RNA Sequencing

MMG 835, SPRING 2016
Eukaryotic Molecular Genetics

George I. Mias

Department of Biochemistry and Molecular Biology
gmias@msu.edu

RNA-Sequencing

- High throughput method
- Count the number of reads (transcripts)
- Can probe entire genome for expression (non-targeted approach)
- Splicing information
- Alternative splicing and novel isoforms
- Possible to phase (algorithmically/new longer read technology)
- Can get variant information
- Now cheaper (almost comparative to microarray, cost mainly library prep)

RNA-Sequencing

RNA-Seq Advantages

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Wang, Gerstein & Snyder, Nature Review Genetics 10:57-63 (2009)

RNA-Sequencing

Keep in mind

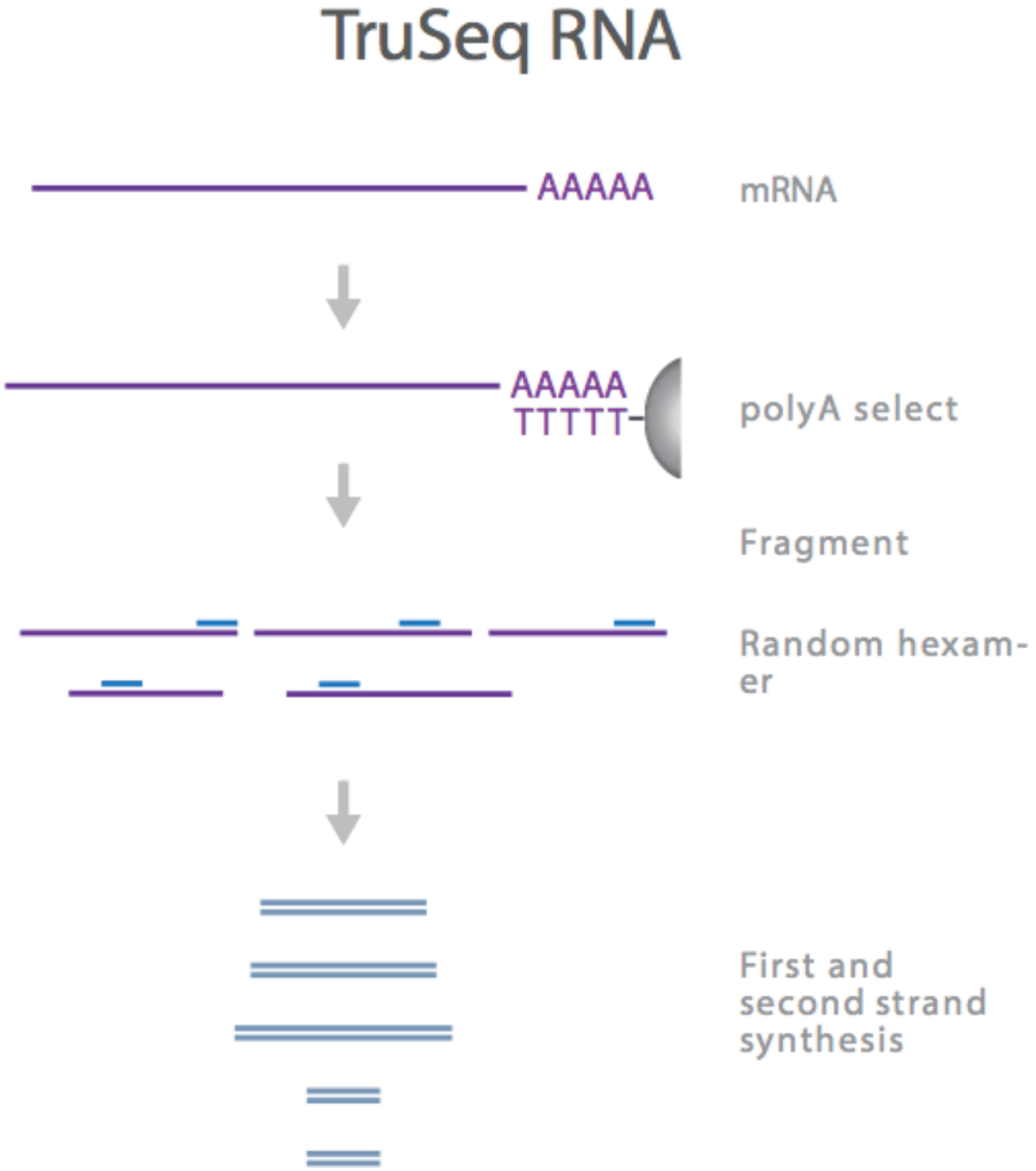
- Quantitation Issues
- Biases
- Transcribed region Complexity
- Sample quality (degradation)
- Aligner dependent alignment

Data Considerations

Steps

- Experiment Design
- Library Prep
- Sequencing
- Data
- Quality Control (QC)
- Assembly
- Annotation
- Quantification
- Differential Expression
- Biological Significance

