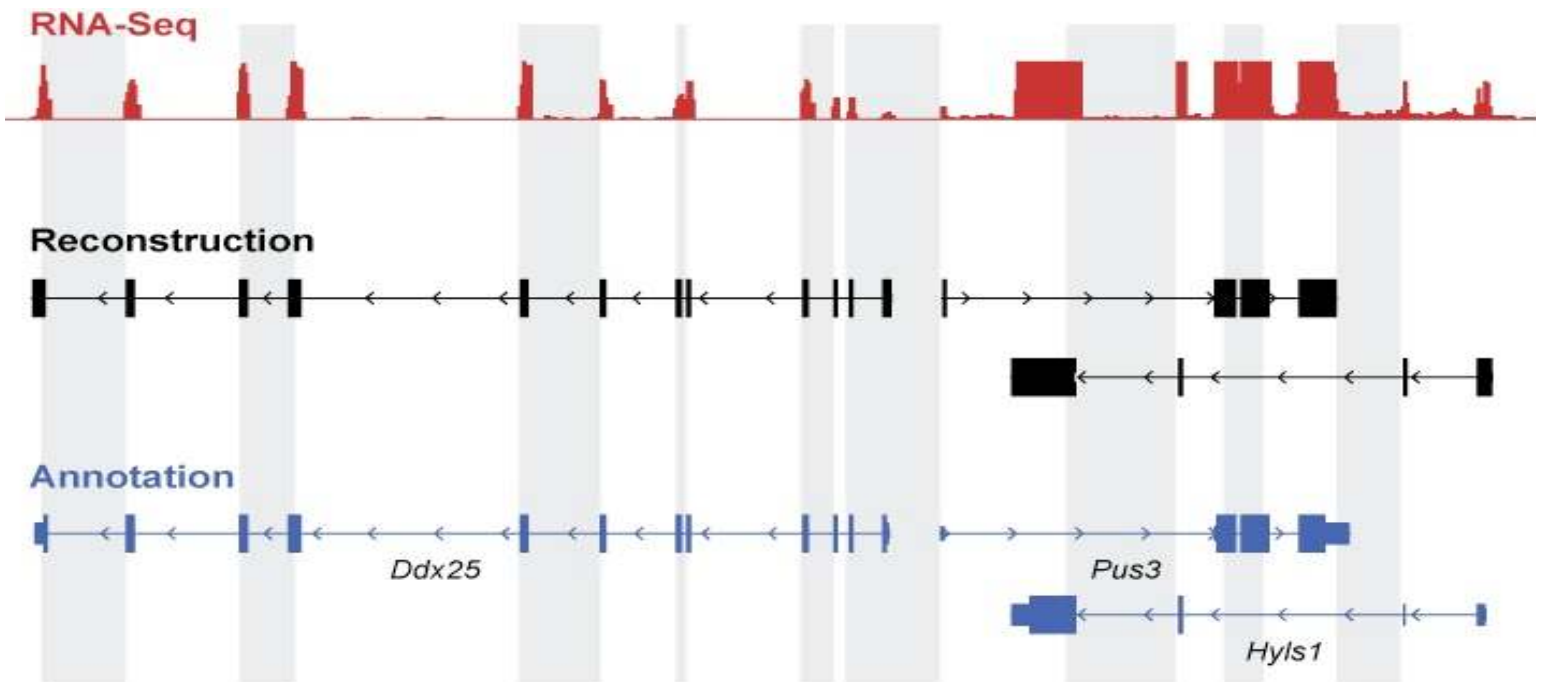
Reconstructing the transcriptome of mouse cell types

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



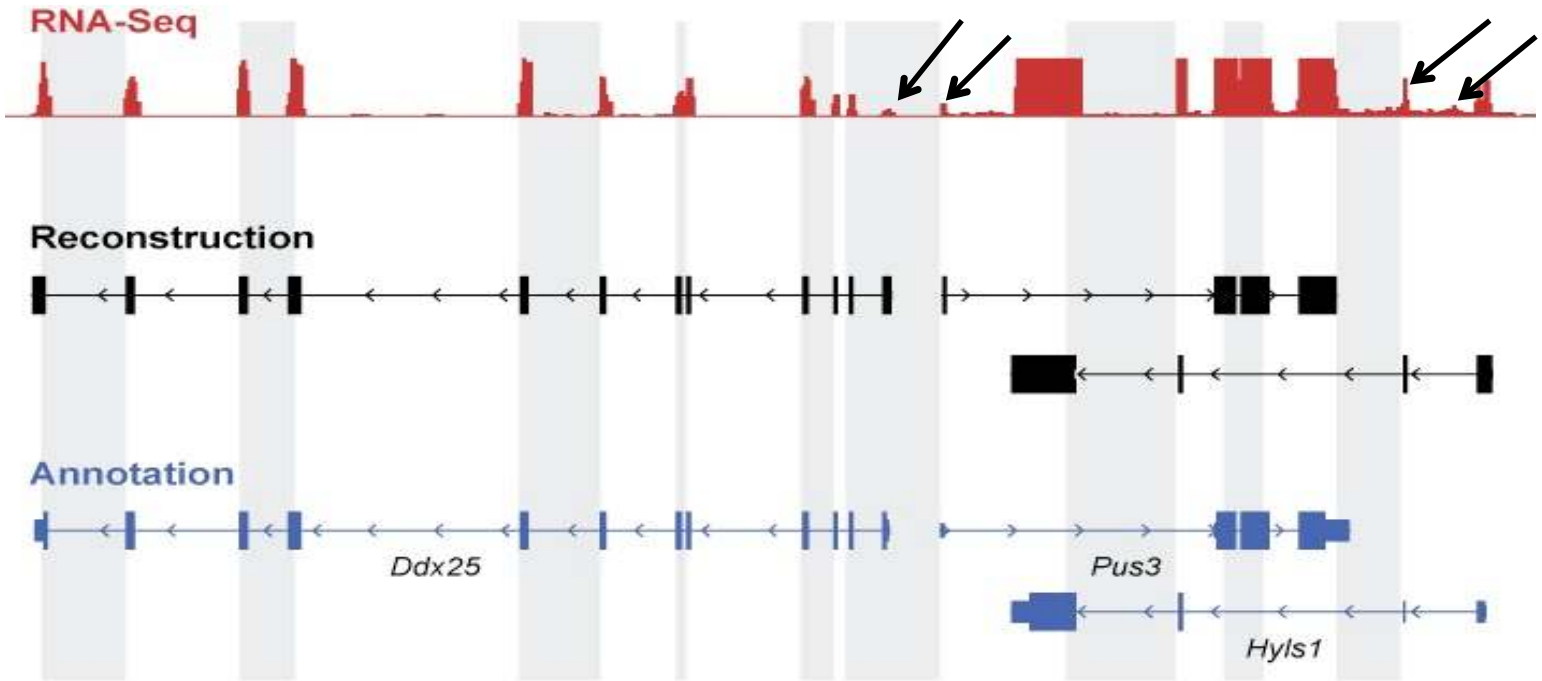
Reconstructing the transcriptome of mouse cell types

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

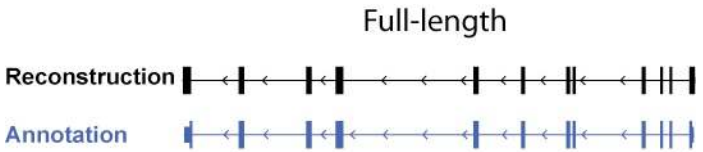


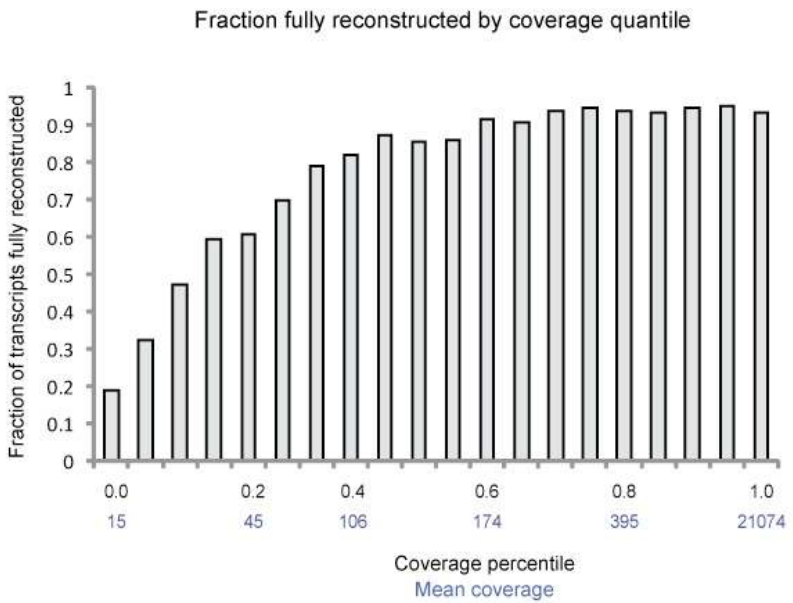
Reconstructing the transcriptome of mouse cell types

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

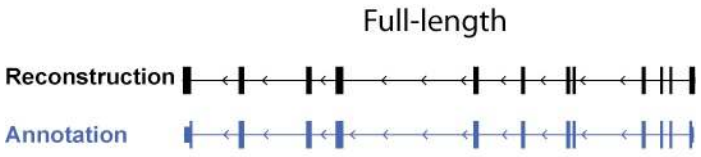


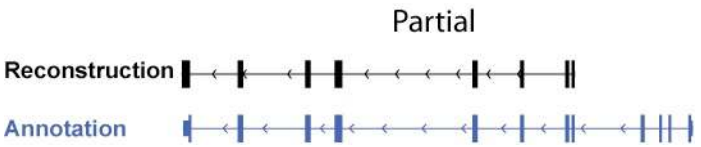
Sensitivity across expression levels

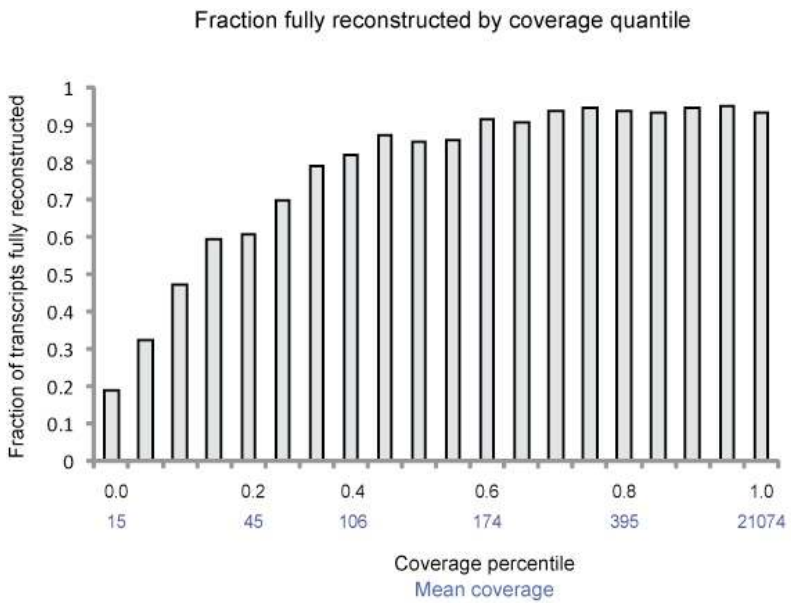




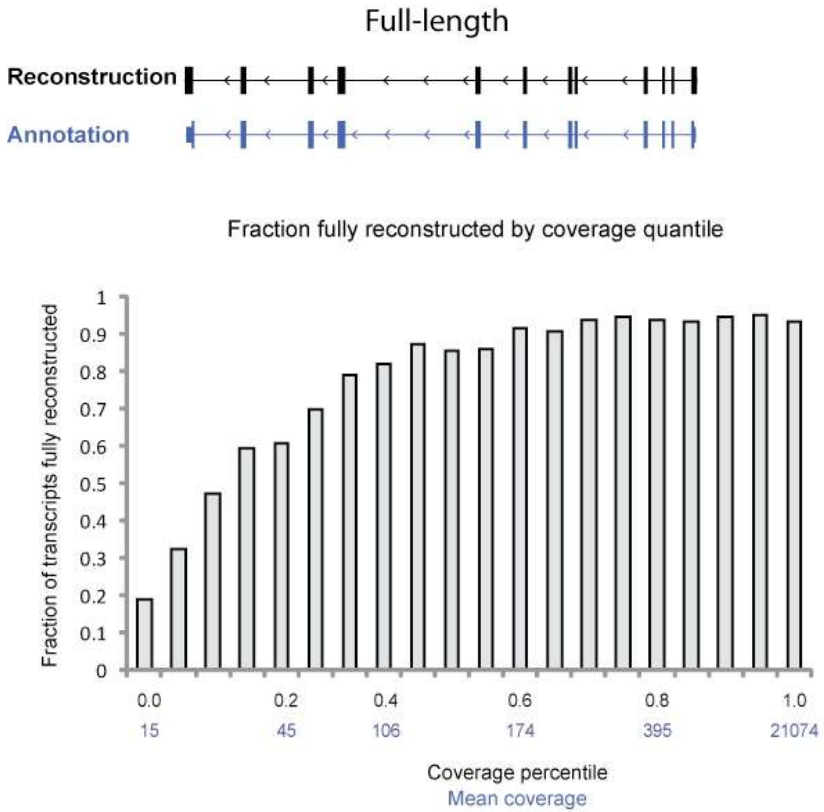
Sensitivity across expression levels

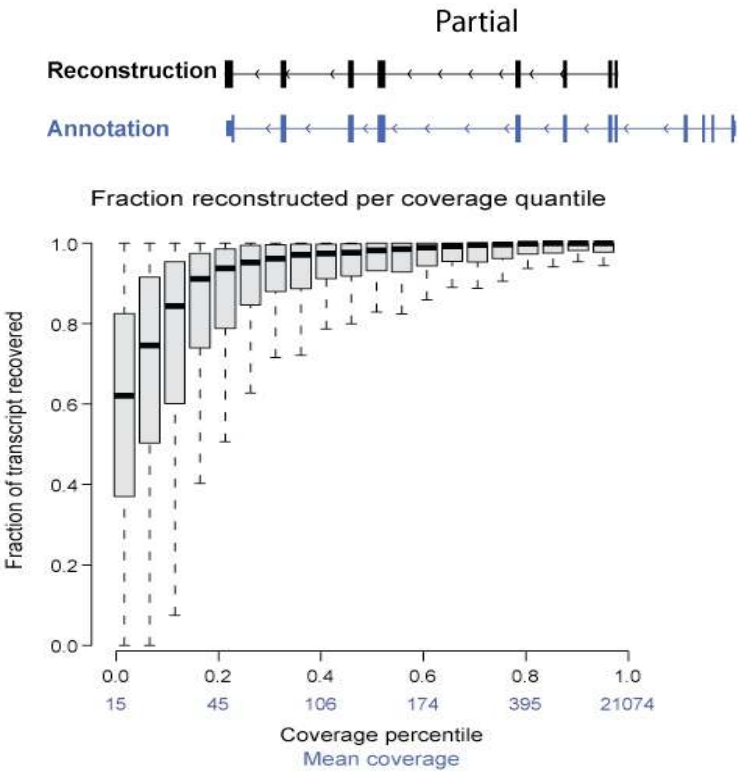




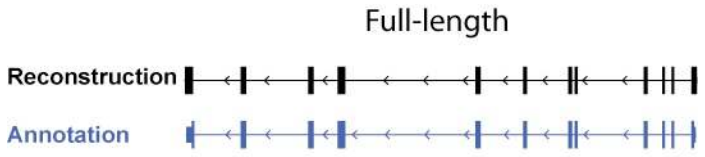


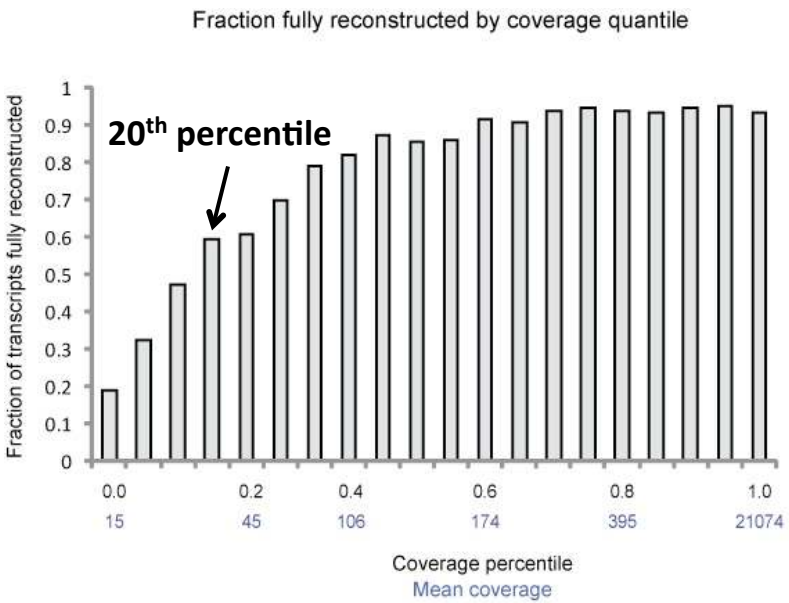
Sensitivity across expression levels

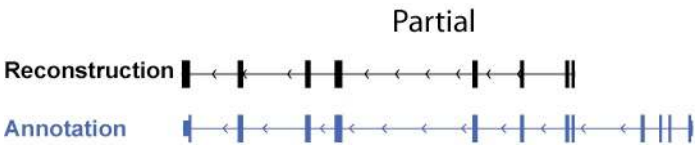


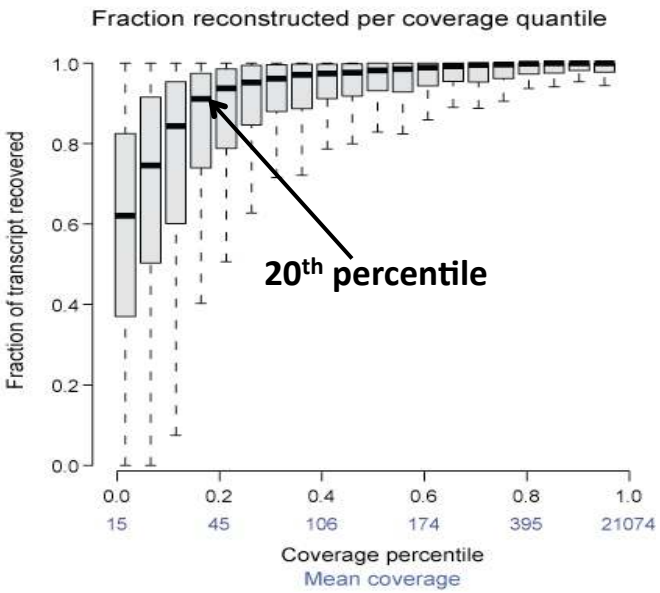


Sensitivity across expression levels



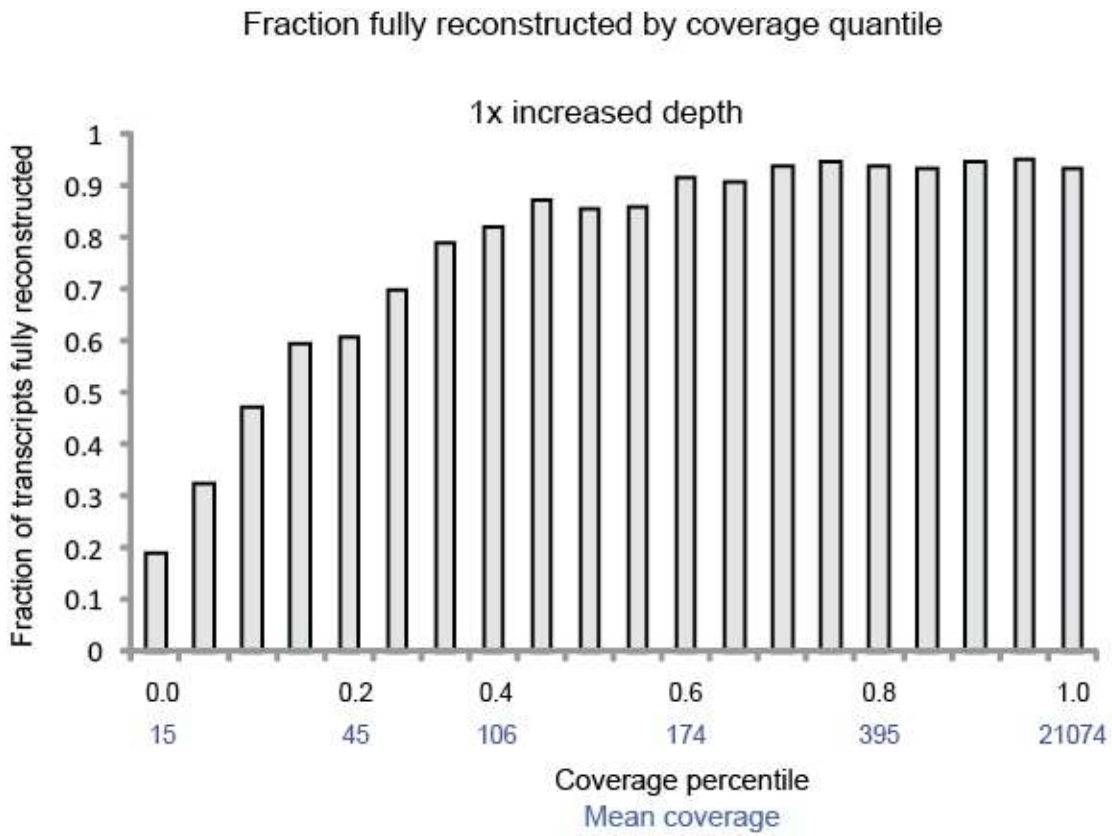




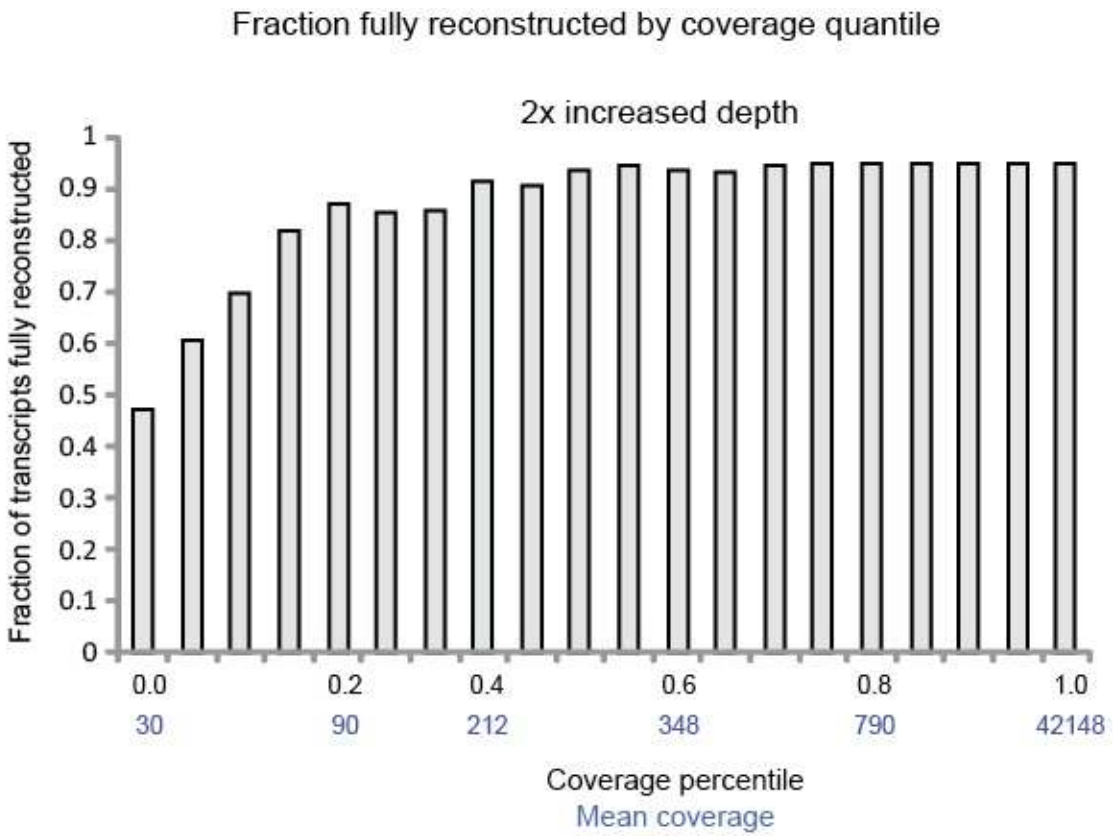


Even at low expression (20th percentile), we have:
average coverage of transcript is ~95% and 60% have full coverage

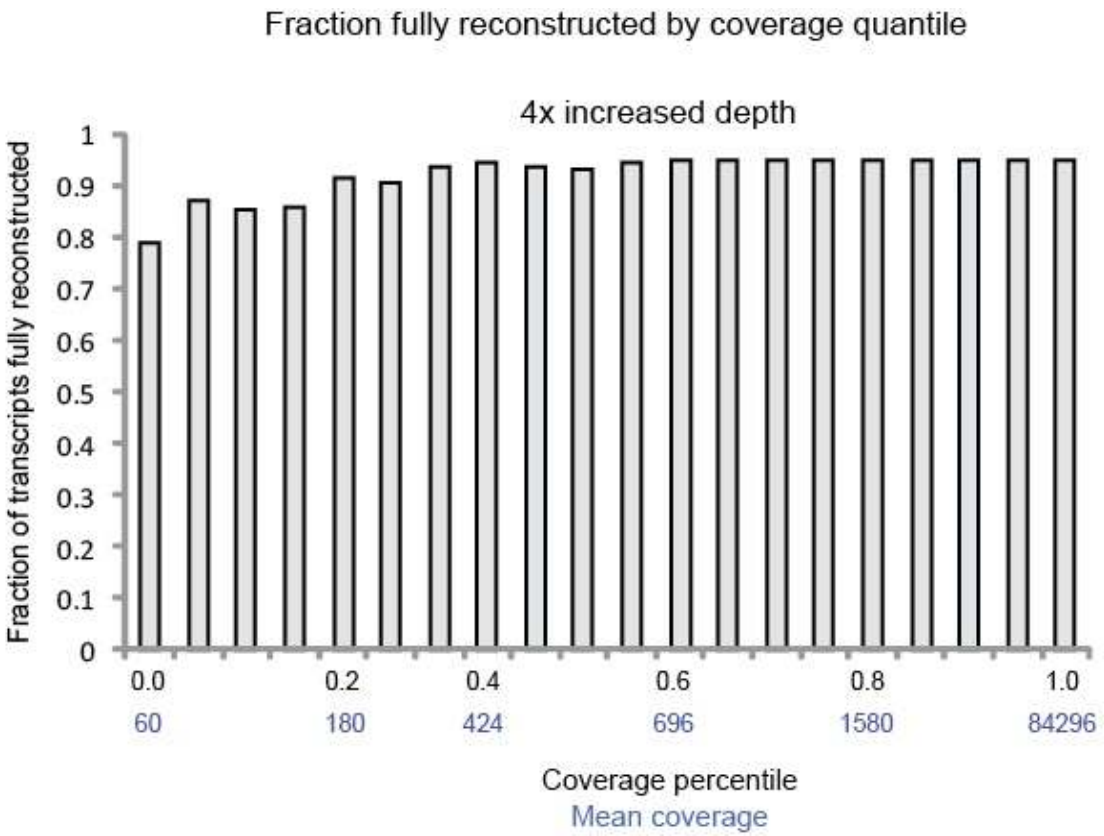
Sensitivity at low expression levels improves with depth



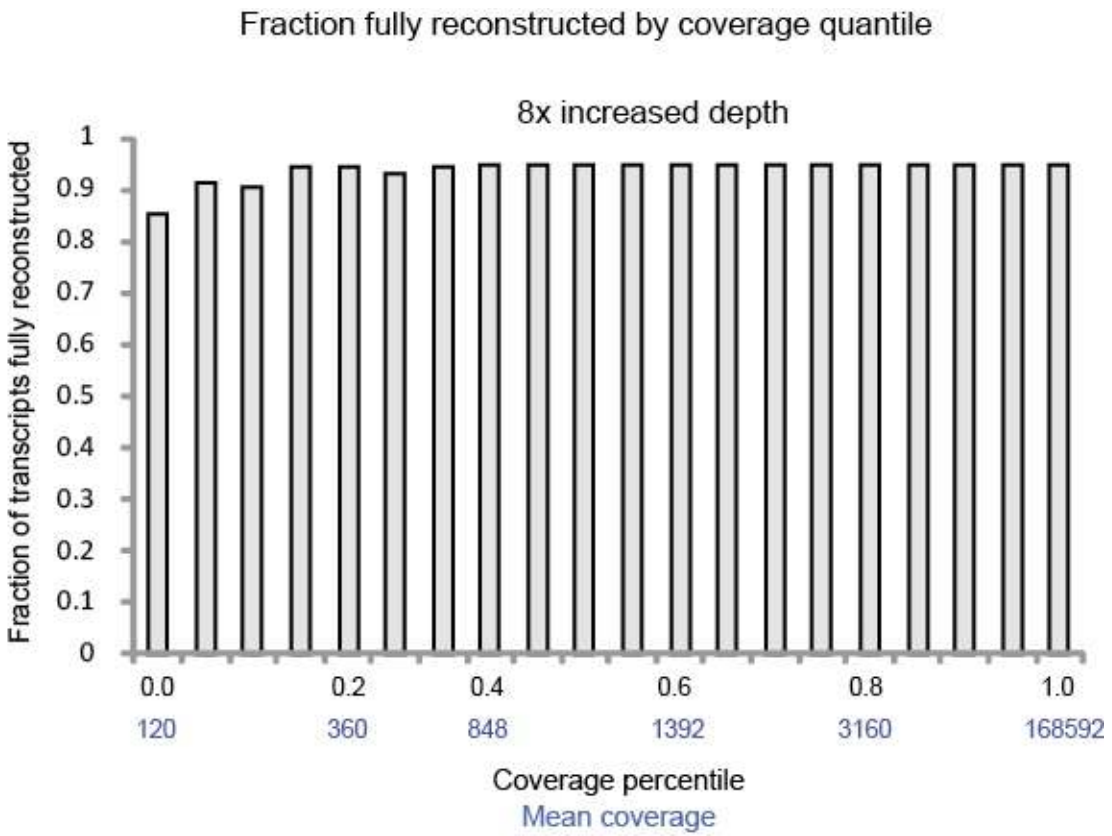
Sensitivity at low expression levels improves with depth



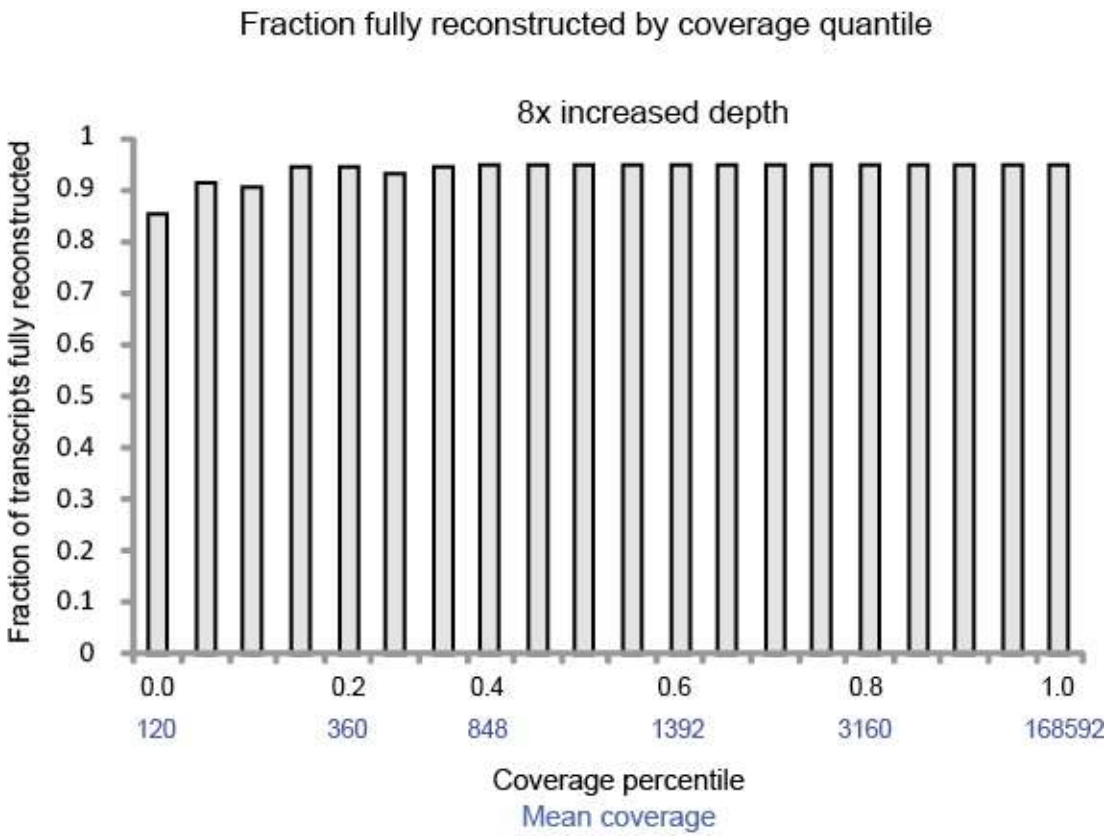
Sensitivity at low expression levels improves with depth



Sensitivity at low expression levels improves with depth



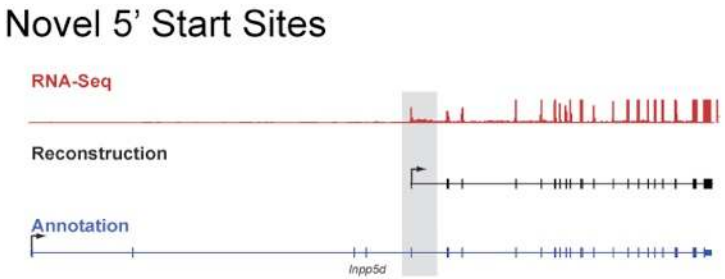
Sensitivity at low expression levels improves with depth



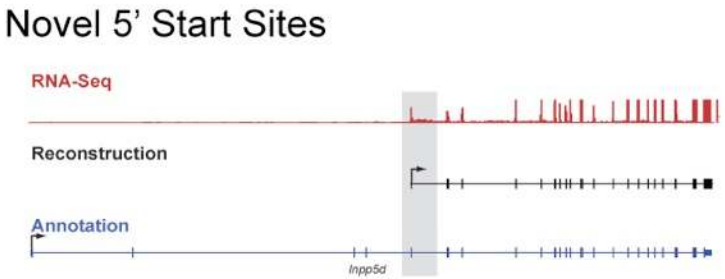
As coverage increases we are able to fully reconstruct a larger percentage of known protein-coding genes

