RNASeq Analysis

RNA-seq can be used to build a complete map of the transcriptome across all cell types, perturbations and states.

Computational methods that will be reported are indepent of the choice of library construtuction protocols. Note that, the «paired-end» sequencing is an additional features crucial to provide valuable information.

Computational methods needed to address RNA-seq analysis core challenges

RNASeq Analysis

GOAL: *map RNA-seq reads to a reference transcriptome*

methods to align reads directly to a reference transcriptome or genome (‘read mapping’).

- **unspliced read aligners**:
  - **seed methods**
  - **Burrows-Wheeler transform methods**
- **spliced aligners:**
  - **exon first**
  - **seed andextend**.

GOAL: *reconstruct the transcriptome*

methods to identify expressed genes and isoforms (‘transcriptome reconstruction’).

- ***genome-guided***

  exon identification

  genome-guided assembly

- ***genome- independent***’

GOAL*:* *quantify gene expression*

methods for estimation of gene and isoform abundance, as well as methods for the analysis of differential expression across samples (‘expression quantification’).

- ***exon intersection method***
- ***exon union method’***

Garber et al Nature Methods 2011

Mapping short RNA-seq reads

RNA-seq reads:

- are short (~36–125 bases)
- error rates are considerable
- many reads span exon-exon junctions.
- the number of reads per experiment is increasingly large, currently as many as hundreds of millions.

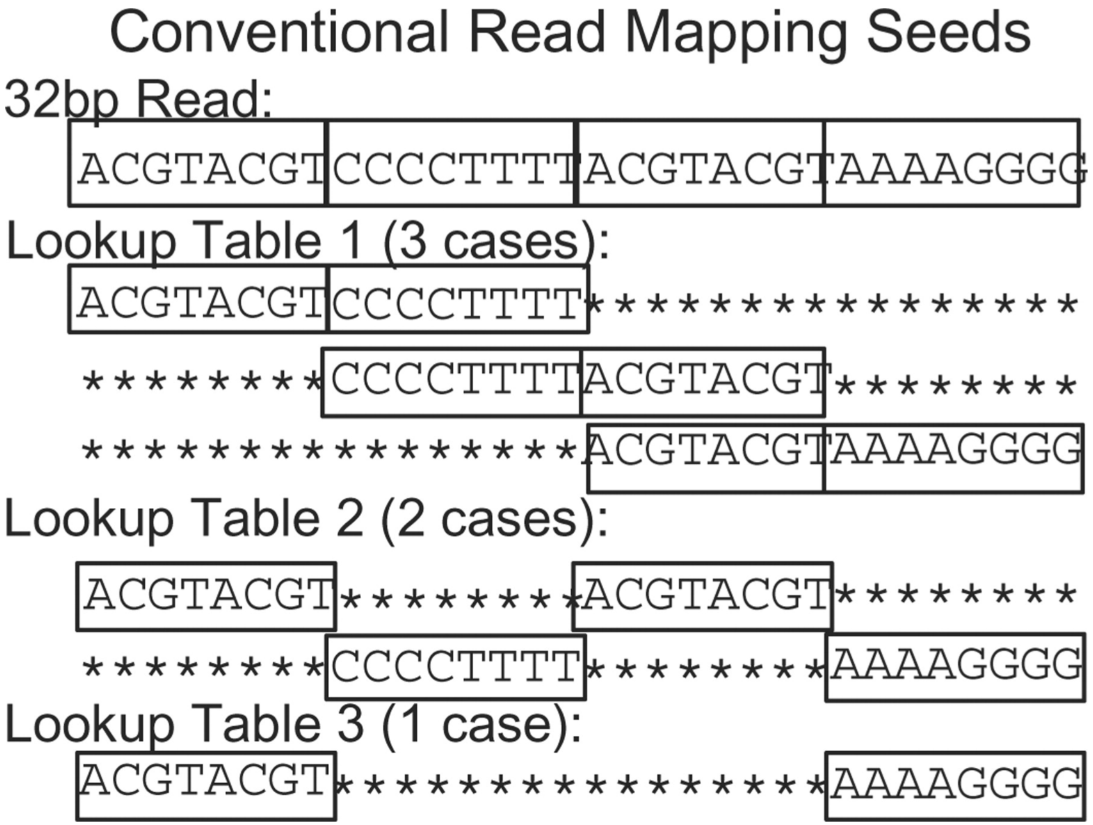
There are two major algorithmic approaches to map RNA-seq reads to a **reference transcriptome**.

The first: **unspliced read aligners** align reads to a reference *without allowing any large gaps*.
The unspliced read aligners fall into two main categories:

‘**seed methods**’
‘**Burrows-Wheeler transform methods**’.

Mapping short RNA-seq reads – unspliced reads, seed idea

Seed methods such as mapping and assembly with quality find matches for short subsequences, termed ‘**seeds**’, assuming that at least one seed in a read will perfectly match the reference. Each seed is used to narrow candidate regions where more sensitive methods (such as Smith-Waterman) can be applied to extend seeds to full alignments.



Conventional Read Mapping Seeds

32bp Read:

```
ACGTACGTCCCCTTTTACGTACGTAAAAGGGG
```

Lookup Table 1 (3 cases):

```
ACGTACGTCCCCTTTT****************
********CCCCTTTTACGTACGT********
****************ACGTACGTAAAAGGGG
```

Lookup Table 2 (2 cases):

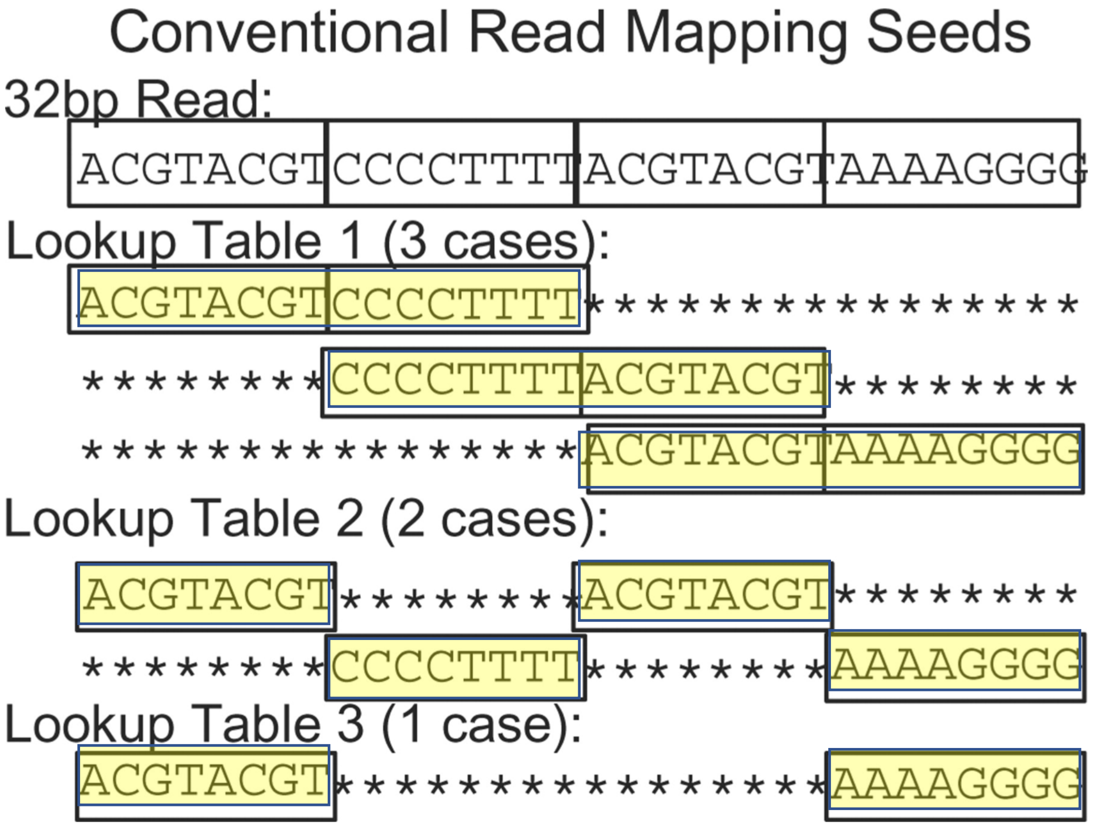
```
ACGTACGT********ACGTACGT********
********CCCCTTTT********AAAAGGGG
```

Lookup Table 3 (1 case):

```
ACGTACGT****************AAAAGGGG
```

Mapping short RNA-seq reads – unspliced reads, seed idea

Seed methods such as mapping and assembly with quality find matches for short subsequences, termed ‘**seeds**’, assuming that at least one seed in a read will perfectly match the reference. Each seed is used to narrow candidate regions where more sensitive methods (such as Smith-Waterman) can be applied to extend seeds to full alignments.



Conventional Read Mapping Seeds

32bp Read:

ACGTACGTCCCCTTTTACGTACGTAAAAGGGG

Lookup Table 1 (3 cases):

ACGTACGTCCCCTTTT****************

********CCCCTTTTACGTACGT********

****************ACGTACGTAAAAGGGG

Lookup Table 2 (2 cases):

ACGTACGT********ACGTACGT********

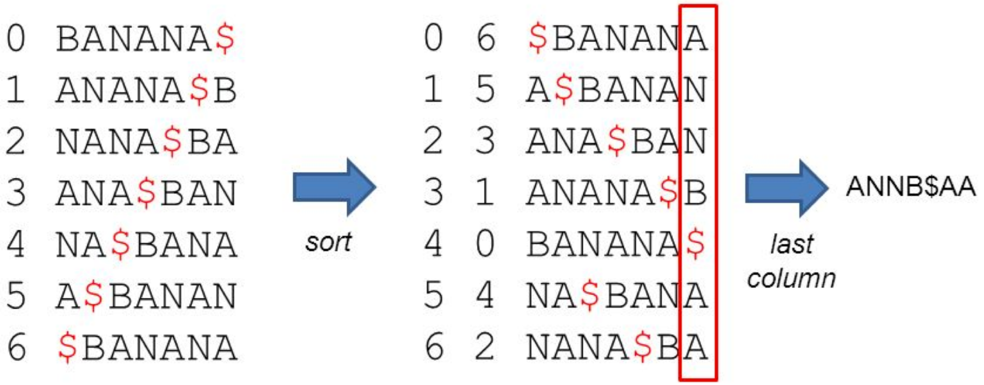
********CCCCTTTT********AAAAGGGG

Lookup Table 3 (1 case):

ACGTACGT****************AAAAGGGG

Mapping short RNA-seq reads – unspliced reads, BWT

Second approach includes **Burrows-Wheeler transform methods** such as Burrows-Wheeler alignment Bowtie, which compact the genome into a data structure that is very efficient when searching *for perfect matches*. When allowing mismatches, the performance of Burrows-Wheeler transform methods decreases exponentially with the number of mismatches as they iteratively perform perfect searches

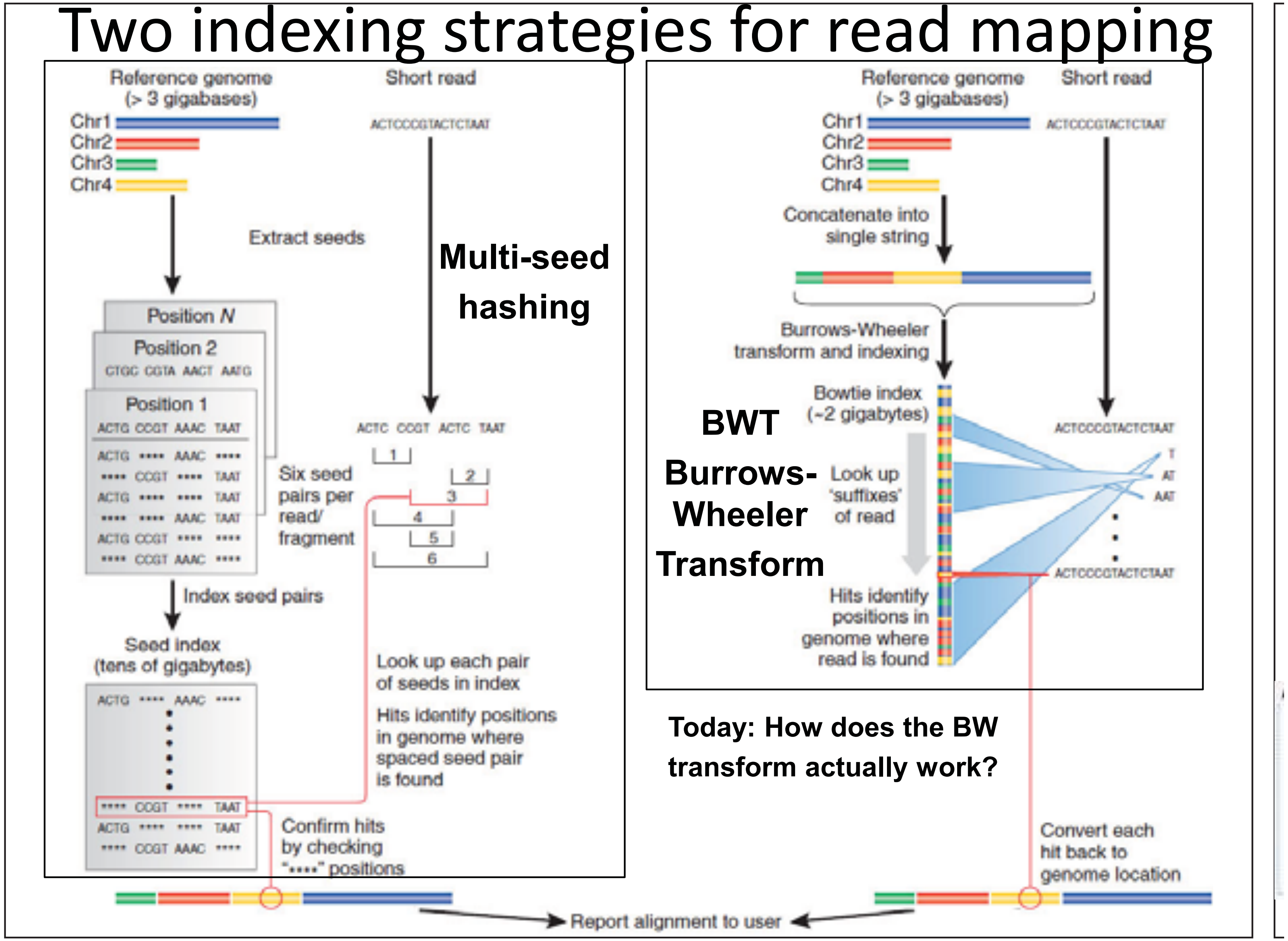


```
0  BANANA$          0  6  $BANANA
1  ANANA$B          1  5  A$BANAN
2  NANA$BA          2  3  ANA$BAN
3  ANA$BAN   sort   3  1  ANANA$B   last column   ANNB$AA
4  NA$BANA          4  0  BANANA$
5  A$BANAN          5  4  NA$BANA
6  $BANANA          6  2  NANA$BA
```

BWT(“BANANA$”) = “ANNB$AA”

Mapping short RNA-seq reads – seed idea VS BWT

Two indexing strategies for read mapping



Multi-seed hashing

BWT

Burrows-Wheeler Transform

Today: How does the BW transform actually work?

Mapping short RNA-seq reads – unspliced reads

Unspliced read aligners are ideal for mapping reads against a reference databases for quantification purposes

BWT

If the exact reference transcriptome is available, Burrows-Wheeler methods are faster than seed-based methods

SEED

when only the reference transcriptome of a distant species is available, 'seed methods' can result in a large increase in sensitivity.

Similarly, an increase in sensitivity using seed methods has been observed when aligning reads to polymorphic regions in a species for quantification of allele-specific gene expression

Unspliced read aligners are limited to identifying known exons and junctions, and do not allow for the identification of splicing events involving new exons (since they used the transcriptome as reference).