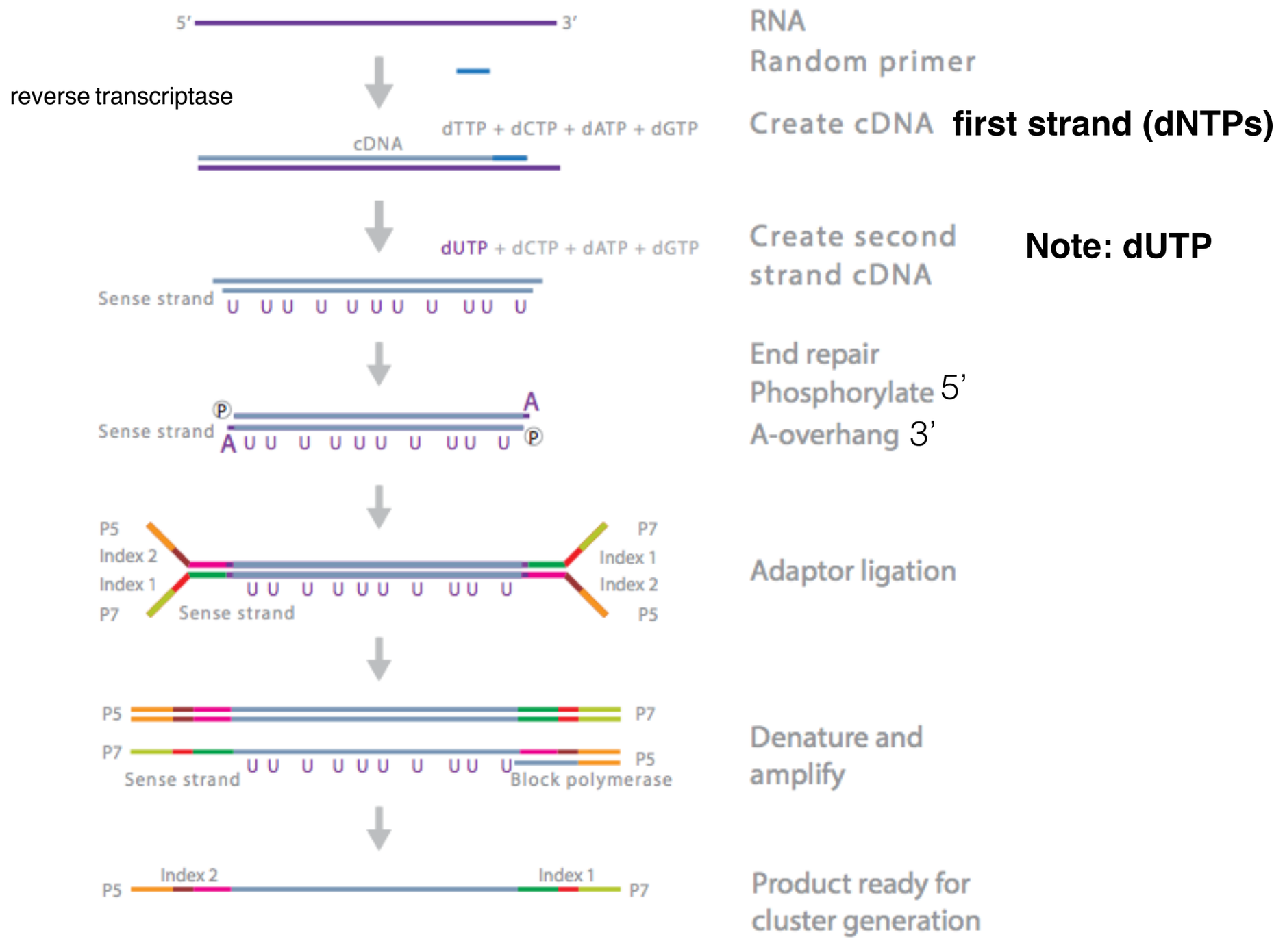
Rna-Seq Protocols



illumina.com

Rna-Seq Protocols

TruSeq RNA Stranded



Preserve strandness information

Novel transcripts/ annotation

TSS

illumina.com

Rna-Seq Protocols

Possibilities

- **mRNA Poly(A)-selection**
  - hybridization to oligo-dT beads
    - N.B. 3’ bias
- **rRNA depletion**
- **small RNA**
  - size selection
- **Paired end**
- **Low RNA amounts - amplification**

Rna-Seq Protocols

Normalization

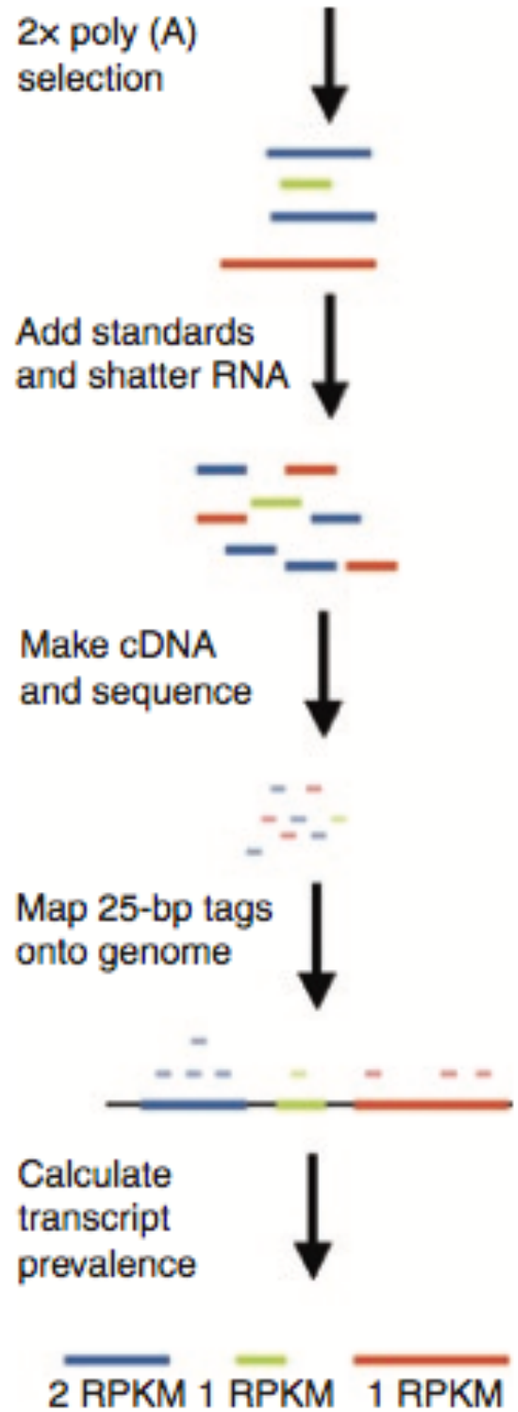
- Long transcripts
- Coverage differences