Novel variation in protein-coding genes

ES cells



Novel variation in protein-coding genes



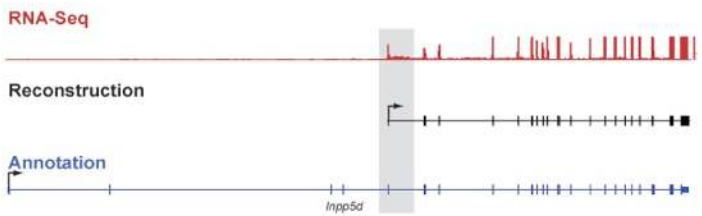
ES cells

1,804

Novel variation in protein-coding genes

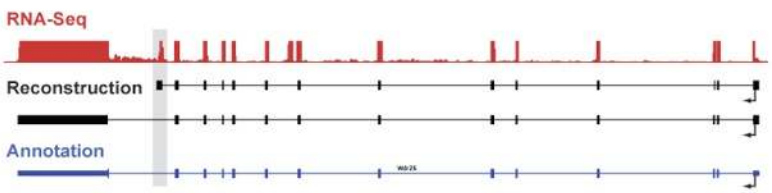
ES cells

Novel 5' Start Sites



1,804

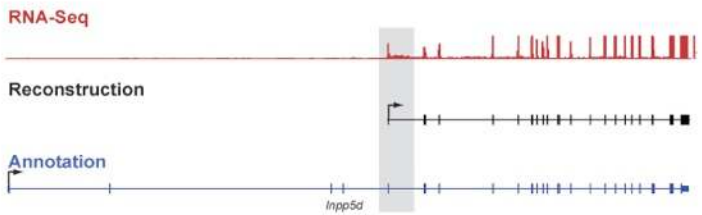
Novel 3' End



Novel variation in protein-coding genes

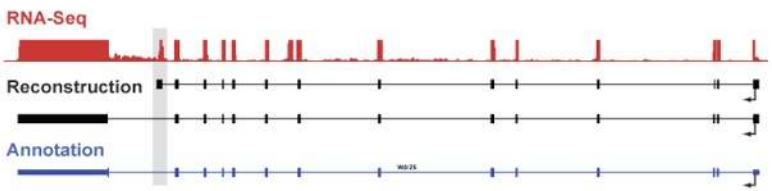
ES cells

Novel 5' Start Sites



1,804

Novel 3' End

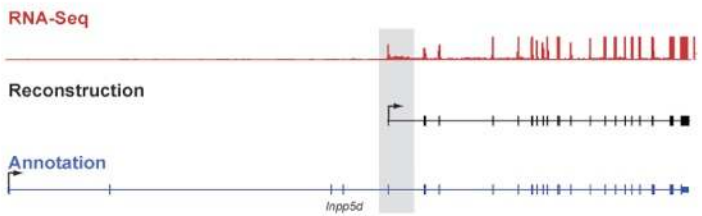


1,310

Novel variation in protein-coding genes

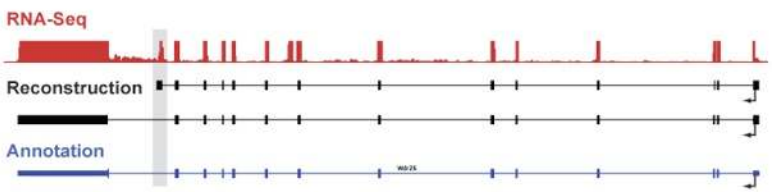
ES cells

Novel 5' Start Sites



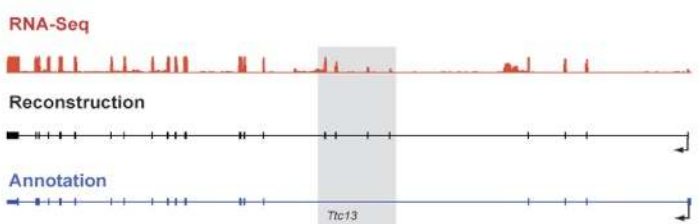
1,804

Novel 3' End



1,310

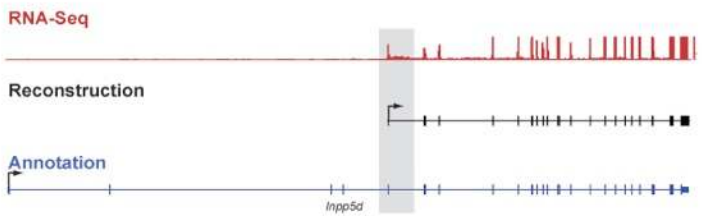
Novel Coding Exons



Novel variation in protein-coding genes

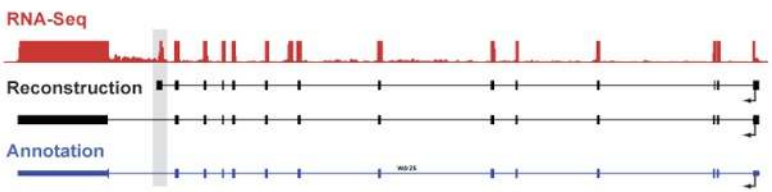
ES cells

Novel 5' Start Sites



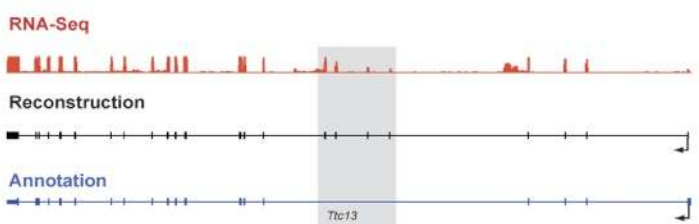
1,804

Novel 3' End



1,310

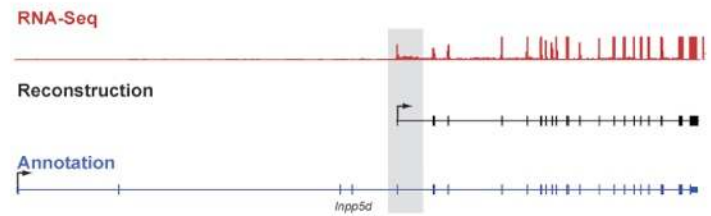
Novel Coding Exons

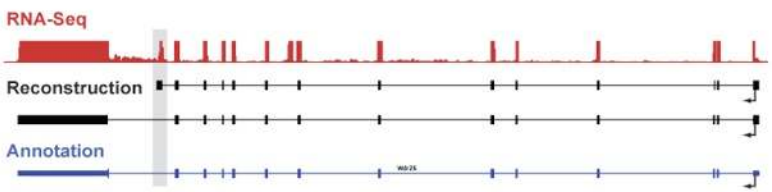


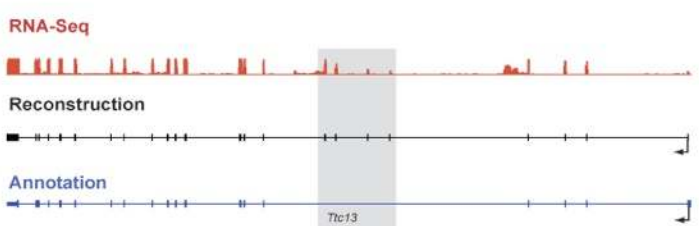
588

Novel variation in protein-coding genes

	ES cells	3 cell types
Novel 5' Start Sites	1,804	3,137
Novel 3' End	1,310	2,477
Novel Coding Exons	588	903

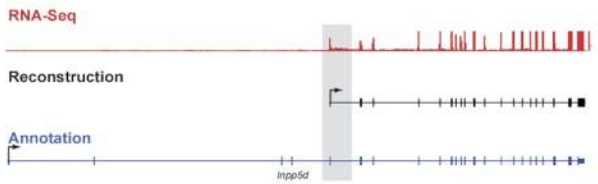




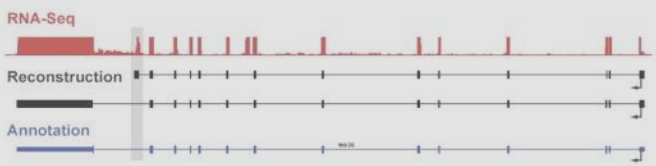


Novel variation in protein-coding genes

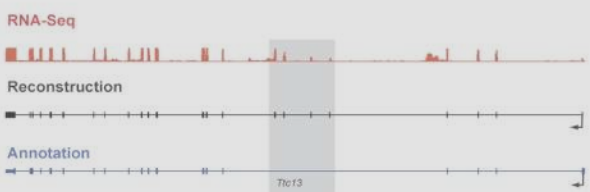
Novel 5' Start Sites

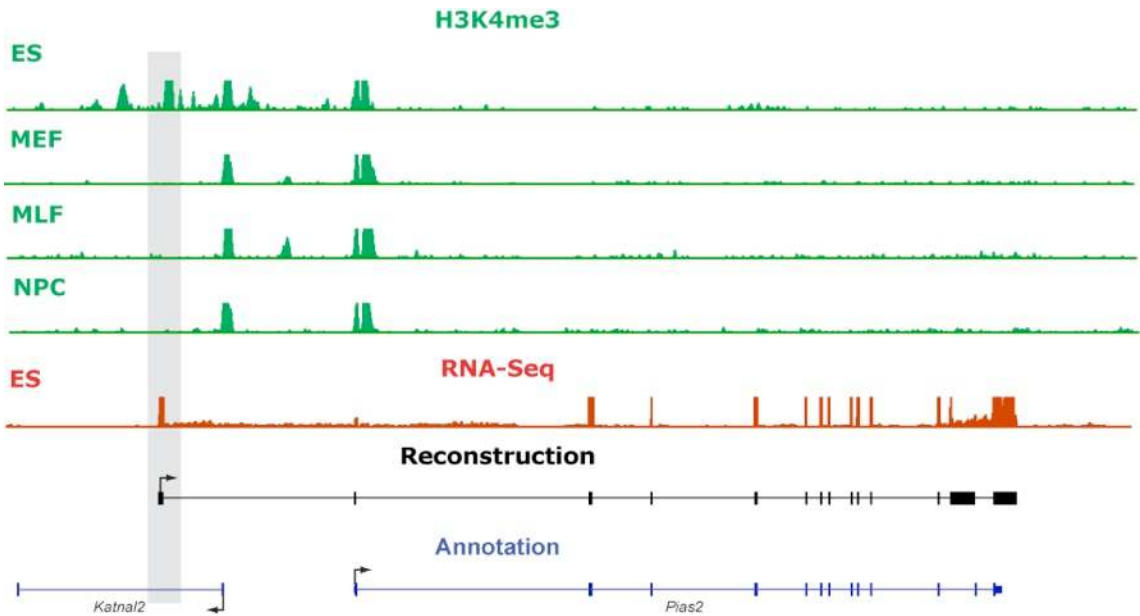


Novel 3' End



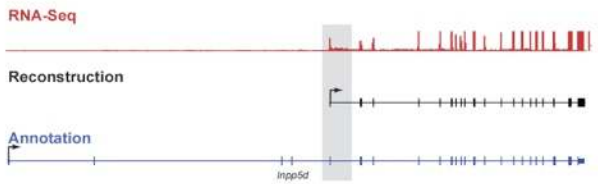
Novel Coding Exons



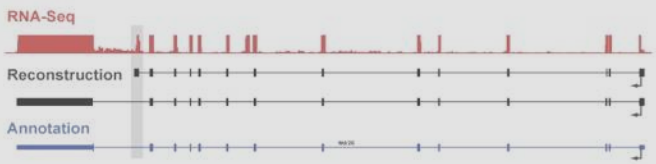


Novel variation in protein-coding genes

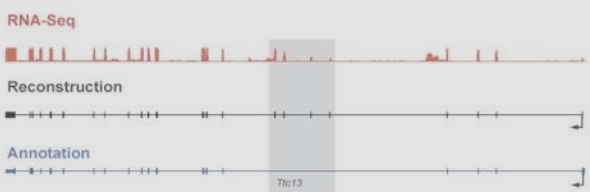
Novel 5' Start Sites

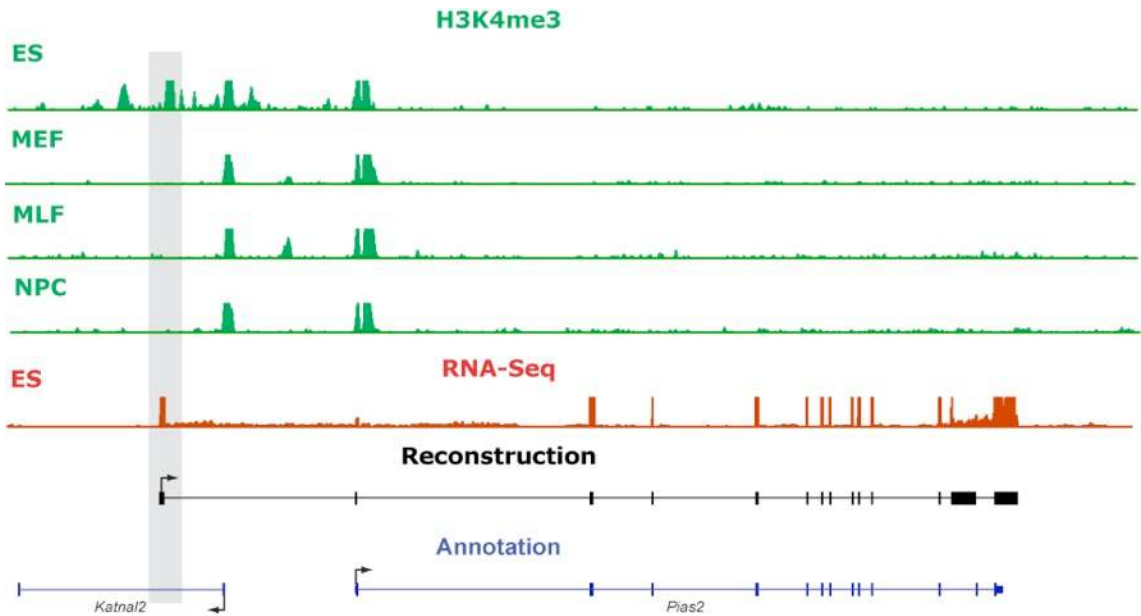


Novel 3' End



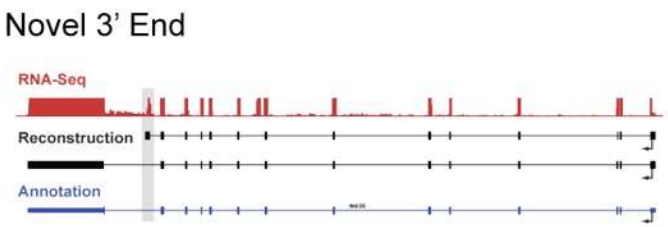
Novel Coding Exons

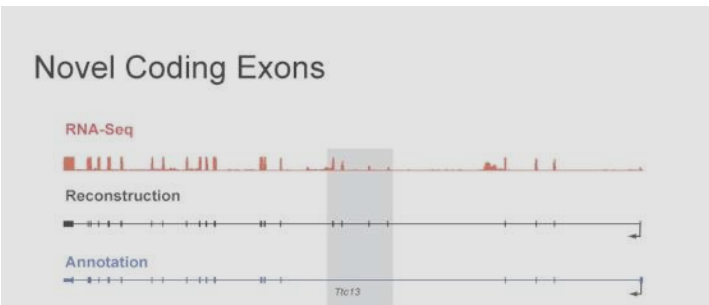




~85% overlap K4me3

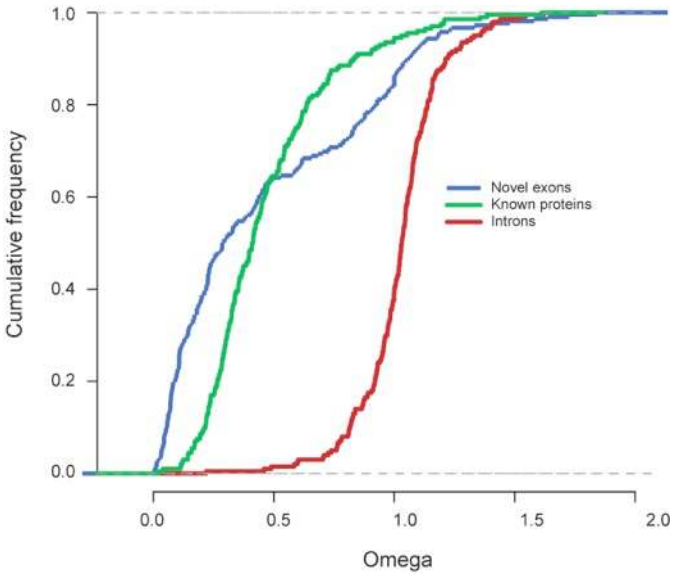
Novel variation in protein-coding genes

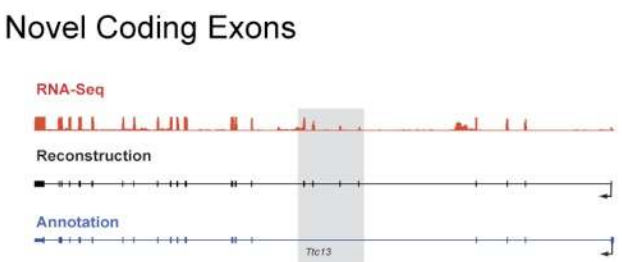




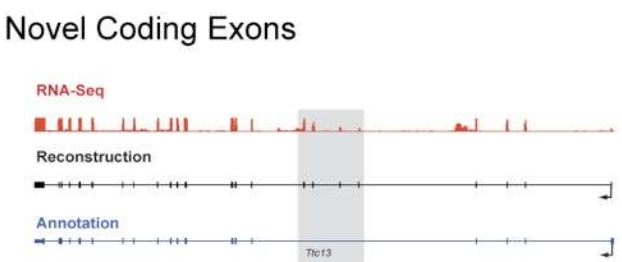
~50% contain polyA motif Compared to ~6% for random

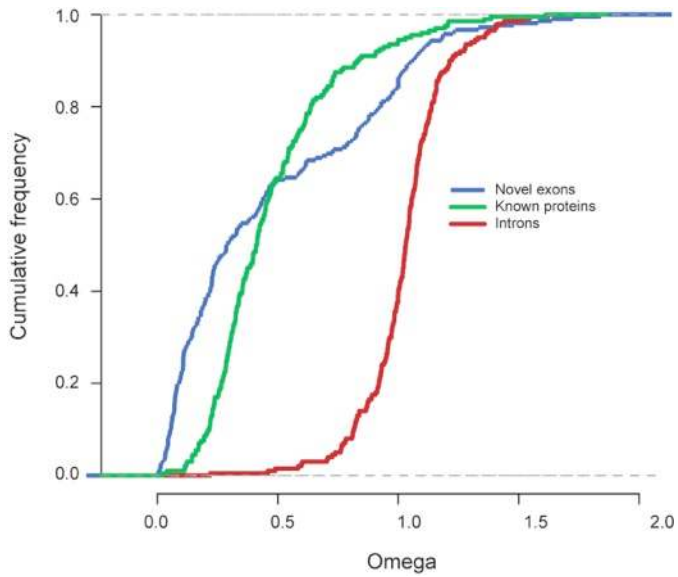
Novel variation in protein-coding genes





Novel variation in protein-coding genes





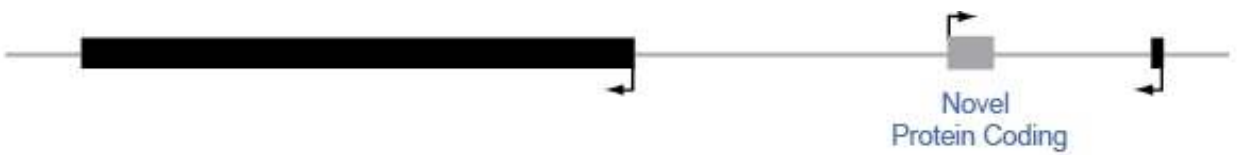
~80% retain ORF

What about novel genes?

Class 1: Overlapping ncRNA

Class 2: Large Intergenic ncRNA (lincRNA)

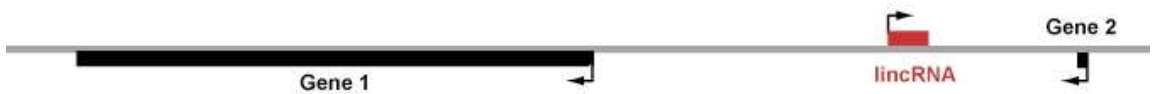
Class 3: Novel protein-coding genes



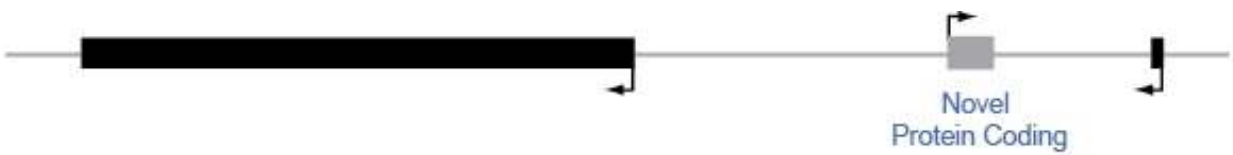
What about novel genes?

Class 1: Overlapping ncRNA

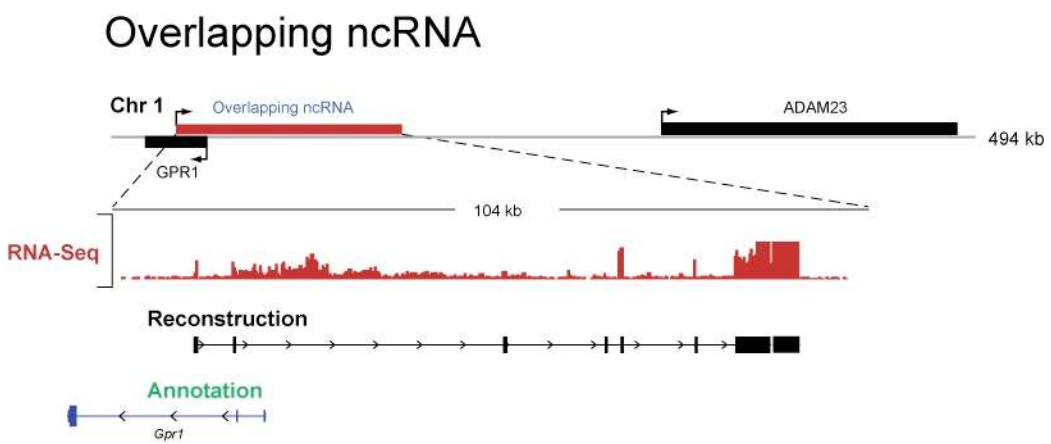
Class 2: Large Intergenic ncRNA (lincRNA)



Class 3: Novel protein-coding genes



Class 1: Overlapping ncRNA

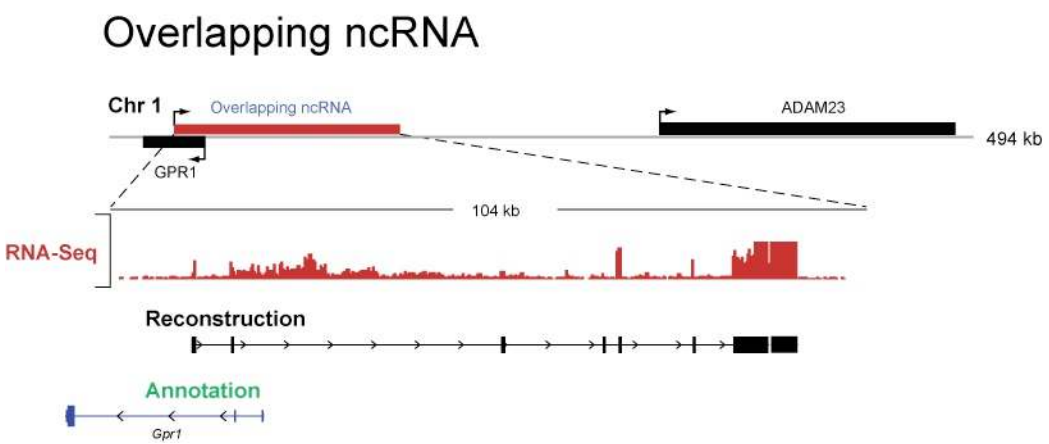


ES cells

3 cell types

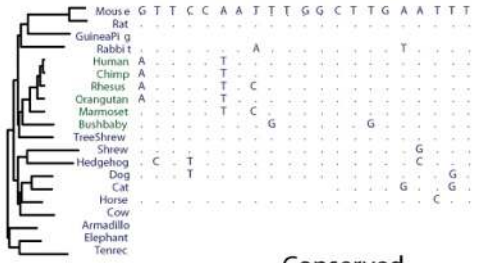
201

Class 1: Overlapping ncRNA

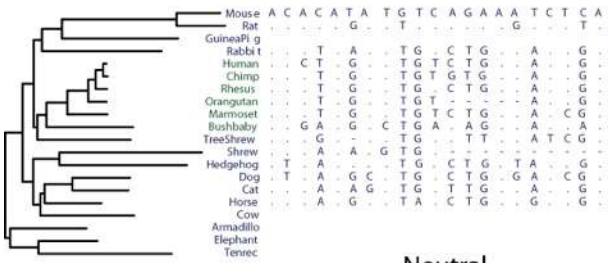


ES cells	3 cell types
201	**446**

Overlapping ncRNAs: Assessing their evolutionary conservation



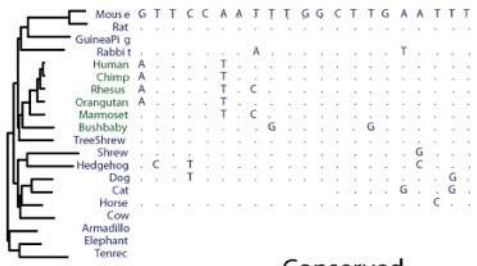
Conserved



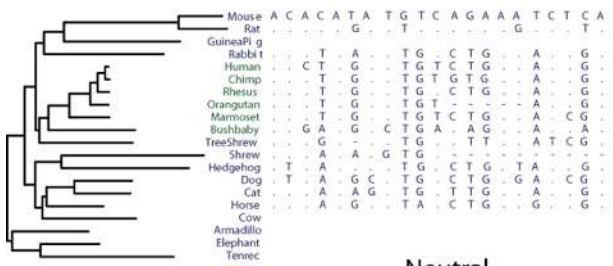
Neutral

SiPhy-Garber et al. 2009

Overlapping ncRNAs: Assessing their evolutionary conservation

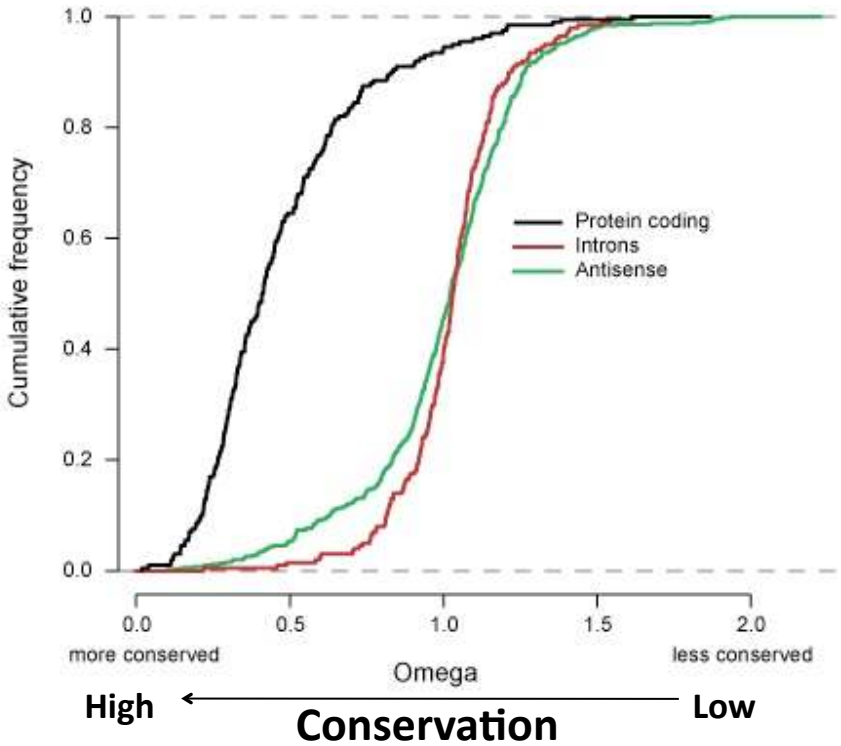


Conserved



Neutral

SiPhy-Garber et al. 2009



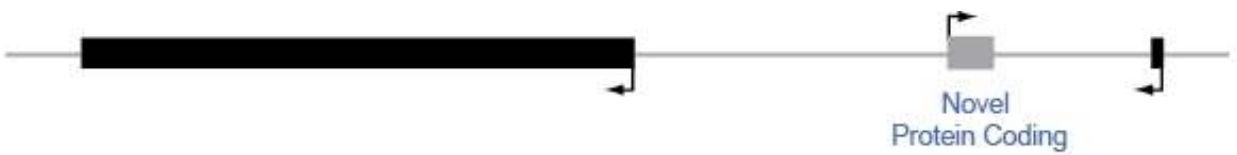
Overlapping ncRNAs show little evolutionary conservation

What about novel genes?

Class 1: Overlapping ncRNA

Class 2: Large Intergenic ncRNA (lincRNA)

Class 3: Novel protein-coding genes

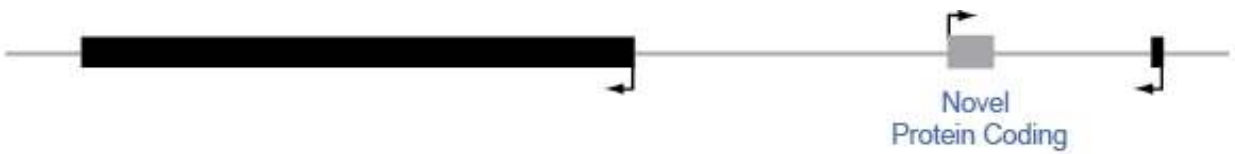


What about novel genes?