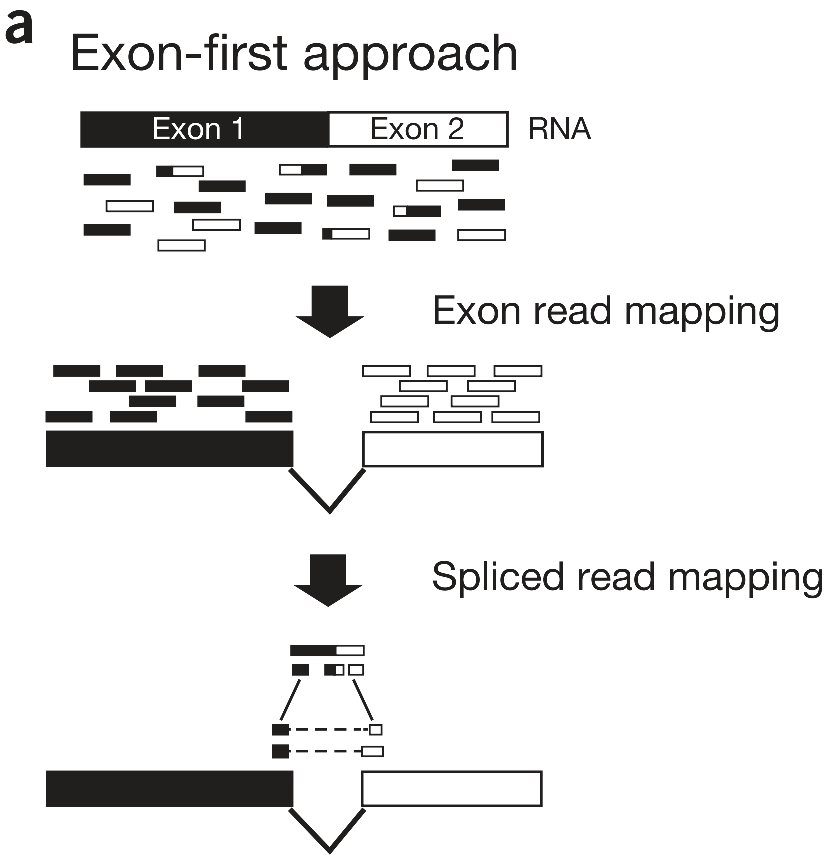
Mapping short RNA-seq reads – spliced aligners – Exon first

Alternatively, **reads can be aligned to the entire genome**, including intron-spanning reads that require large gaps for proper placement. Several methods exist, collectively referred to as '**spliced aligners**', that fall into two main categories: '**exon first**' and '**seed and extend**'.

Exon-first methods such as TopHat use a two-step process.

1. First, they map reads continuously to the genome using the unspliced read aligners.
2. Second, unmapped reads are split into shorter segments and aligned independently.

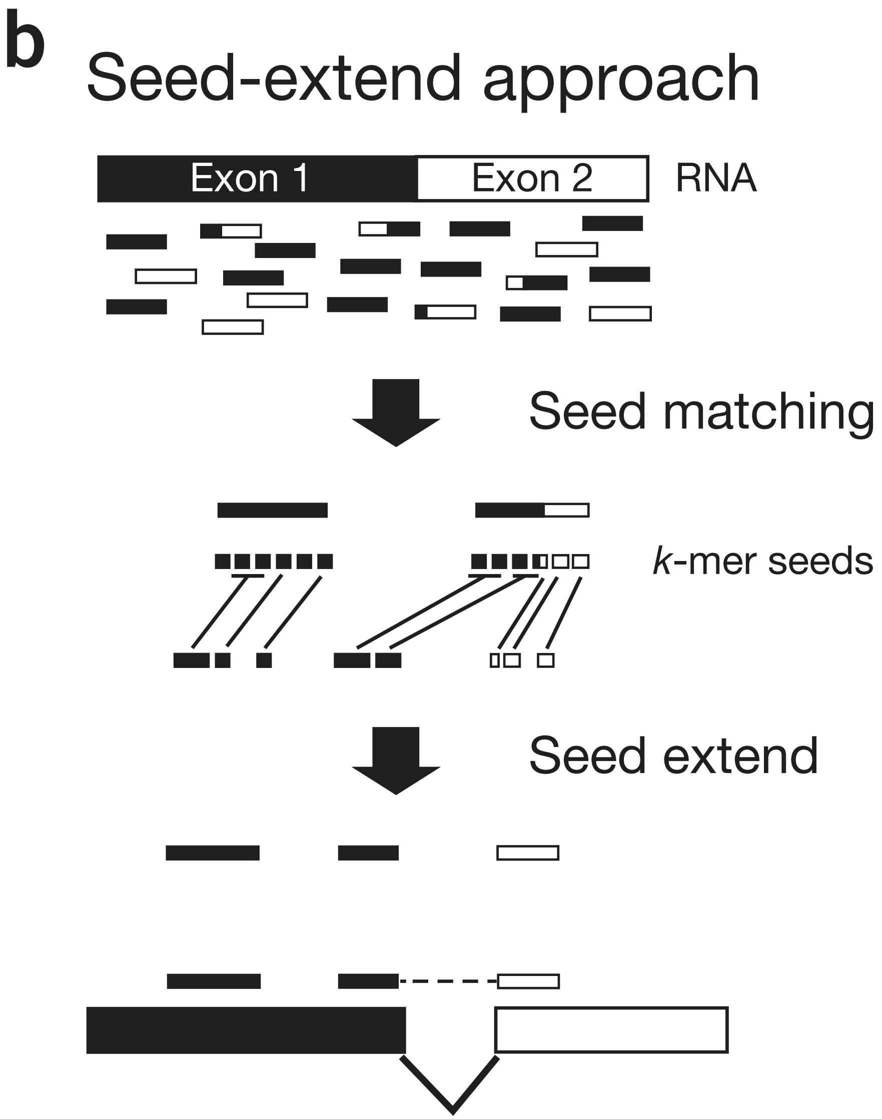
Exon- first aligners are very efficient when only a small portion of the reads require the more computationally intensive second step.



Mapping short RNA-seq reads – spliced aligners – seed-extend

Seed-extend methods:

1) break reads into short seeds, which are placed onto the genome to localize the alignment.
2) candidate regions are then examined with more sensitive methods:
   i. such as the local alignment (i.e. Smith-Waterman algorithm)
   ii. iterative extension and merging of initial seeds to determine the exact spliced alignment for the read .

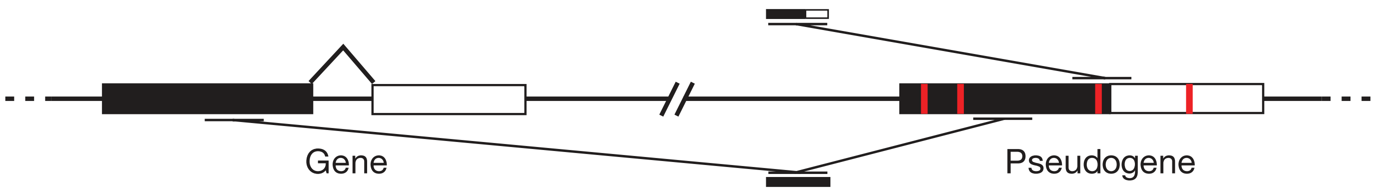


Mapping short RNA-seq reads – spliced aligners

Comparison

A seed- extend method takes ~8× longer (~340 CPU hours) than an exon-first method resulting in ~1.5× more spliced reads. However, the biological meaning of these additional splice junctions has not been demonstrated.

c Potential limitations of exon-first approaches



A potential disadvantage of exon-first approaches illustrated for a gene and its associated retrotransposed pseudogene. Mismatches compared to the gene sequence are indicated in red. Exonic reads will map to both the gene and its pseudogene, preferring gene placement owing to lack of mutations, but a spliced read could be incorrectly assigned to the pseudogene as it appears to be exonic, preventing higher-scoring spliced alignments from being pursued.

Computational methods needed to address RNA-seq analysis core challenges

RNASeq Analysis

GOAL: *map RNA-seq reads to a reference transcriptome*

methods to align reads directly to a reference transcriptome or genome ('read mapping').

- **unspliced read aligners**:
  - **seed methods**
  - **Burrows-Wheeler transform methods**
- **spliced aligners:**
  - **exon first**
  - **seed andextend**.

GOAL: *reconstruct the transcriptome*

methods to identify expressed genes and isoforms ('transcriptome reconstruction').

- ***genome-guided***

  exon identification

  genome-guided assembly

- ***genome- independent***'

GOAL*:* *quantify gene expression*

methods for estimation of gene and isoform abundance, as well as methods for the analysis of differential expression across samples ('expression quantification').

- ***exon intersection method***
- ***exon union method'***

Garber et al Nature Methods 2011

Transcriptome reconstruction

Defining a precise map of **all transcripts** and isoforms that are expressed in a particular sample requires the assembly of these reads or read alignments into transcription units. This process is **transcriptome reconstruction**.

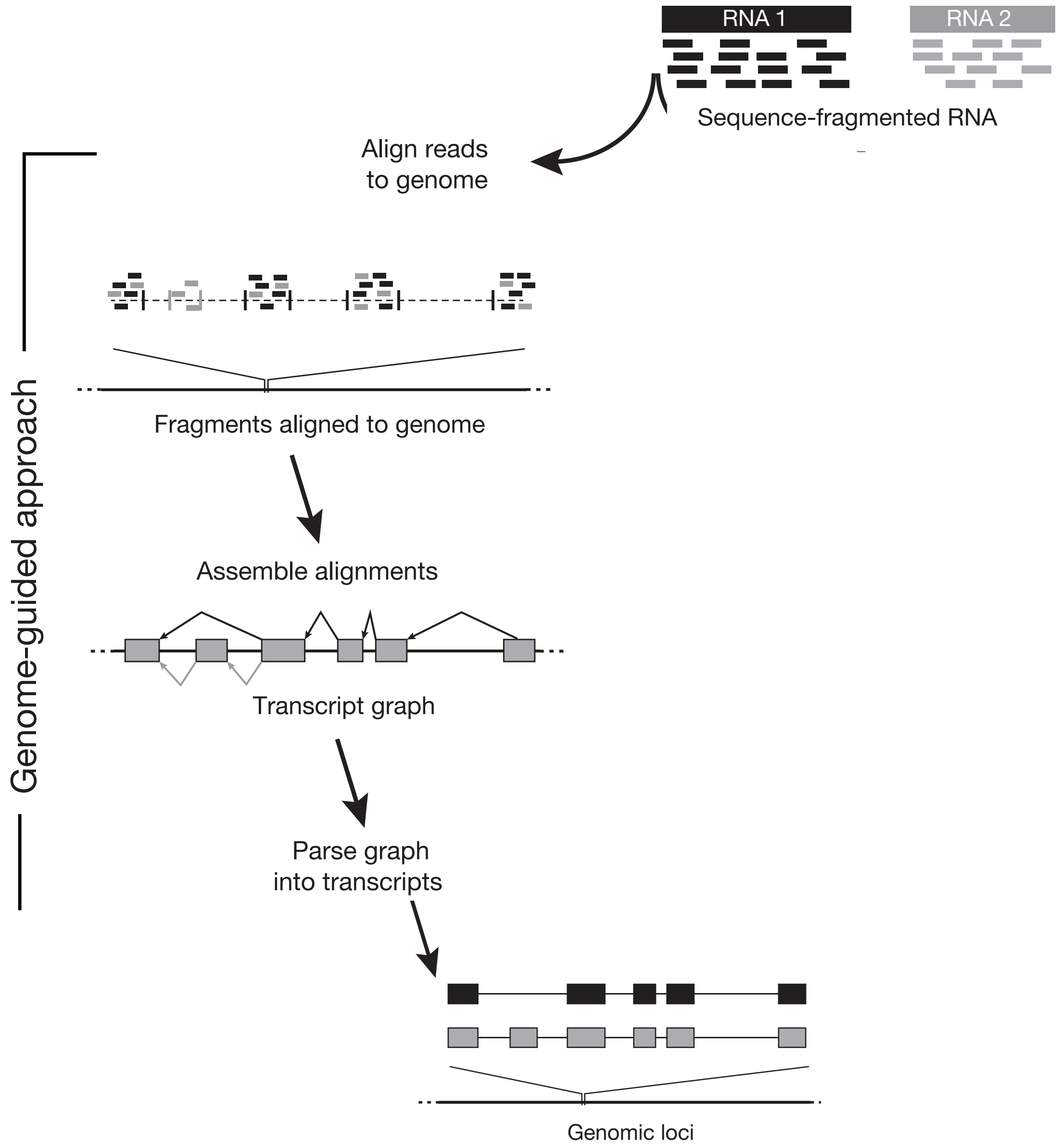
Transcriptome reconstruction is a difficult computational task for three main reasons.

1. Gene expression spans several orders of magnitude, with some genes represented by only a few reads.
2. Reads originate from the mature mRNA (exons only) as well as from the incompletely spliced precursor RNA (containing intronic sequences), making it difficult to identify the mature transcripts.
3. Reads are short, and genes can have many isoforms, making it challenging to determine which isoform produced each read.

Several methods exist to reconstruct the transcriptome, and they fall into two main classes: '***genome-guided***' and '***genome- independent***'

Transcriptome reconstruction – genome guided, exon identification

Genome-guided reconstruction. Existing genome-guided methods can be classified in two main categories: *exon identification* and *genome-guided* assembly approaches.



Transcriptome reconstruction – genome guided, exon identification

Genome-guided reconstruction. Existing genome-guided methods can be classified in two main categories: *exon identification* and *genome-guided* assembly approaches.

Exon identification methods were developed early when reads were short (~36 bases) and few aligned to exon-exon junctions.
They first define putative **exons as coverage islands**, and then use spliced reads that span across these coverage islands to define exon boundaries and to establish connections between exons.

Performance: they are underpowered to identify full- length structures of lowly expressed, long and alternatively spliced genes.

Transcriptome reconstruction – genome guided, genome guided

Genome-guided assembly methods such as Cufflinks and Scripture have been developed. These methods use spliced reads directly to reconstruct the transcriptome

Two approaches:

Scripture Transform genome into a graph topology, which represents all possible connections of bases in the transcriptome either when they occur consecutively or when they are connected by a spliced read. This graph topology to reduce the transcript reconstruction problem to a statistical segmentation problem of identifying **significant transcript paths across the graph.**

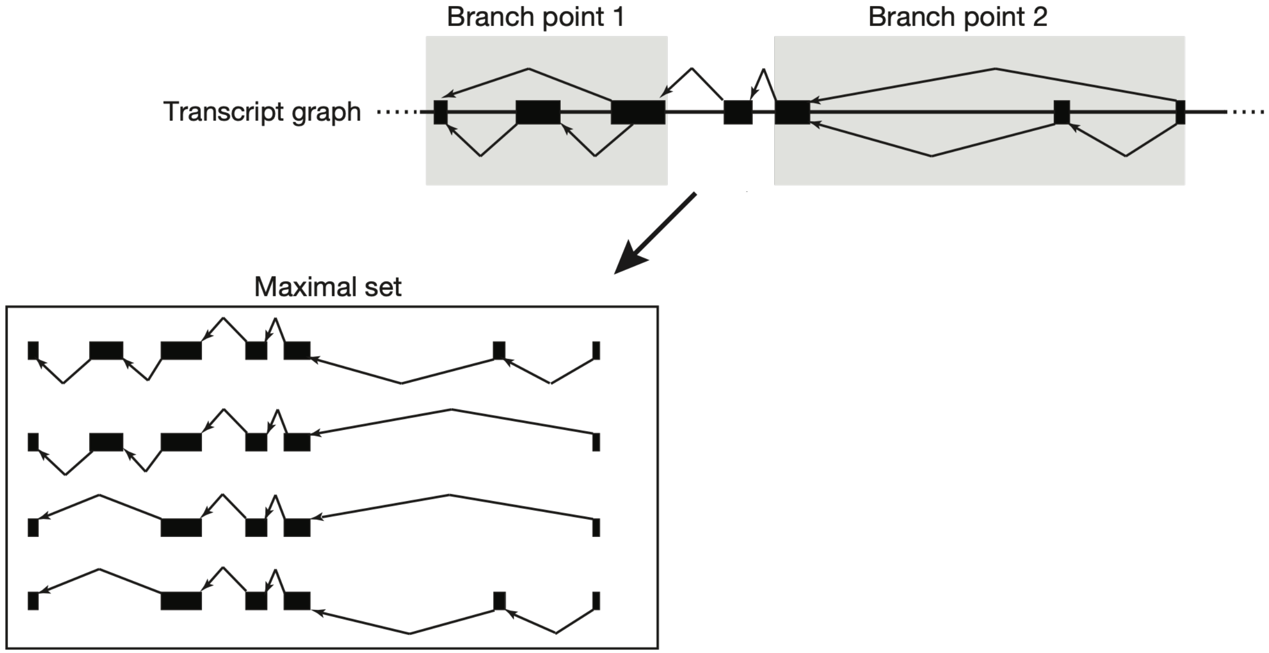
Scripture provides increased sensitivity to identify transcripts expressed at **low** levels by working with significant paths, rather than significant exons

Cufflinks Transform genome into a graph topology, which connect aligned reads into an overlap graph.