Rna-Seq Protocols

RPKM, Reads **p**er **K**ilobase per **M**illion mapped reads

FPKM, Fragments **p**er **K**ilobase per **M**illion mapped reads

(paired-end)



Mortazavi A., et al. Nature Methods 5, 621 - 628 (2008)

errors in first strand cDNA

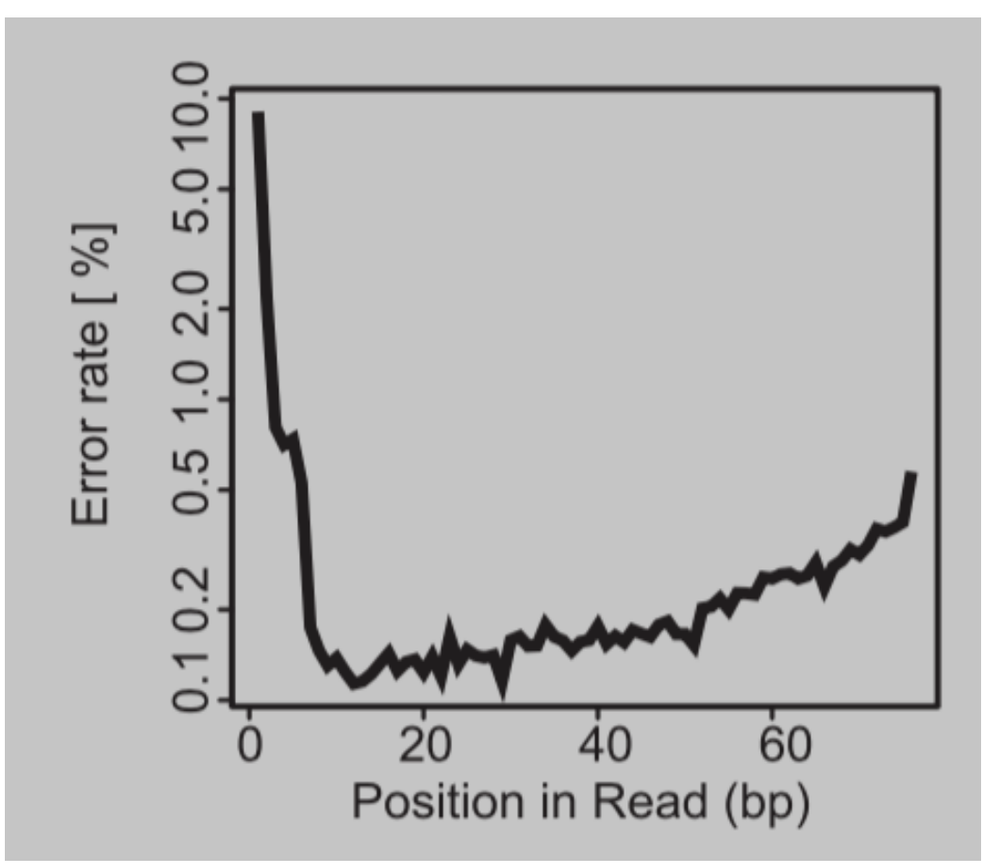
Mismatches

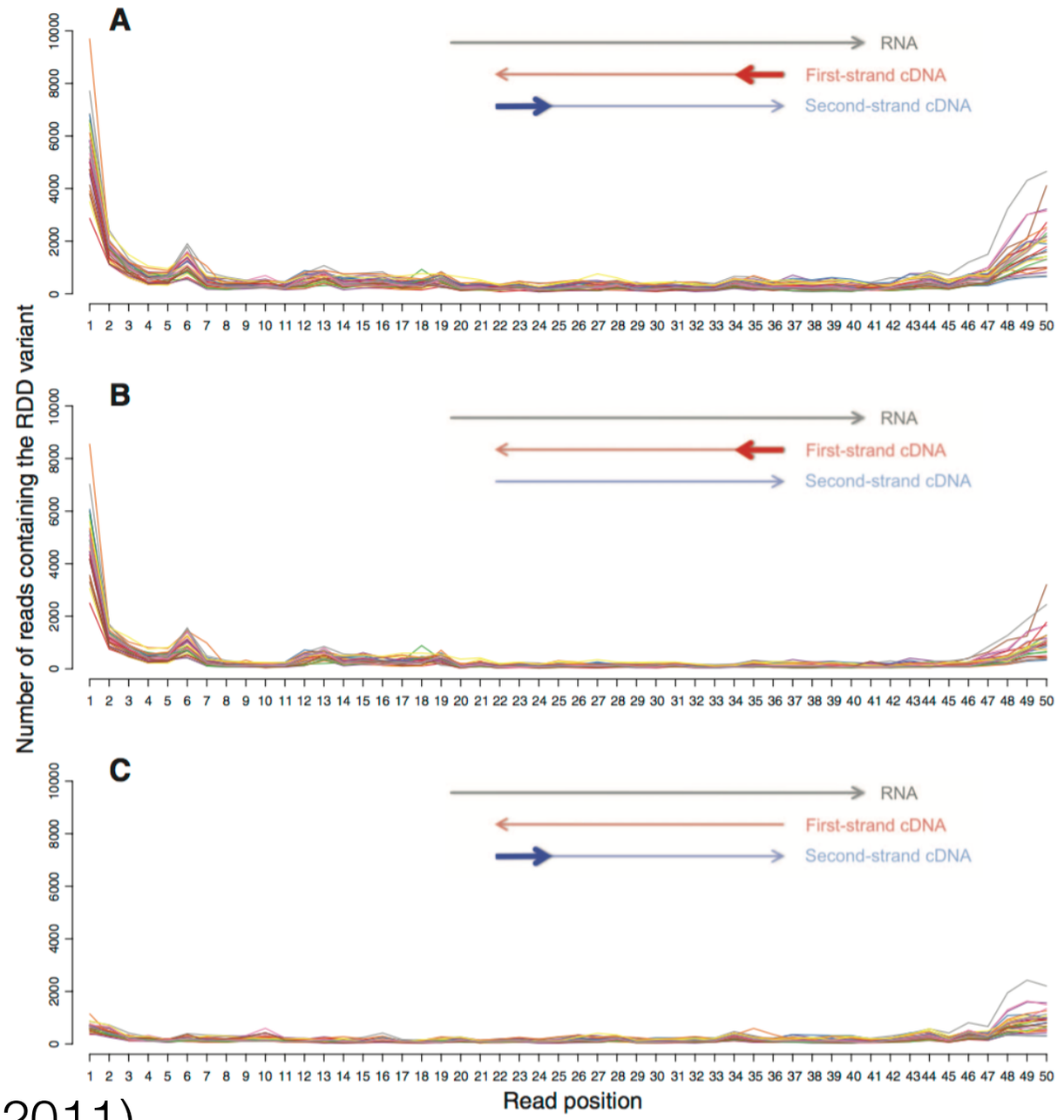
- biological variation
  - genomic DNA
  - RNA-editing
- errors in library preparation
  - hexamer mispriming
  - PCR errors
- sequencing errors [id of bases]

van Gurp, McIntyre & Verhoeven, PLoS ONE 8(12): e85583 (2013)