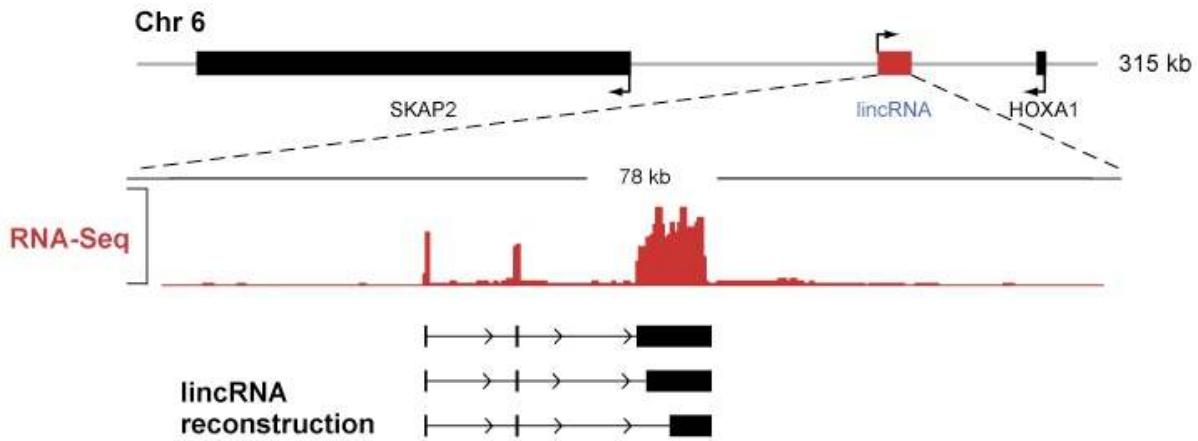
Class 1: Overlapping ncRNA

Class 2: Large Intergenic ncRNA (lincRNA)

Class 3: Novel protein-coding genes



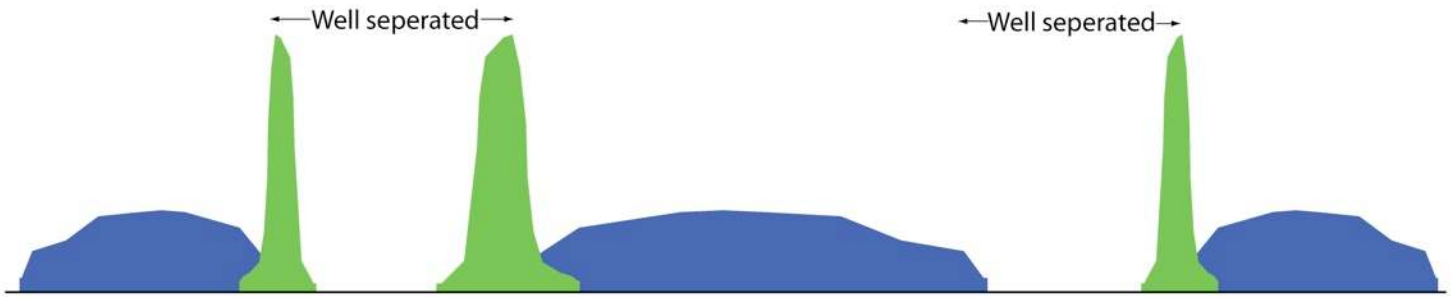
Class 2: Intergenic ncRNA (lincRNA)



ES cells	3 cell types
~500	**~1500**

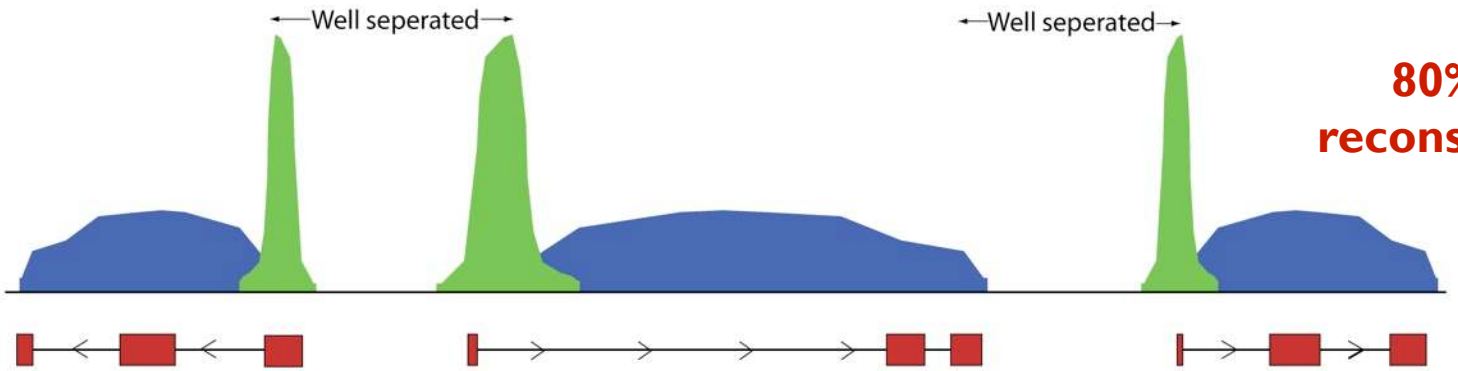
lincRNAs: Comparison to K4-K36

Chromatin Approach



lincRNAs: Comparison to K4-K36

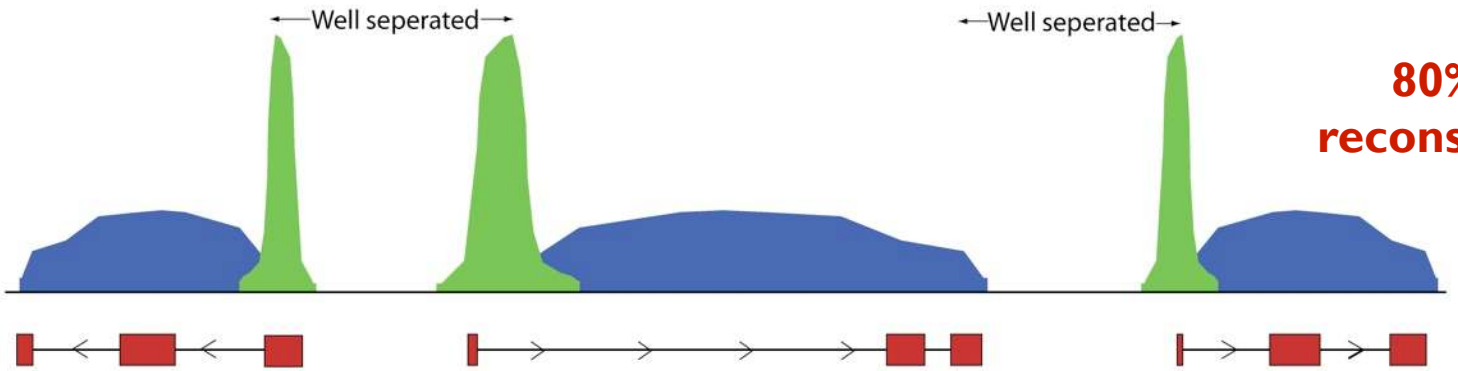
Chromatin Approach



80% have reconstructions

lincRNAs: Comparison to K4-K36

Chromatin Approach

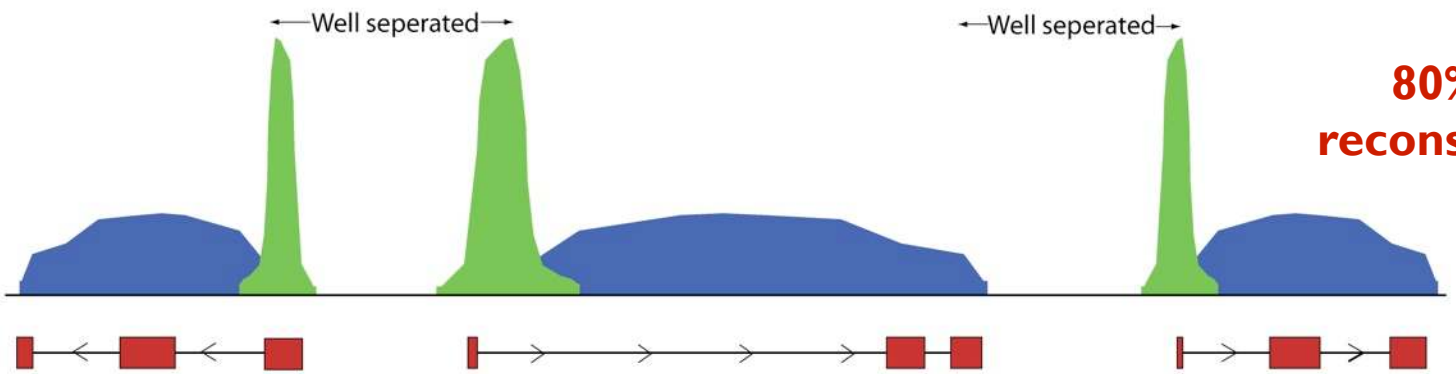


80% have reconstructions

RNA-Seq First Approach

lincRNAs: Comparison to K4-K36

Chromatin Approach

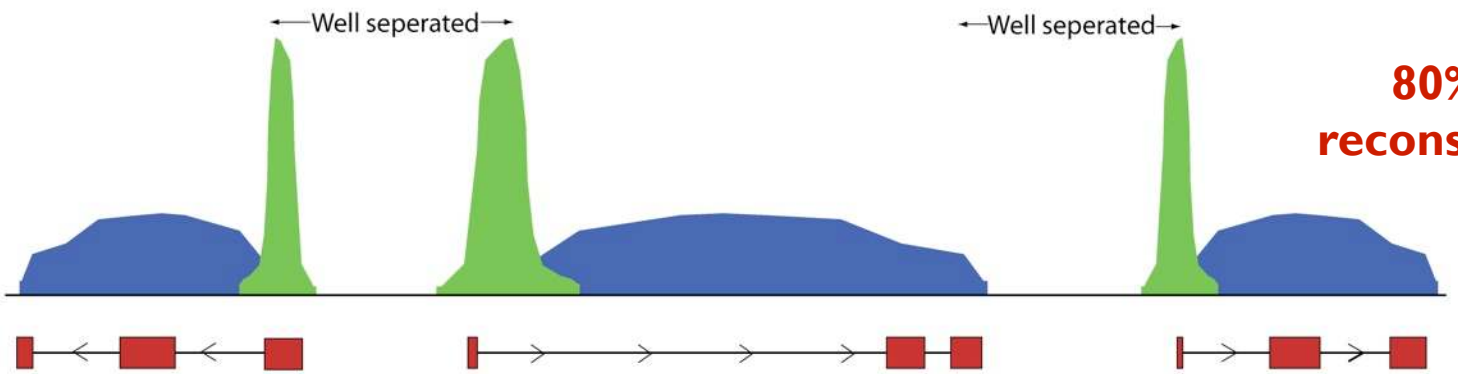


80% have reconstructions

RNA-Seq First Approach

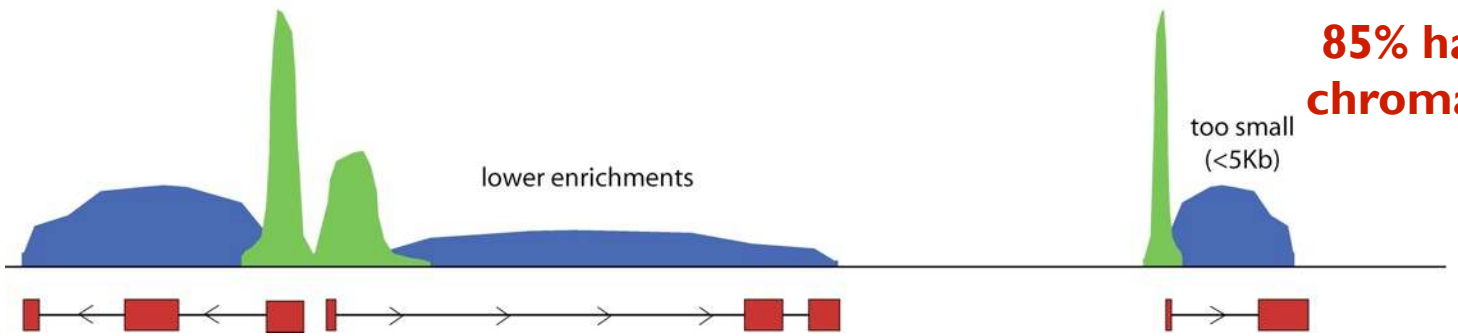
lincRNAs: Comparison to K4-K36

Chromatin Approach



80% have reconstructions

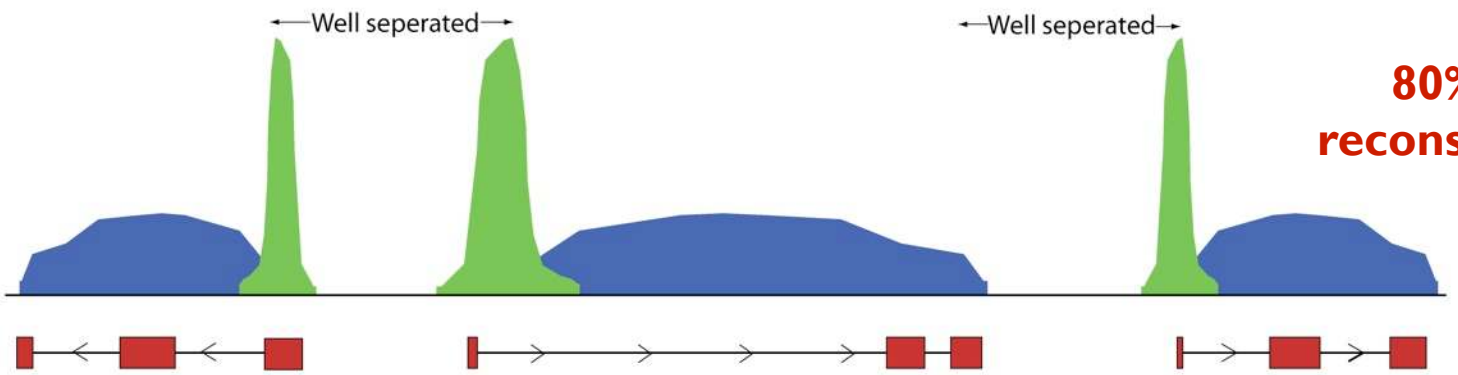
RNA-Seq First Approach



85% have chromatin

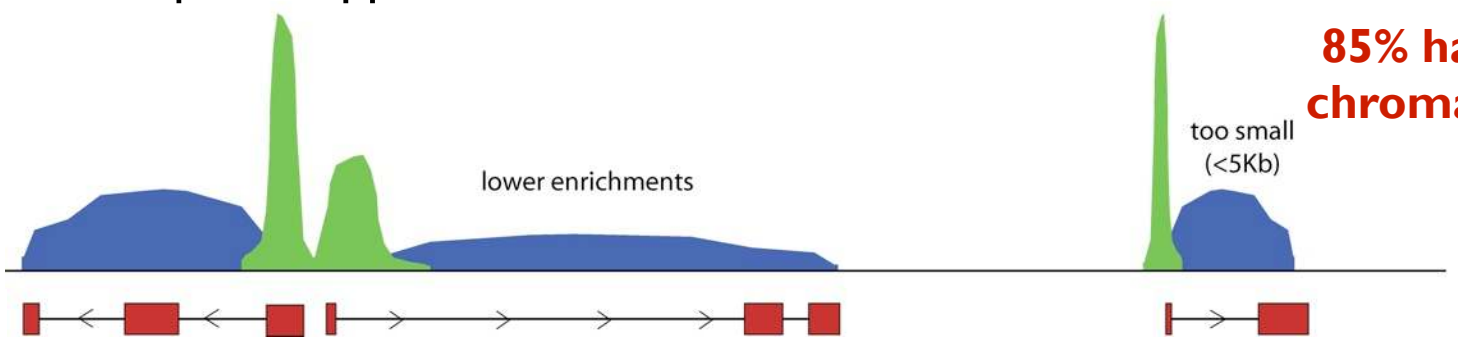
lincRNAs: Comparison to K4-K36

Chromatin Approach



80% have reconstructions

RNA-Seq First Approach

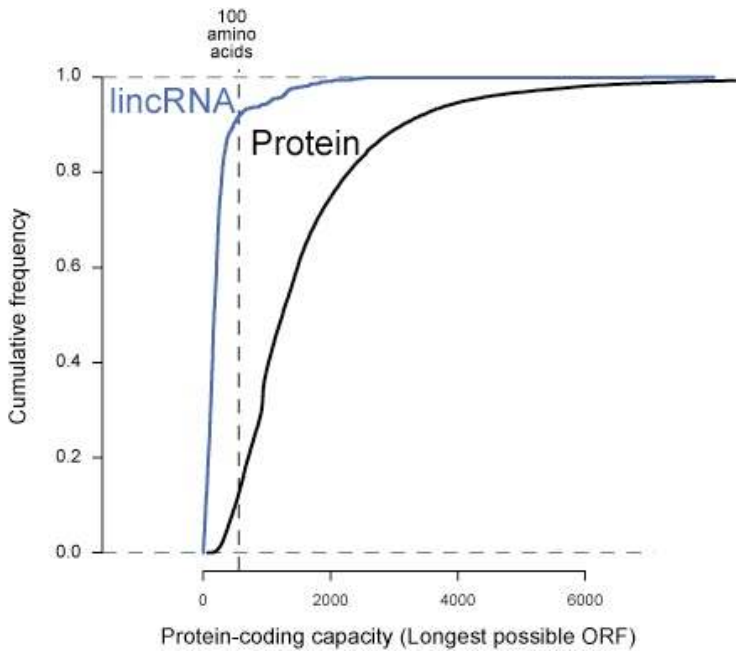


85% have chromatin

RNA-Seq reconstruction and chromatin signature synergize to identify lincRNAs

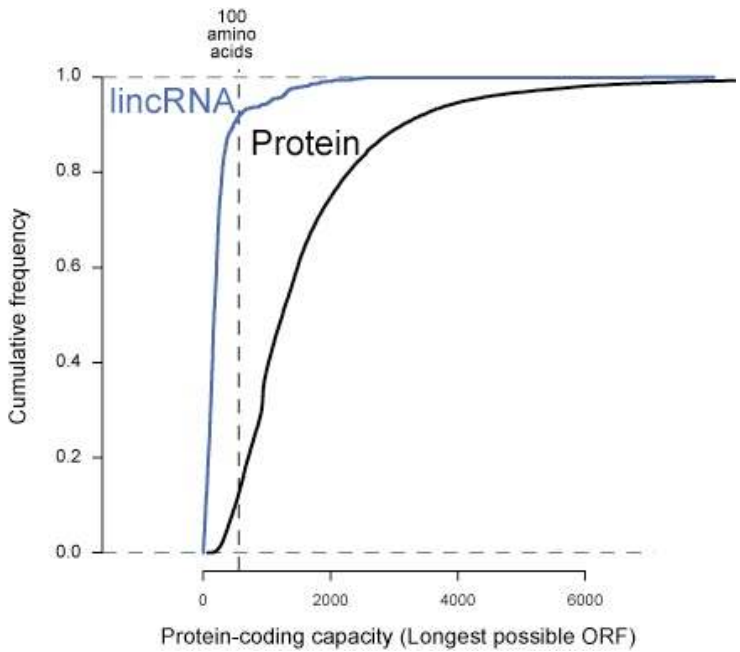
lincRNAs: How do we know they are non-coding?

ORF Length

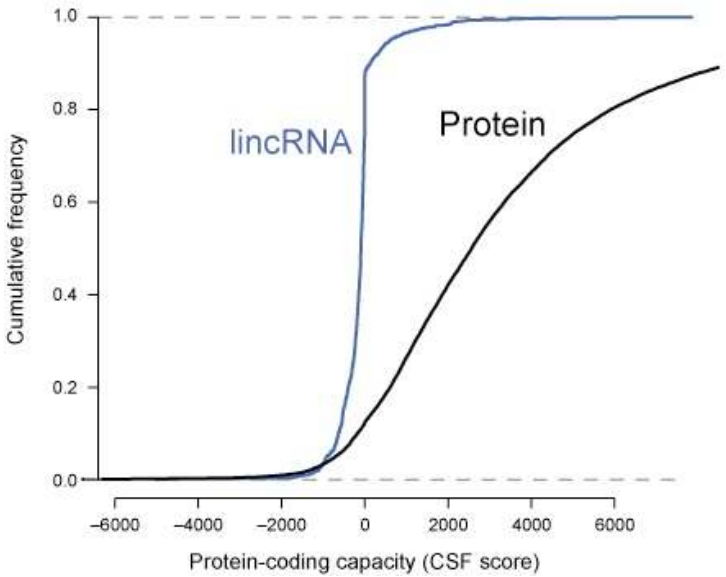


lincRNAs: How do we know they are non-coding?

ORF Length

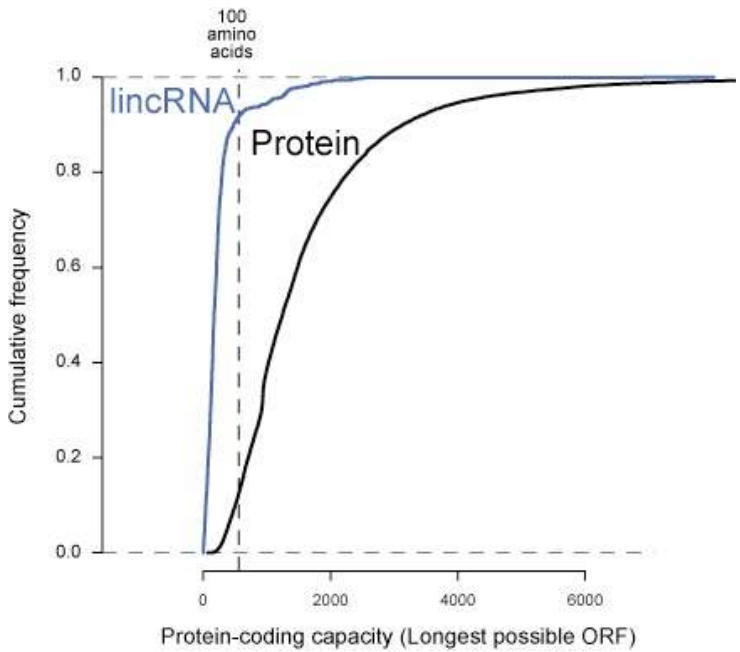


CSF (ORF Conservation)

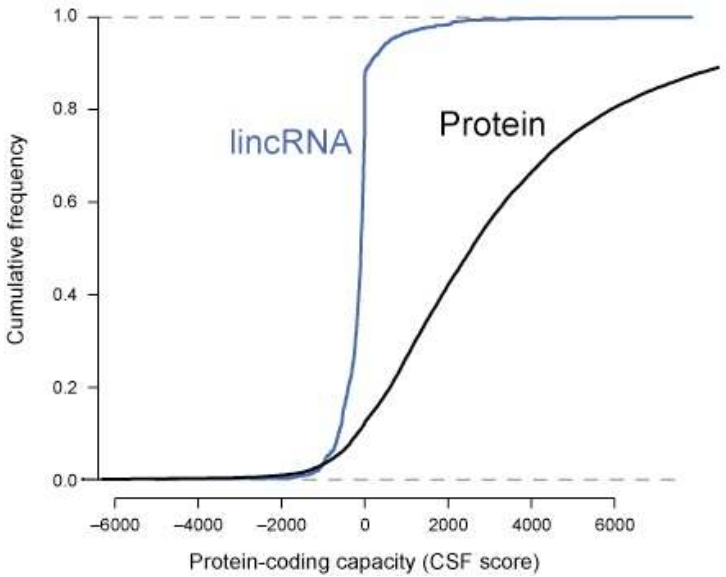


lincRNAs: How do we know they are non-coding?

ORF Length

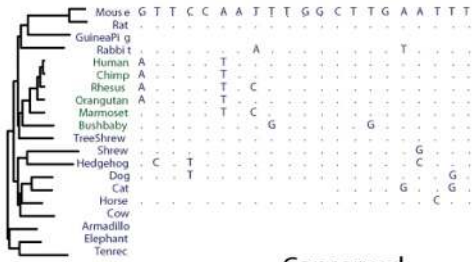


CSF (ORF Conservation)

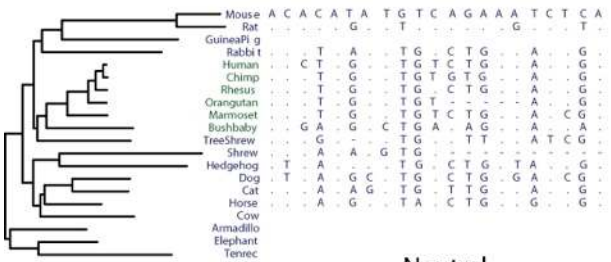


>95% do not encode proteins

lincRNAs: Assessing their evolutionary conservation

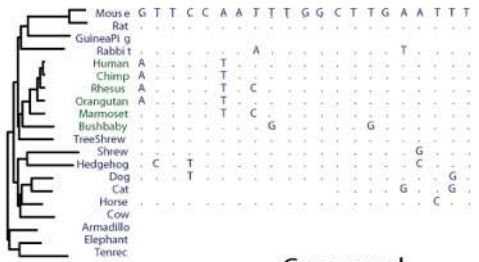


Conserved

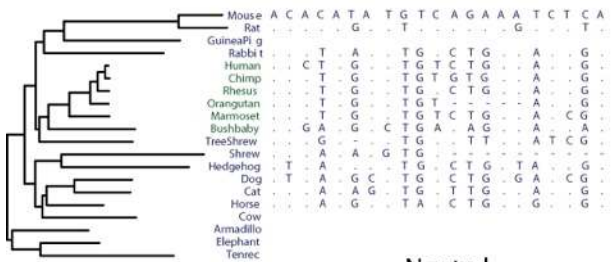


Neutral

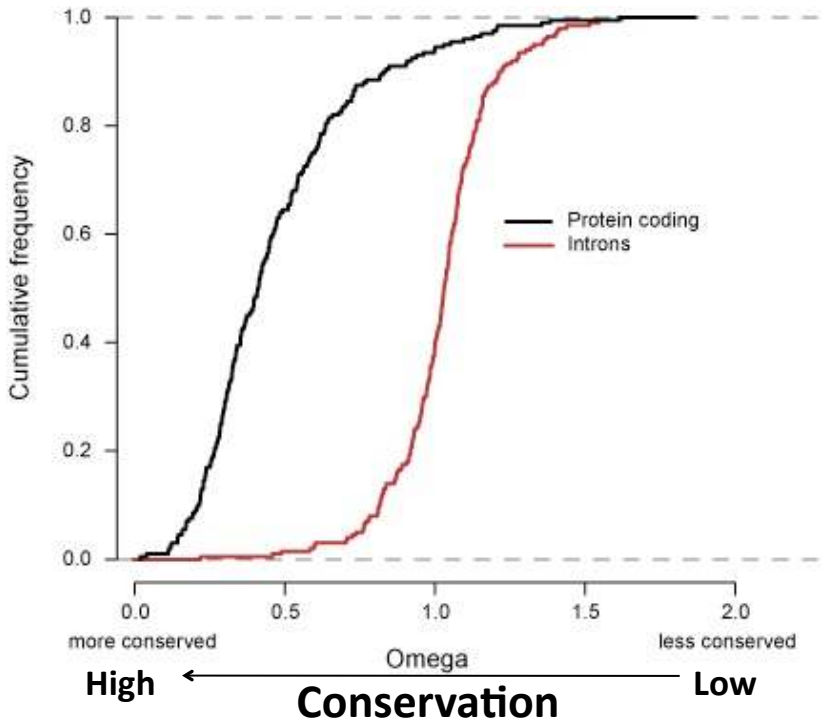
lincRNAs: Assessing their evolutionary conservation



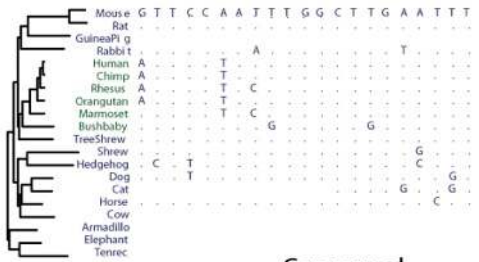
Conserved



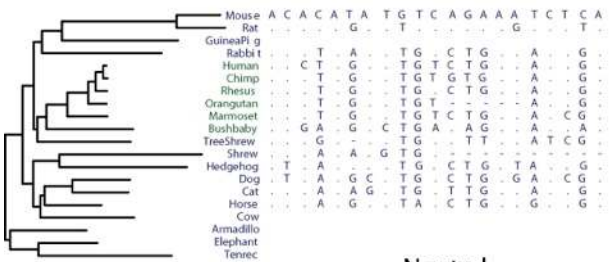
Neutral



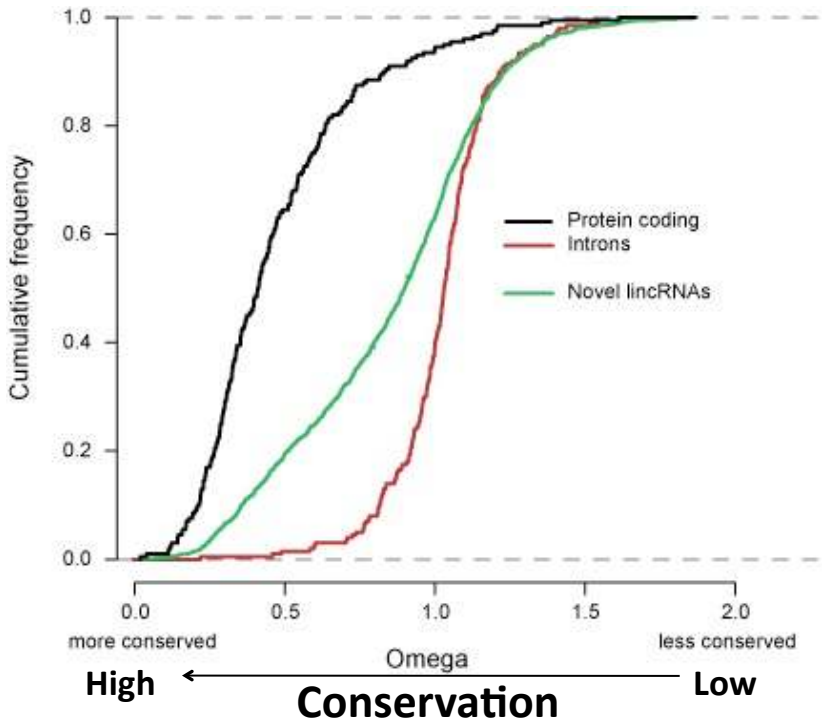
lincRNAs: Assessing their evolutionary conservation



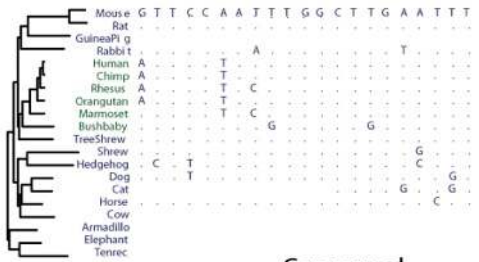
Conserved



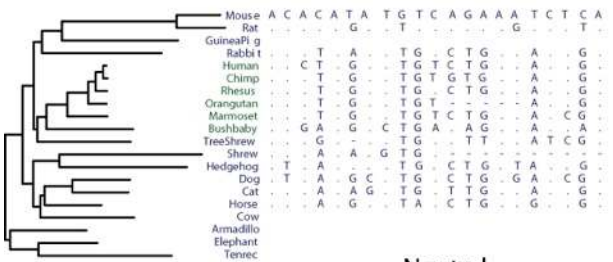
Neutral



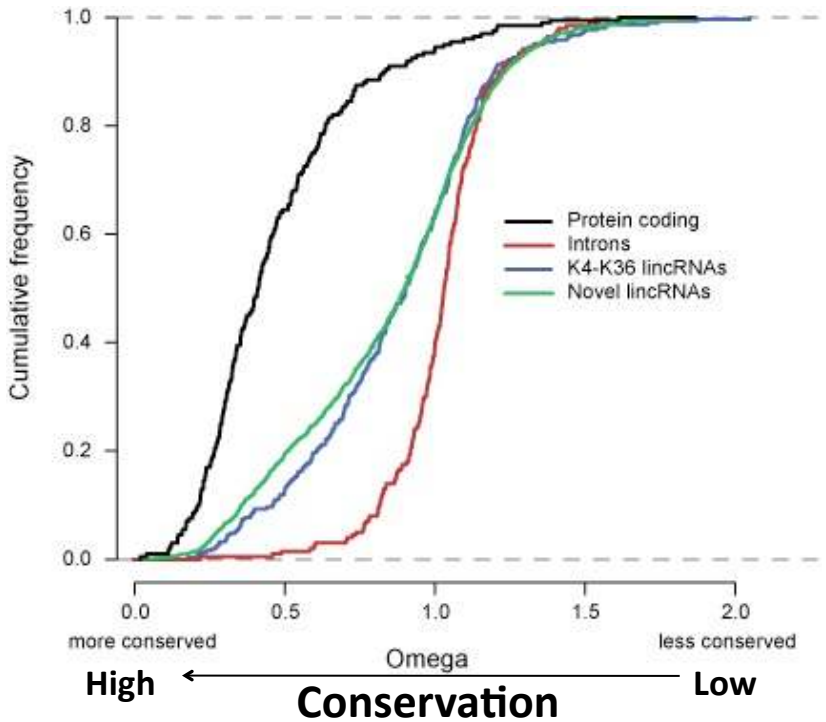
lincRNAs: Assessing their evolutionary conservation



Conserved



Neutral

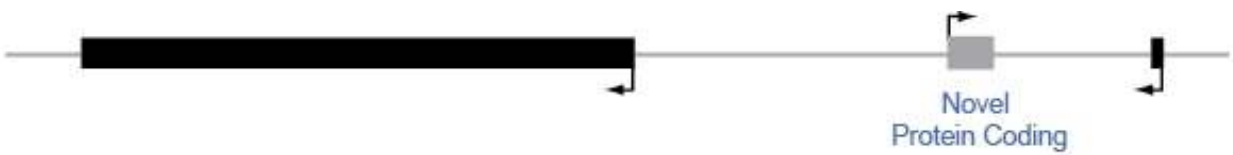


What about novel coding genes?

Class 1: Overlapping ncRNA

Class 2: Large Intergenic ncRNA (lincRNA)

Class 3: Novel protein-coding genes

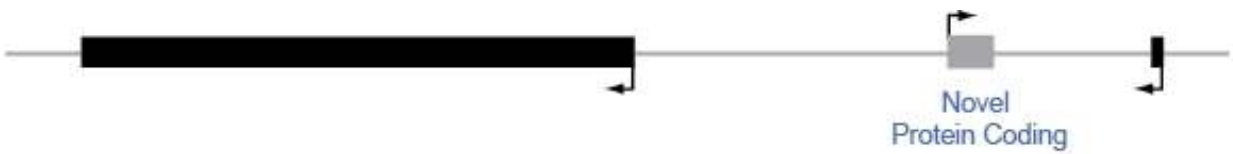


What about novel coding genes?