Transcriptome reconstruction – genome guided, genome guided

Scripture reports all isoforms that *are compatible with the read data* (maximum sensitivity), whereas ***Cufflinks*** reports the minimal number of *compatible isoforms* (maximum precision)

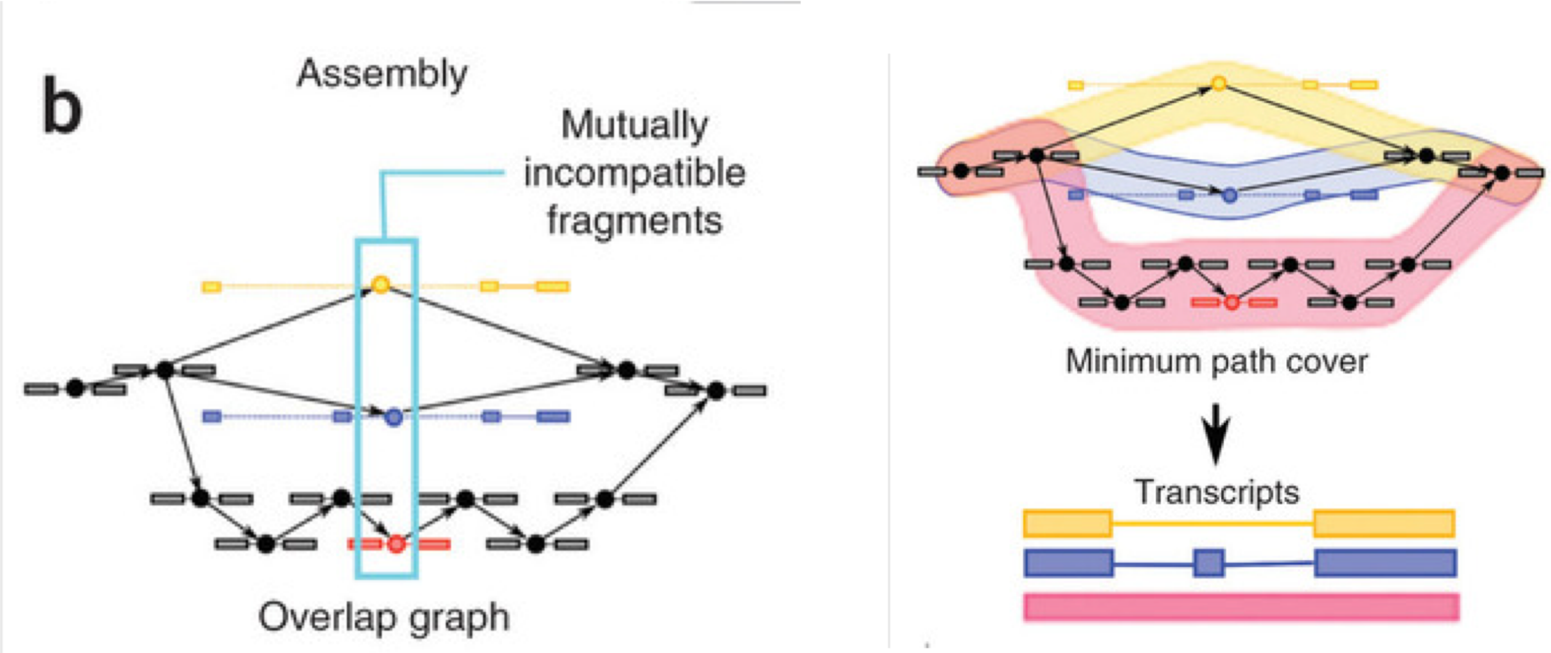
Scripture enumerates **all possible paths** through the assembly graph that are consistent with the spliced reads. Its maximum sensitivity at the transcript level derived from its strategy of predicting a near-exhaustive list of all possible splice variants for a given gene.



Transcriptome reconstruction – genome guided, genome guided

Scripture reports all isoforms that *are compatible with the read data* (maximum sensitivity), whereas ***Cufflinks*** reports the minimal number of *compatible isoforms* (maximum precision)

Cufflinks chooses a minimal set of paths through the **graph such that all reads are included in at least one path**. Each path defines an isoform, so this minimal set of paths is a minimal assembly of reads. As there can be **many minimal sets of isoforms**, Cufflinks uses **read coverage** across each path to decide which combination of paths is most likely to originate from the same RNA.

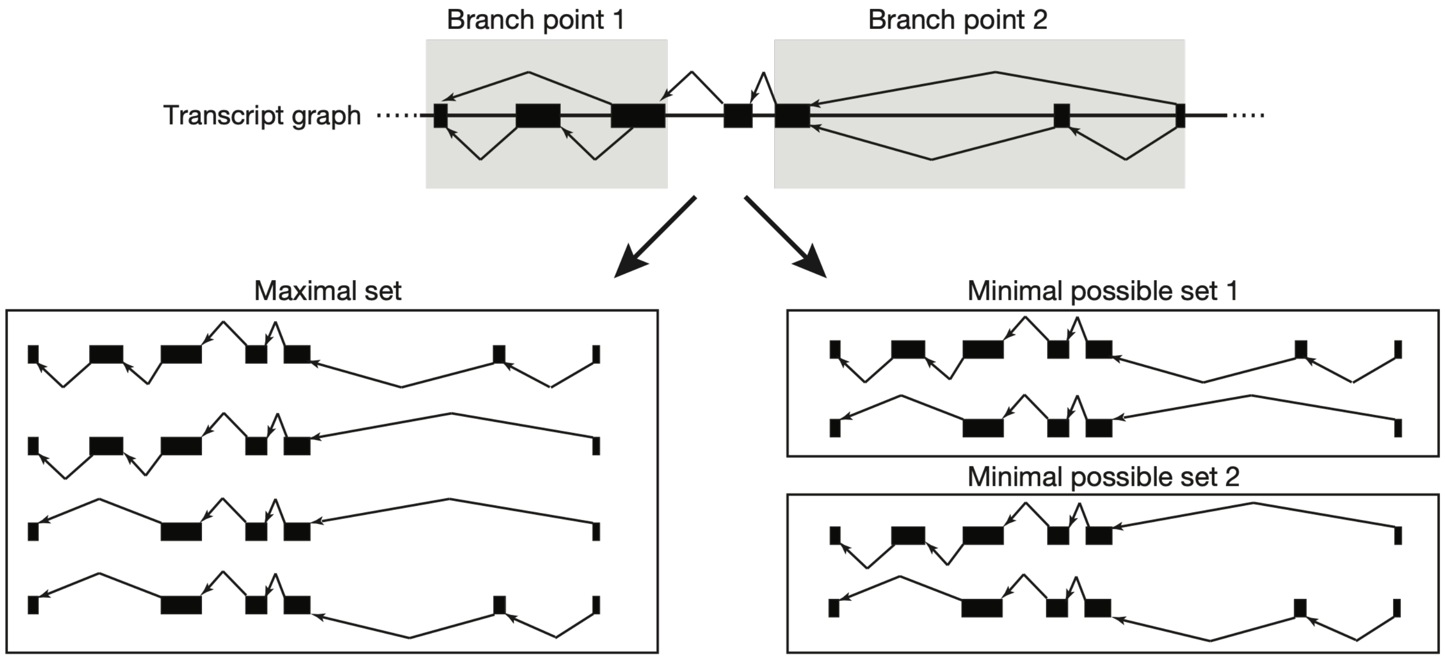


Defining a minimum path cover of G, meaning that every fragment node is contained in some path in the cover, and the cover contained as few paths as possible.
Each path in the cover corresponded to a set of mutually compatible fragments overlapping each other on the left and right.

Transcriptome reconstruction – genome guided, genome guided

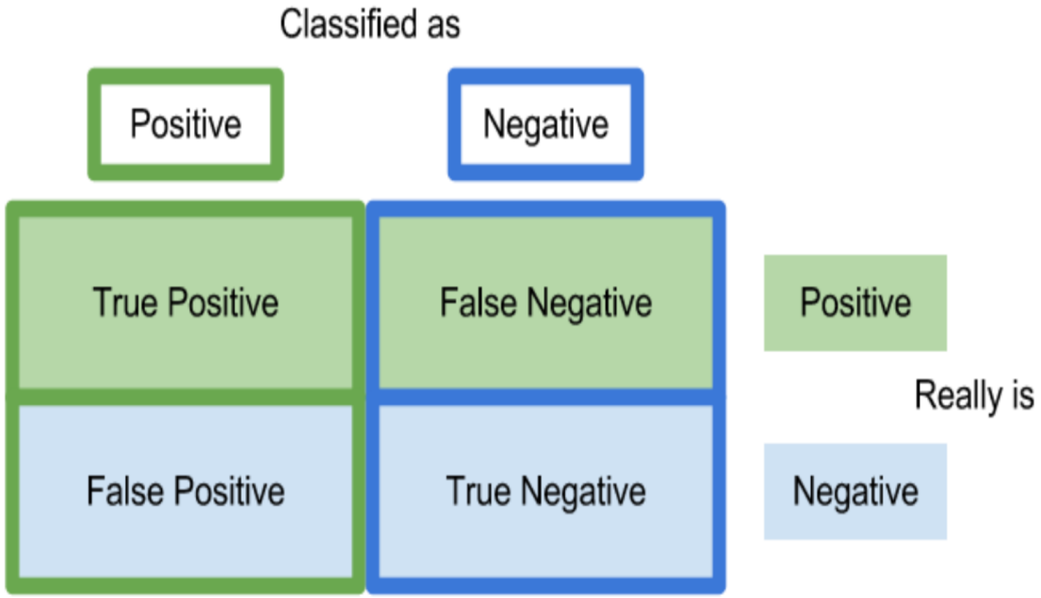
Scripture reports all isoforms that *are compatible with the read data* (maximum sensitivity), whereas ***Cufflinks*** reports the minimal number of *compatible isoforms* (maximum precision)

Cufflinks chooses a minimal set of paths through the **graph such that all reads are included in at least one path**. Each path defines an isoform, so this minimal set of paths is a minimal assembly of reads. As there can be **many minimal sets of isoforms**, Cufflinks uses **read coverage** across each path to decide which combination of paths is most likely to originate from the same RNA.



Defining a minimum path cover of G, meaning that every fragment node is contained in some path in the cover, and the cover contained as few paths as possible.
Each path in the cover corresponded to a set of mutually compatible fragments overlapping each other on the left and right.

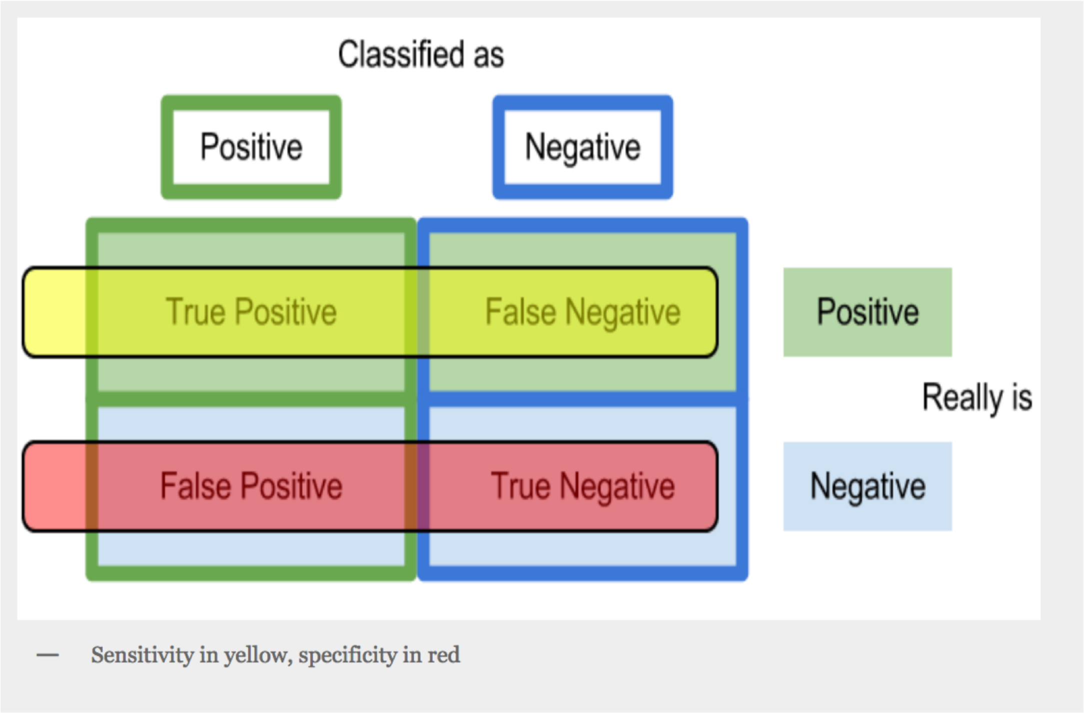
Transcriptome reconstruction – genome guided, genome guided



Classified as

	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

Really is



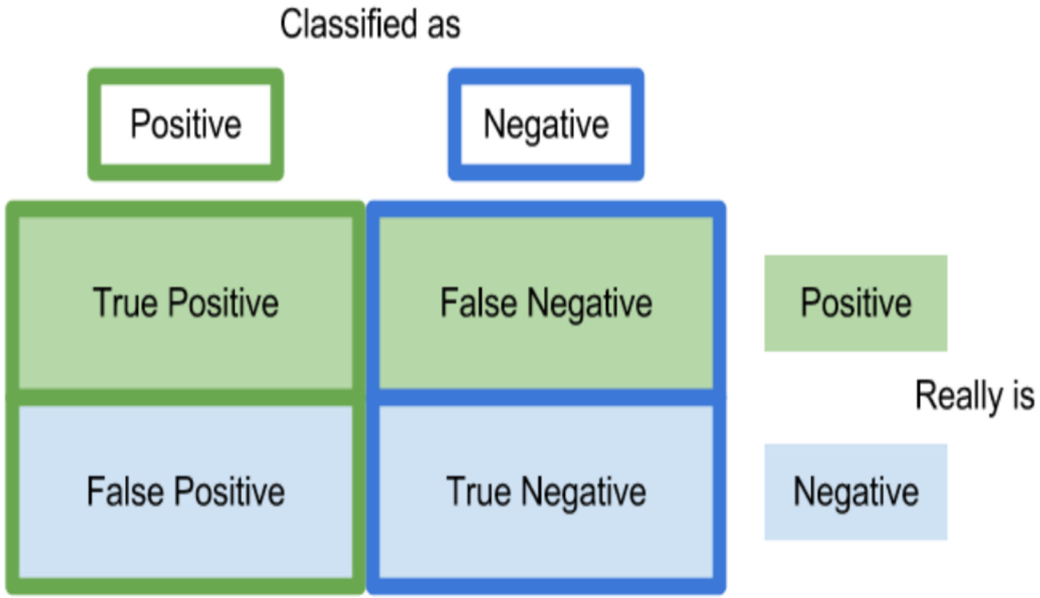
Classified as

	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

Really is

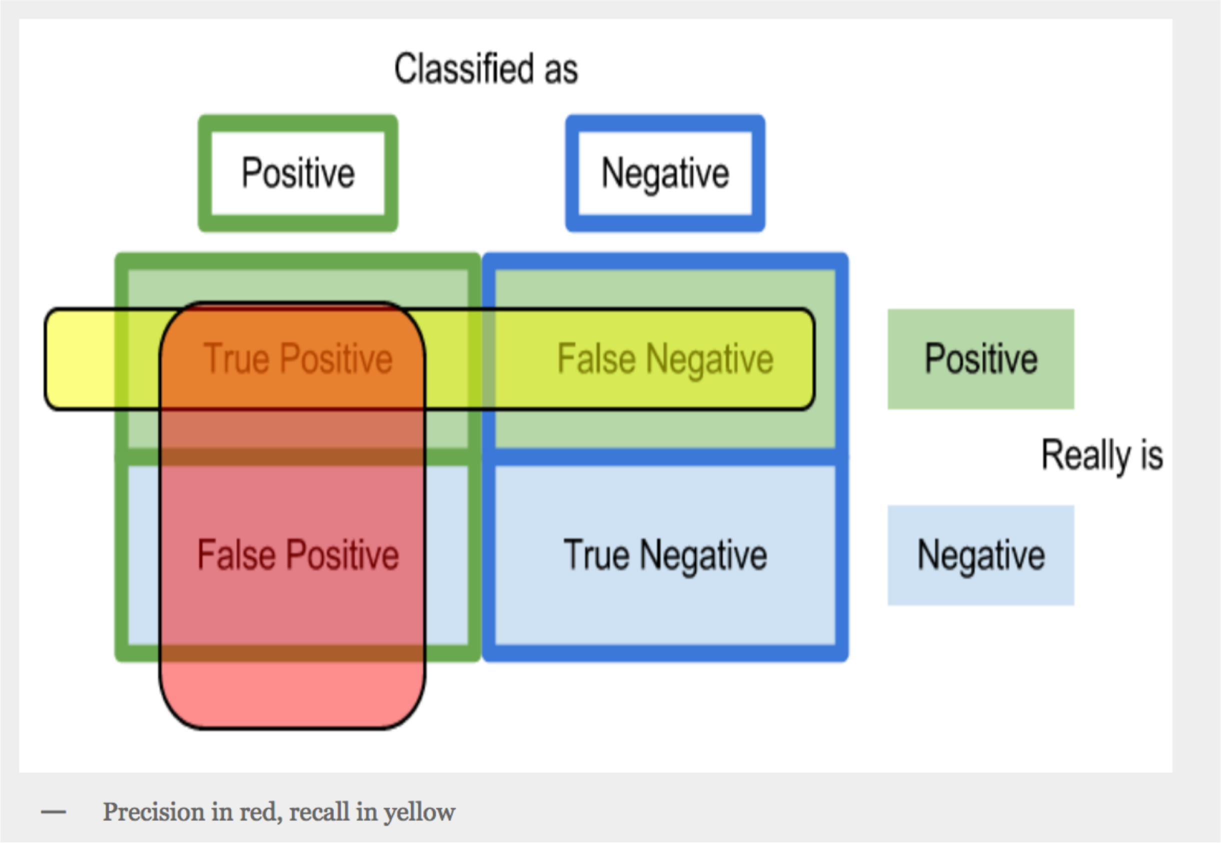
— Sensitivity in yellow, specificity in red