errors in first strand cDNA

- mismatches mainly in the first seven nucleotides of first strand cDNA

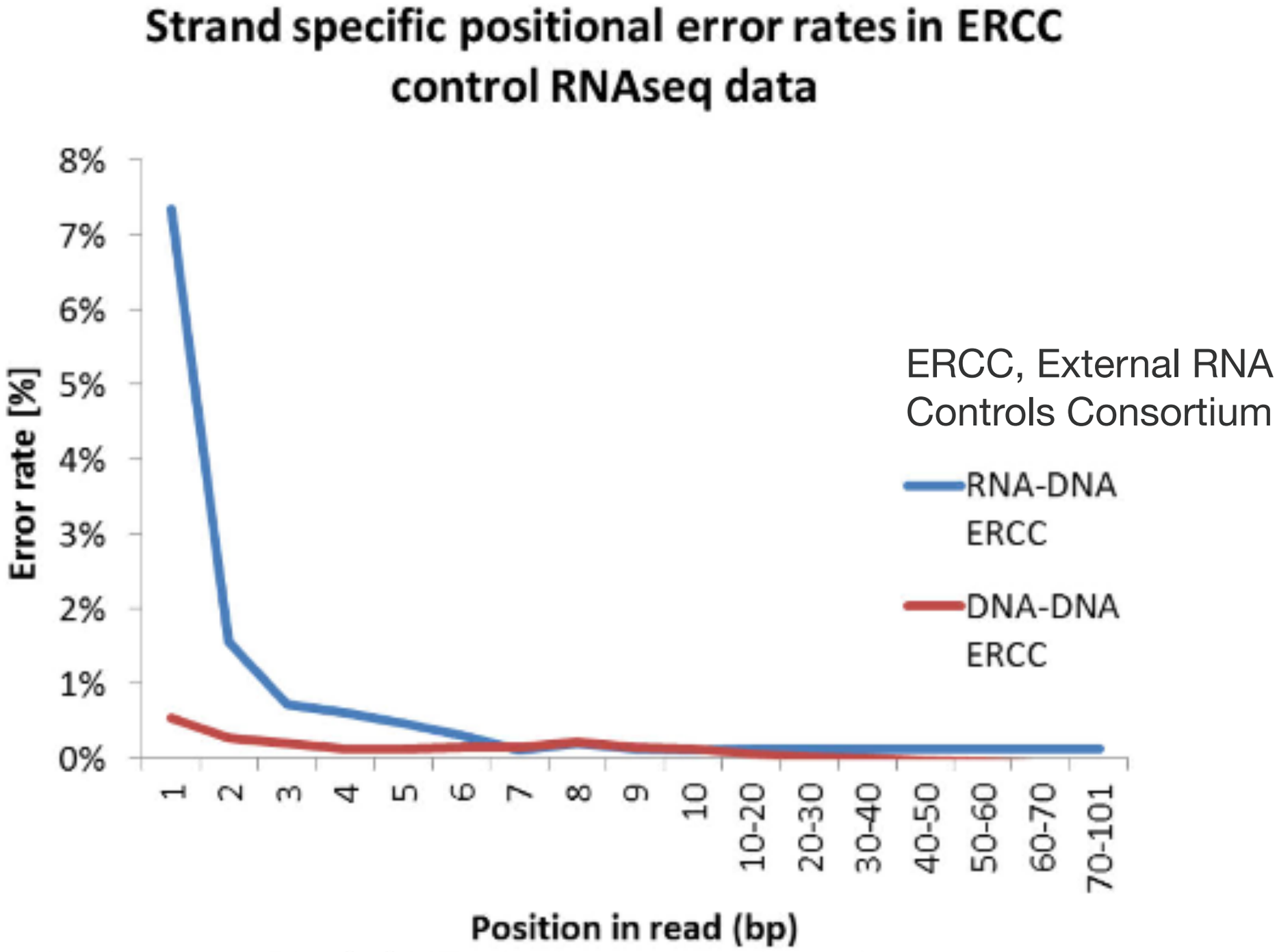




Jiang L et al., Genome Res 21: 1543– 1551 (2011)