Class 1: Overlapping ncRNA

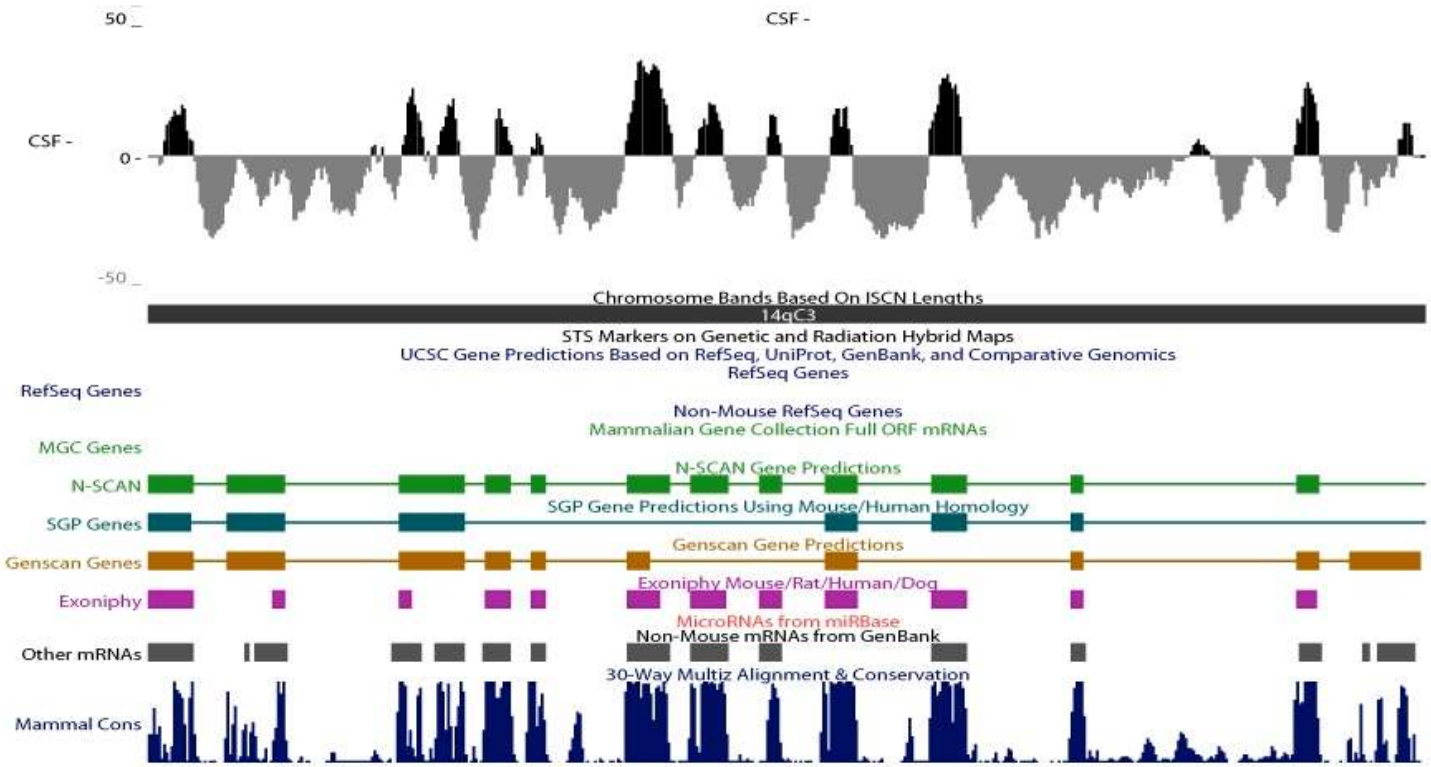
Class 2: Large Intergenic ncRNA (lincRNA)

Class 3: Novel protein-coding genes

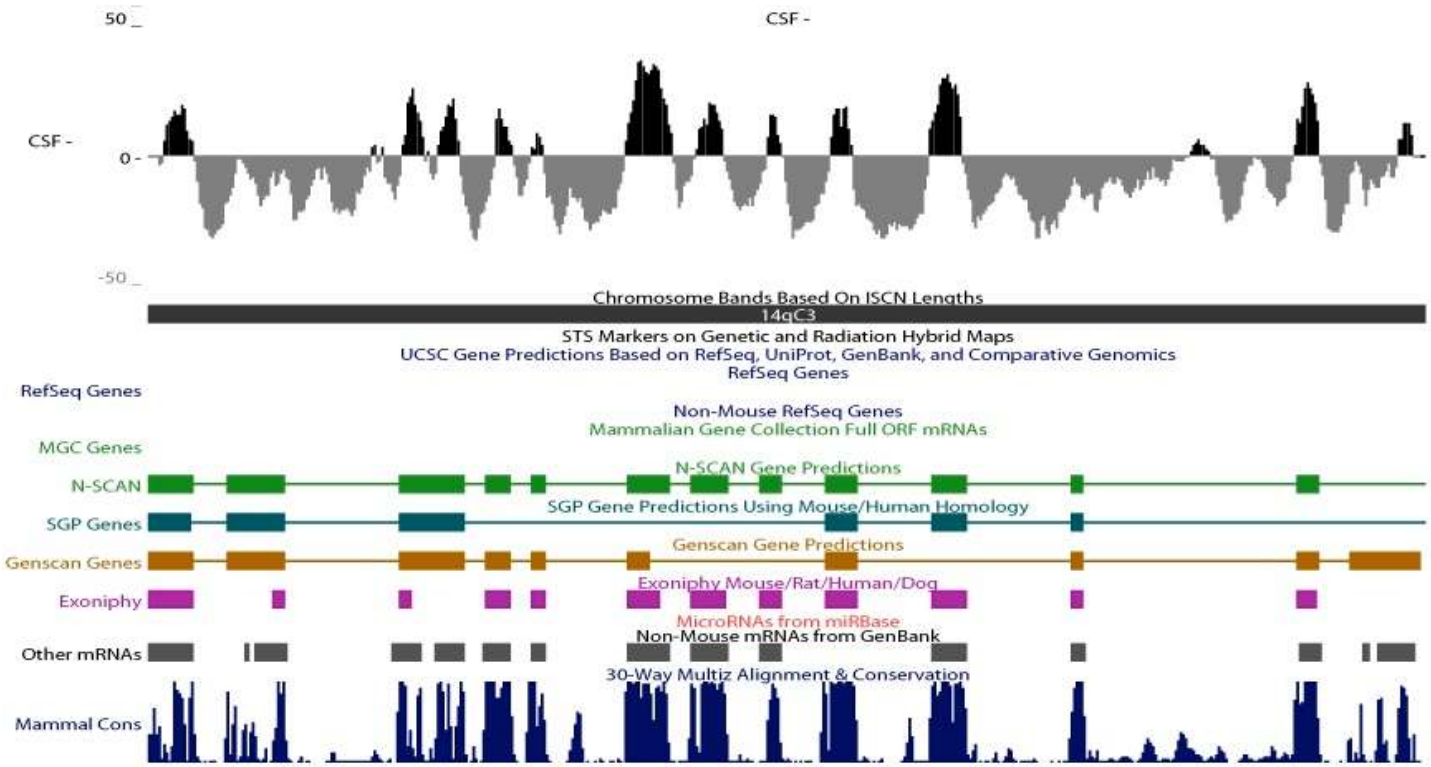


Finding novel coding genes

Finding novel coding genes



Finding novel coding genes

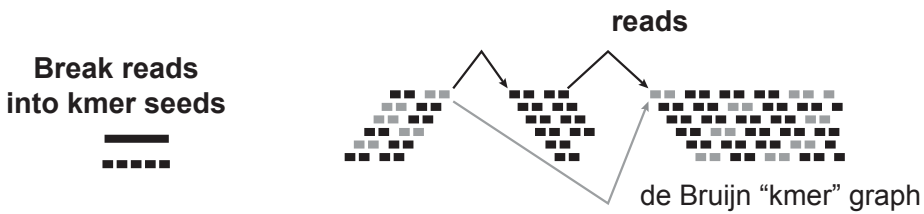


~40 novel protein-coding genes

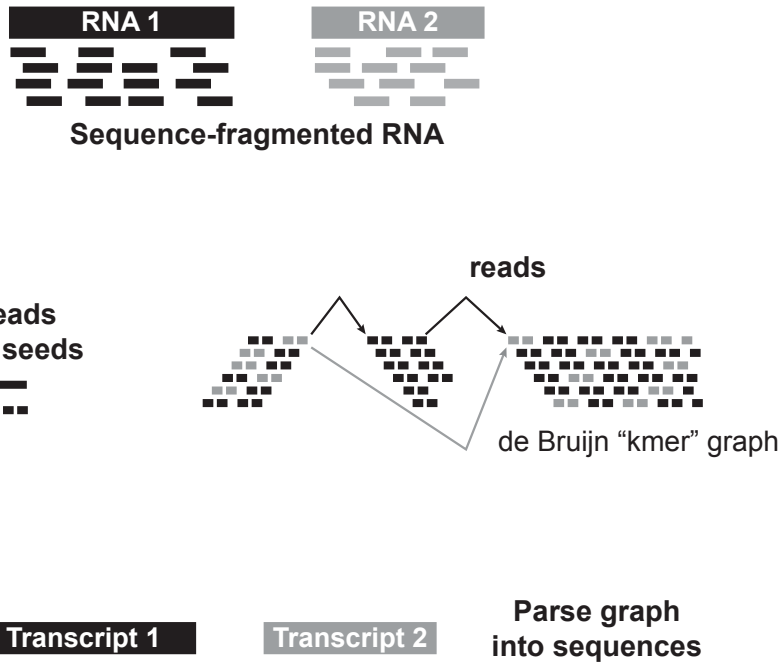
Other transcript reconstruction methods

Method I: Direct assembly

Method I: Direct assembly



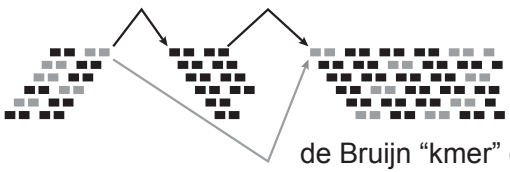
Method I: Direct assembly



Method I: Direct assembly

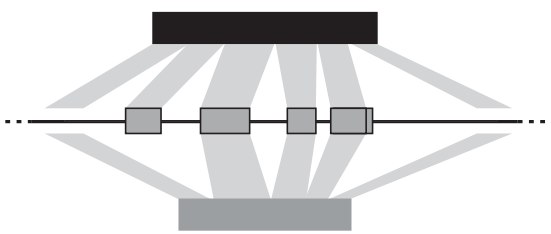
Sequence-fragmented RNA

Break reads into kmer seeds



de Bruijn “kmer” graph

Parse graph into sequences

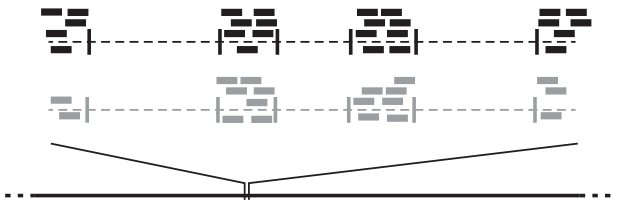


Align sequences to genome

Method II: Genome-guided

Method II: Genome-guided

Sequence-fragmented RNA

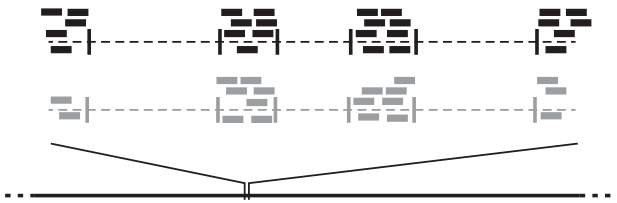


Align reads to genome

Fragments aligned to genome

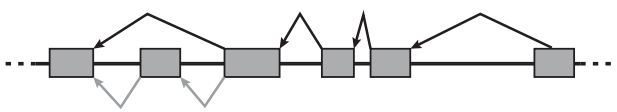
Method II: Genome-guided

Sequence-fragmented RNA



Align reads to genome

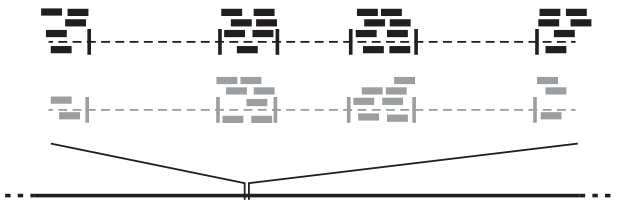
Fragments aligned to genome



ranscript graph

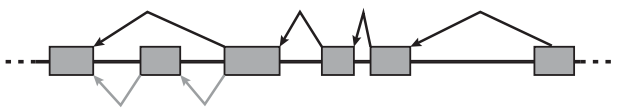
Method II: Genome-guided

Sequence-fragmented RNA

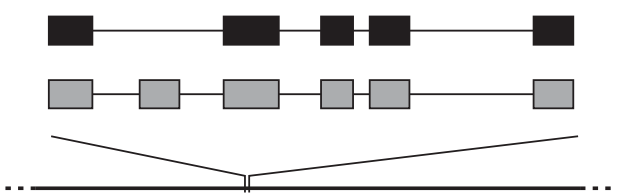


Align reads to genome

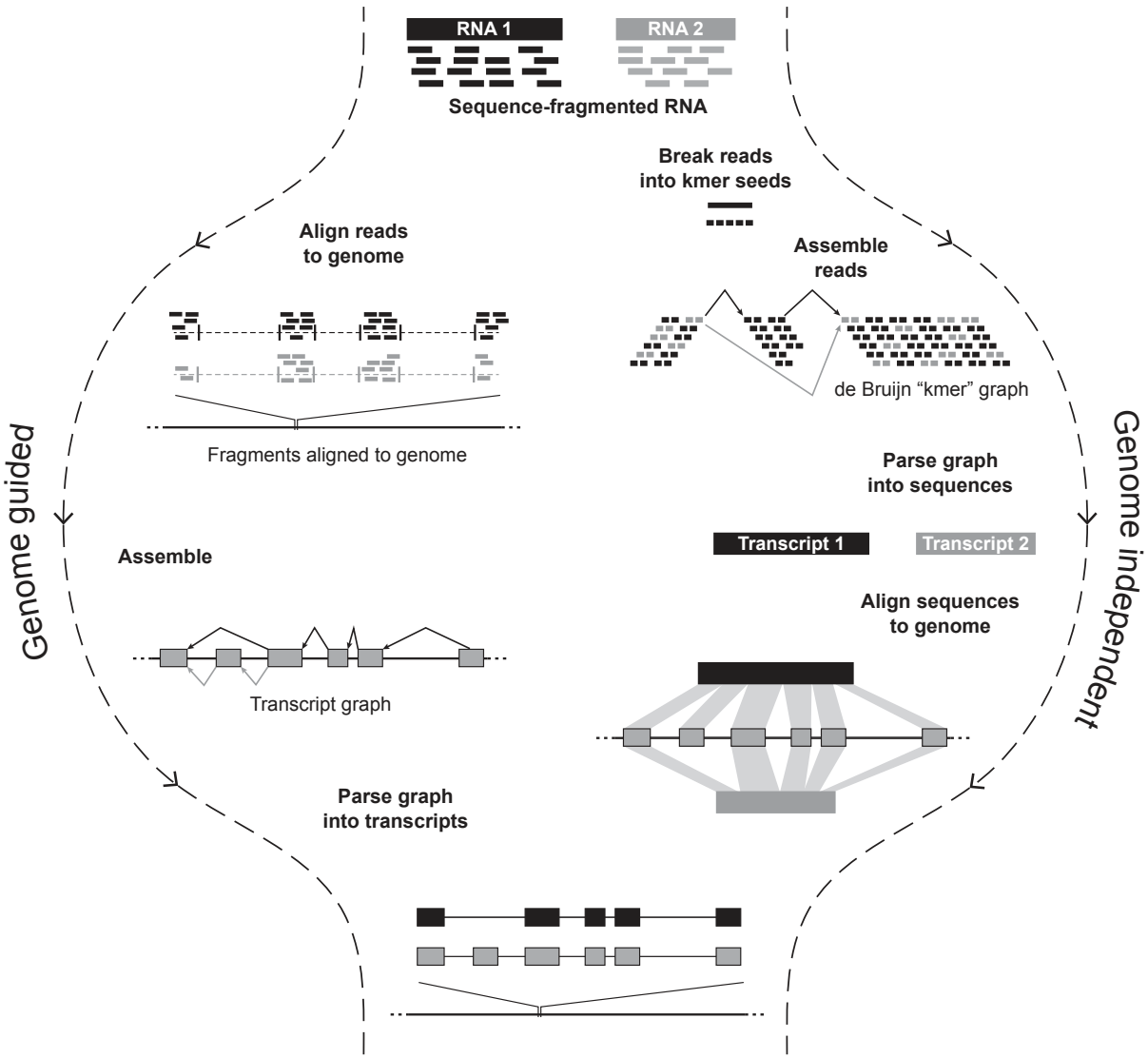
Fragments aligned to genome



ranscript graph



Transcriptome reconstruction method summary



Pros and cons of each approach

Pros and cons of each approach

Pros and cons of each approach

Pros and cons of each approach

Pros and cons of each approach

Pros and cons of each approach

- Transcript assembly methods are the obvious choice for organisms without a reference sequence.

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.

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.

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.

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	Cufflinks
Trans-ABySS	Scripture
Trinity	

Differences between Cufflinks and Scripture

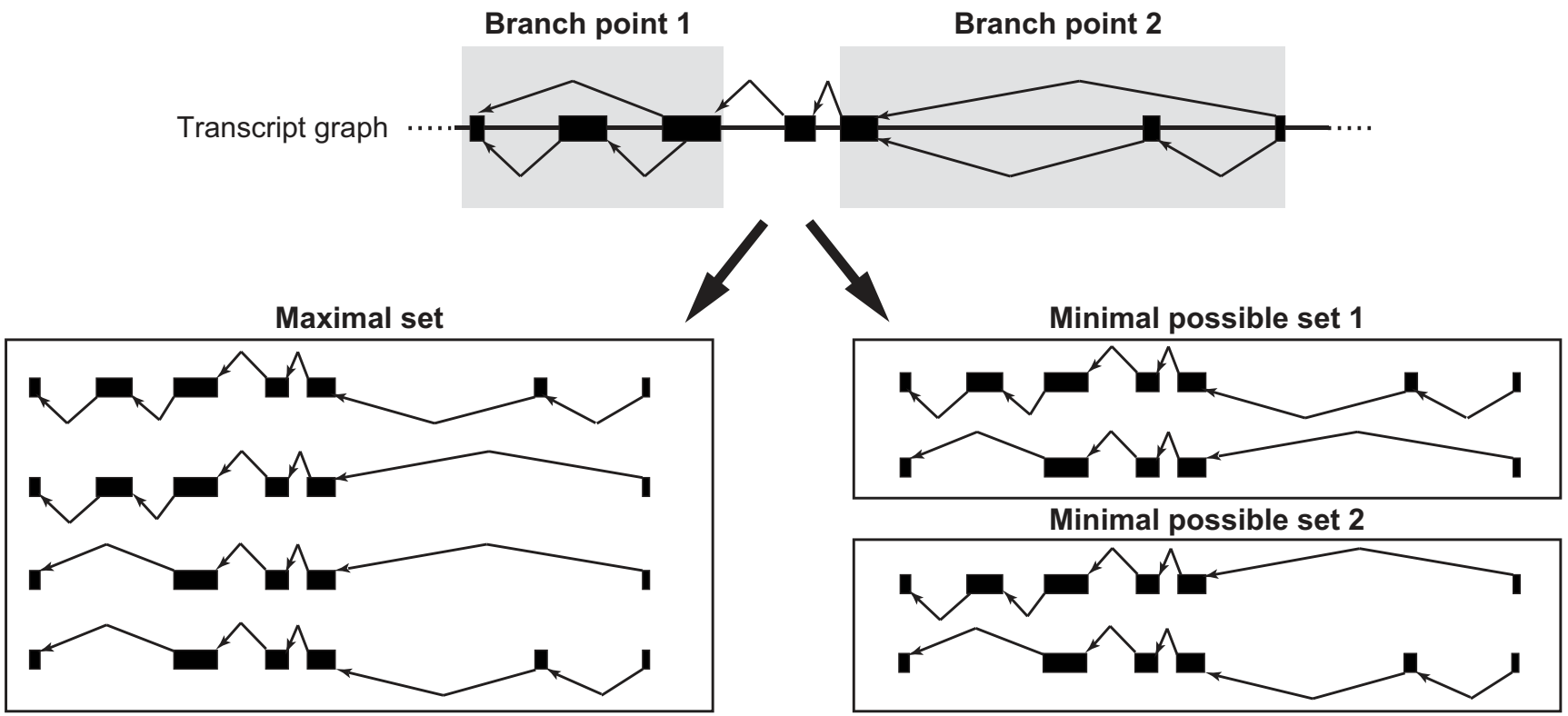
Differences between Cufflinks and Scripture

- Scripture was designed with annotation in mind. It reports all possible transcripts that are *significantly expressed* given the aligned data (*Maximum sensitivity*).

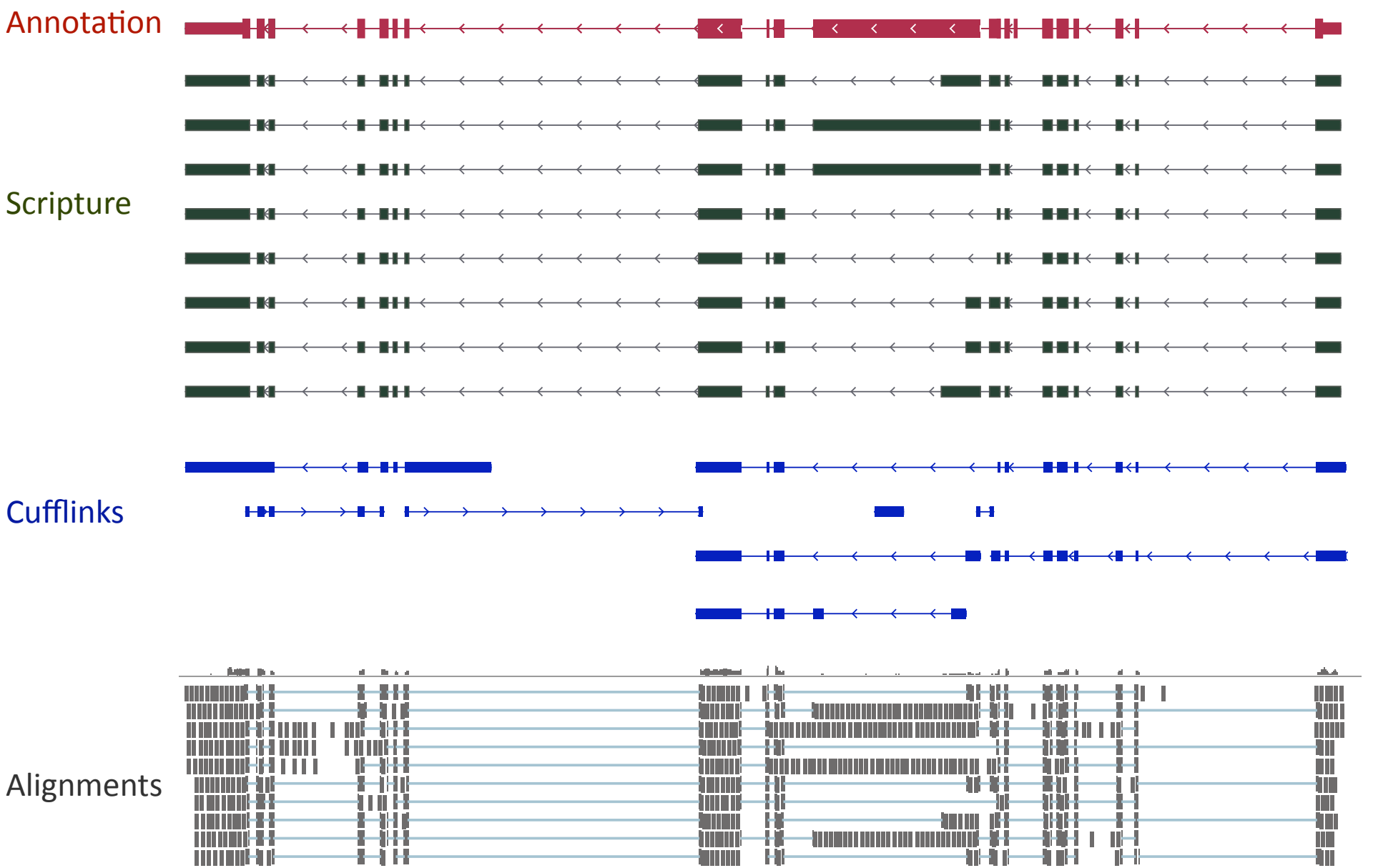
Differences between Cufflinks and Scripture

- 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

Comparing reconstructions

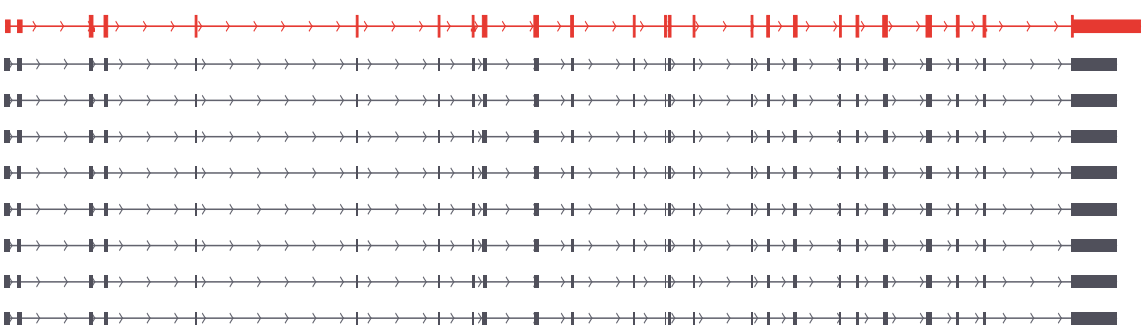
	CPU Hours Total Memory	Genes fully reconstructed	Mean isoforms per reconstruction	Mean fragments per known annotation	Number of fragments predicted
Cufflinks	10	1.4 G	5,994	1.2	1.4	**159,856**
Scripture	16	3.5 G	6,221	**1.6**	1.3	61,922
Trans- Abyss	650	120 G[4]	3,330	4.7	2.6	3,117,238