Transcriptome reconstruction – genome guided, genome guided



Classified as

Really is	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

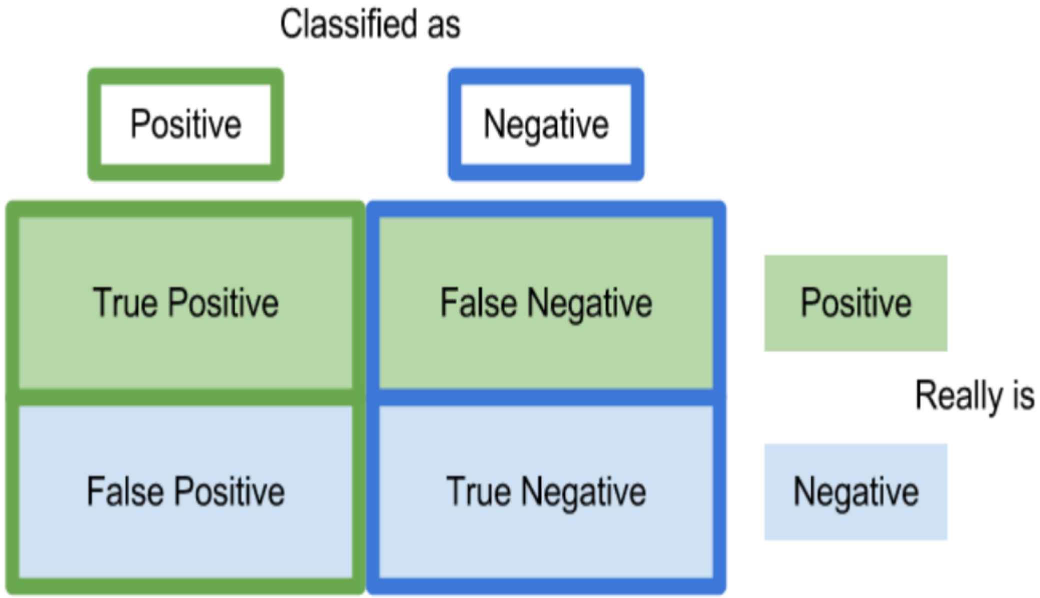


Classified as

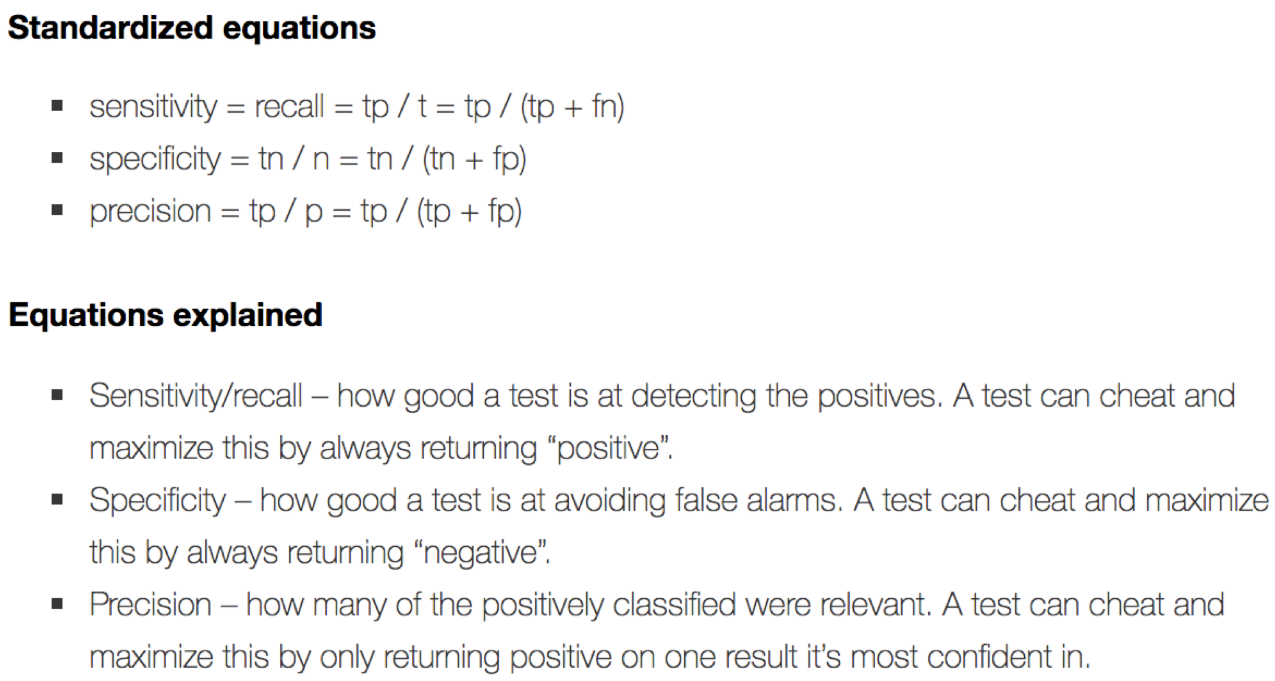
Really is	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

— Precision in red, recall in yellow

Transcriptome reconstruction – genome guided, genome guided



		Classified as	
		Positive	Negative
Really is	Positive	True Positive	False Negative
	Negative	False Positive	True Negative



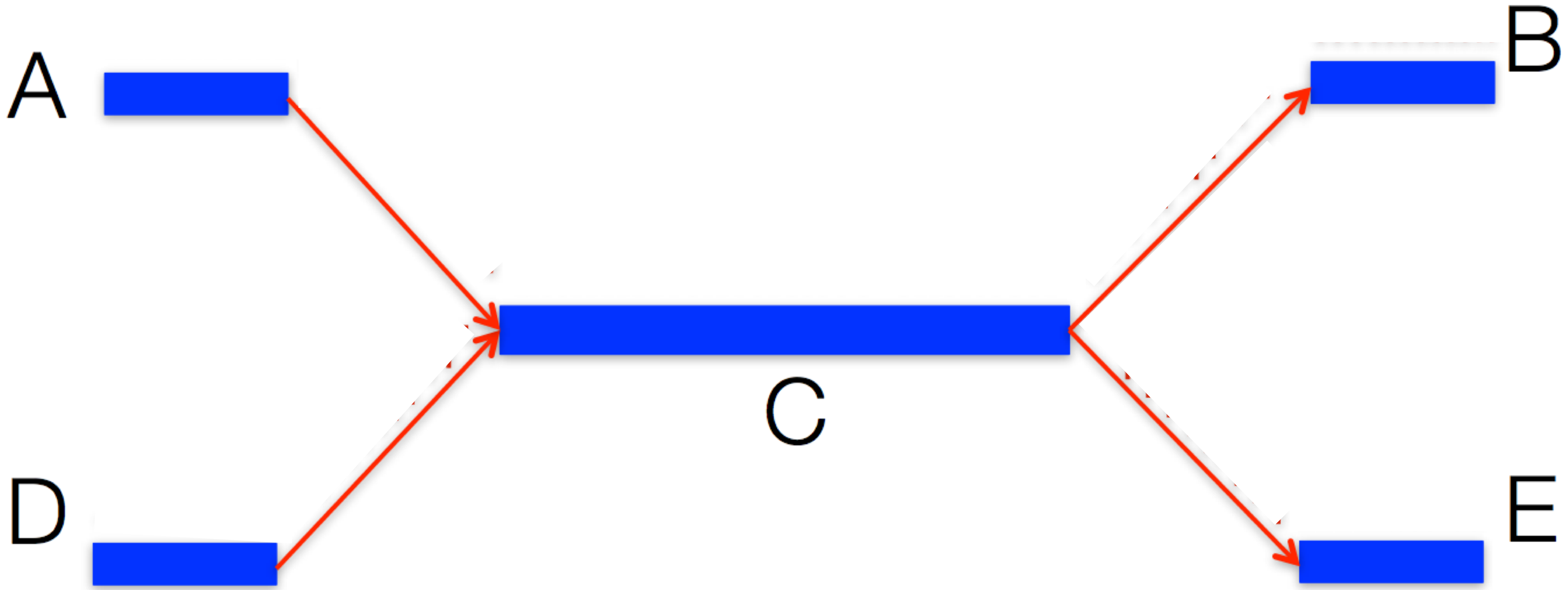
Standardized equations

- sensitivity = recall = tp / t = tp / (tp + fn)
- specificity = tn / n = tn / (tn + fp)
- precision = tp / p = tp / (tp + fp)

Equations explained

- Sensitivity/recall – how good a test is at detecting the positives. A test can cheat and maximize this by always returning “positive”.
- Specificity – how good a test is at avoiding false alarms. A test can cheat and maximize this by always returning “negative”.
- Precision – how many of the positively classified were relevant. A test can cheat and maximize this by only returning positive on one result it’s most confident in.

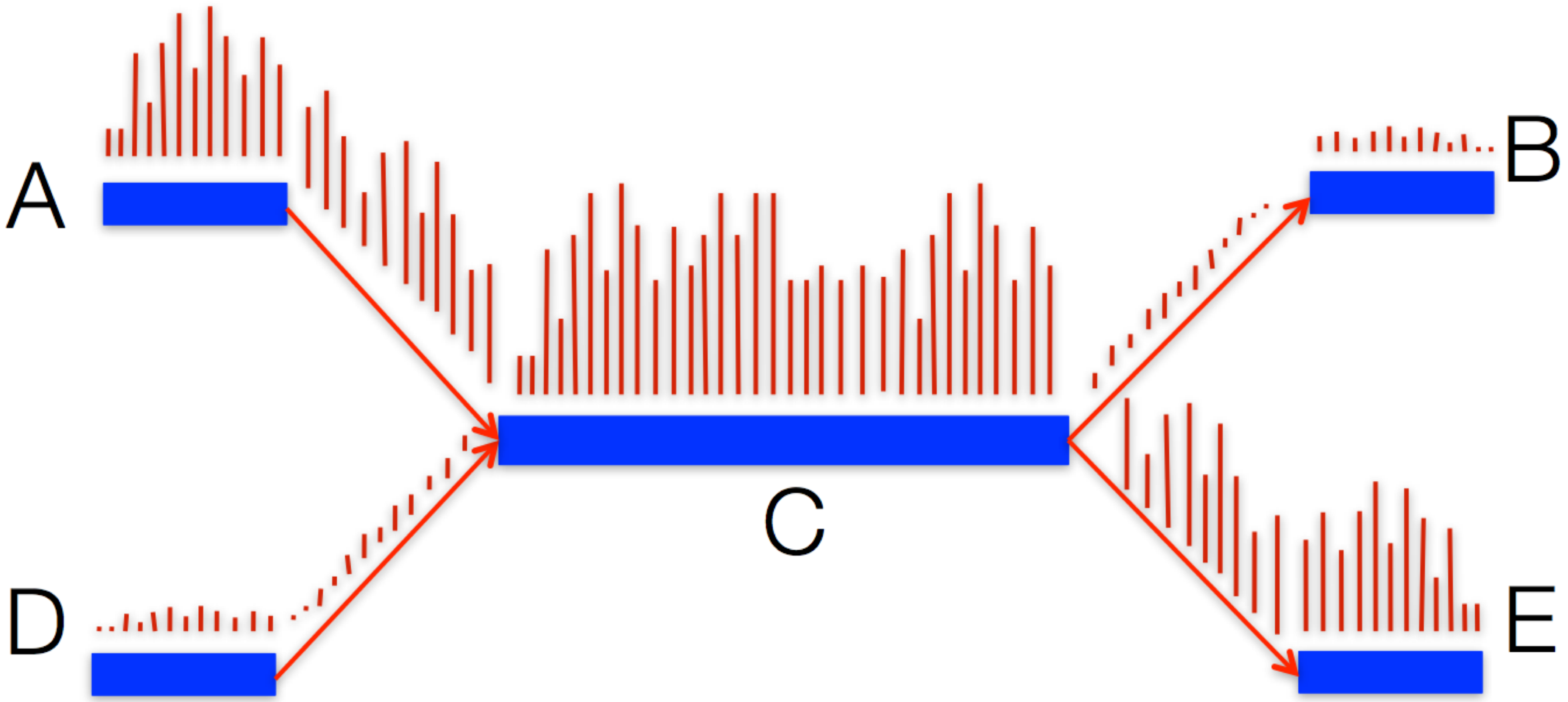
Transcriptome reconstruction – genome guided, genome guided



A, C, B
A,C,E
D,C,B
D,C,E

are equiprobable transcripts

Transcriptome reconstruction – genome guided, genome guided



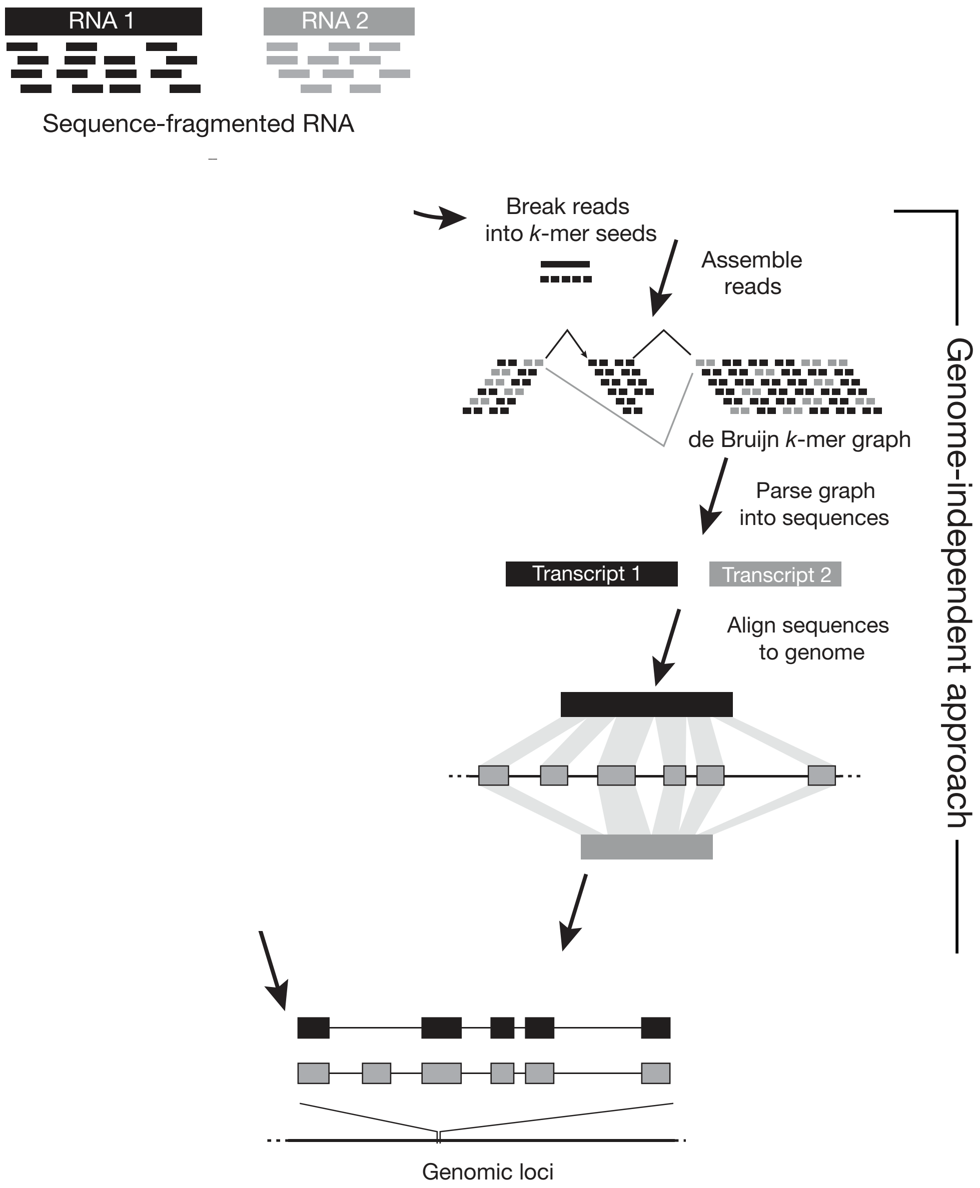
Cuffilinks uses read coverage across each path to decide which combination of paths is most likely to originate from the same RNA.