Pickrell, Gilad & Pritchard, Science 335: 1302–authorreply1302 (2012).

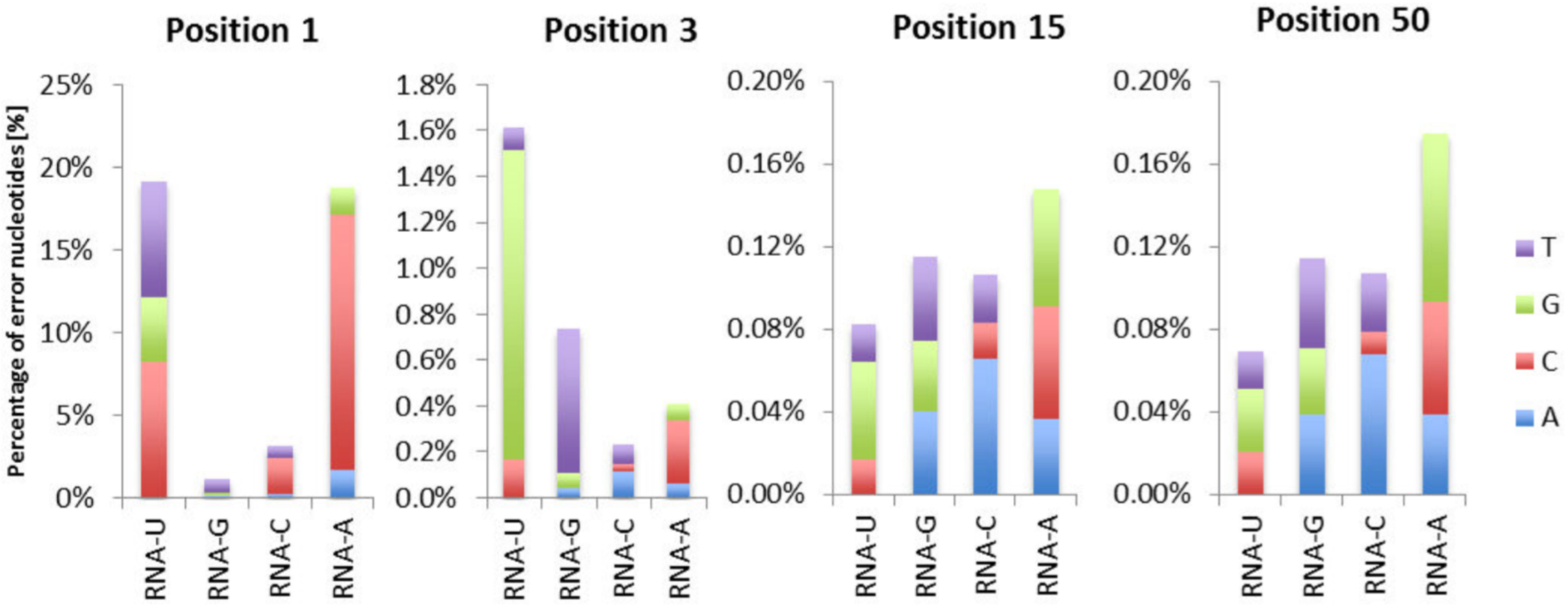
errors in first strand cDNA

random hexamer mispriming



van Gurp, McIntyre & Verhoeven, PLoS ONE 8(12): e85583 (2013)