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

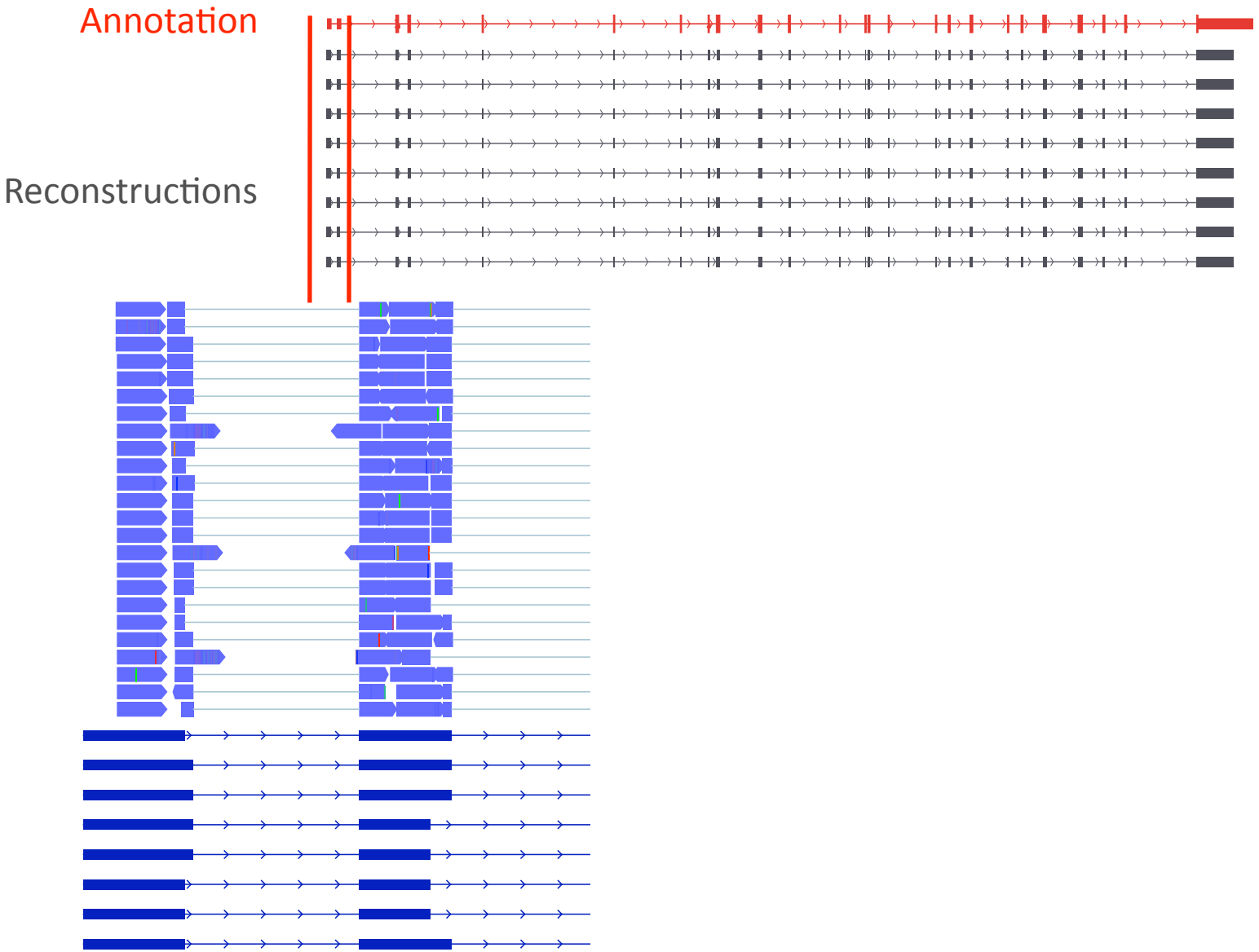
Why so many isoforms

Annotation

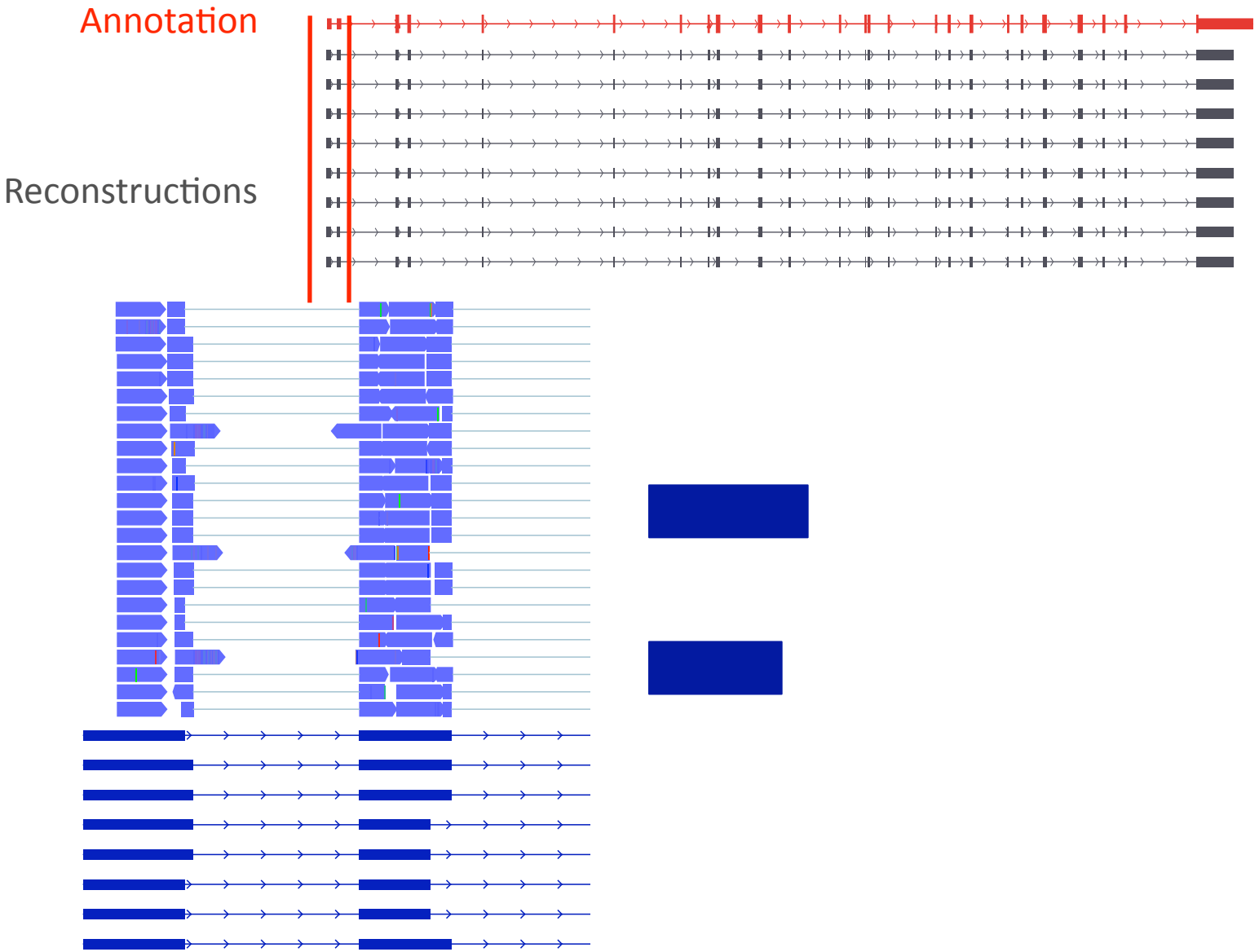
Reconstructions



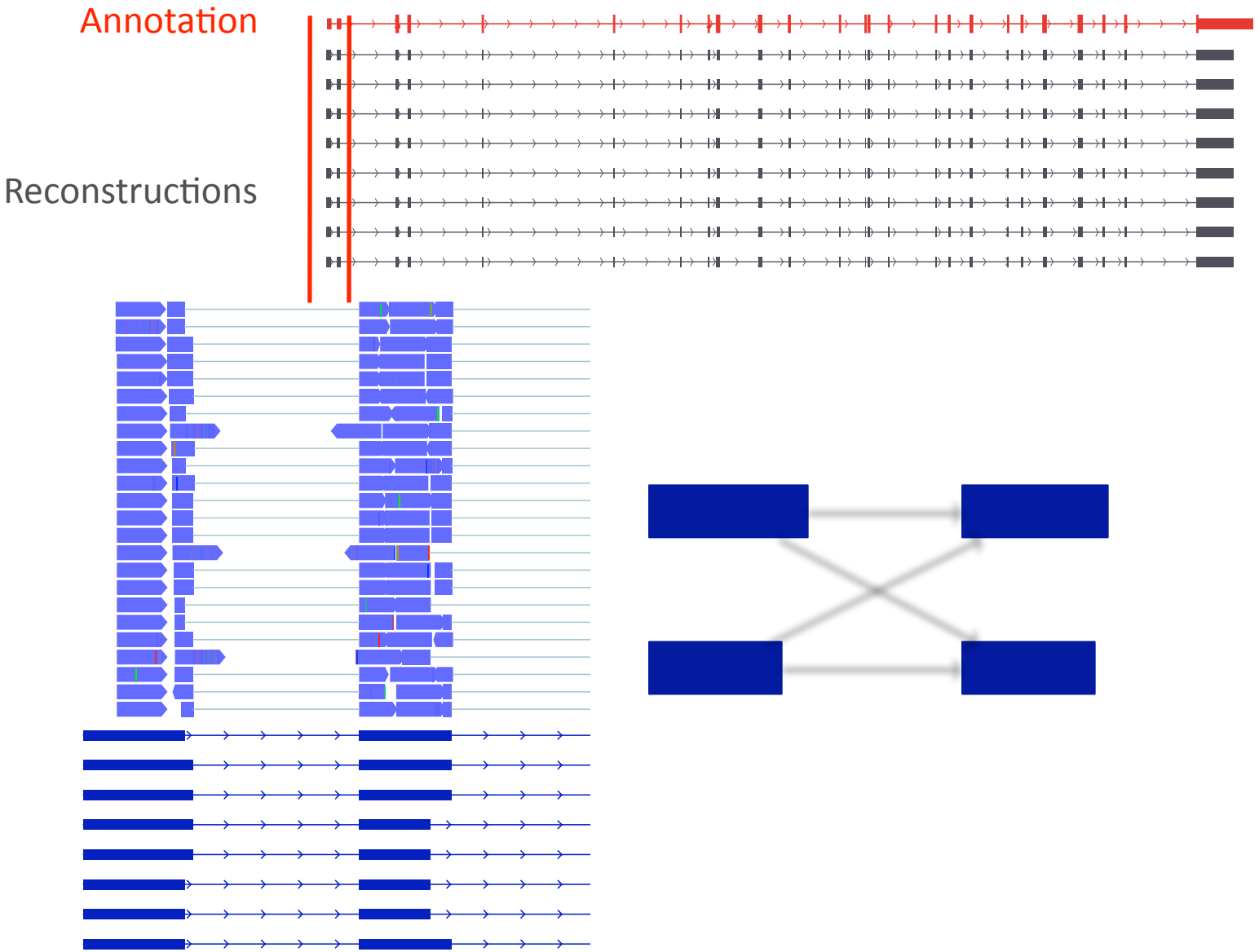
Why so many isoforms



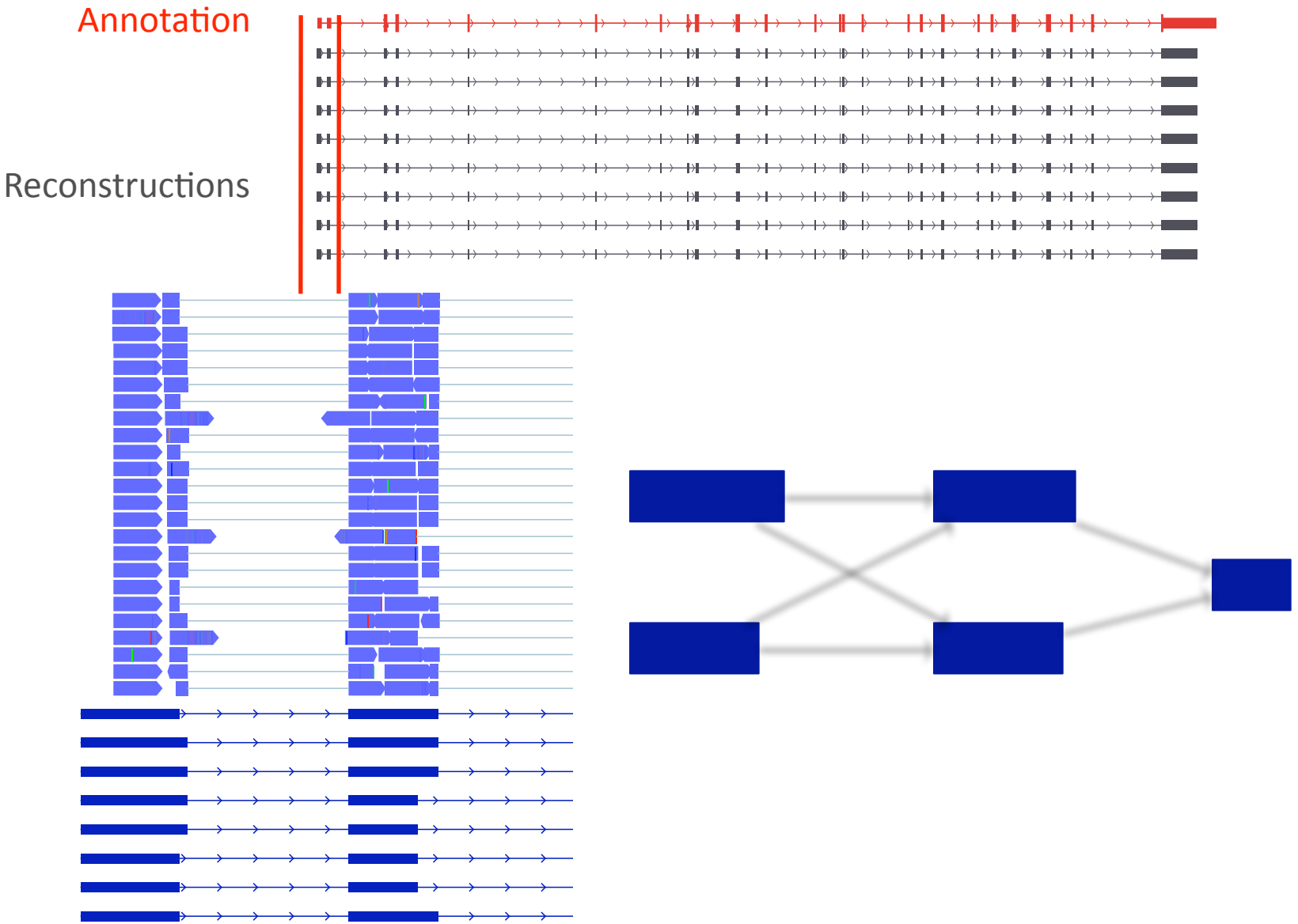
Why so many isoforms



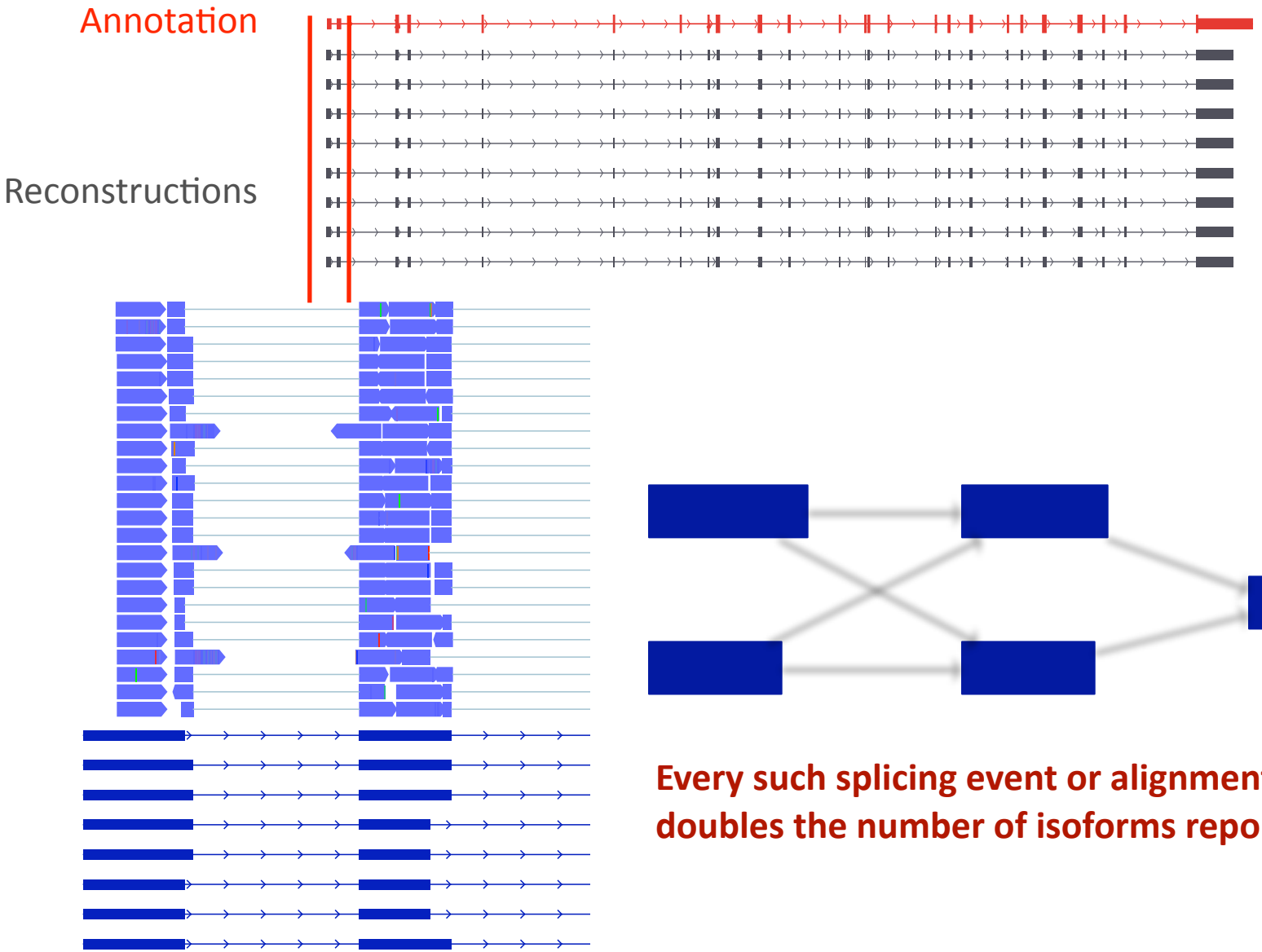
Why so many isoforms



Why so many isoforms

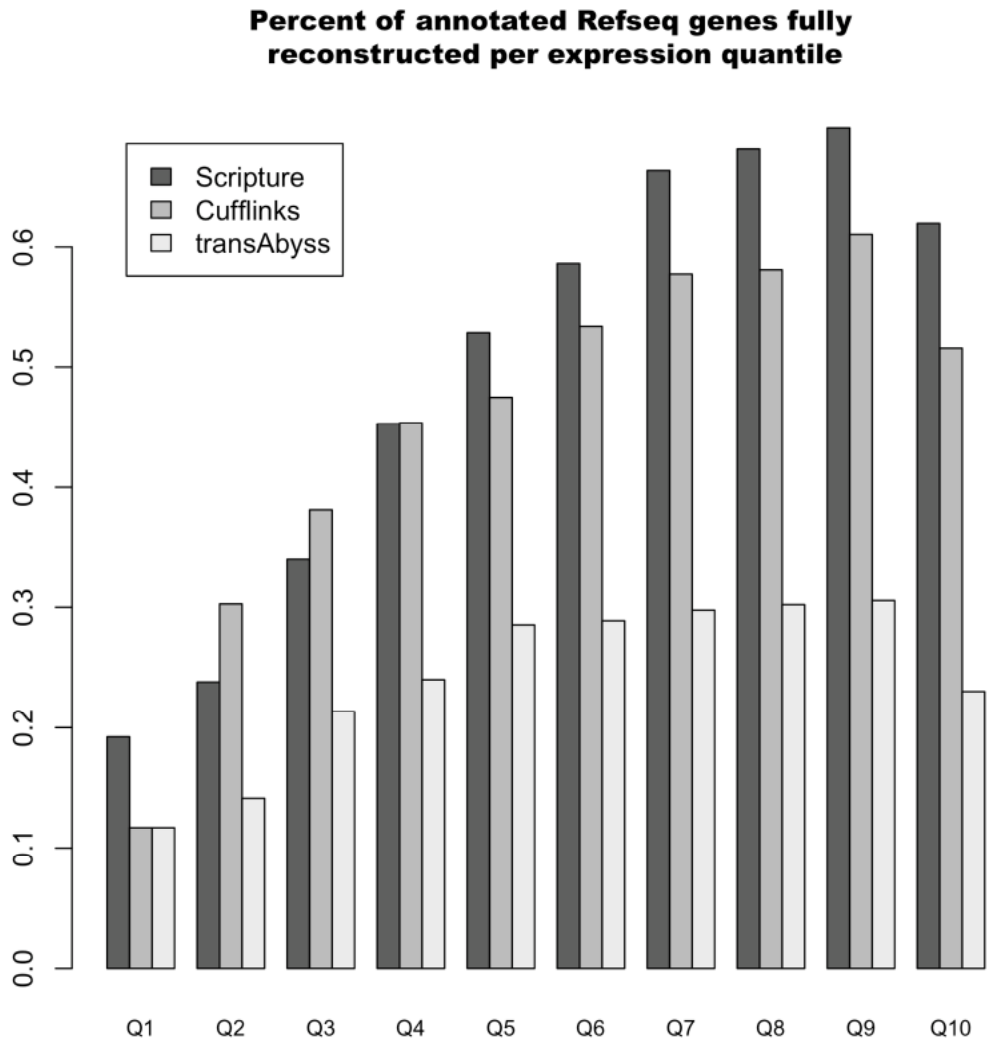


Why so many isoforms

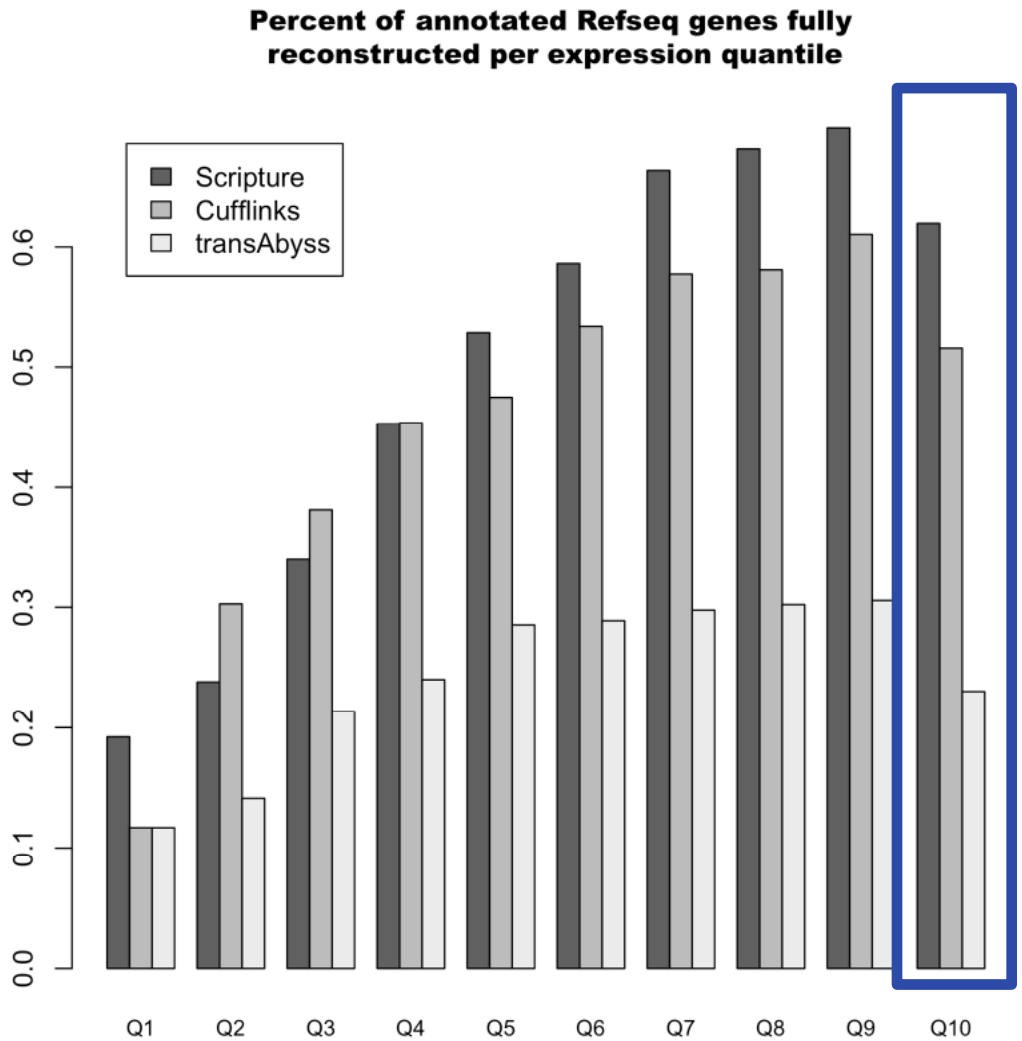


Every such splicing event or alignment artifact doubles the number of isoforms reported

Reconstruction comparison

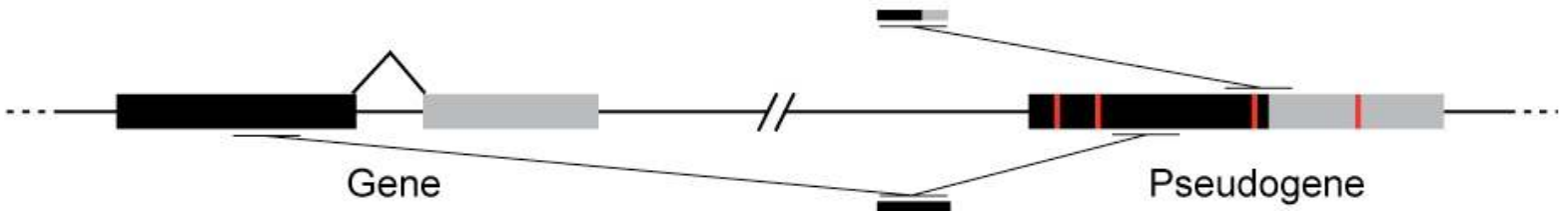


Reconstruction comparison

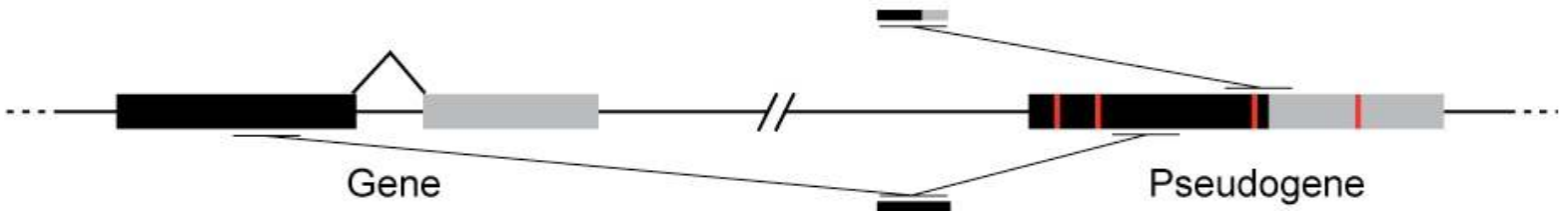


Alignment revisited — spliced alignment is still work in progress

Exon-first aligners are faster but at cost

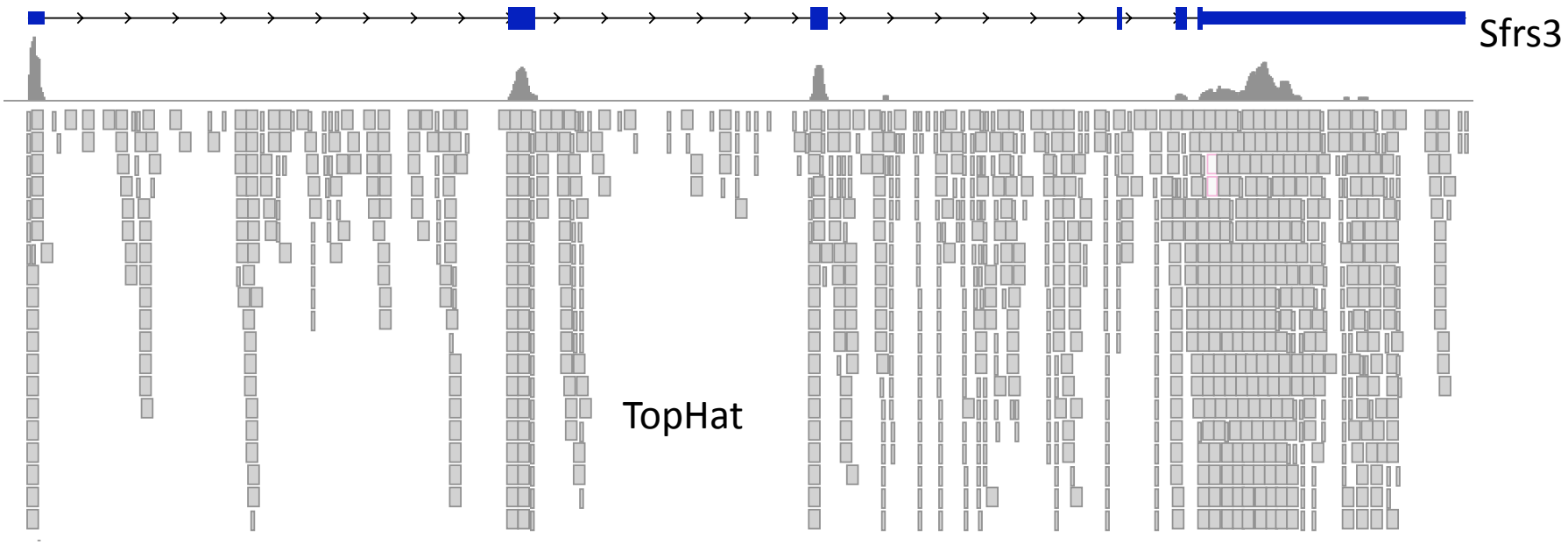


Exon-first aligners are faster but at cost



Alignment artifacts can also decrease sensitivity

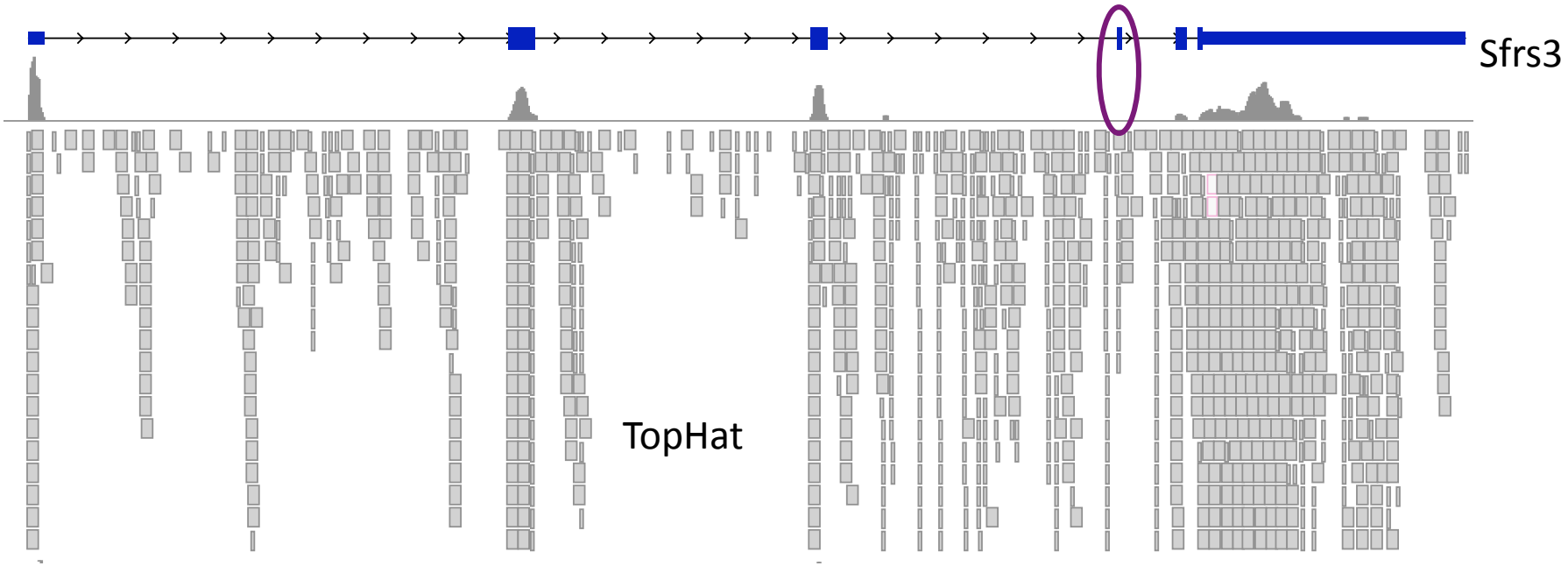
Missing spliced reads for highly expressed genes



Read mapped uniquely

Read ambiguously mapped

Missing spliced reads for highly expressed genes



Read mapped uniquely

Read ambiguously mapped

Can more sensitive alignments overcome this problem?

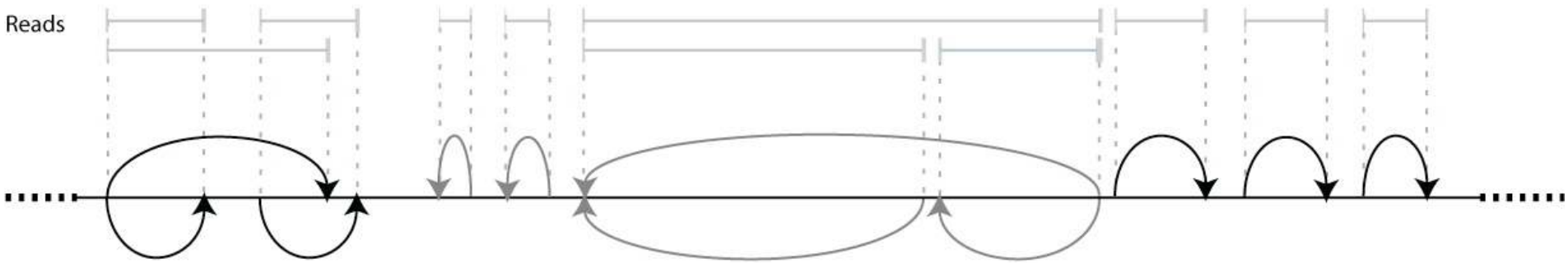
Can more sensitive alignments overcome this problem?

Can more sensitive alignments overcome this problem?

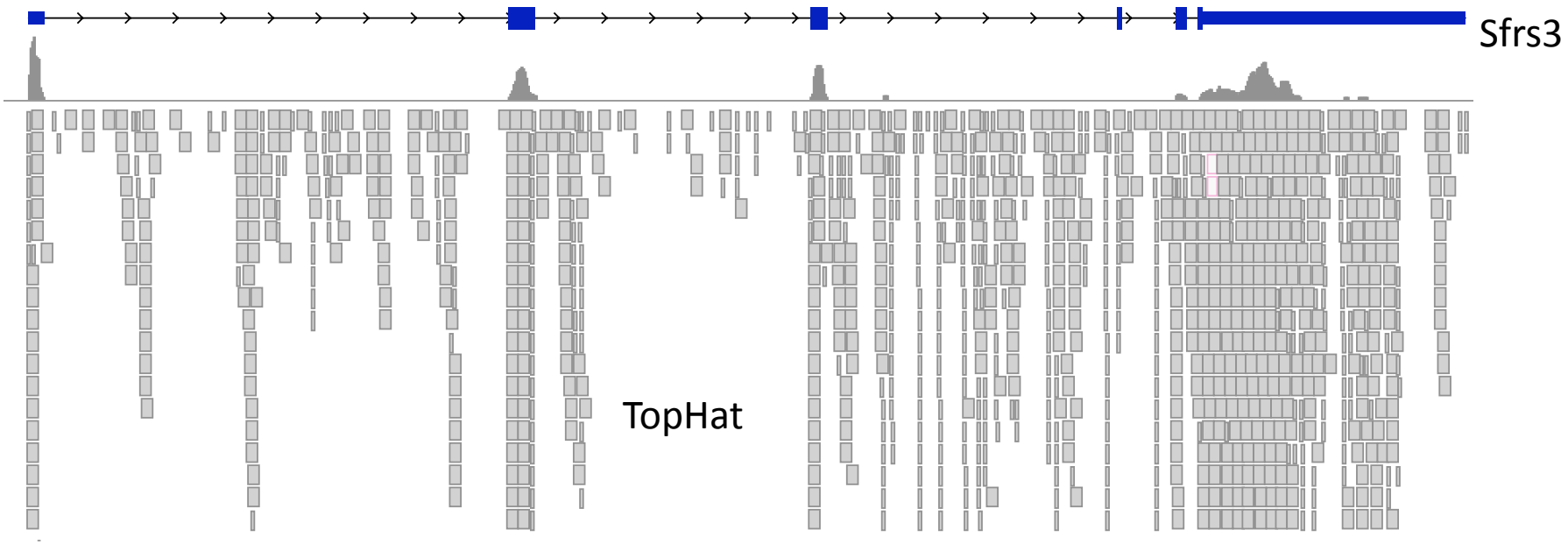
- Use gapped aligners (e.g. BLAT) to map reads
  - Align all reads with BLAT
  - Filter hits and build candidate junction "database" from BLAT hits (Scripture light).
  - Use a short read aligner (Bowtie) to map reads against the connectivity graph inferred transcriptome
  - Map transcriptome alignments to the genome

Can more sensitive alignments overcome this problem?

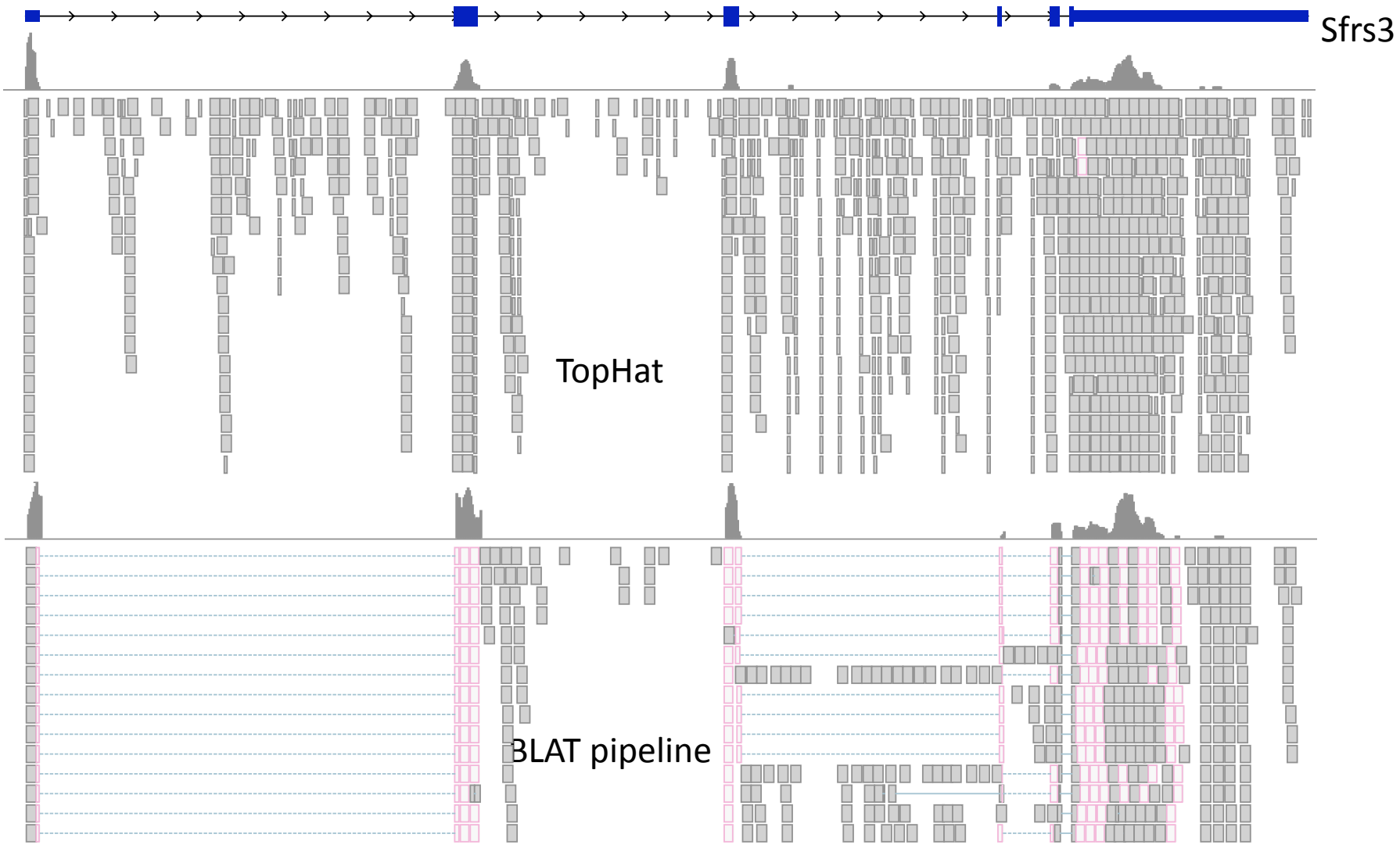
- Use gapped aligners (e.g. BLAT) to map reads
  - Align all reads with BLAT
  - Filter hits and build candidate junction "database" from BLAT hits (Scripture light).
  - Use a short read aligner (Bowtie) to map reads against the connectivity graph inferred transcriptome
  - Map transcriptome alignments to the genome



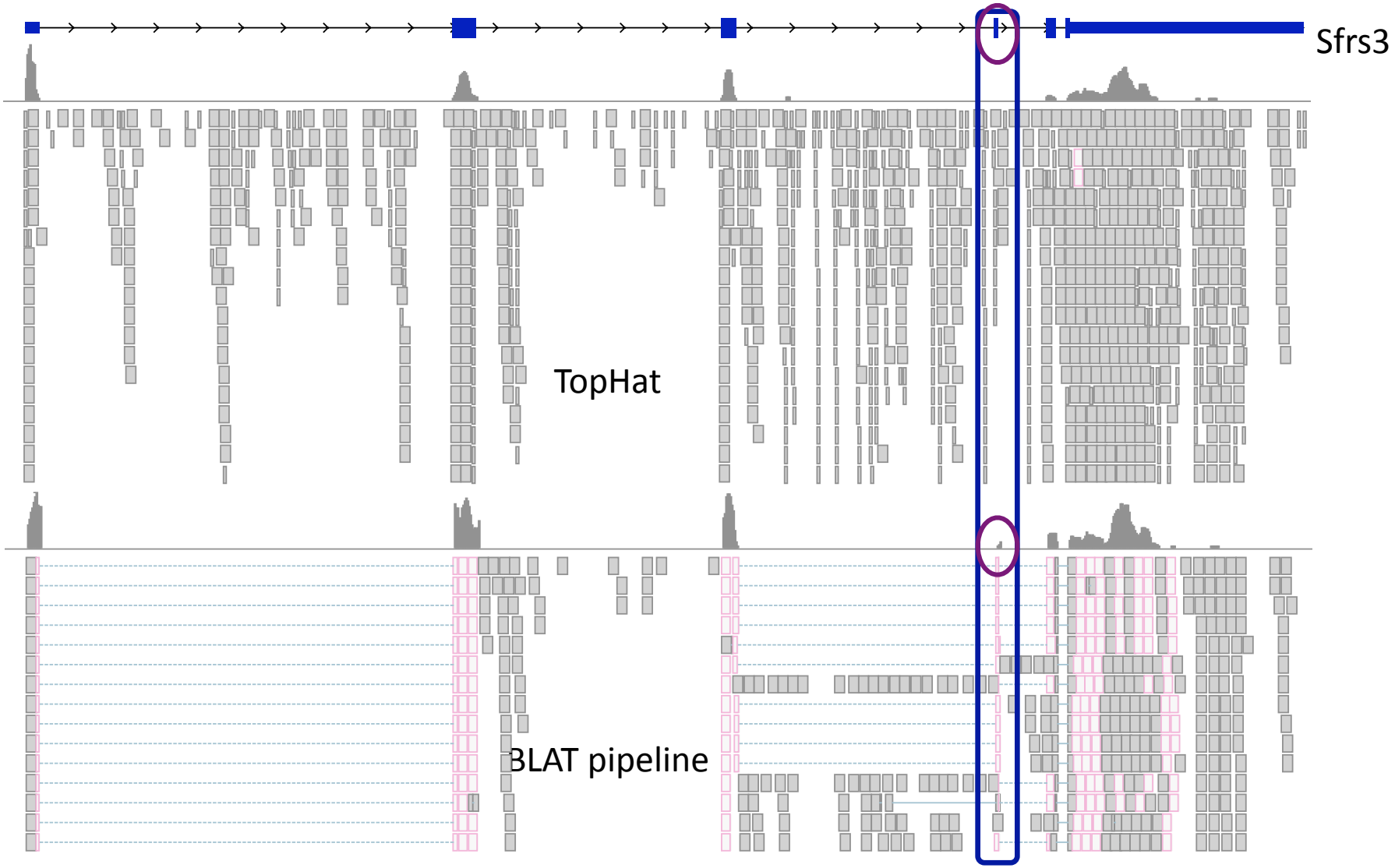
ScriptAlign: Can increase alignment across junctions



ScriptAlign: Can increase alignment across junctions



ScriptAlign: Can increase alignment across junctions



ScriptAlign: Can increase alignment across junctions

ScriptAlign: Can increase alignment across junctions

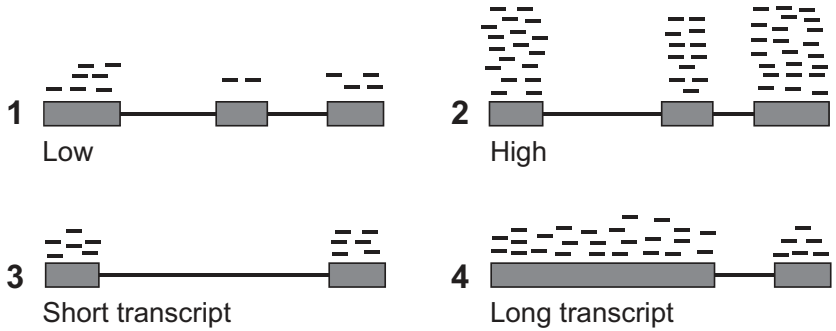
"Map first" reconstruction approaches directly benefit with mapping improvements
We even get more uniquely aligned reads (not just spliced reads)

Overview of the session

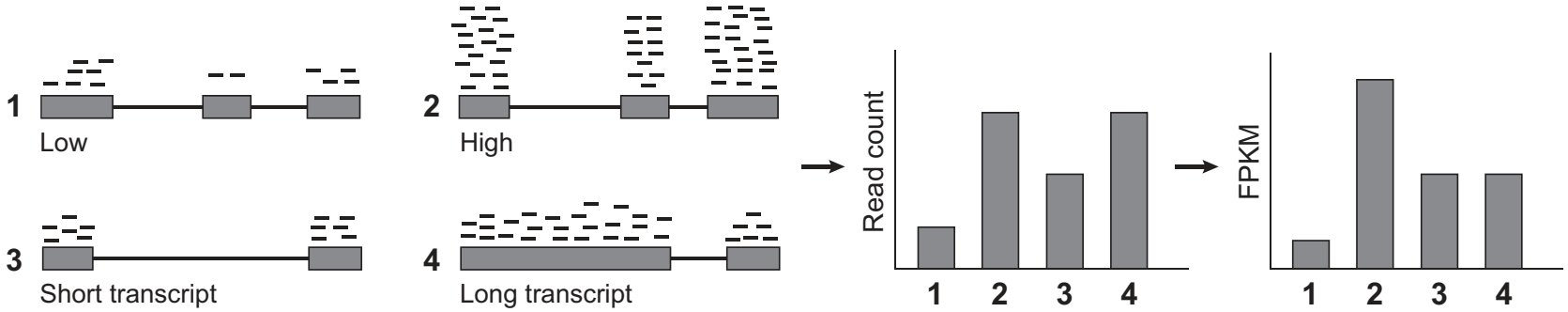
The 3 main computational challenges of sequence analysis for *counting applications*:

- Read mapping: Placing short reads in the genome
- Reconstruction: Finding the regions that originate the reads
- Quantification:
  - Assigning scores to regions
  - Finding regions that are differentially represented between two or more samples.