A, C, E
D,C,B

are the transcripts that match with the coverage and exon quantification data

Transcriptome reconstruction – genome independent

Genome-independent reconstruction. Rather than mapping reads to a reference sequence first, genome-independent transcriptome reconstruction algorithms use the reads to *directly build consensus transcripts*. Consensus transcripts can then be mapped to a genome or aligned to a gene or protein database for annotation purposes. The central challenge for genome-independent approaches is to partition reads into disjoint components, which represent all isoforms of a gene.


RNA 1
RNA 2
Sequence-fragmented RNA
Break reads into *k*-mer seeds
Assemble reads
de Bruijn *k*-mer graph
Parse graph into sequences
Transcript 1
Transcript 2
Align sequences to genome
Genomic loci
Genome-independent approach

Transcriptome reconstruction – genome independent

A commonly used strategy is to first build a de Bruijn graph, which models overlapping subsequences, termed ‘*k*-mers’ (*k* consecutive nucleotides), rather than reads. This reduces the complexity associated with handling millions of reads to a fixed number of possible *k*-mers. The overlaps of *k* – 1 bases between these *k*-mers constitute the graph of all possible sequences that can be constructed. Next, **paths** are traversed in the graph, guided by **read and paired-end coverage** levels, **eliminating false branch** points introduced by *k*-mers that are shared by different transcripts but not supported by reads and paired ends. Each remaining path through the graph is then reported as a separate transcript

Limitations: distinguishing sequencing errors from variation, and finding the optimal balance between sensitivity and graph complexity.

To eliminate artifacts, genome- independent methods look at the coverage of different paths

Smaller values of *k* result in a larger number of overlapping nodes and a more complex graph, whereas larger values of *k* reduce the number of overlaps and results in a simpler graph structure. An **optimal choice** of ***k*** **depends on coverage**:

For low coverage, long or short kmers?

Transcriptome reconstruction – genome independent

A commonly used strategy is to first build a de Bruijn graph, which models overlapping subsequences, termed ‘*k*-mers’ (*k* consecutive nucleotides), rather than reads. This reduces the complexity associated with handling millions of reads to a fixed number of possible *k*-mers. The overlaps of *k* – 1 bases between these *k*-mers constitute the graph of all possible sequences that can be constructed. Next, **paths** are traversed in the graph, guided by **read and paired-end coverage** levels, **eliminating false branch** points introduced by *k*-mers that are shared by different transcripts but not supported by reads and paired ends. Each remaining path through the graph is then reported as a separate transcript

Limitations: distinguishing sequencing errors from variation, and finding the optimal balance between sensitivity and graph complexity.

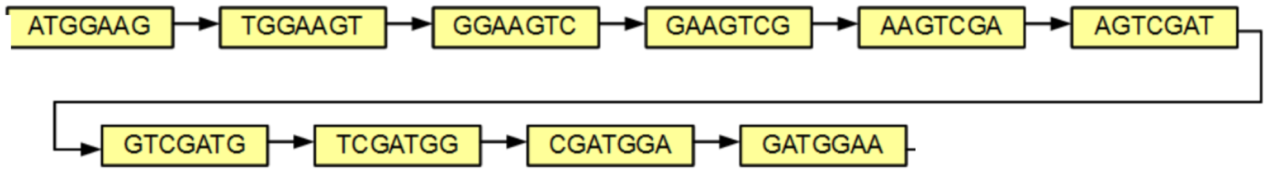
To eliminate artifacts, genome- independent methods look at the coverage of different paths

Smaller values of *k* result in a larger number of overlapping nodes and a more complex graph, whereas larger values of *k* reduce the number of overlaps and results in a simpler graph structure. An **optimal choice** of ***k*** **depends on coverage**: when **coverage is low**, **small values of *k*** are preferable because they increase the number of overlapping reads contributing *k*-mers to the graph. But when **coverage is large**, **small values of *k* are overly** sensitive to sequencing errors and other artifacts, yielding very complex graph structures

Transcriptome reconstruction – genome independent

ATGGAAGTCGATGGAAG

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGA
AGTCGAT
GTCGATG
TCGATGG
CGATGGA
GATGGAA
ATGGAAG

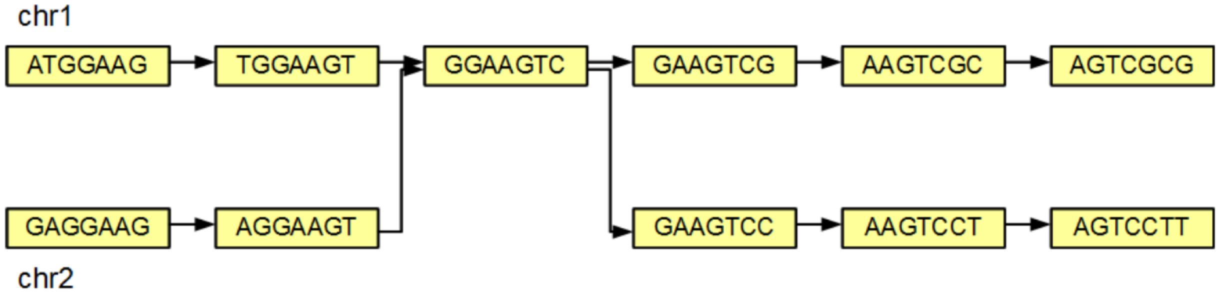


ATGGAAG → TGGAAGT → GGAAGTC → GAAGTCG → AAGTCGA → AGTCGAT → GTCGATG → TCGATGG → CGATGGA → GATGGAA

chr1 ATGGAAGTCGCG

chr2 GAGGAAGTCCTT

To understand which are the correct path we can be guided by **read and paired-end coverage** levels, **eliminating false branch** points



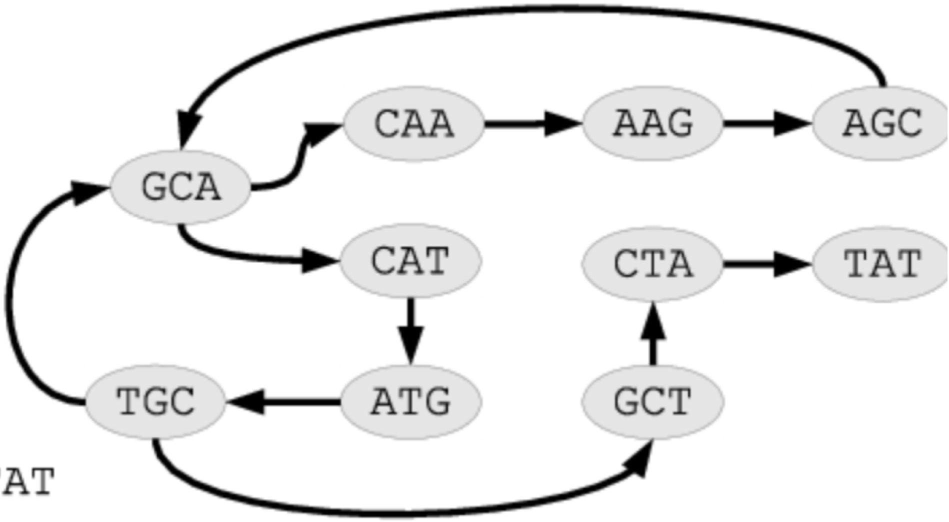
chr1

ATGGAAG → TGGAAGT → GGAAGTC → GAAGTCG → AAGTCGC → AGTCGCG

GAGGAAG → AGGAAGT → GGAAGTC → GAAGTCC → AAGTCCT → AGTCCTT

chr2

Transcriptome reconstruction – genome independent



Overlap graph guided by the set of input reads

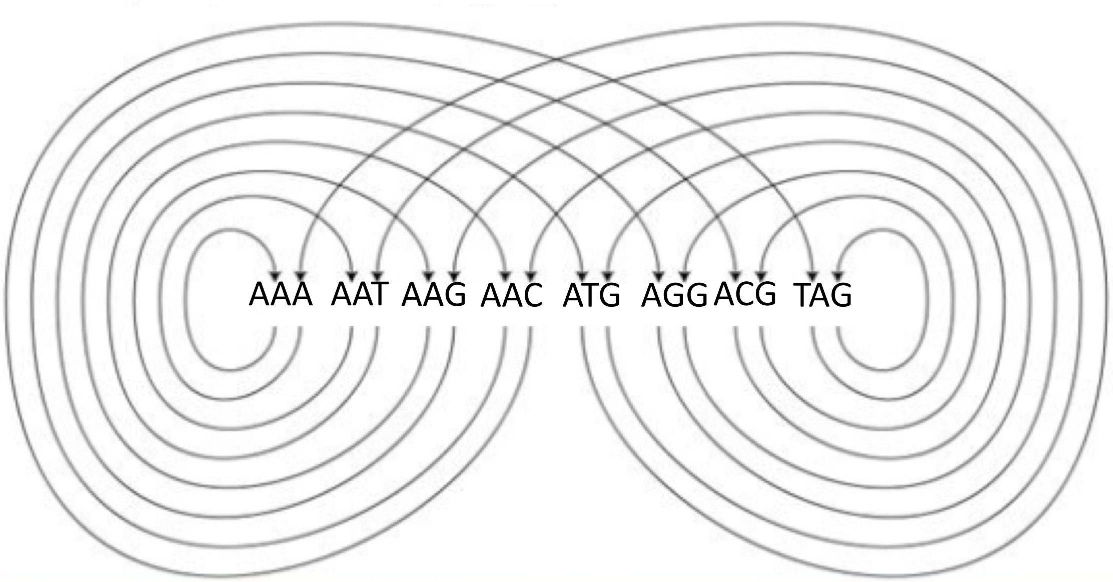
R_1: TGCAAGCA

R_2: AAGCATGC

R_3: CATGCTAT

Example: *k*=3,

Overlap graph



Computational methods needed to address RNA-seq analysis core challenges

RNASeq Analysis

GOAL: *map RNA-seq reads to a reference transcriptome*

methods to align reads directly to a reference transcriptome or genome (‘read mapping’).

- **unspliced read aligners:**
  - **seed methods**
  - **Burrows-Wheeler transform methods**
- **spliced aligners:**
  - **exon first**
  - **seed andextend**.

GOAL: *reconstruct the transcriptome*

methods to identify expressed genes and isoforms (‘transcriptome reconstruction’).

- ***genome-guided***

  exon identification
  genome-guided assembly

- ***genome- independent’***

GOAL: *quantify gene and isoform expression*

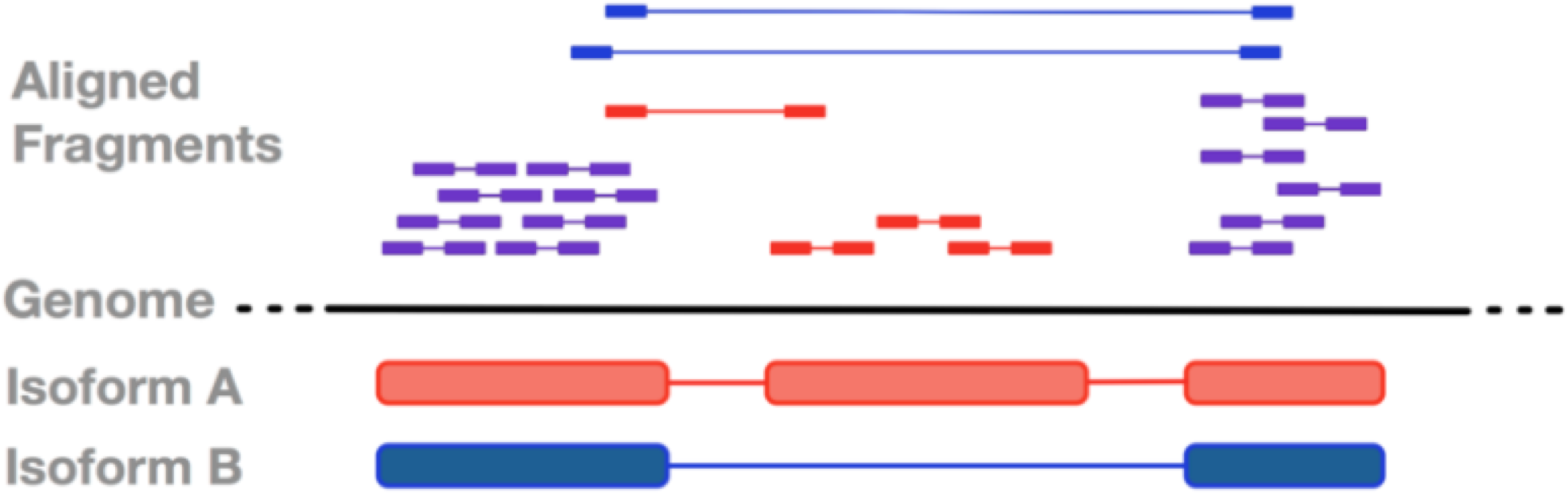
methods for estimation of gene and isoform abundance, as well as methods for the analysis of differential expression across samples (‘expression quantification’).

- ***exon intersection method***
- ***exon union method’***

Garber et al Nature Methods 2011

Estimating transcript expression levels

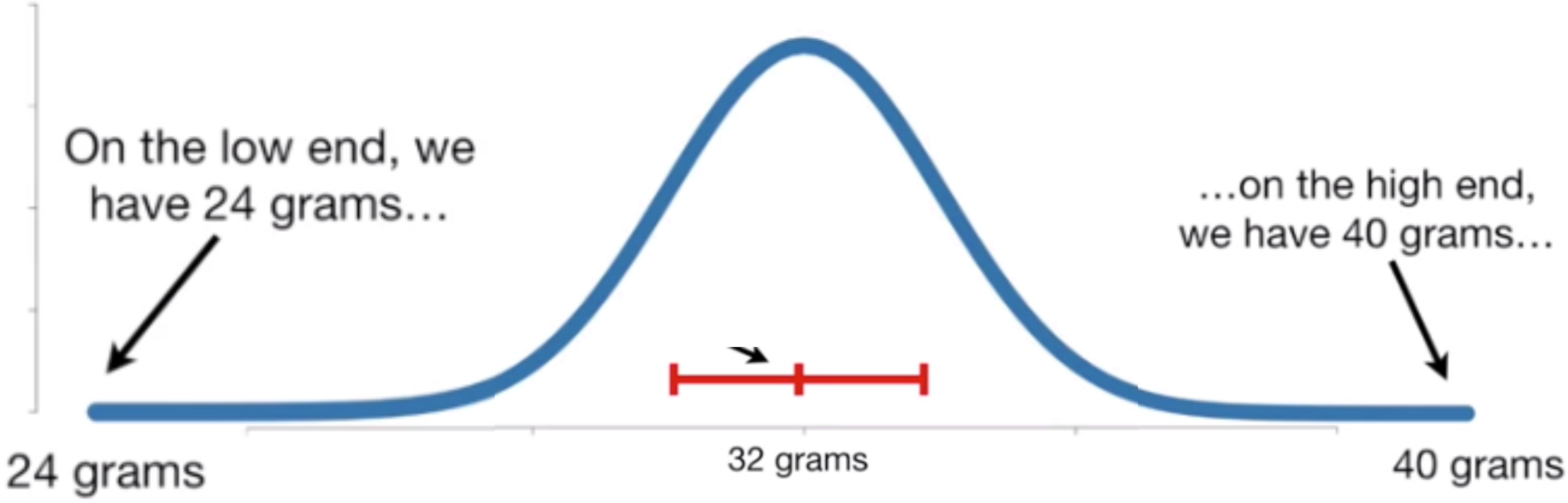
How perfom the correct assignment? Some reads cannot be assigned unequivocally to a transcript.