errors in first strand cDNA



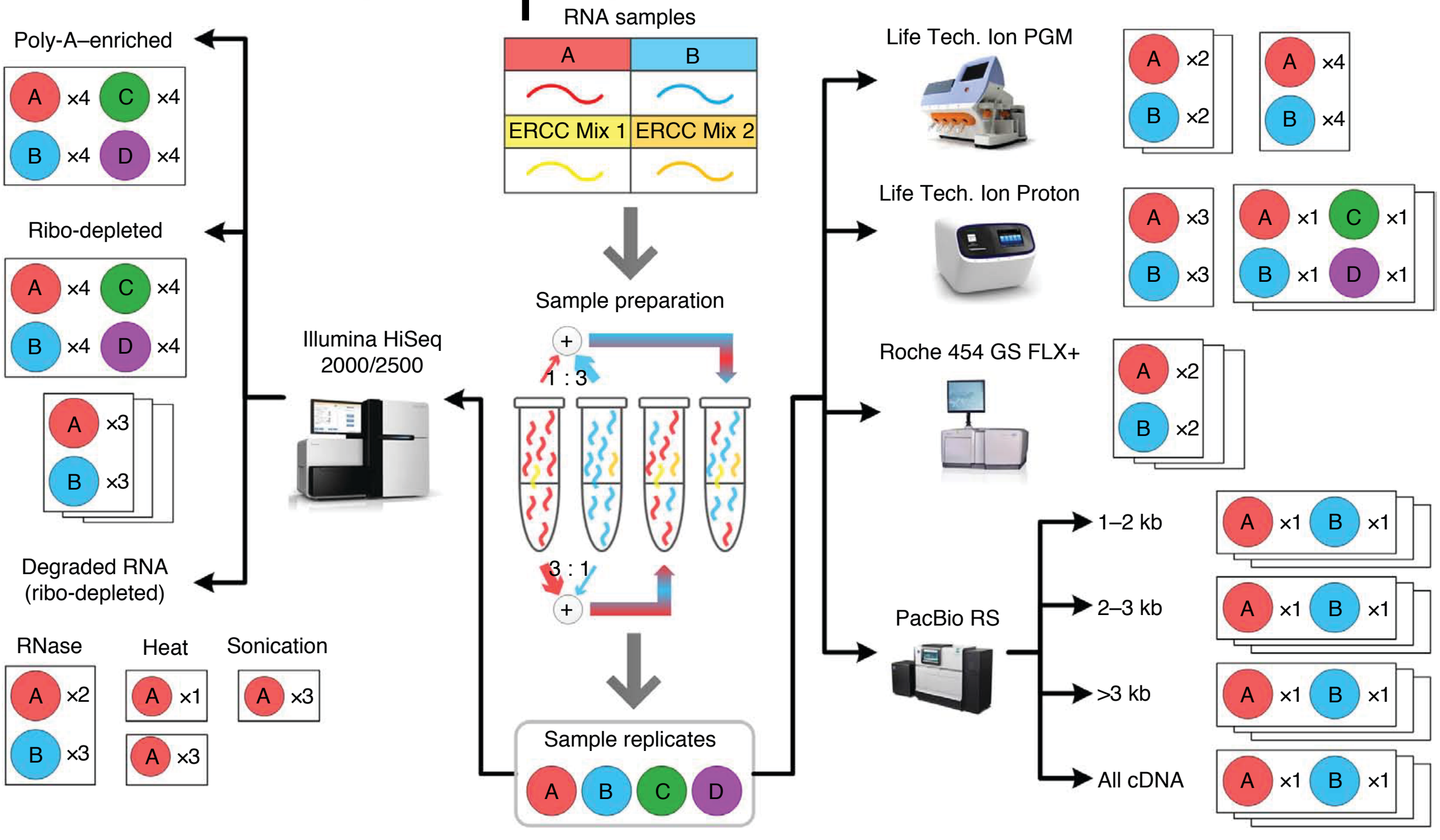
Starting with T or A errors

RNA-U vs DNA-G

mispriming

van Gurp, McIntyre & Verhoeven, PLoS ONE 8(12): e85583 (2013)

Rna-Seq Considerations



Li, Tighe et al., Nature Biotechnology 12(9), p. 915 (2014)

Data Considerations

Mapping Reference

- Genome
- Transcriptome

Alignment

- Short read alignment
- Long read alignment
- Spliced Vs. Non Spliced Alignment
- Mismatches & Duplications
- Non-Uniqueness
- Variant Calls

De Novo Assembly

- Genome based
- non-genome based

.fastq format

Four lines repeating:

1. @title and *optional* description / Sequence Identifier
2. Sequence
3. + (and *optional* repeat of title line)
4. Quality scores corresponding to Sequence (2)

Pearson & Lipman Proc. Natl Acad. Sci. USA, 85, 2444–2448 (1988)
Cock et al., Nucleic Acids Research, Vol. 38, No. 6 1767–1771 (2010)
http://support.illumina.com/content/dam/illumina-support/help/BaseSpaceHelp_v2/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm

.fastq format

Four lines repeating:

1. @title and *optional* description / Sequence Identifier
2. Sequence
3. + (and *optional* repeat of title line)
4. Quality scores corresponding to Sequence (2)

Example:

@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=

Pearson & Lipman Proc. Natl Acad. Sci. USA, 85, 2444–2448 (1988)
Cock et al., Nucleic Acids Research, Vol. 38, No. 6 1767–1771 (2010)
http://support.illumina.com/content/dam/illumina-support/help/BaseSpaceHelp_v2/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm

.fastq format

@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is filtered>:<control number>:<sample number>

Element	Requirements	Description
@	@	Each sequence identifier line starts with @
<instrument>	Characters allowed: a–z, A–Z, 0–9 and underscore	Instrument ID
<run number>	Numerical	Run number on instrument
<flowcell ID>	Characters allowed: a–z, A–Z, 0–9	
<lane>	Numerical	Lane number
<tile>	Numerical	Tile number
<x_pos>	Numerical	Run number on instrument
<y_pos>	Numerical	X coordinate of cluster
<read>	Numerical	Read number. 1 can be single read or Read 2 of paired-end
<is filtered>	Y or N	Y if the read is filtered (did not pass), N otherwise
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. On HiSeq X systems, control specification is not performed and this number is always 0.
<sample number>	Numerical	Sample number from sample sheet

http://support.illumina.com/content/dam/illumina-support/help/BaseSpaceHelp_v2/Content/Vault/Informatics/Sequencing_Analysis/BS/swSEQ_mBS_FASTQFiles.htm

Phred Scores

- Phred Algorithm
- Used with Sanger Data.
- Algorithm assigns probability of error in calling a base

Quality score (Q-Score)

$Q = -10 \times \log(\text{error probability})$

$P = 10^{-Q/10}$

Ewing et al., Genome Research 8: 175-185 (1998).
Ewing & Green, Genome Research 8: 186-194 (1998).