Quantification: only one isoform



Quantification: only one isoform

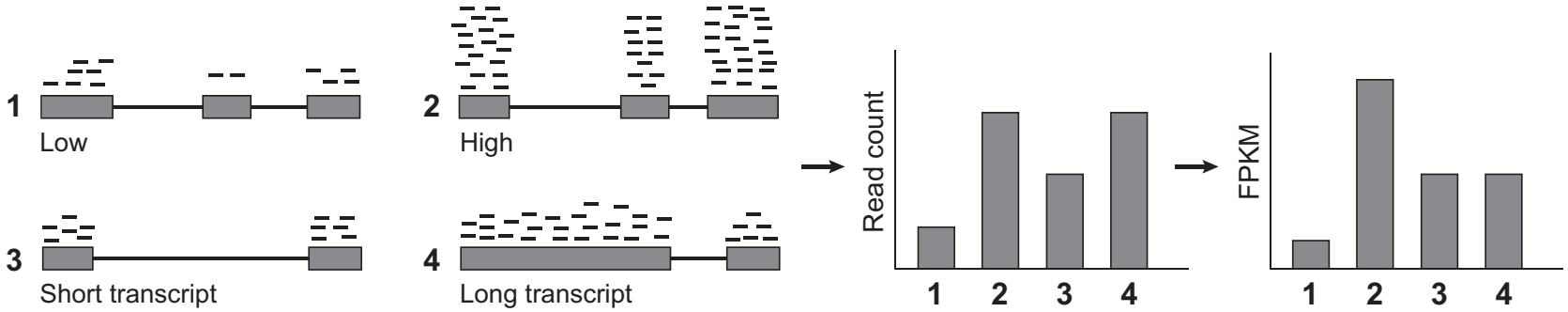


$$RPKM = 10^9 \frac{\#reads}{length \times TotalReads}$$

Reads per kilobase of exonic sequence per million mapped reads

(*Mortazavi* et al Nature methods 2008)

Quantification: only one isoform

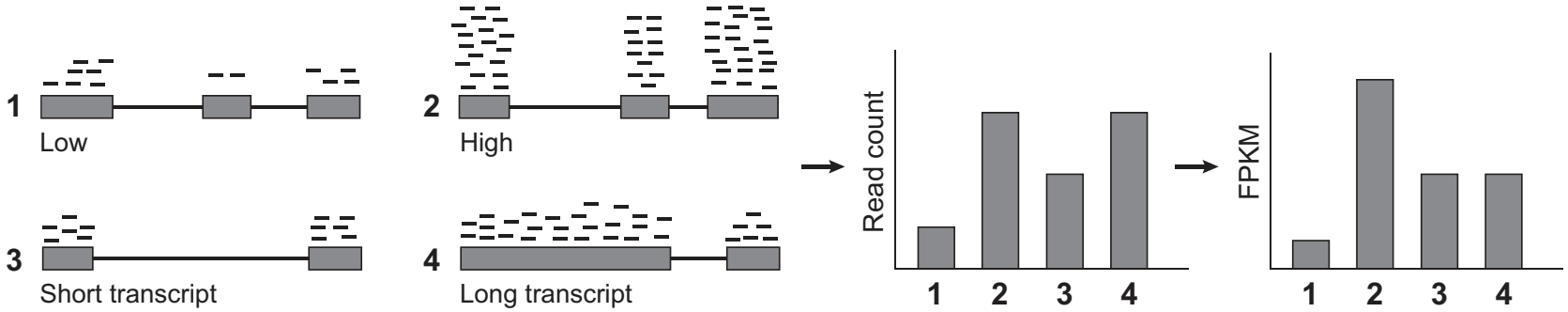


$$RPKM = 10^9 \frac{\#reads}{length \times TotalReads}$$

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 may be required for cross lane comparisons

Quantification: only one isoform



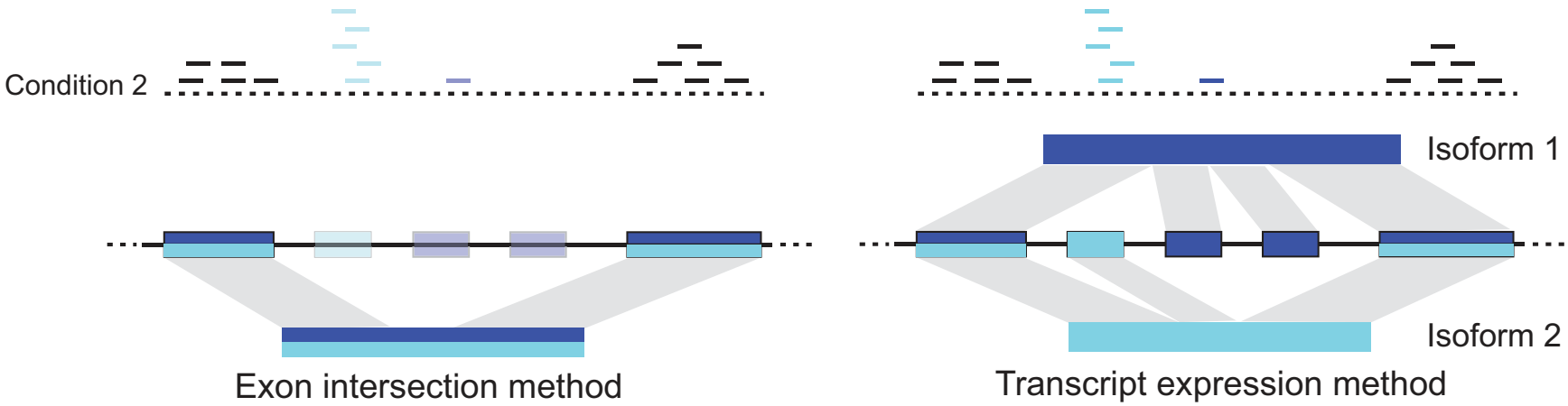
$$RPKM = 10^9 \frac{\#reads}{length \times TotalReads}$$

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 may be required for cross lane comparisons

This is all good when genes have one isoform.

Quantification: gene expression with multiple isoforms

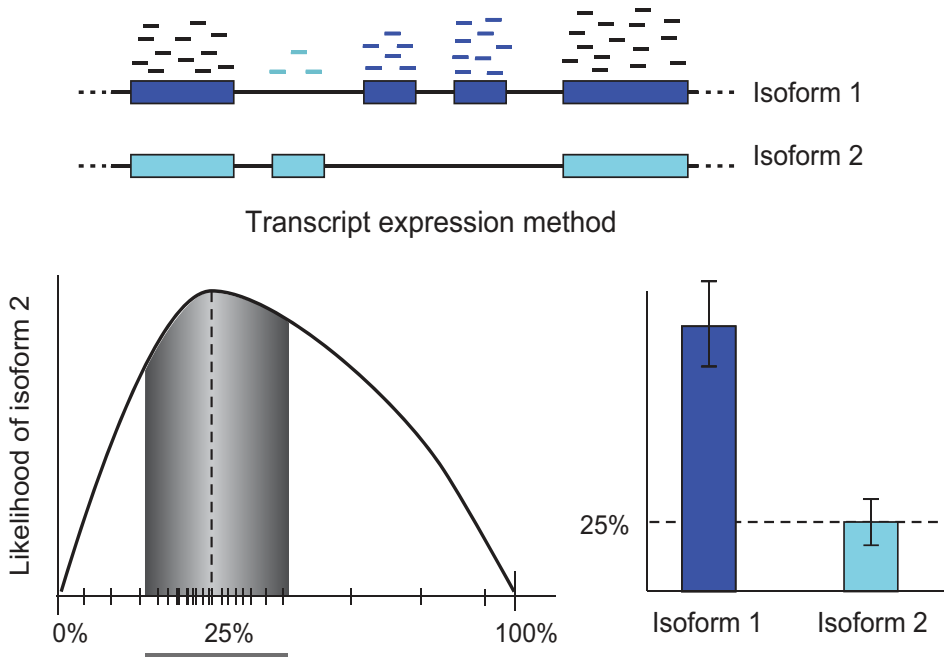


Exon intersection model: Score constituent exons

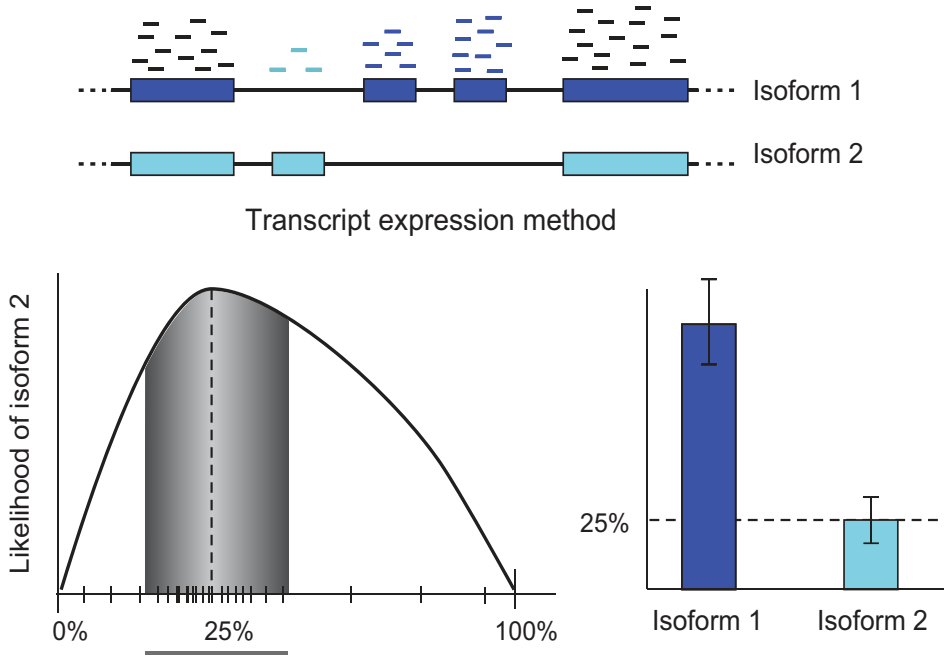
Exon union model: Score the the “merged” transcript

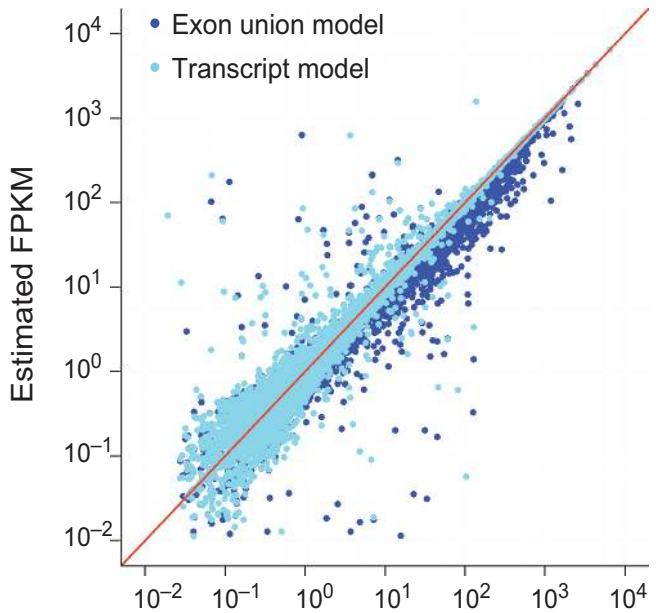
Transcript expression model: Assign reads uniquely to different isoforms. *Not a trivial problem!*

Quantification: read assignment method

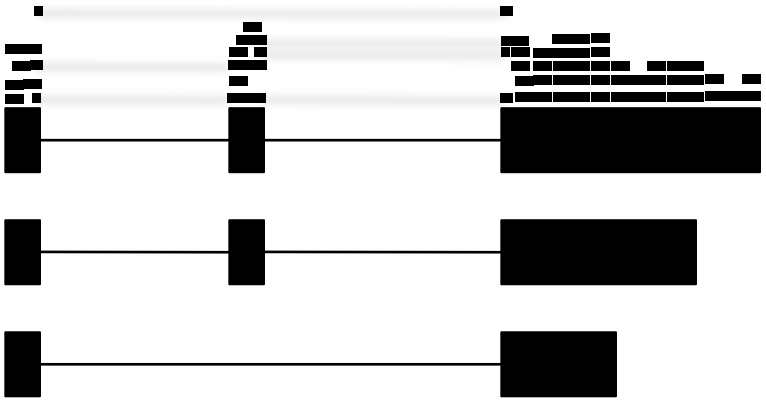


Quantification: read assignment method

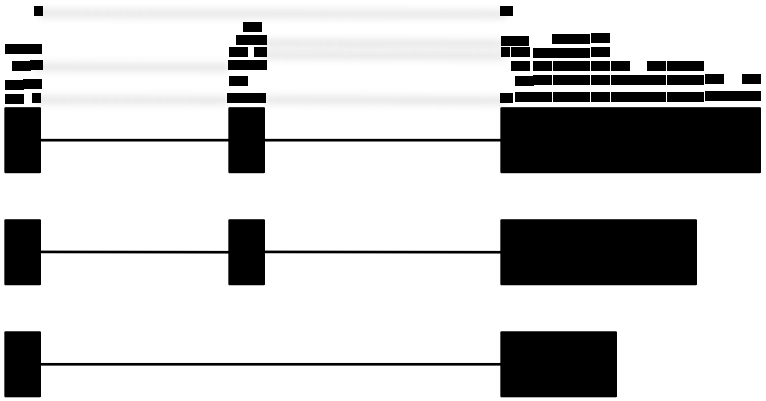




Quantification with multiple isoforms

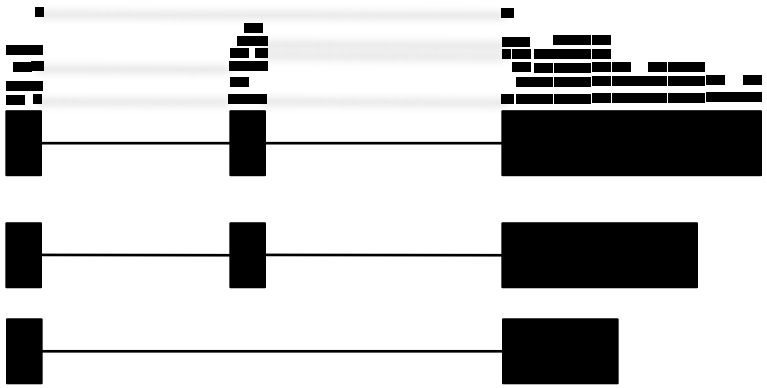


Quantification with multiple isoforms

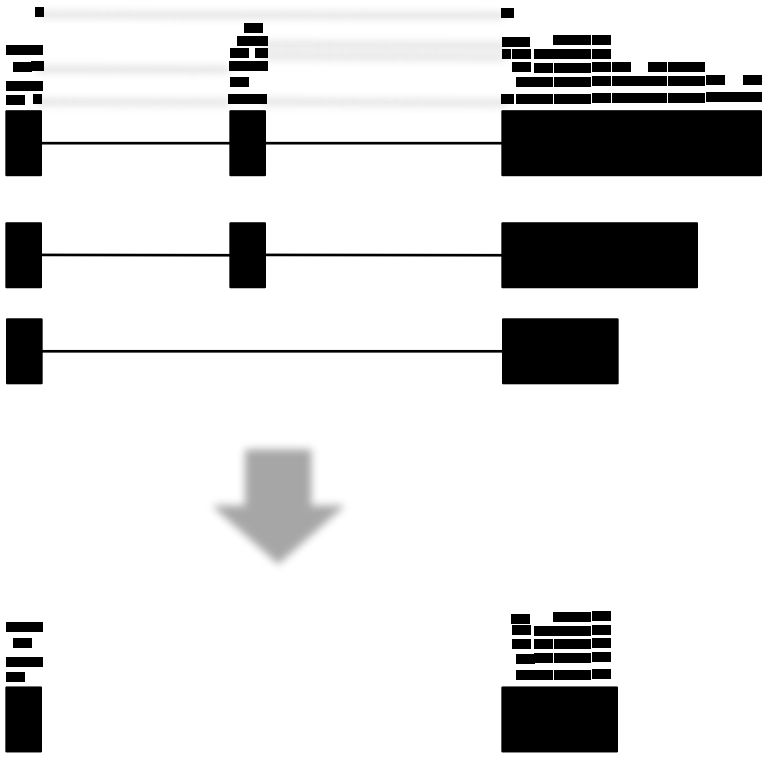


How do we define the gene expression?
How do we compute the expression of each isoform?

Computing gene expression

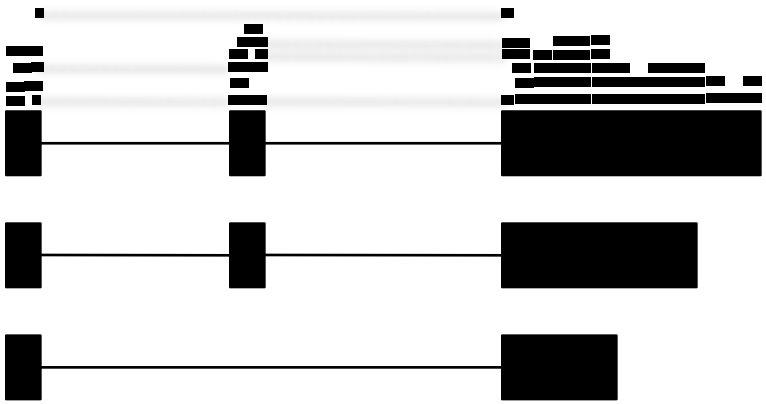


Computing gene expression

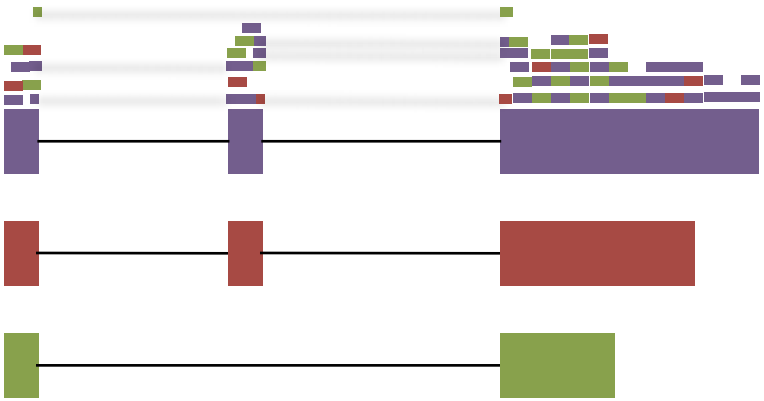


Idea1: RPKM of the constitutive reads (Neuma, Alexa-Seq, Scripture)

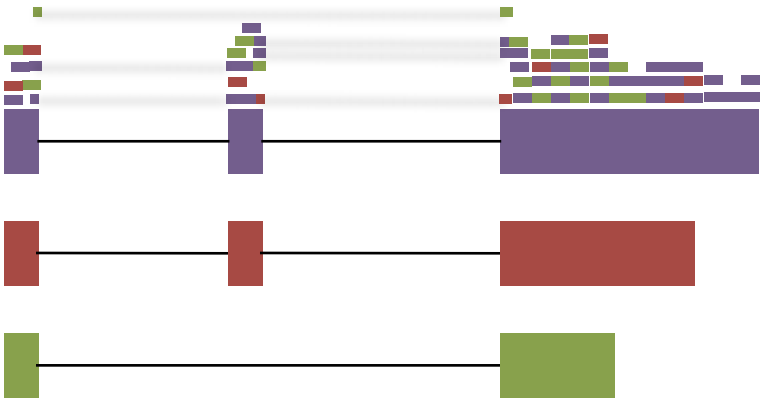
Computing gene expression — isoform deconvolution



Computing gene expression — isoform deconvolution

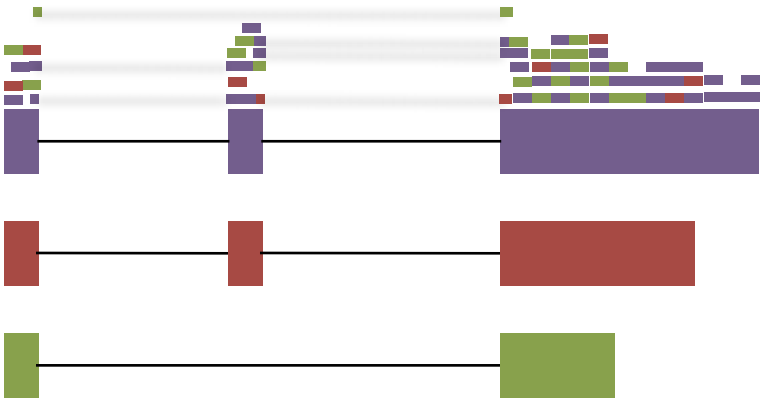


Computing gene expression — isoform deconvolution



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.

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.

$$E = RPKM_1 + RPKM_2 + RPKM_3$$

Programs to measure transcript expression

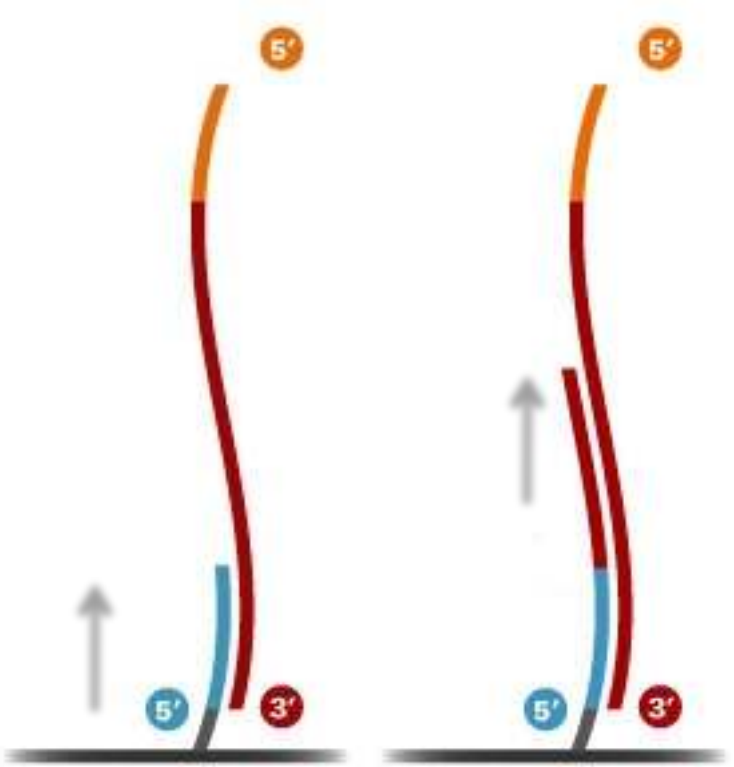
	Implemented method
Alexa-seq	Gene expression by constitutive exons
ERANGE	Gene expression by using all Exons
Scripture	Gene expression by constitutive exons
Cufflinks	**Transcript deconvolution by solving the maximum likelihood problem**
MISO	**Transcript deconvolution by solving the maximum likelihood problem**
RSEM	**Transcript deconvolution by solving the maximum likelihood problem**

Impact of library construction methods

Library construction improvements — Paired-end sequencing

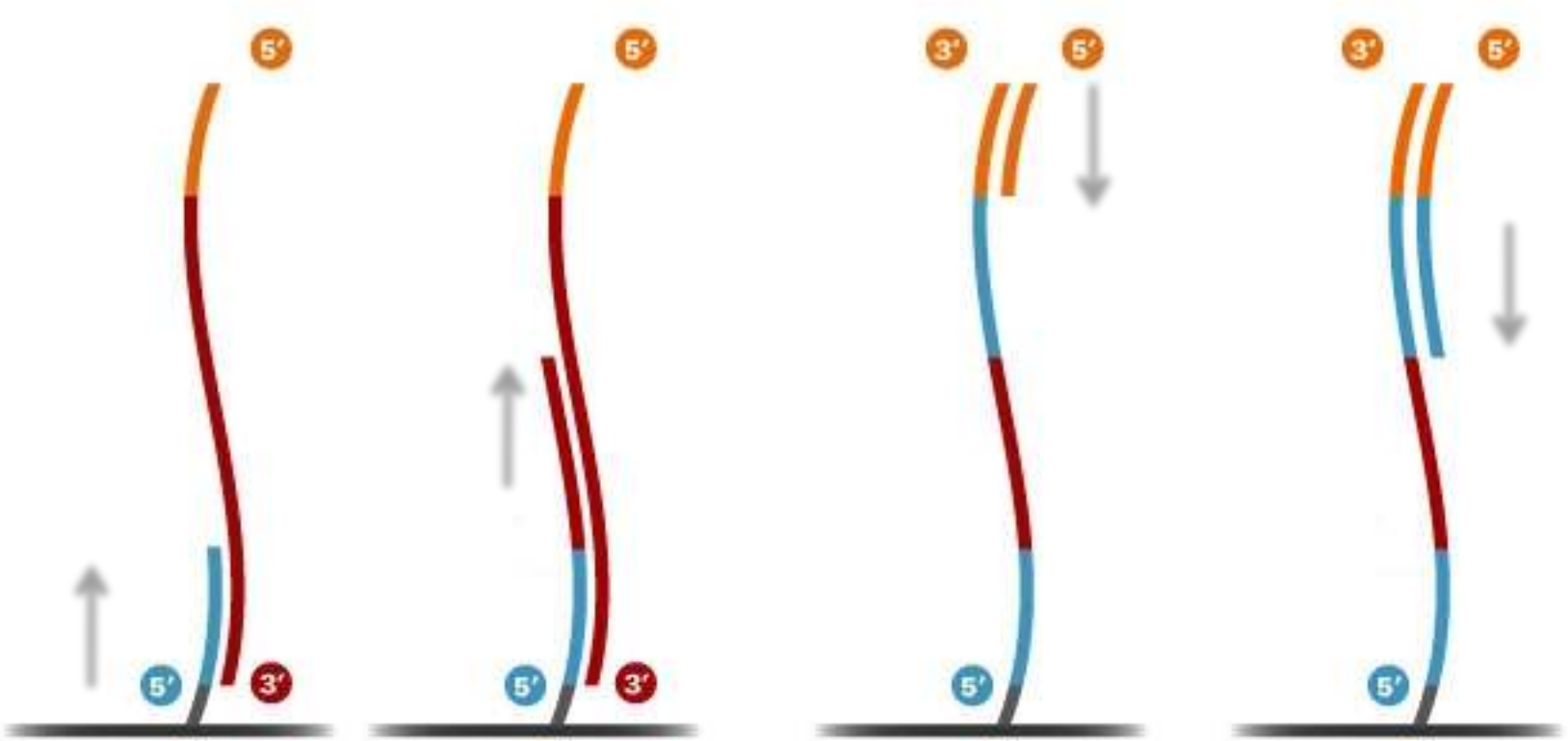
Adapted from the Helicos website

Library construction improvements — Paired-end sequencing



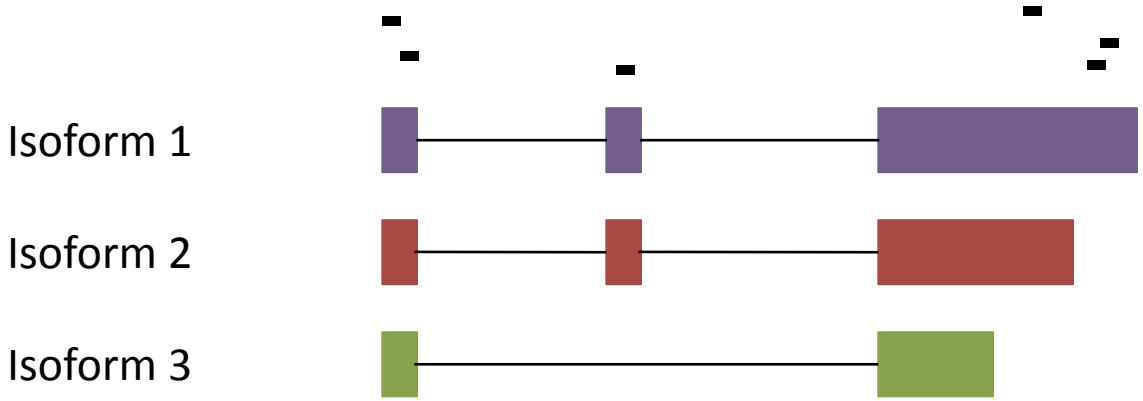
Adapted from the Helicos website

Library construction improvements — Paired-end sequencing

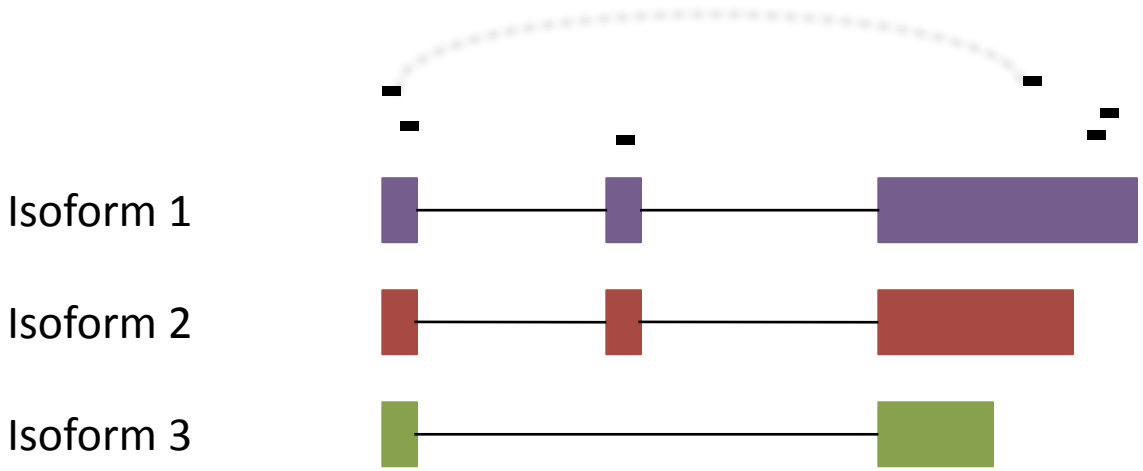


Adapted from the Helicos website

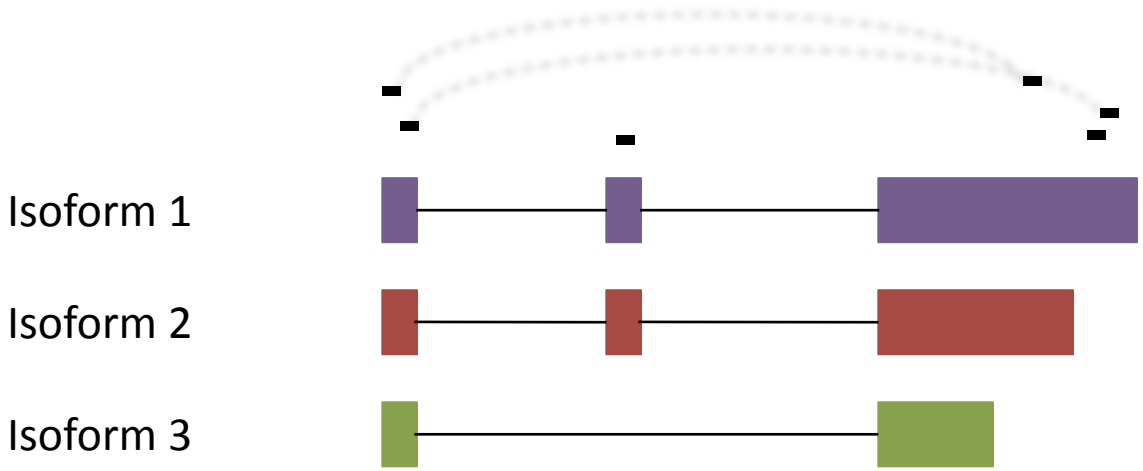
Paired-end reads are easier to associate to isoforms



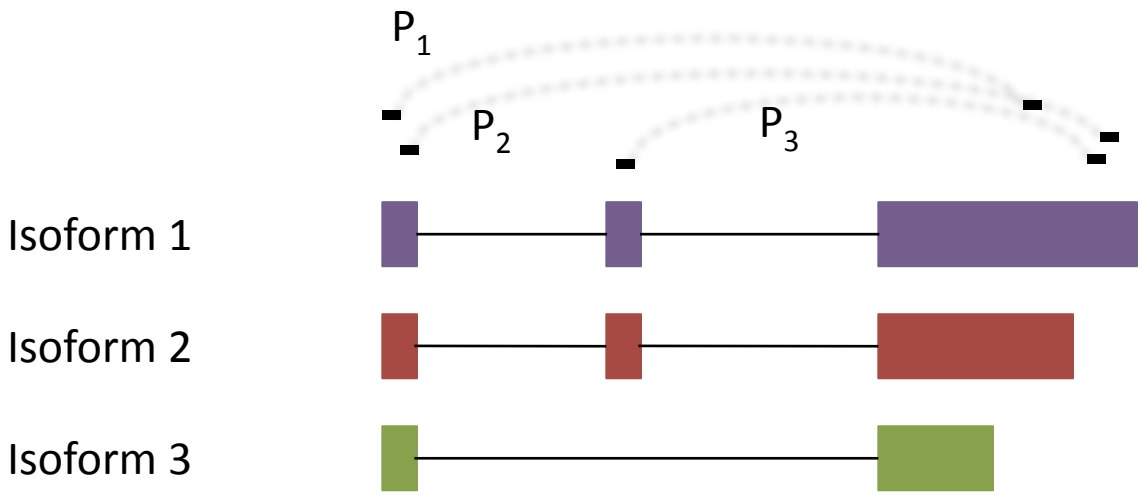
Paired-end reads are easier to associate to isoforms



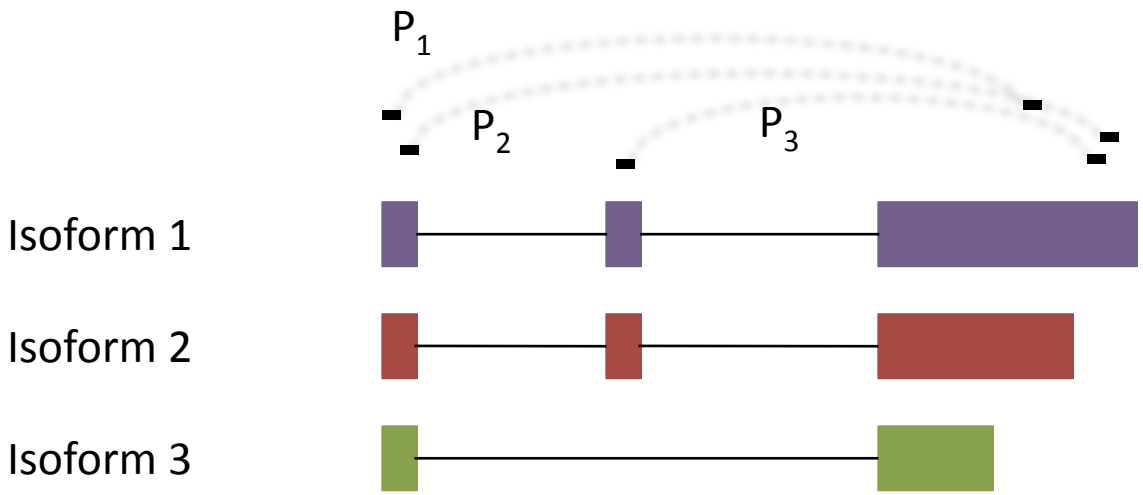
Paired-end reads are easier to associate to isoforms



Paired-end reads are easier to associate to isoforms

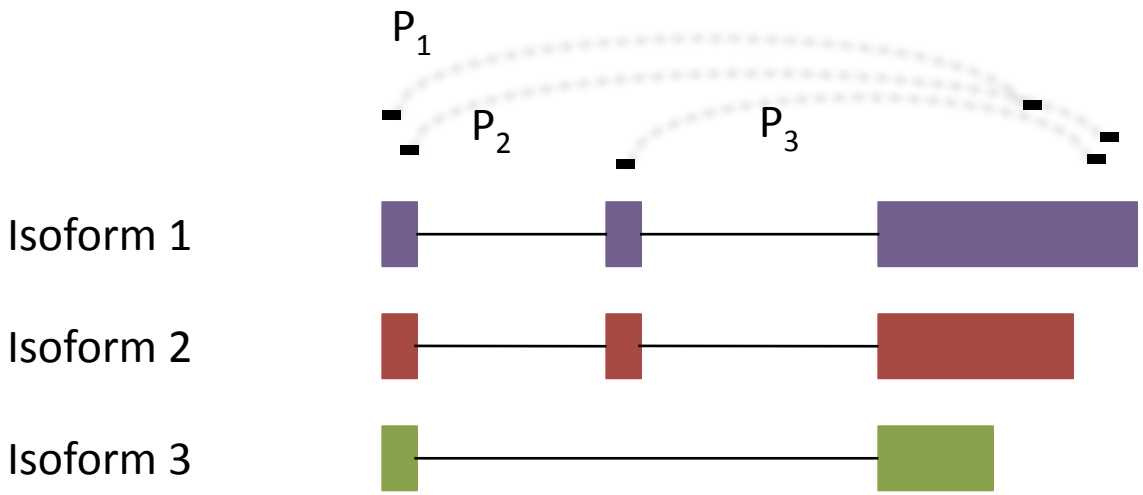


Paired-end reads are easier to associate to isoforms



Paired ends increase isoform deconvolution confidence

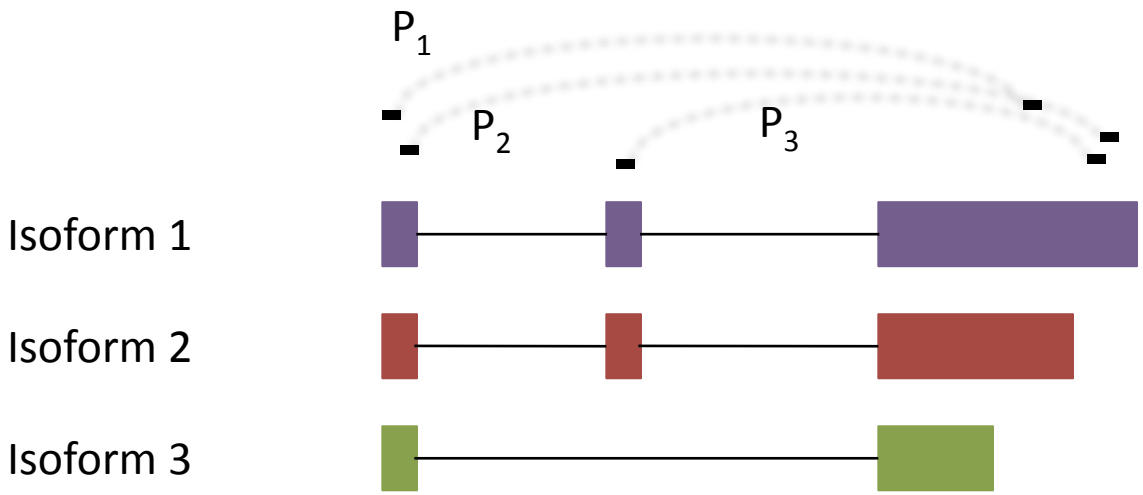
Paired-end reads are easier to associate to isoforms



Paired ends increase isoform deconvolution confidence

- P_1 originates from isoform 1 or 2 but not 3.