The ‘isoform-expression methods’ such as Cufflinks, handle uncertainty by constructing a ‘***likelihood function***’ that models the sequencing process and identifies isoform abundance estimates that best explain the reads obtained in the experiment.

Probability versus Likelihood

In this case, let's imagine that this is a distribution of mouse weights.



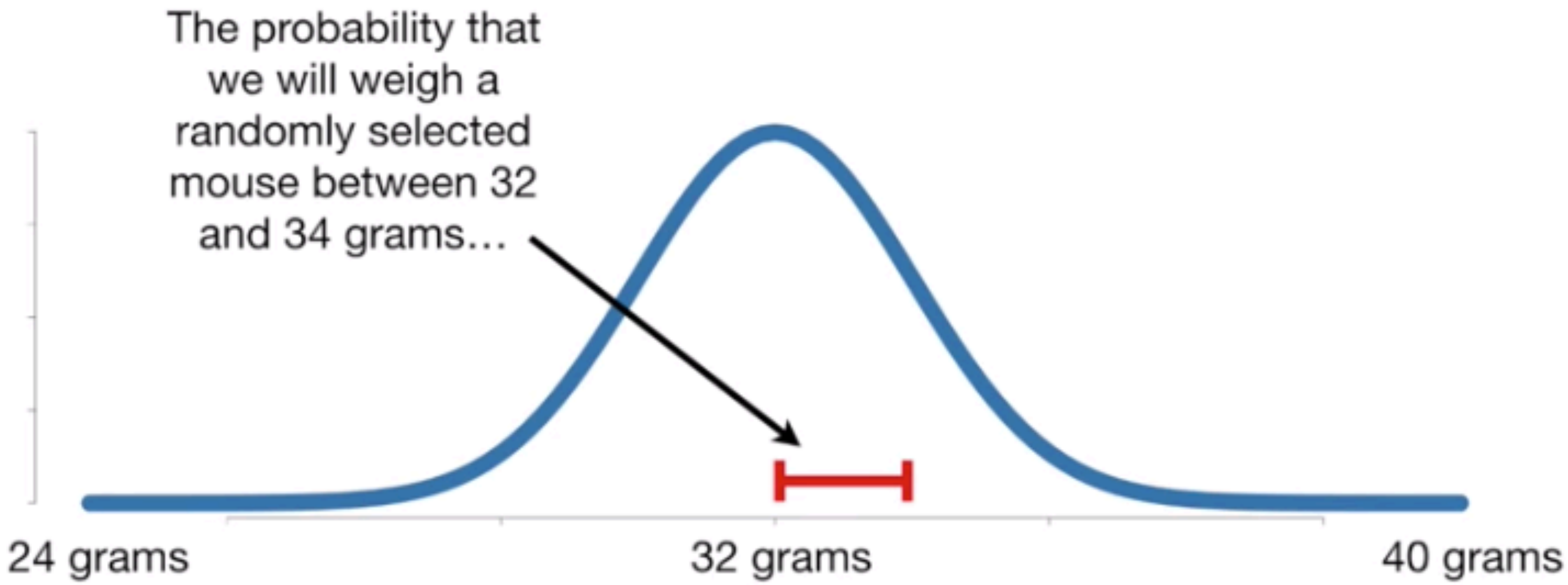
On the low end, we have 24 grams…

…on the high end, we have 40 grams…

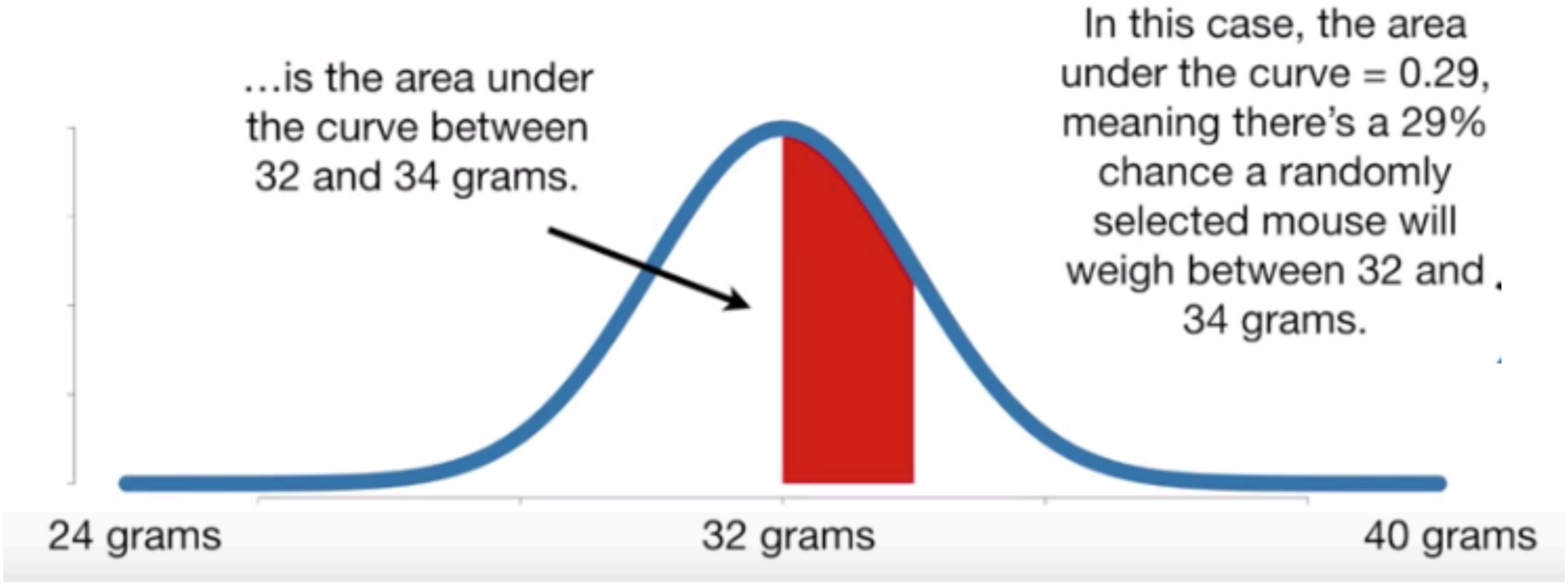
It has a mean of 32 grams…

…and a standard deviation of 2.5.

Probability



The probability that we will weigh a randomly selected mouse between 32 and 34 grams…

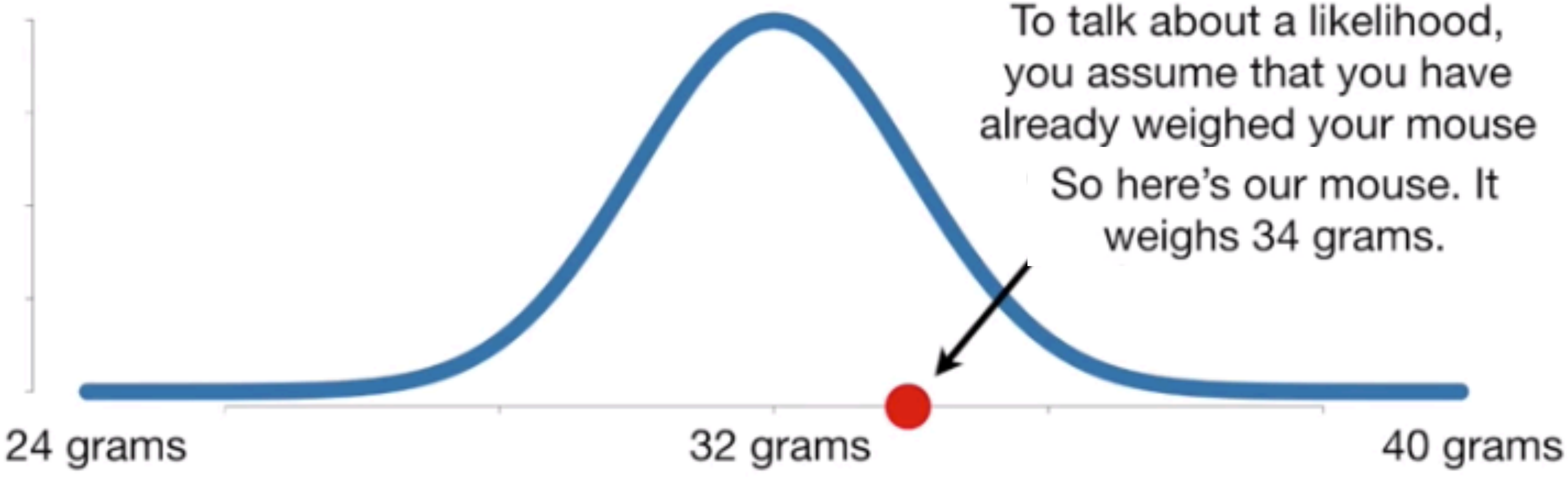


…is the area under the curve between 32 and 34 grams.

In this case, the area under the curve = 0.29, meaning there's a 29% chance a randomly selected mouse will weigh between 32 and 34 grams.

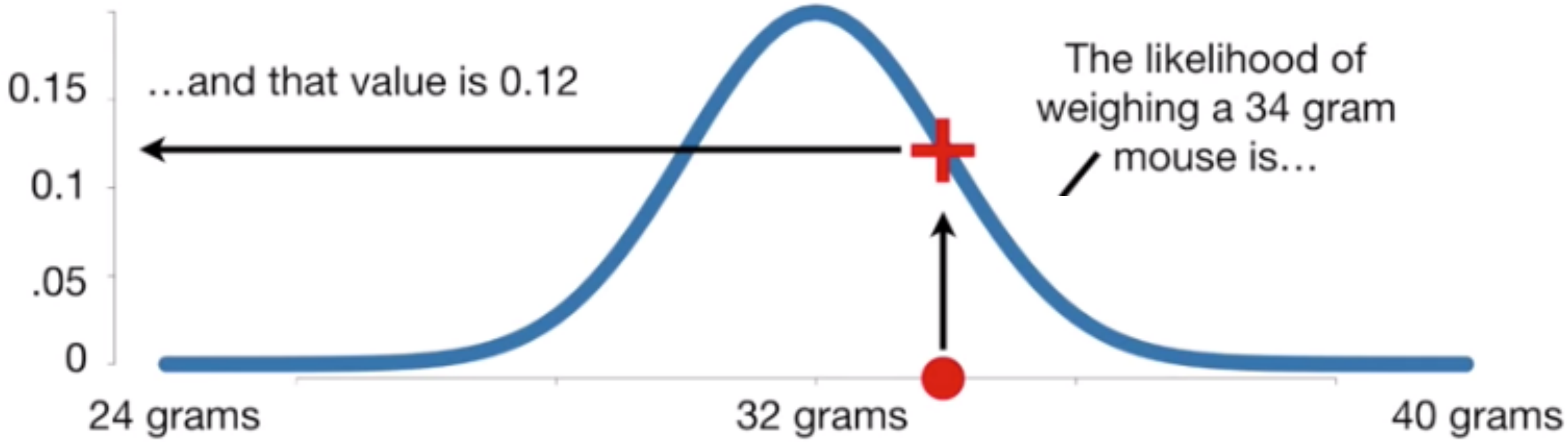
pr(weight between 32 and 34 grams | mean = 32 and standard deviation = 2.5)

Likelihood



To talk about a likelihood, you assume that you have already weighed your mouse

So here's our mouse. It weighs 34 grams.

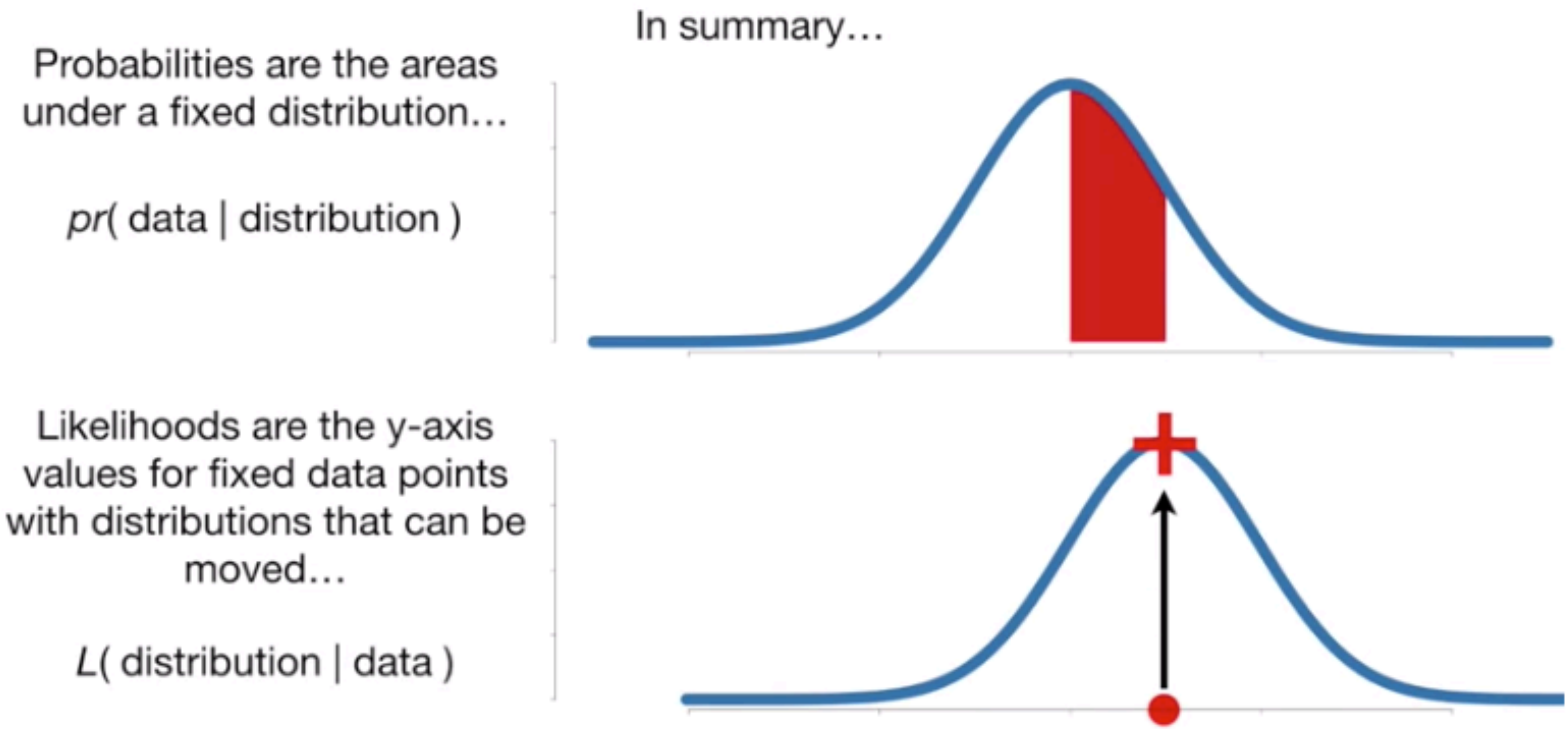


The likelihood of weighing a 34 gram mouse is…

…and that value is 0.12

L(mean = 32 and standard deviation = 2.5 | mouse weighs 34 grams)