Q-Scores

Quality Score	Error Probability
Q40	0.0001 (1 in 10,000)
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)

Quality score (Q-Score)

Q= -10 x log(error probability)

$P=10^{-Q/10}$

http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_understanding_quality_scores.pdf

Q-Scores

Encode Q-Score to characters Using ASCII

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

(ASCII, American Standard Code for Information Interchange)

https://simple.wikipedia.org/wiki/ASCII

Q-Scores

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
 ........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                    3....9.............................40
  0.2.....................26...31.......41
```

```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

ASCII Conversion for different schemes

https://en.wikipedia.org/wiki/FASTQ_format#Encoding

Aligners

- Unspliced
  - Fast
  - Known exons and splice junctions
  - Reference required
    - c.f. DNA alignment
      - Borrows Wheeler Transform based (BWT)
        - BWA
        - Bowtie
      - SHRiMP
      - Stampy

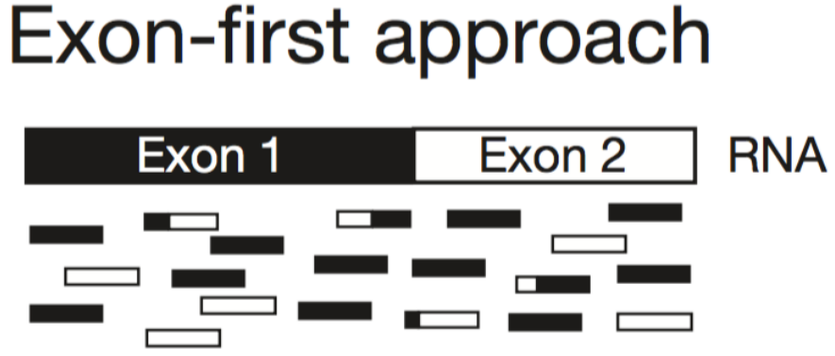
Garber et al., Nature Methods 8, 469–477 (2011)

Aligners

- Align to whole genome, including intron-spanning reads that allow large gaps
- Exon first
  - TopHat, MapSplice
  - Unspliced alignment first
  - Unmapped reads after
- Seed-extend
  - GSNAP, QPALMA
  - Short seeds mapped first
- *K-mer mappers*

Rna-Seq Alignment

Example 2 Exons

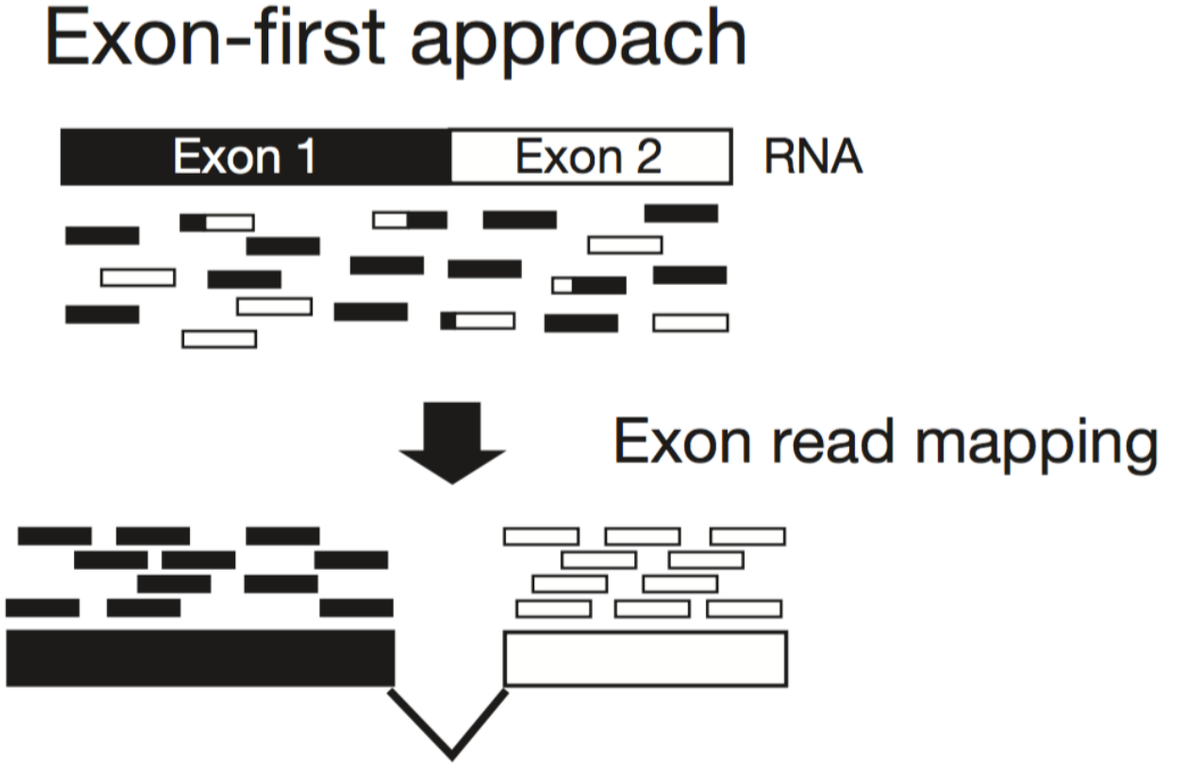


Garber et al., Nature Methods 8, 469–477 (2011)