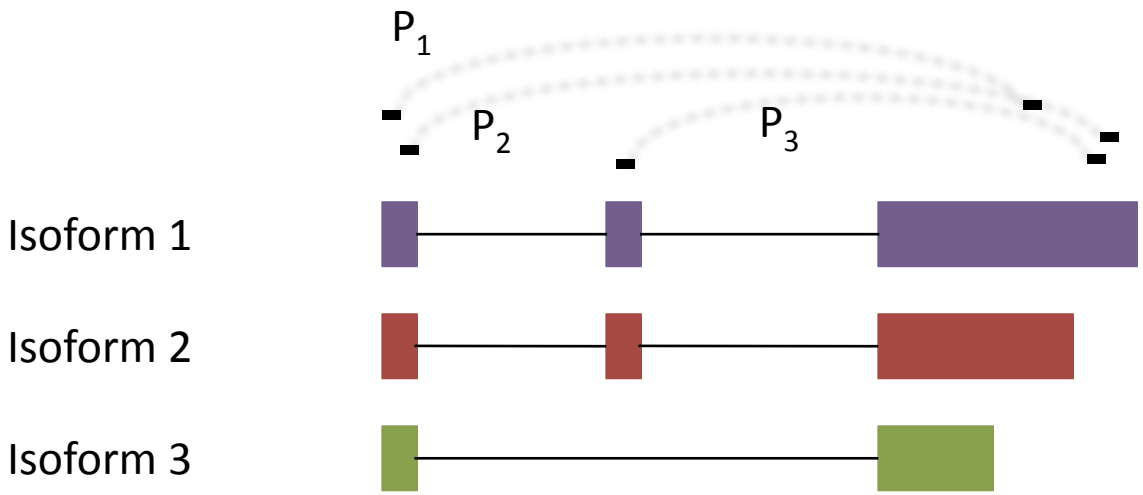
Paired-end reads are easier to associate to isoforms



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

Paired-end reads are easier to associate to isoforms

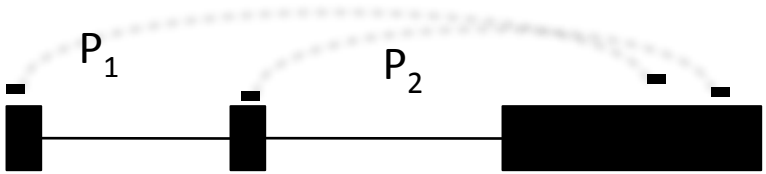


Paired ends increase isoform deconvolution confidence

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

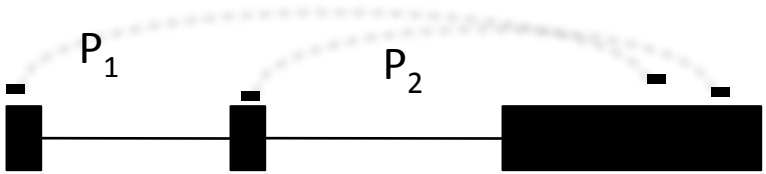
Do paired-end reads also help identifying reads originating in isoform 3?

We can estimate the insert size distribution



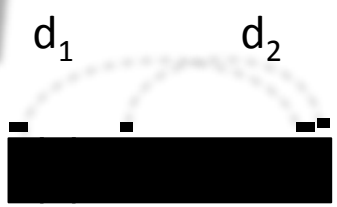
Get all single isoform reconstructions

We can estimate the insert size distribution

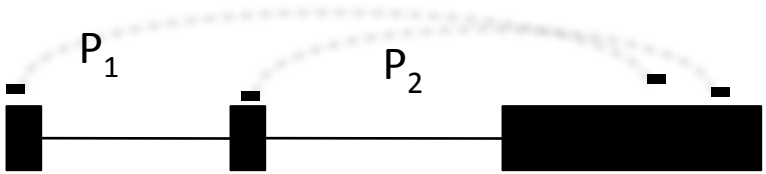


Get all single isoform reconstructions

Splice and compute insert distance

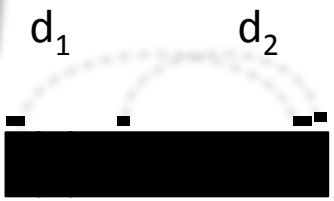


We can estimate the insert size distribution



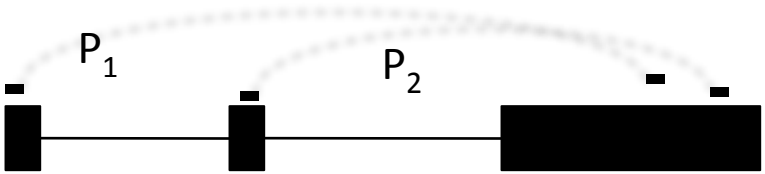
Get all single isoform reconstructions

Splice and compute insert distance



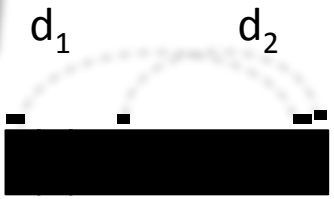
Estimate insert size empirical distribution

We can estimate the insert size distribution

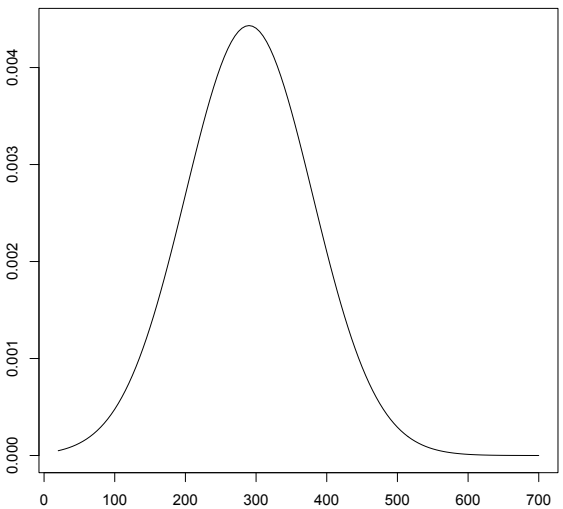


Get all single isoform reconstructions

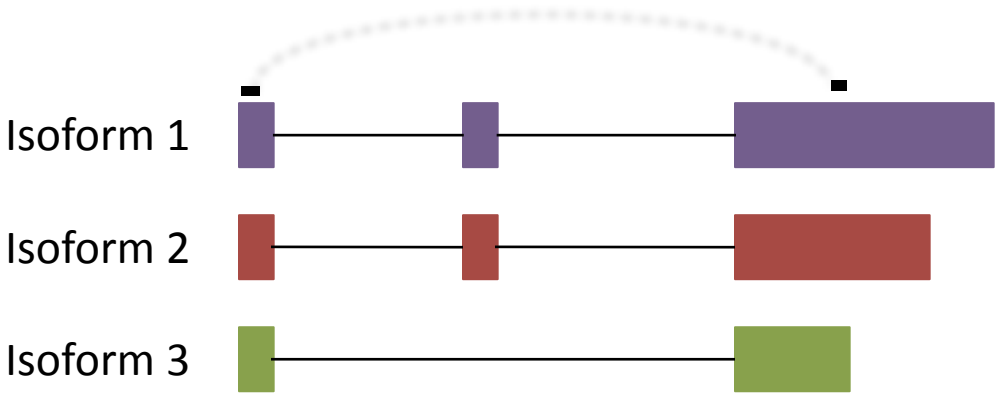
Splice and compute insert distance



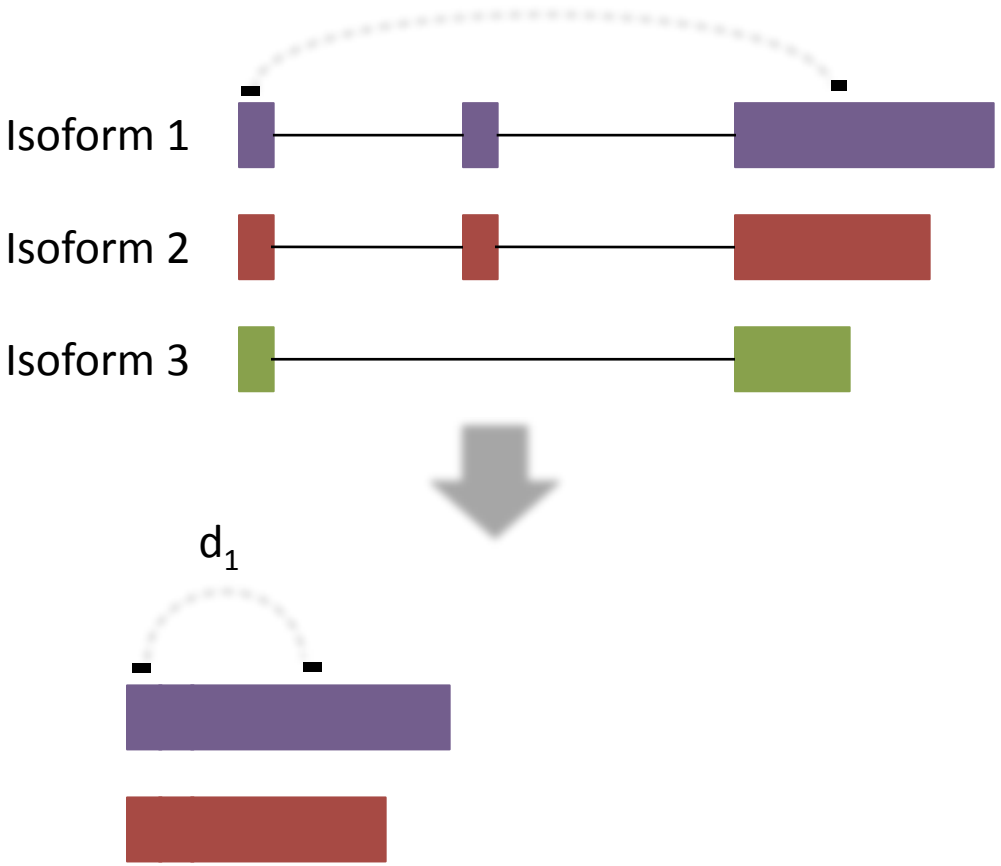
Estimate insert size empirical distribution



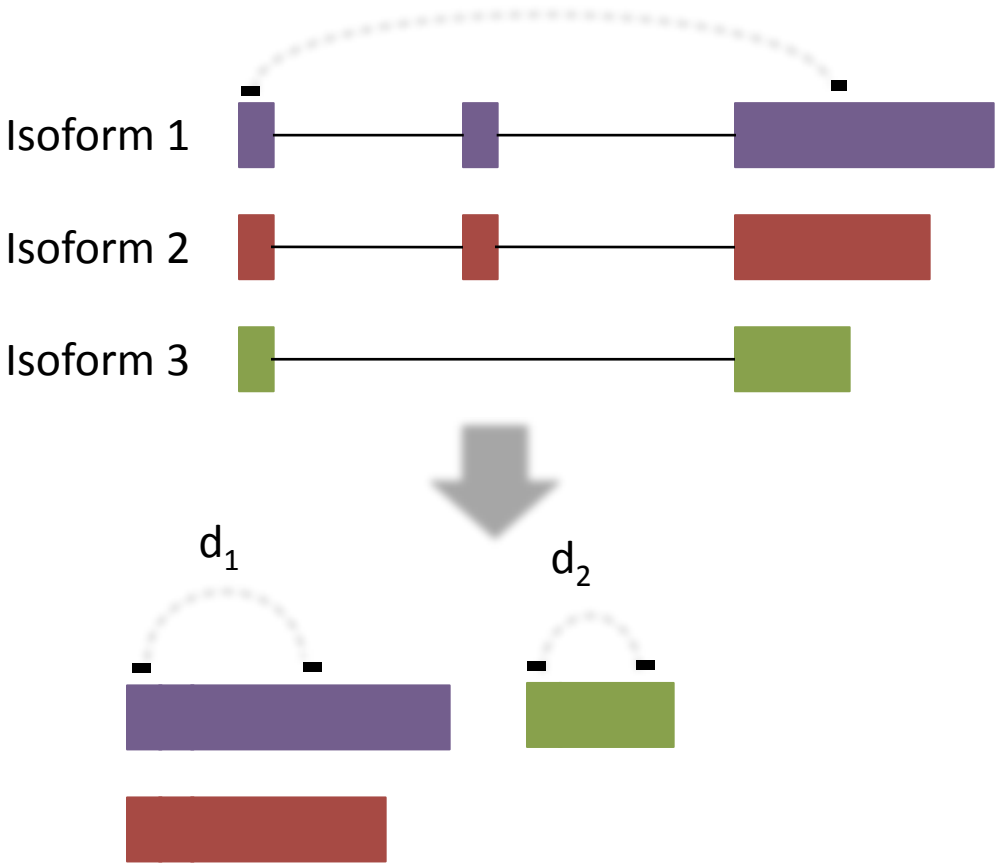
… and use it for probabilistic read assignment



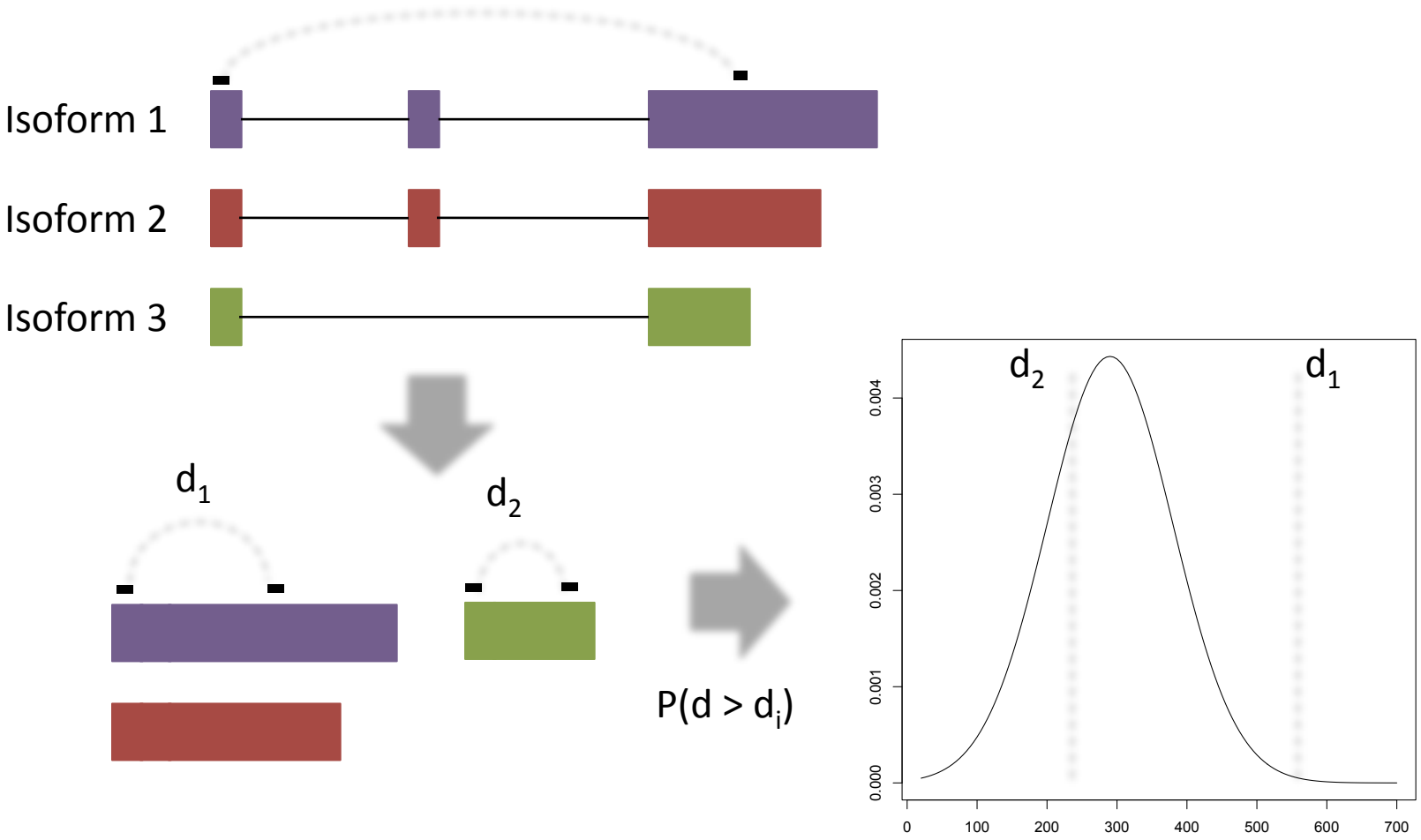
… and use it for probabilistic read assignment



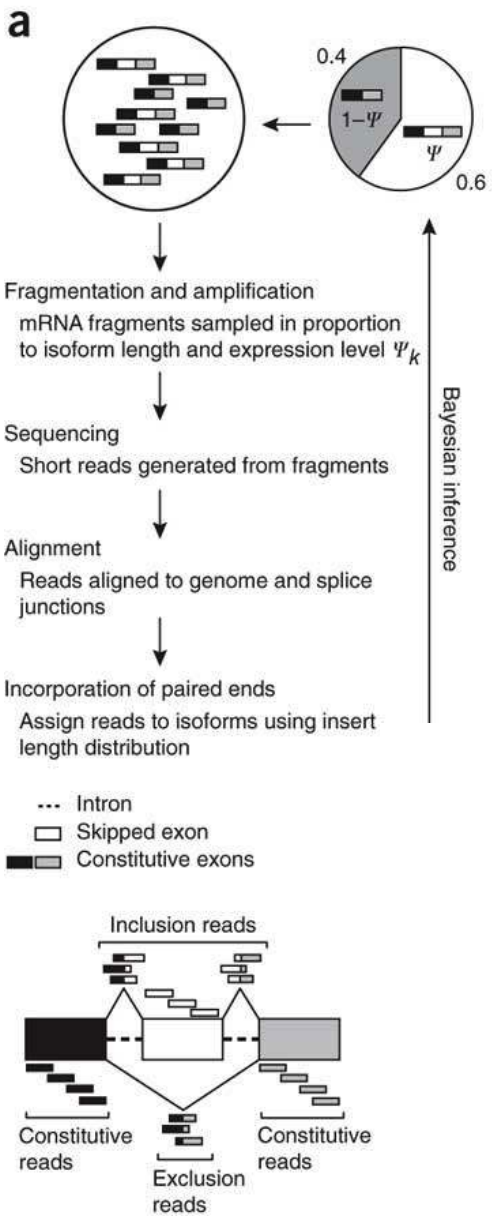
… and use it for probabilistic read assignment



… and use it for probabilistic read assignment

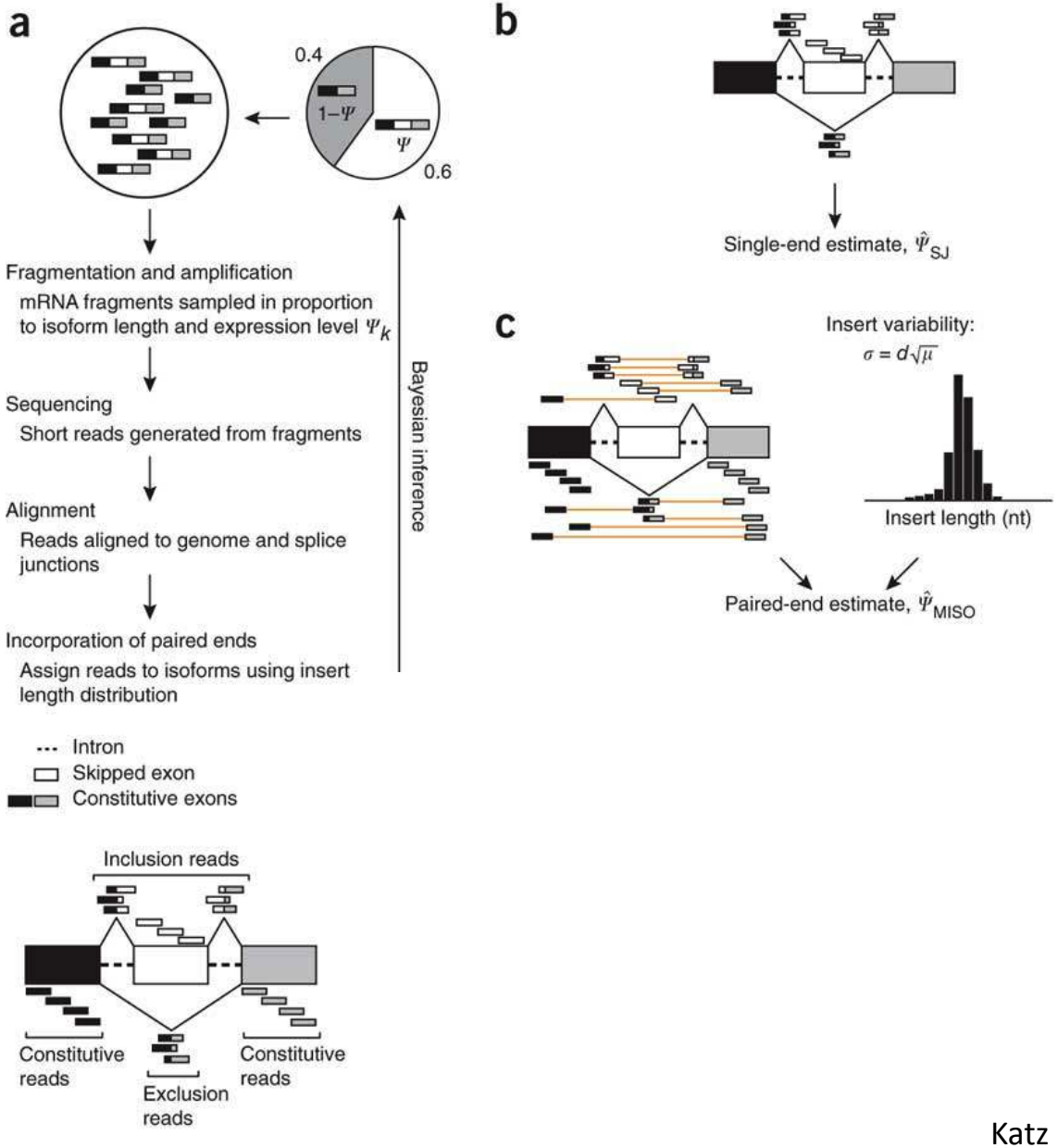


And improve quantification



Katz et al Nature Methods 2008

And improve quantification

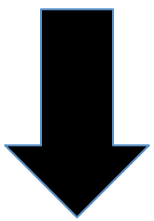


Katz et al Nature Methods 2008

Quantification with paired ends (FPKM)

Cufflinks leverages paired ends to quantify fragments rather than raw reads. The extension of RPKM.

$$RPKM = 10^9 \frac{\#\,reads}{length \times Total\,\mathrm{Re}\,ads}$$



$$FPKM = 10^9 \frac{\#\,fragments}{length \times TotalFragments}$$

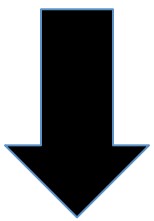
Fragments per kilobase of exonic sequence per million mapped fragments

(*Trapnel* et al Nature Biotechnology 2010)

Quantification with paired ends (FPKM)

Cufflinks leverages paired ends to quantify fragments rather than raw reads. The extension of RPKM.

$$RPKM = 10^9 \frac{\#\,reads}{length \times Total\,\mathrm{Re}\,ads}$$

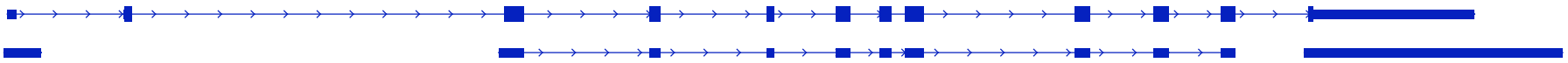


$$FPKM = 10^9 \frac{\#\,fragments}{length \times TotalFragments}$$

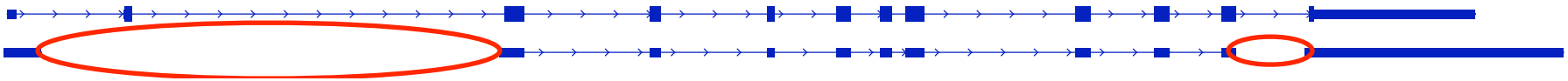
Fragments per kilobase of exonic sequence per million mapped fragments
(*Trapnel* et al Nature Biotechnology 2010)