Probability versus Likelihood



In summary…

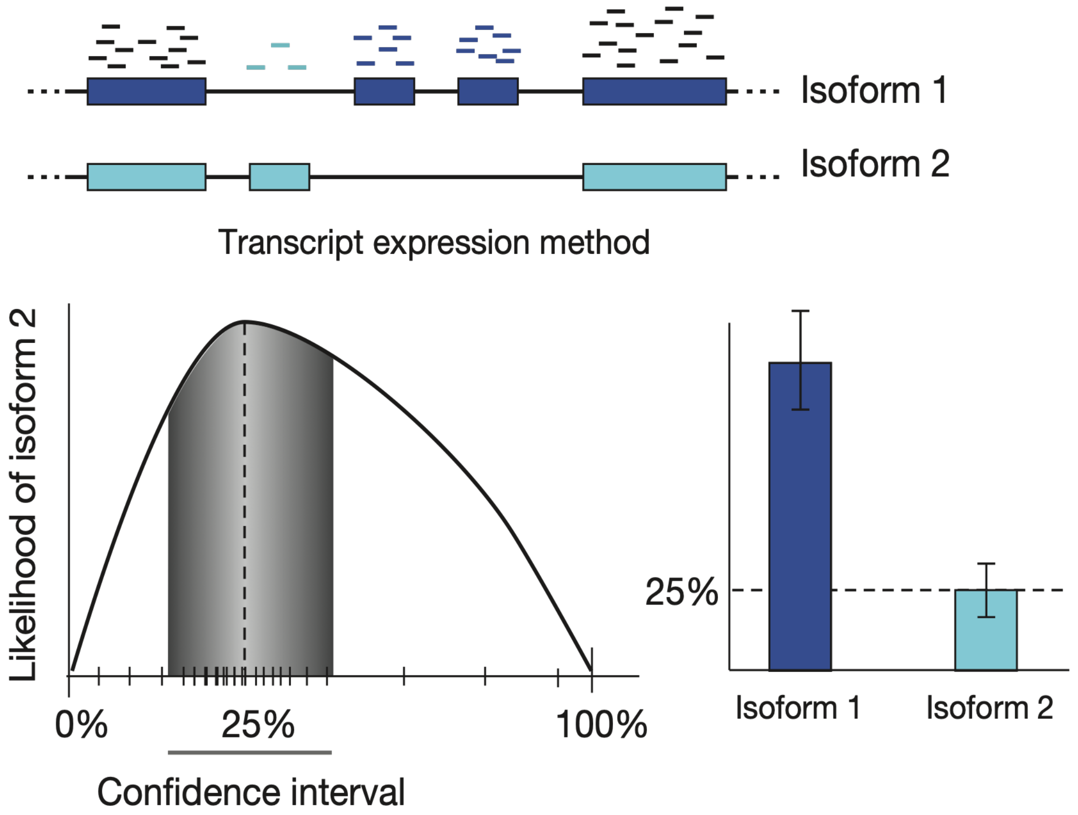
Probabilities are the areas under a fixed distribution…

pr(data | distribution)

Likelihoods are the y-axis values for fixed data points with distributions that can be moved…

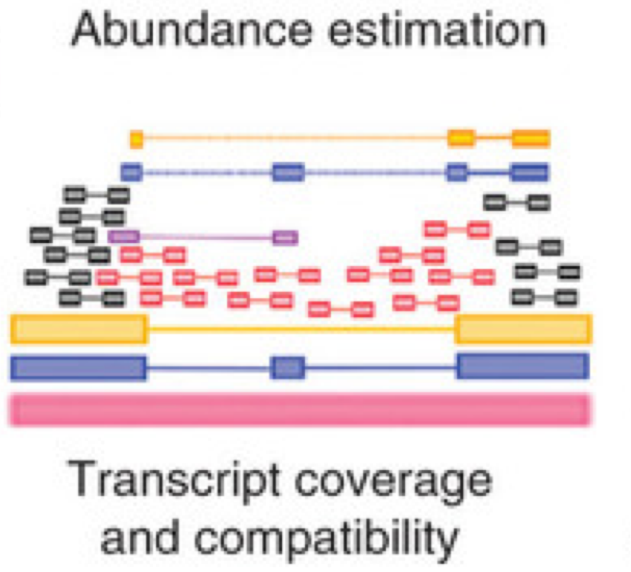
L(distribution | data)

Estimating transcript expression levels



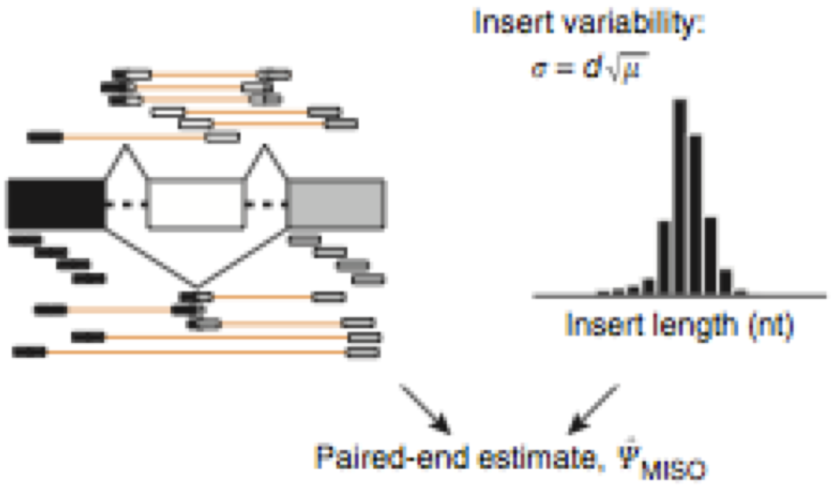
Reads from alternatively spliced genes may be attributable to a single isoform or more than one isoform. Reads are color-coded when their isoform of origin is clear. Black reads indicate reads with uncertain origin. 'Isoform expression methods' estimate isoform abundances that best explain the observed read counts under a generative model. Samples near the original maximum likelihood estimate (dashed line) improve the robustness of the estimate and provide a confidence interval around each isoform's abundance.

Estimating transcript expression levels



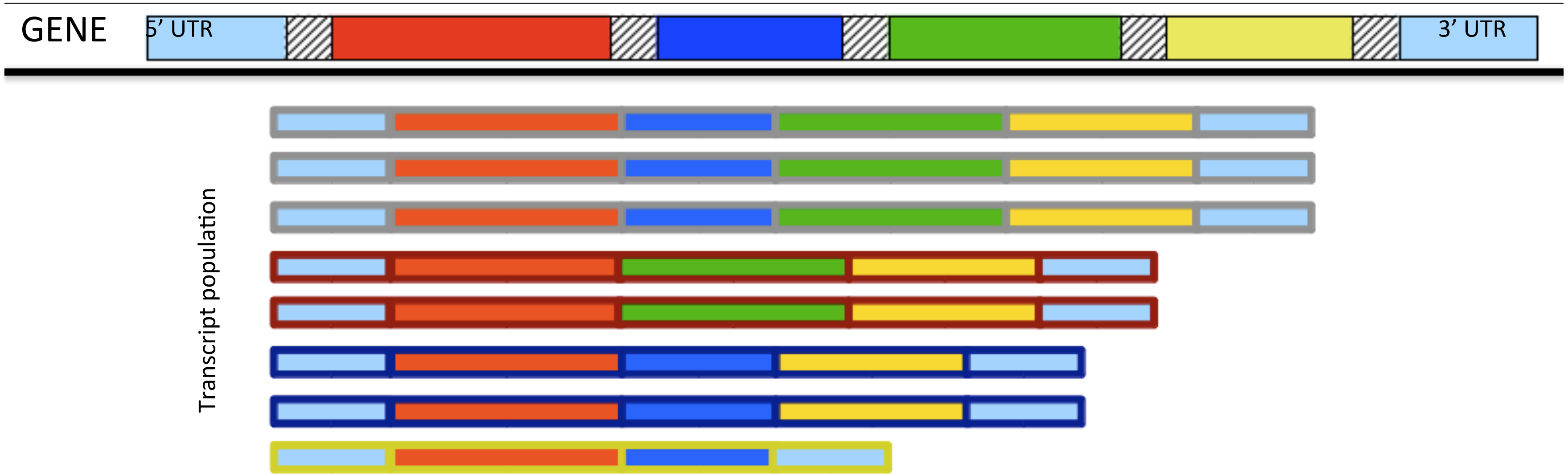
Cufflinks estimates transcript abundances using a statistical model in which the probability of observing each fragment is a linear function of the abundances of the transcripts from which it could have originated.

To estimate the transcripts abundance we need to have two inputs:

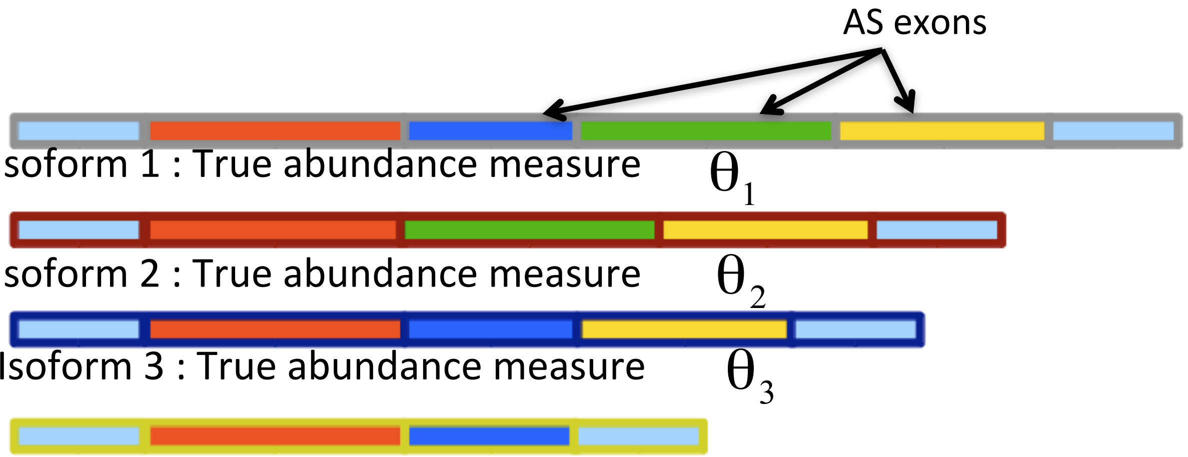


Because only the ends of each fragment are sequenced, the length of each may be unknown. Assigning a fragment to different isoforms often implies a different length for it. Cufflinks can incorporate **the distribution of fragment lengths to help assign fragments to isoforms.**

Estimating transcript expression levels



Suppose we have a gene with 4 isoforms and 3 alternatively spliced (AS) exons as shown above.



Isoform 1 : True abundance measure θ_1

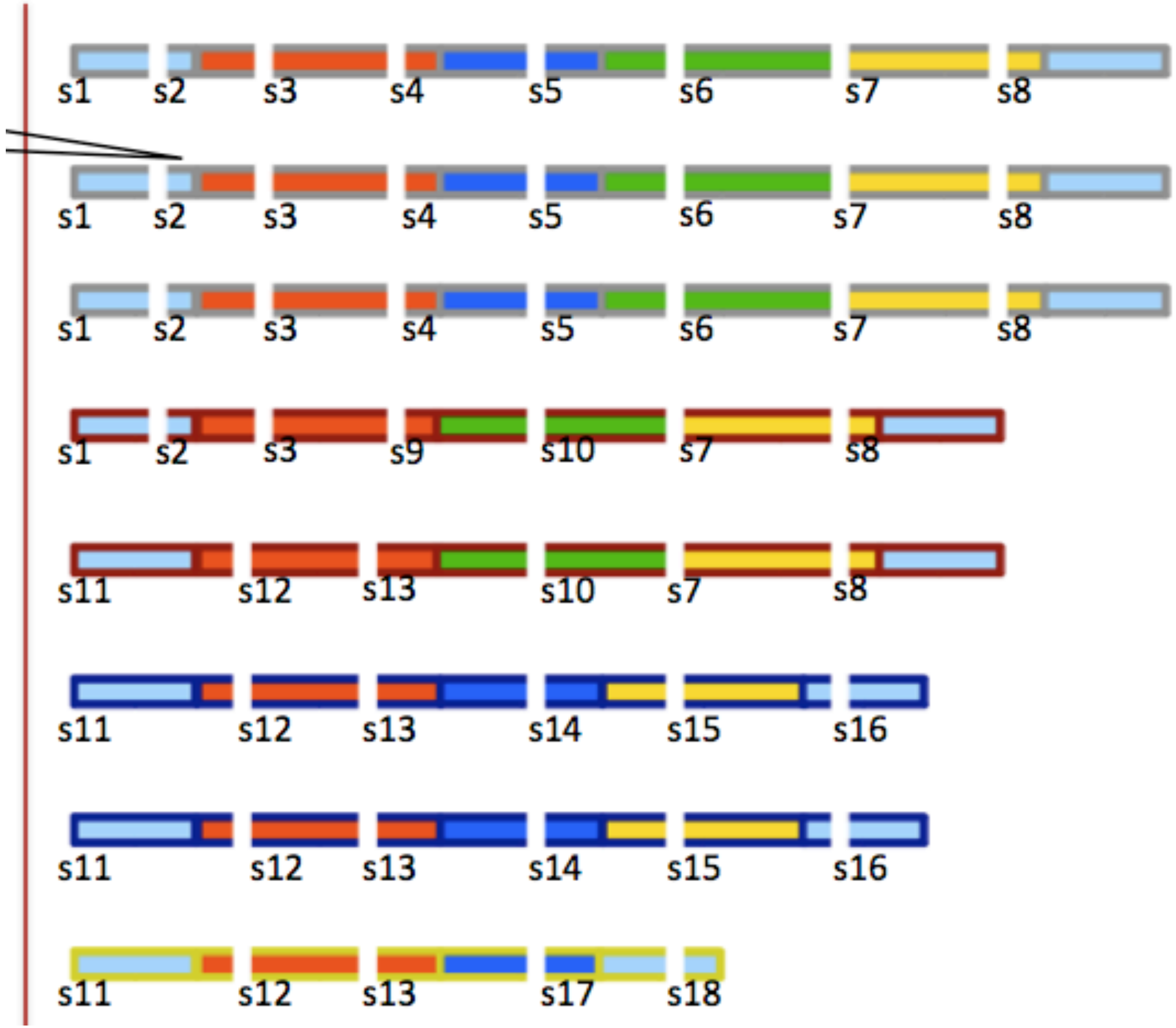
Isoform 2 : True abundance measure θ_2

Isoform 3 : True abundance measure θ_3

Isoform 4 : True abundance measure θ_4

The goal is to estimate the true abundance measure of the 4 isoforms.

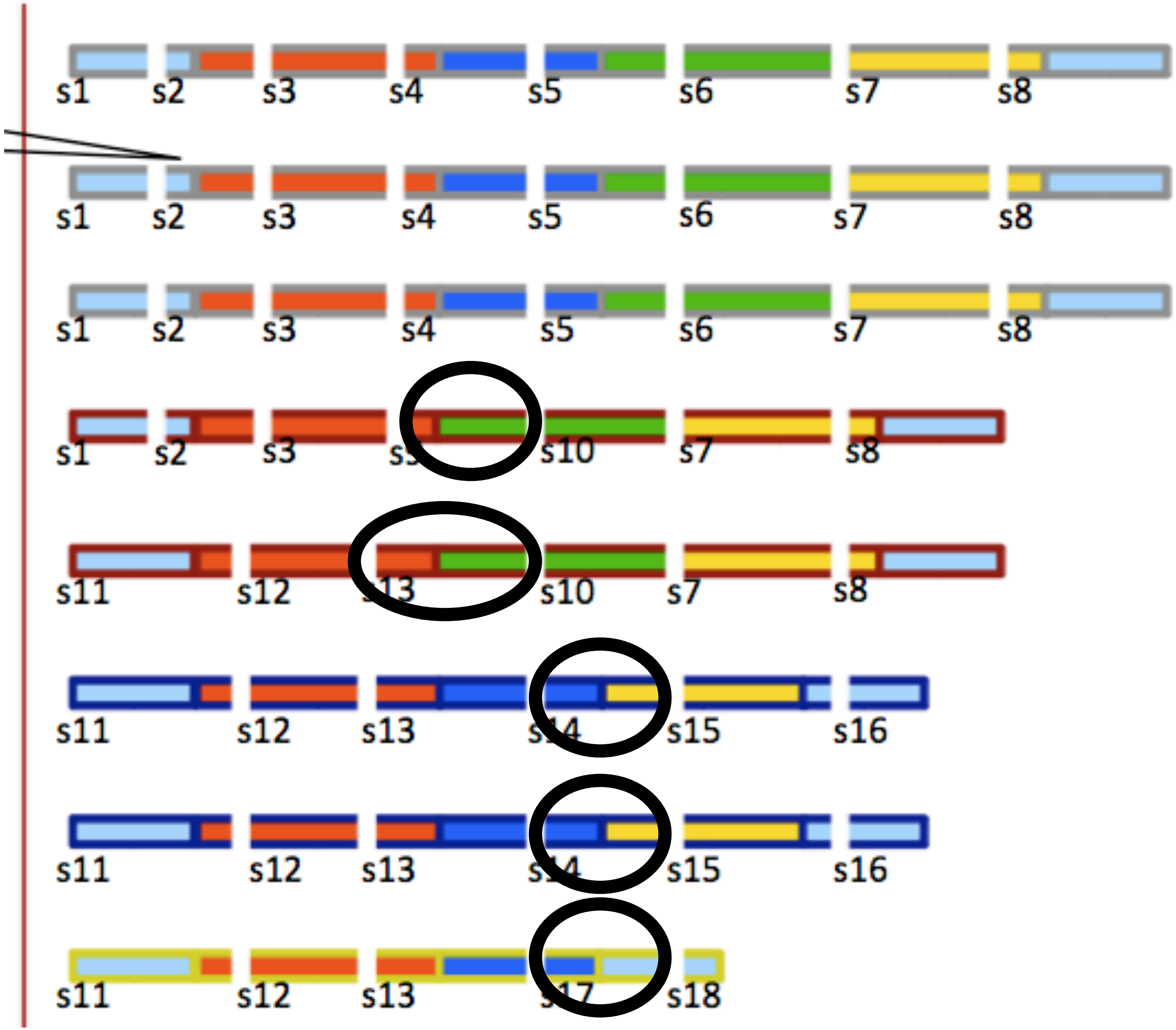
Estimating transcript expression levels



Reads that could have originated from multiple transcripts are informative.