paired-end reads improve quantification accuracy

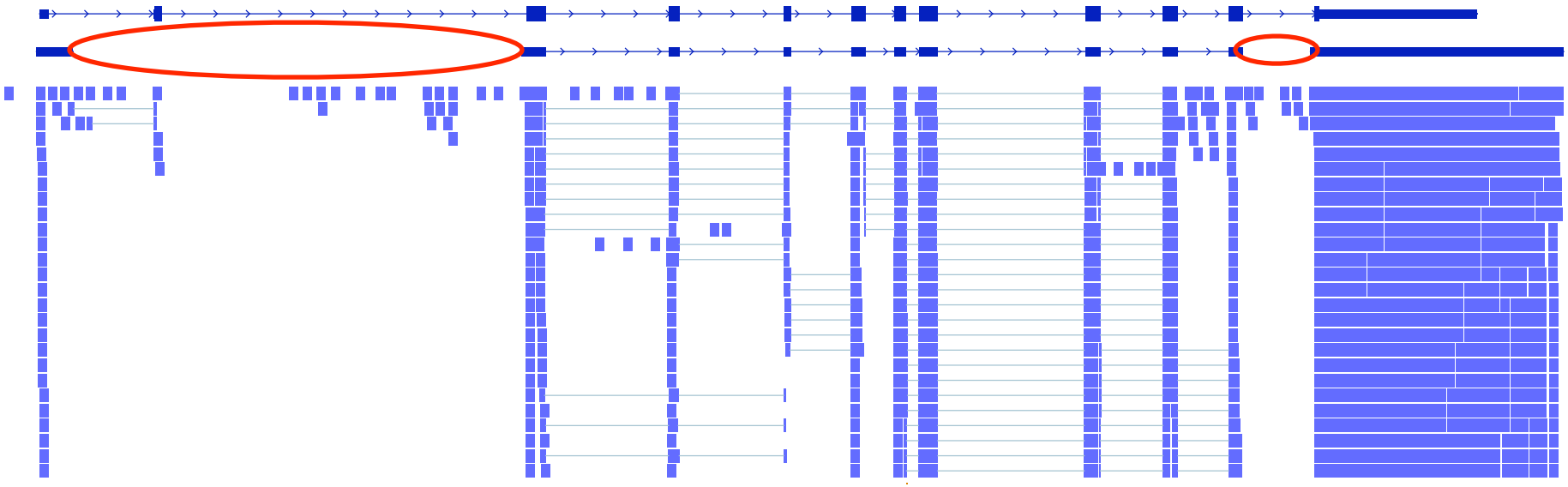
Paired-end improve reconstructions



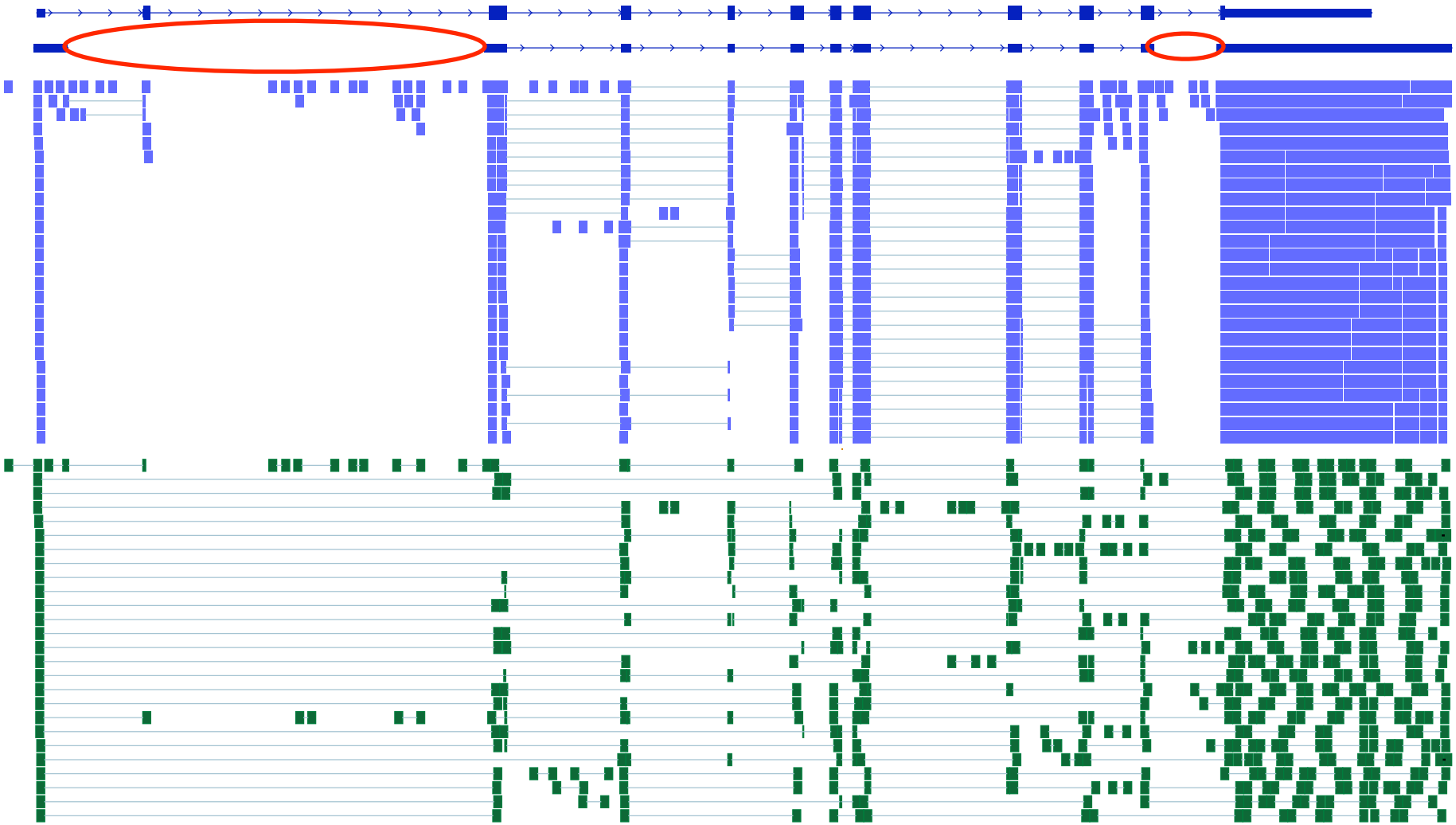
Paired-end improve reconstructions



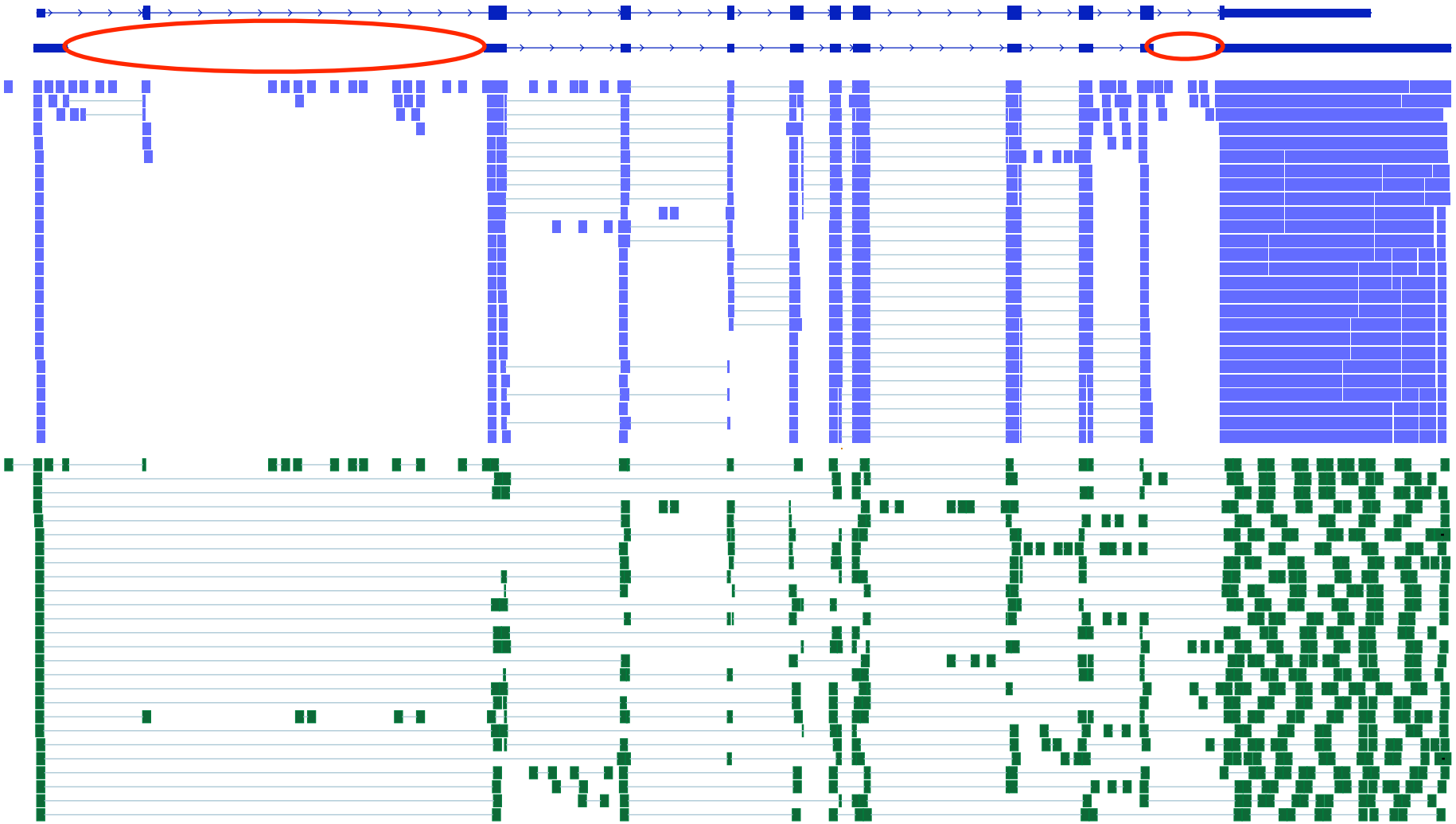
Paired-end improve reconstructions



Paired-end improve reconstructions



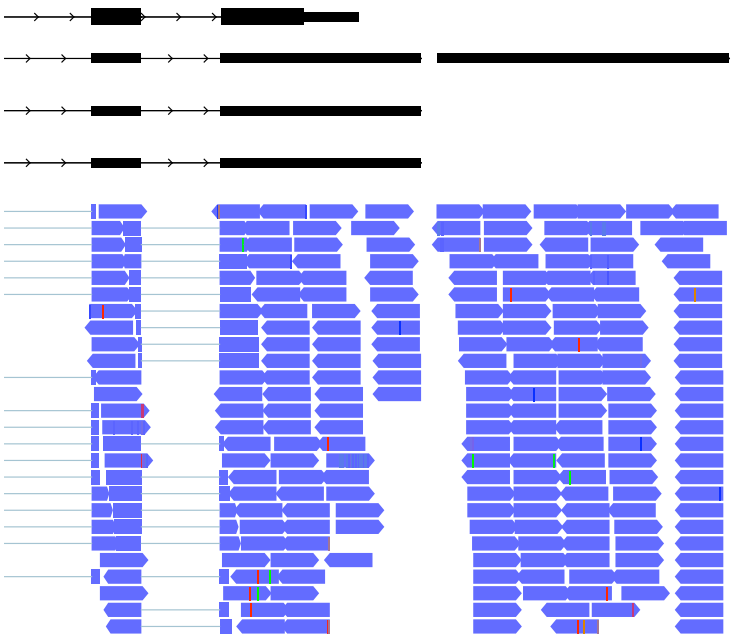
Paired-end improve reconstructions



Paired-end data complements the connectivity graph

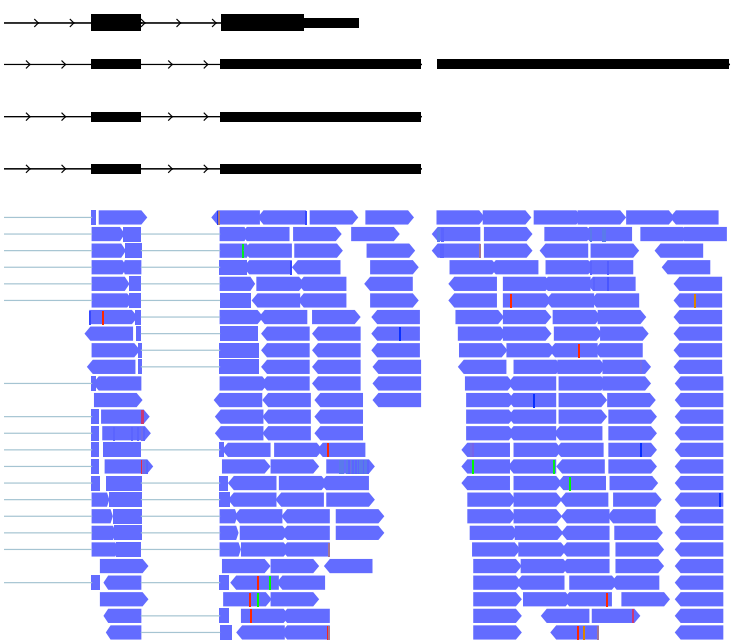
And merge regions

Single reads



And merge regions

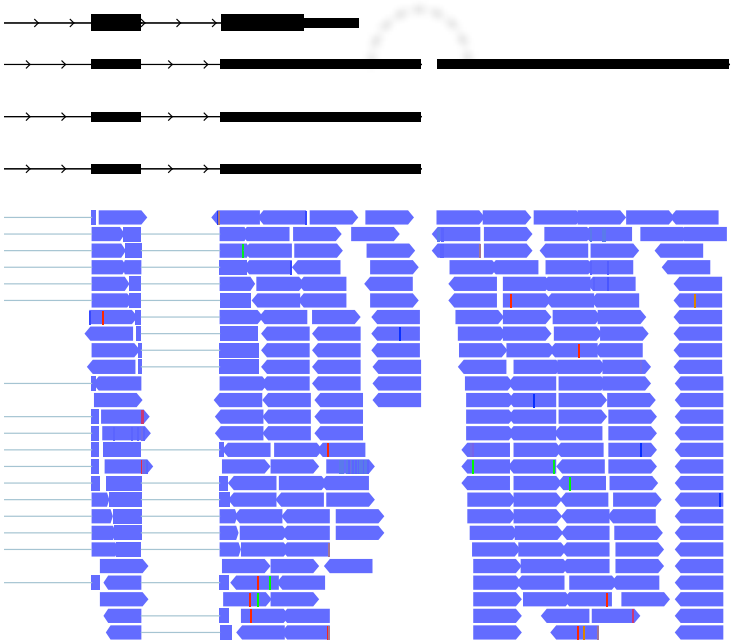
Single reads



Paired reads

And merge regions

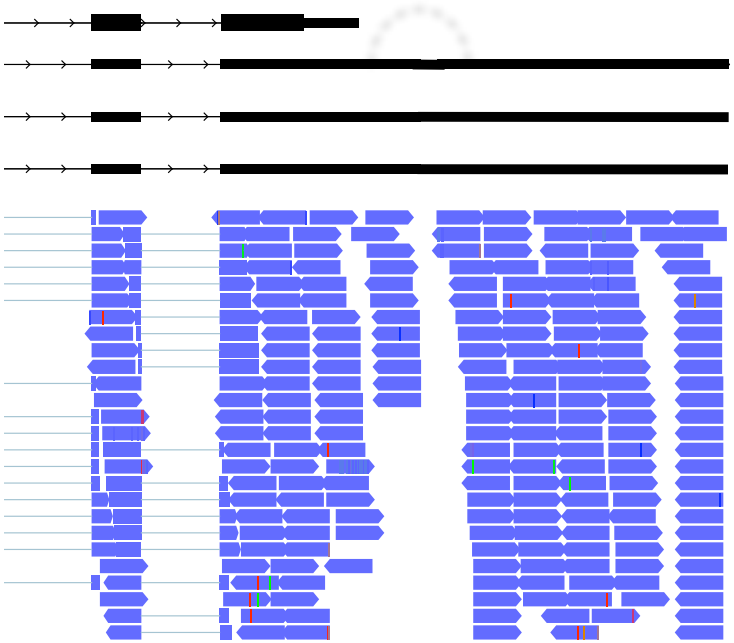
Single reads



Paired reads

And merge regions

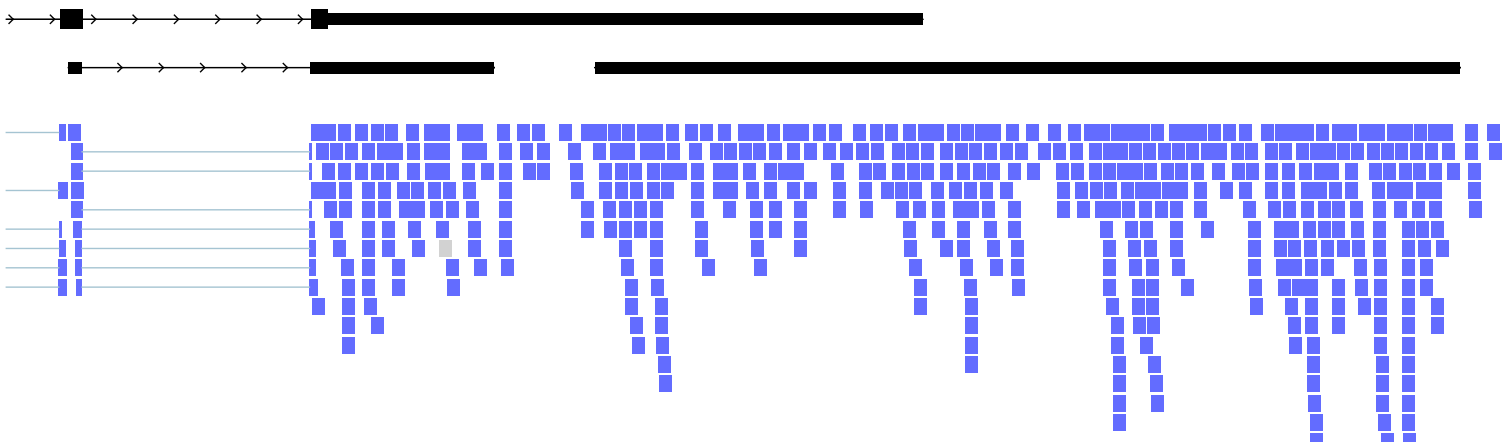
Single reads



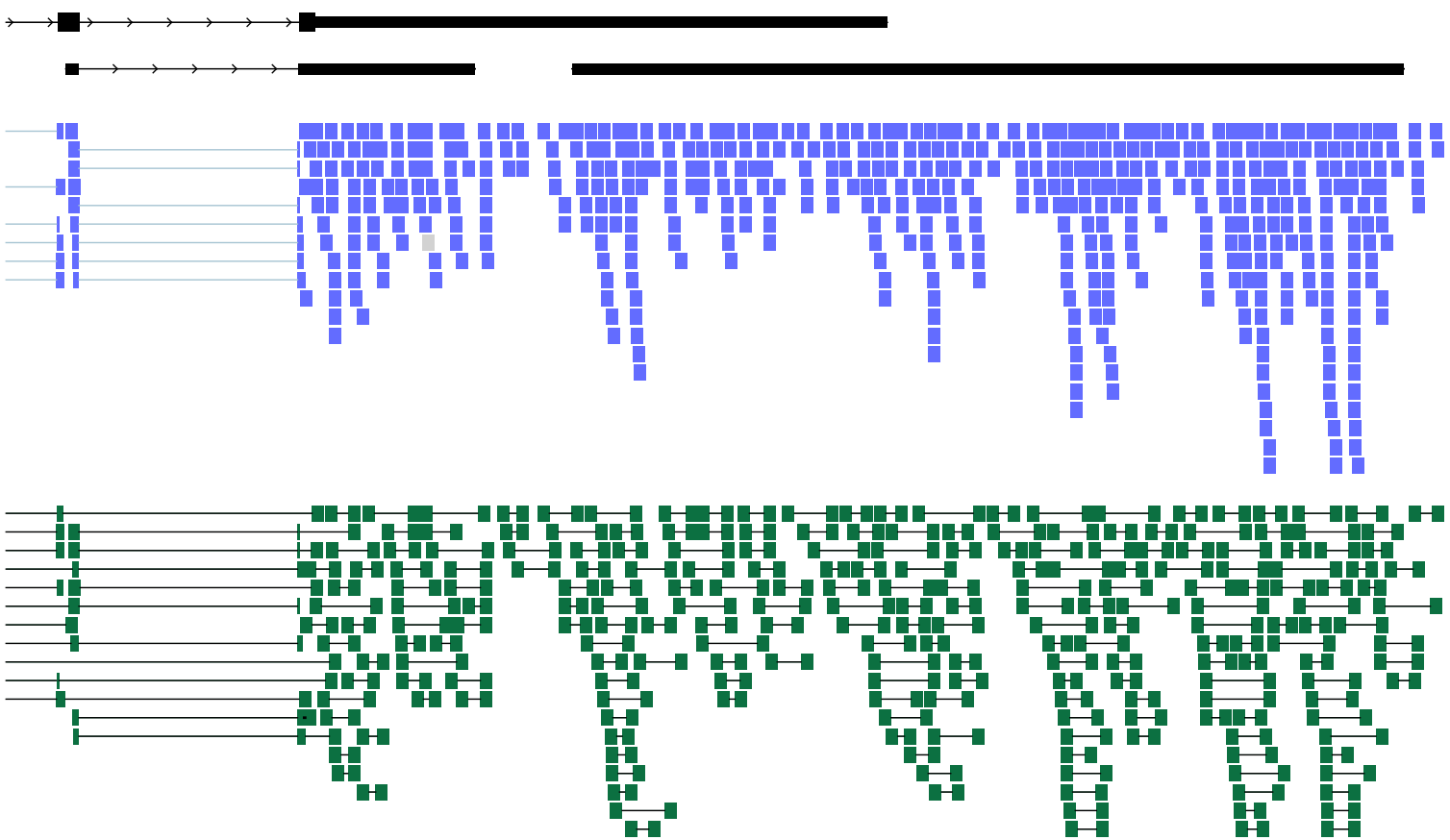
Paired reads

Or split regions

Single reads



Or split regions



Summary

Summary

- Paired-end reads are now routine in Illumina and SOLiD sequencers.

Summary

- Paired-end reads are now routine in Illumina and SOLiD sequencers.
- Paired end alignment is supported by most short read aligners

Summary

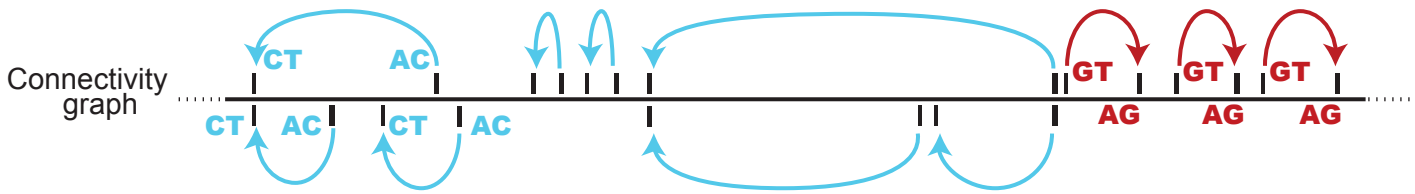
- Paired-end reads are now routine in Illumina and SOLiD sequencers.
- Paired end alignment is supported by most short read aligners
- Transcript quantification depends heavily in paired-end data

Summary

- Paired-end reads are now routine in Illumina and SOLiD sequencers.
- Paired end alignment is supported by most short read aligners
- Transcript quantification depends heavily in paired-end data
- Transcript reconstruction is greatly improved when using paired-ends (work in progress)

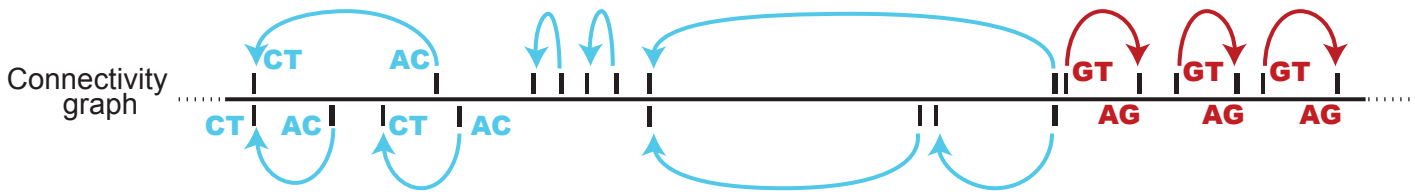
Giving orientation to transcripts — Strand specific libraries

Scripture relies on splice motifs to orient transcripts. It orients every edge in the connectivity graph.



Giving orientation to transcripts — Strand specific libraries

Scripture relies on splice motifs to orient transcripts. It orients every edge in the connectivity graph.



?

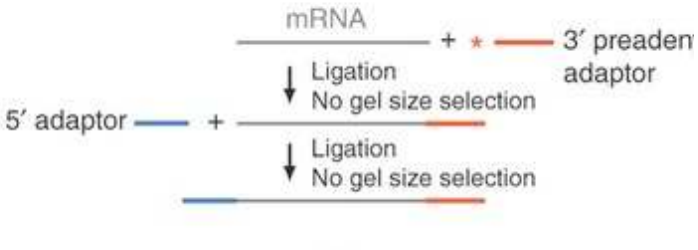
Single exon genes are left unoriented

Strand specific library construction results in oriented reads.

Illumina RNA ligation

3′ preadenylated adaptors and 5′ adaptors ligated sequentially to RNA without cleanup

(S. Luo and G. Schroth, personal communication)



Sequence depends on the adapters ligated

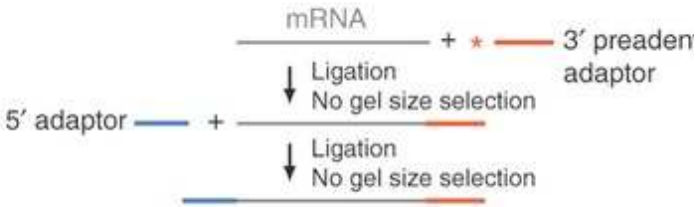
Adapted from Levine et al Nature Methods

Strand specific library construction results in oriented reads.

Illumina RNA ligation

3′ preadenylated adaptors and 5′ adaptors ligated sequentially to RNA without cleanup

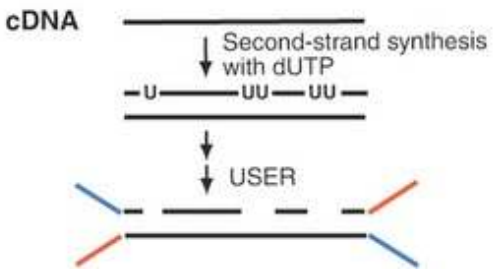
(S. Luo and G. Schroth, personal communication)



Sequence depends on the adapters ligated

dUTP second strand[13]

Second-strand synthesis with dUTP; remove 'U's after adaptor ligation and size selection



The second strand is destroyed, thus the cDNA read is always in reverse orientation to the RNA

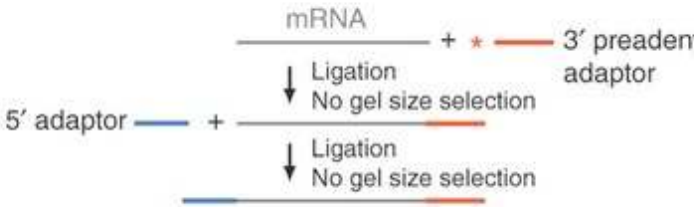
Adapted from Levine et al Nature Methods

Strand specific library construction results in oriented reads.

Illumina RNA ligation

3′ preadenylated adaptors and 5′ adaptors ligated sequentially to RNA without cleanup

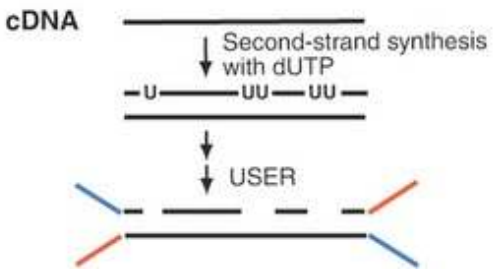
(S. Luo and G. Schroth, personal communication)



Sequence depends on the adapters ligated

dUTP second strand[13]

Second-strand synthesis with dUTP; remove 'U's after adaptor ligation and size selection



The second strand is destroyed, thus the cDNA read is always in reverse orientation to the RNA

Adapted from Levine et al Nature Methods

Scripture & Cufflinks allow the user to specify the orientation of the reads.

The libraries we will work with are strand sepcific

Summary

Summary

- Several methods now exist to build strand sepecific RNA-Seq libraries.

Summary

- Several methods now exist to build strand sepecific RNA-Seq libraries.
- Quantification methods support strand specific libraries. For example Scripture will compute expression on both strand if desired.

Overview of the session

The 3 main computational challenges of sequence analysis for *counting applications:*

- Read mapping: Placing short reads in the genome
- Reconstruction: Finding the regions that originate the reads
- Quantification:
    - Assigning scores to regions
    - Finding regions that are differentially represented between two or more samples.

The problem.

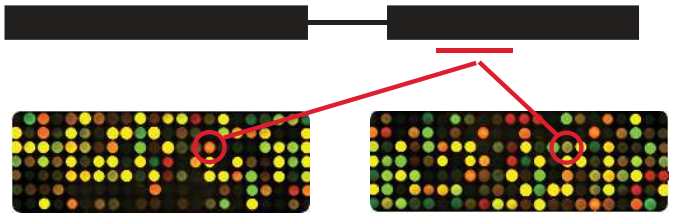
The problem.

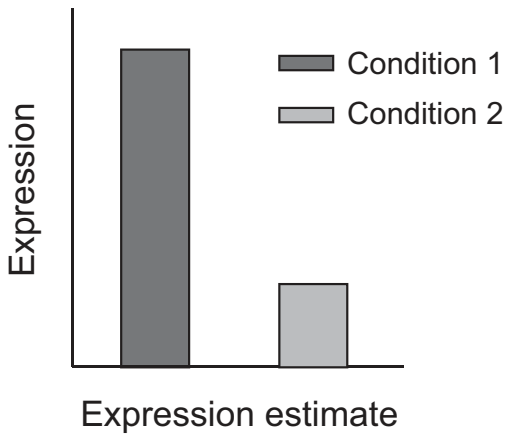
- Finding genes that have different expression between two or more conditions.

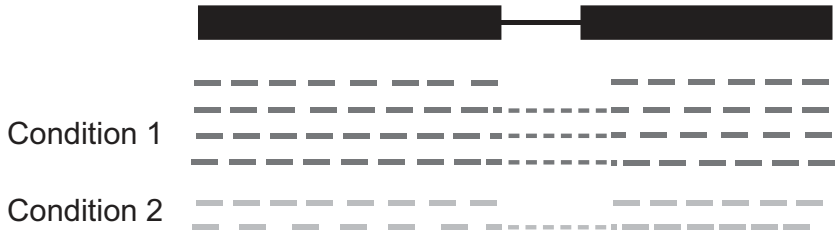
The problem.

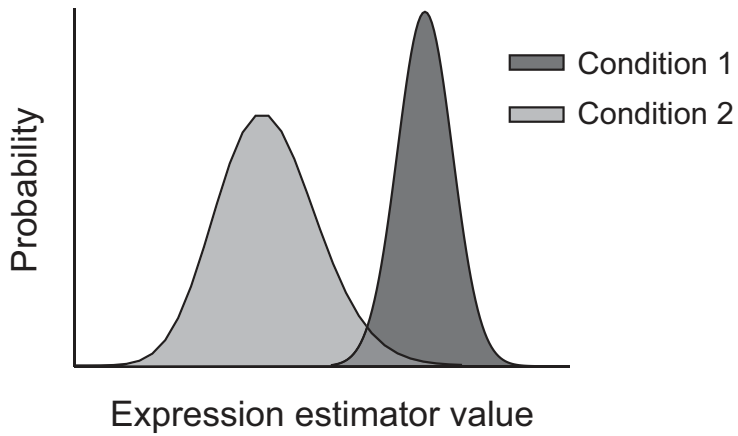
- Finding genes that have different expression between two or more conditions.
- Find gene with isoforms expressed at different levels between two or more conditions.
  - Find differentially used slicing events
  - Find alternatively used transcription start sites
  - Find alternatively used 3' UTRs

Differential gene expression using RNA-Seq

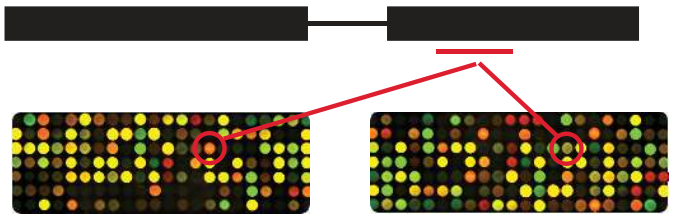


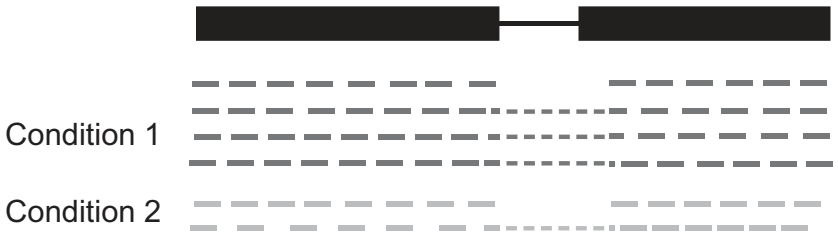


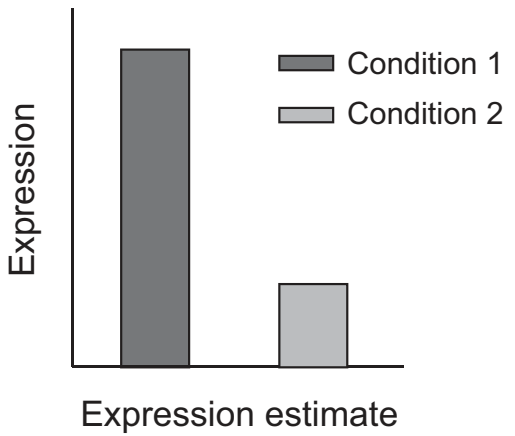


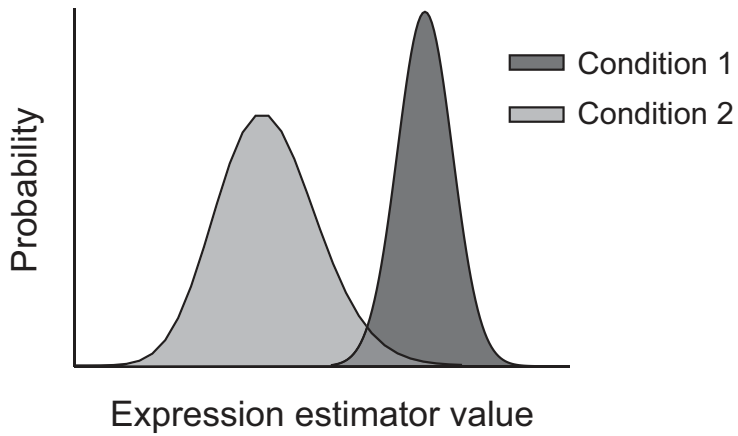


Differential gene expression using RNA-Seq









- (Normalized) read counts ←→ Hybridization intensity

Differential analysis strategies

- Use read counts
  - Standard Fisher exact (no preplicates) or χ^2 test (replicates)

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

Differential analysis strategies

- Use read counts
  - Standard Fisher exact (no preplicates) or χ^2 test (replicates)

	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

Differential analysis strategies

- Use read counts
  - Standard Fisher exact (no preplicates) or χ^2 test (replicates)

	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 a summary statistic and "standard" array

Cufflinks differential issoform ussage

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: $P_1 \mathrel{!}= P_2$ 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	Poisson	Gene read counts table
Myrna	Empirical	Sequence reads and reference transcriptome

Digital gene expression

If all you want is the expression level

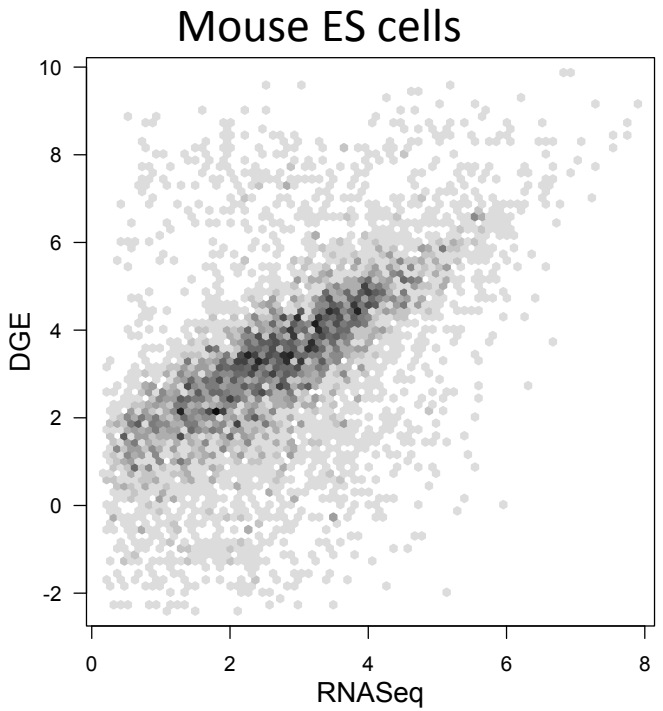
Easy

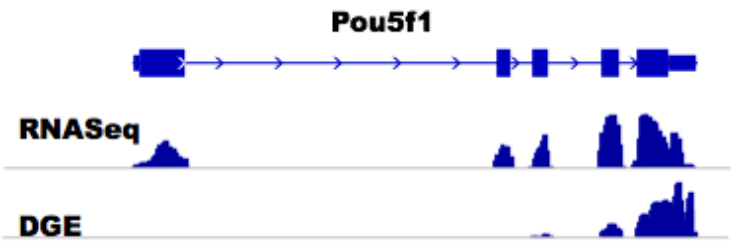
- Fragment RNA (heat)
- PolyA select -> RT -> 2nd strand
- Amplify
- Sequence

Cheap

RNASeq requires 100 mill reads.
DGE requires ~6-10 mill reads.

No size bias





Digital gene expression

If all you want is the expression level

Easy

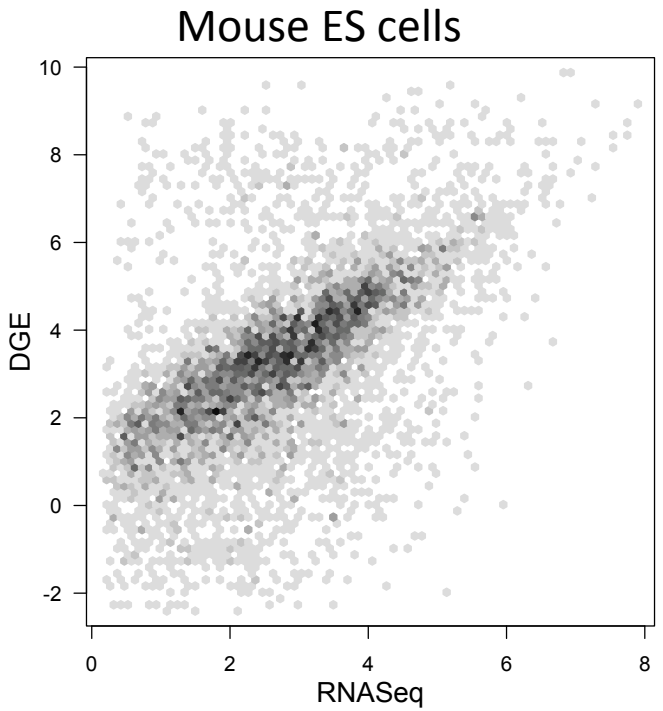
- Fragment RNA (heat)
- PolyA select -> RT -> 2nd strand
- Amplify
- Sequence

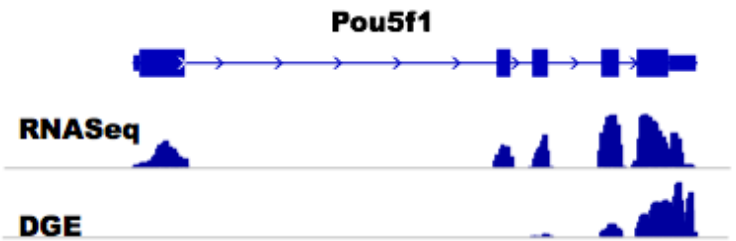
Cheap

RNASeq requires 100 mill reads.

DGE requires ~6-10 mill reads.

No size bias





Replicates will be natural and analysis standard

Acknowledgements

Mitchell Guttman

New Contributors:
Moran Cabili
Hayden Metsky

RNA-Seq analysis:
Cole Trapnel
Manfred Grabher

Dendritic Cells
Ido Amit
Nir Yosef
Raktima Raychowdhury

Could not do it without the support of:
Eric Lander
Aviv Regev