Relative abundance estimation requires "discriminatory reads"

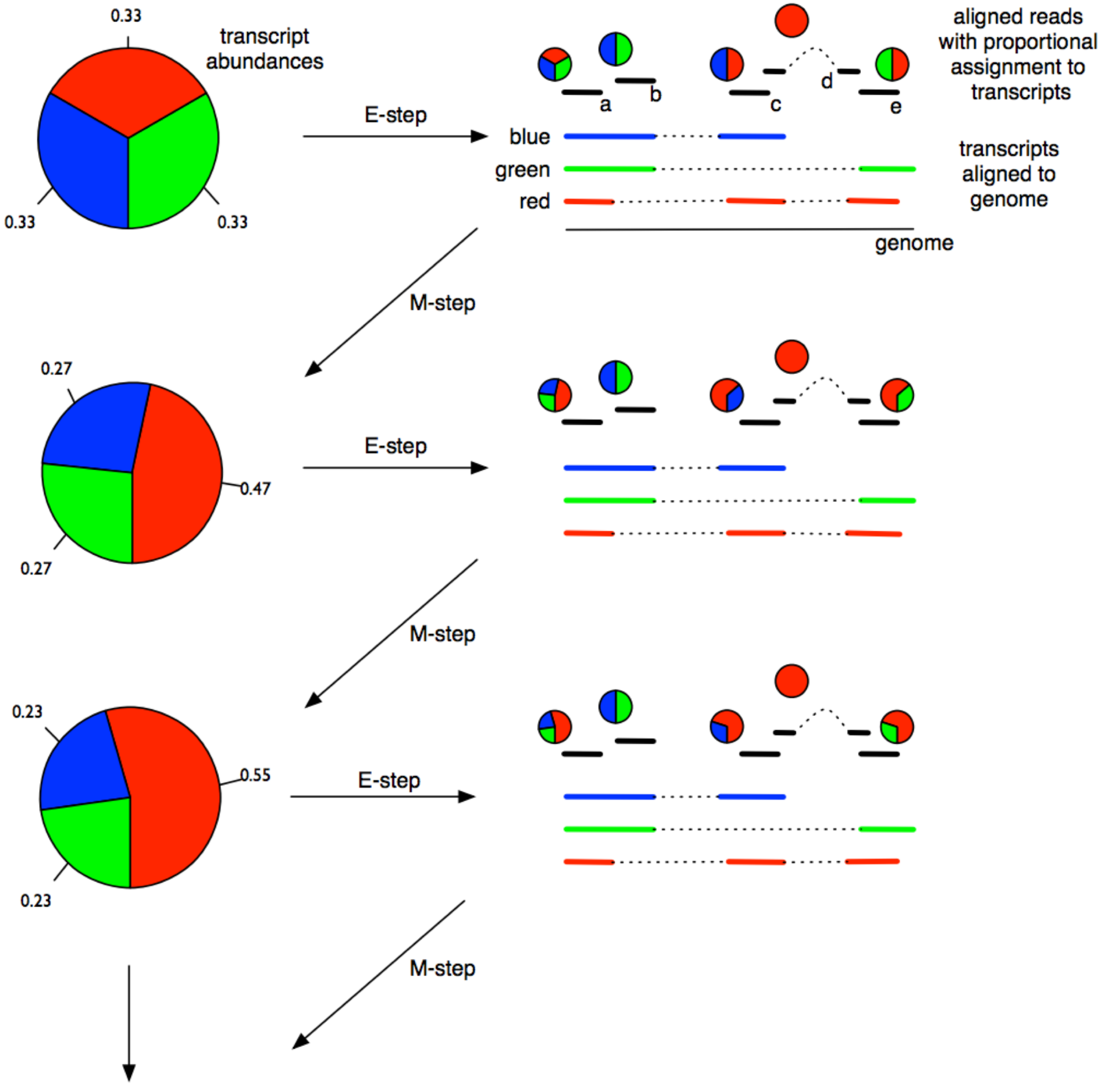
Estimating transcript expression levels



Reads that could have originated from multiple transcripts are informative.

Relative abundance estimation requires "**discriminatory reads**"

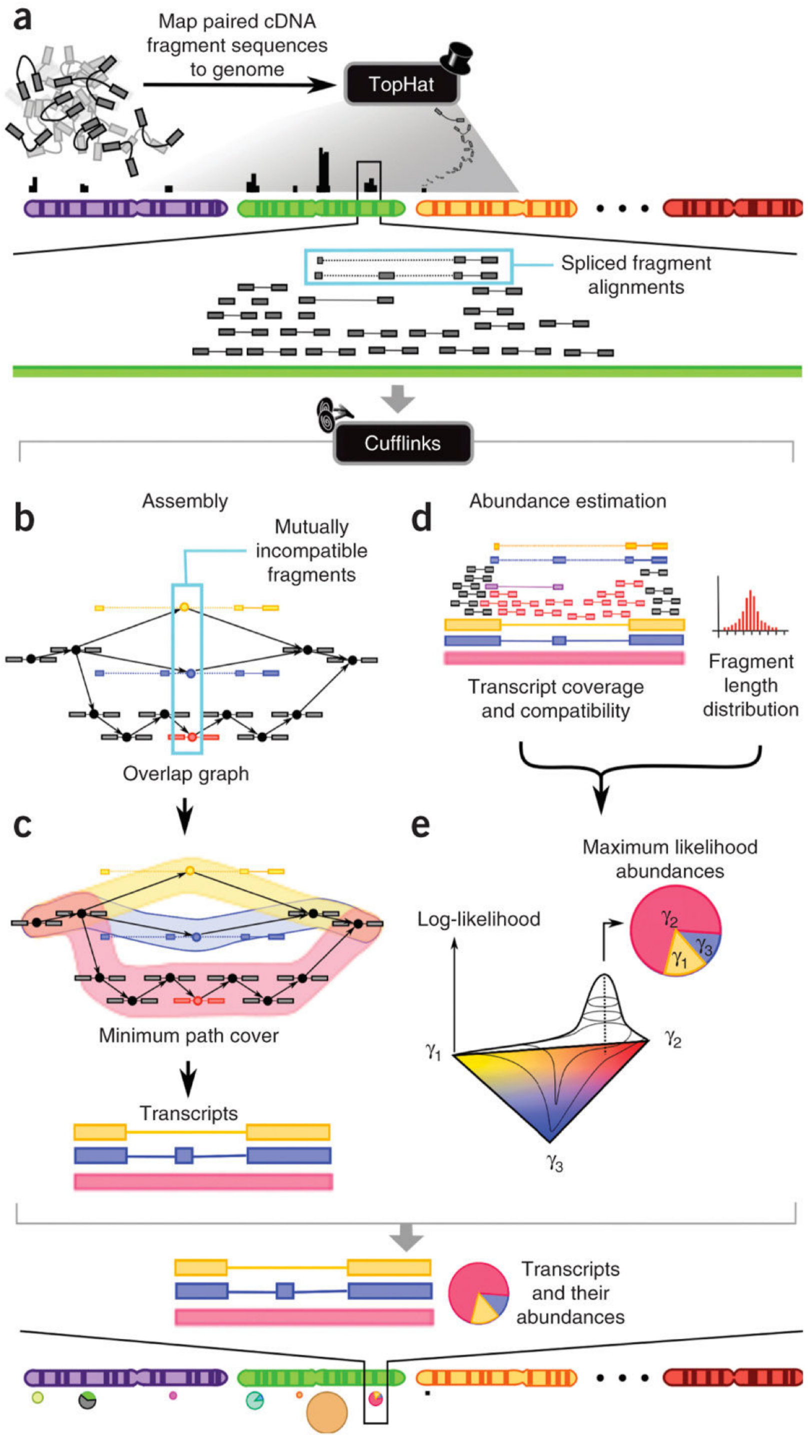
Estimating transcript expression levels



⋮ Until the M and E steps converge

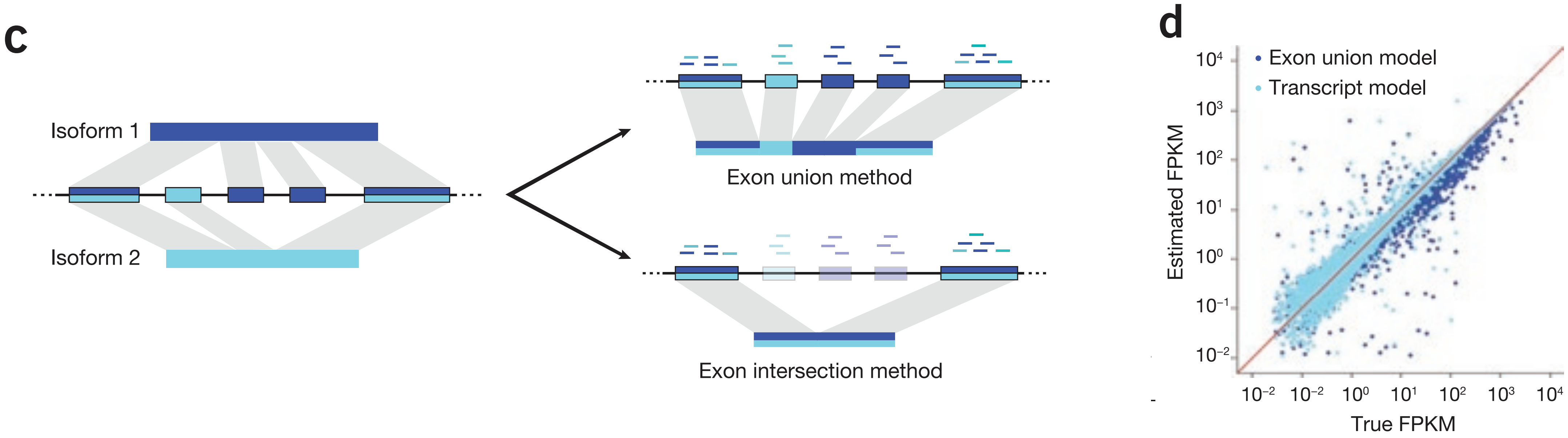
The gene has three isoforms (red, green, blue) of the same length. There are five reads (a,b,c,d,e) mapping to the gene. One maps to all three isoforms, one only to red, and the other three to each of the three pairs of isoforms. Initially every isoform is assigned the same abundance (1 3 , 1 3 , 1 3). During the expectation (E) step reads are proportionately assigned to transcripts according to the isoform abundances. Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts. Thus, for example, the abundance of red after the first M step is estimated by 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33). The number of reads associated with each transcripts as denominator

Pipeline


a
Map paired cDNA fragment sequences to genome
TopHat
Spliced fragment alignments
Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ_1
γ_2
γ_3
Transcripts and their abundances

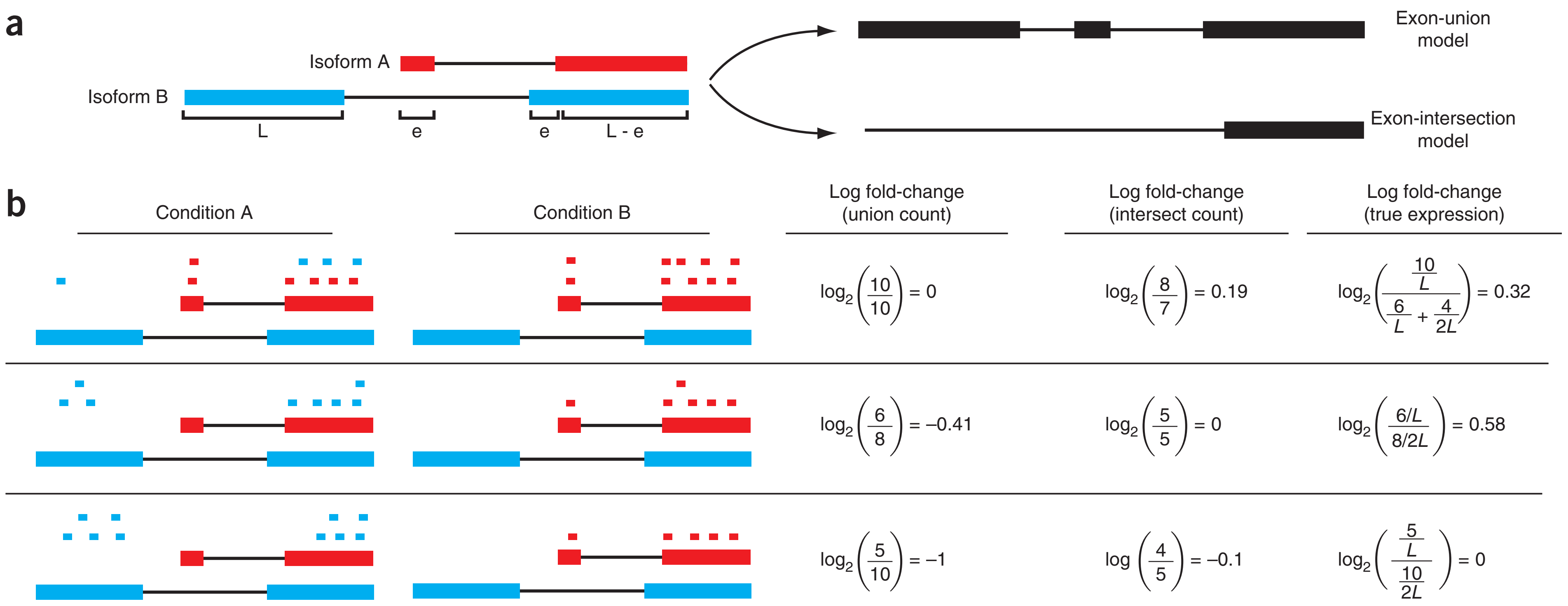
Gene quantification

For ***quantify gene expression***, the two most commonly used counting schemes are: the ***‘exon intersection method***’, which counts reads mapped to its constitutive exons, and the ‘***exon union method’*** which counts all reads mapped to any exon in any of the gene’s isoforms. The exon intersection method is analogous to expression microarrays, which typically probe expression signal in constitutive regions of each gene. Although convenient, these simplified models come at a cost; the exon union model **underestimates** expression for alternatively spliced gene, and the intersection can **reduce power for differential expression analysis**.



Gene quantification

The exon-union model counts reads falling on any of a gene's exons, whereas the exon-intersection model counts only reads on constitutive exons.



a
Isoform A
Isoform B
L
e
e
L - e
Exon-union model
Exon-intersection model
b
Condition A
Condition B
Log fold-change (union count)
Log fold-change (intersect count)
Log fold-change (true expression)
$\log_2\left(\frac{10}{10}\right) = 0$
$\log_2\left(\frac{8}{7}\right) = 0.19$
$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
$\log_2\left(\frac{6}{8}\right) = -0.41$
$\log_2\left(\frac{5}{5}\right) = 0$
$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
$\log_2\left(\frac{5}{10}\right) = -1$
$\log\left(\frac{4}{5}\right) = -0.1$
$\log_2\left(\frac{\frac{5}{L}}{\frac{10}{2L}}\right) = 0$

The true expression is estimated by the sum of the length-normalized isoform read counts.

Both simplified counting schemes register a change in count that does not reflect a change in gene expression.

Trapnell et al Nature Biotech 2013

Gene quantification

In contrast, gene expression levels calculated by isoform deconvolution correlated well with true gene expression even when relative abundance of the isoforms changed between conditions. Thus, identifying accurate, statistically significant expression changes at the **resolution level of genes requires transcript-level calculations**.

Cuffdiff 2 assumes that the expression of a **transcript** in each condition can be measured by counting the number of fragments generated by it. **A change in the expression level of a transcript is measured by comparing its fragment count in each condition**. If the chance of seeing a change in this count is small enough under an appropriate statistical model of the inherent variability in this count, the **transcript is deemed significantly differentially expressed.**