Rna-Seq Alignment

Example 2 Exons

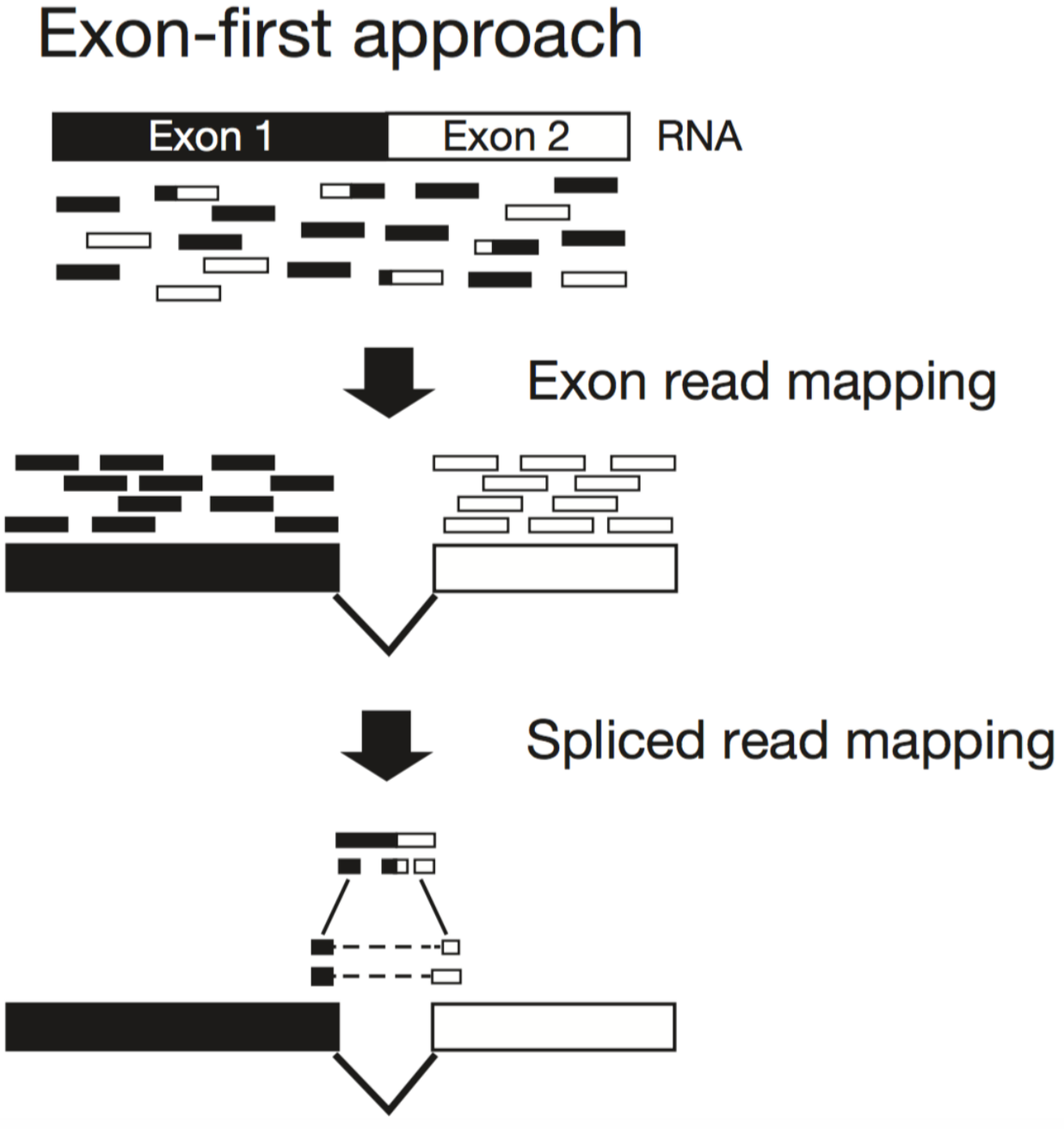
map full, unspliced
reads (exonic reads)



Garber et al., Nature Methods 8, 469–477 (2011)

Rna-Seq Alignment

Example 2 Exons

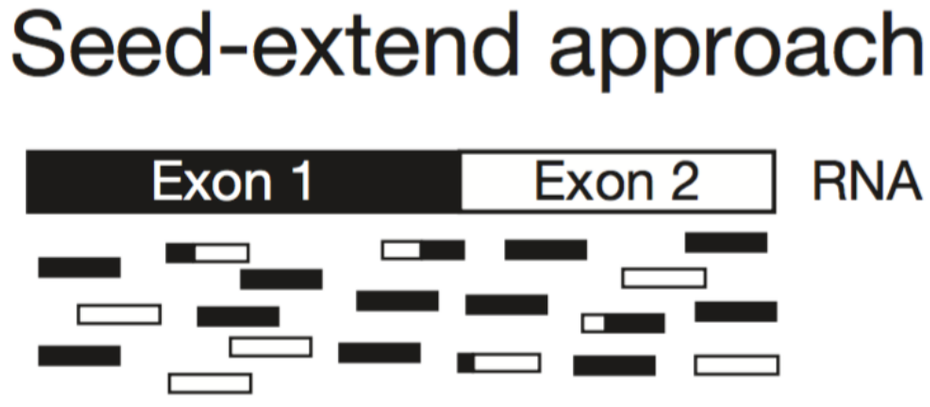


map full, unspliced reads (exonic reads)

remaining reads divided into smaller pieces & mapped to genome

Garber et al., Nature Methods 8, 469–477 (2011)

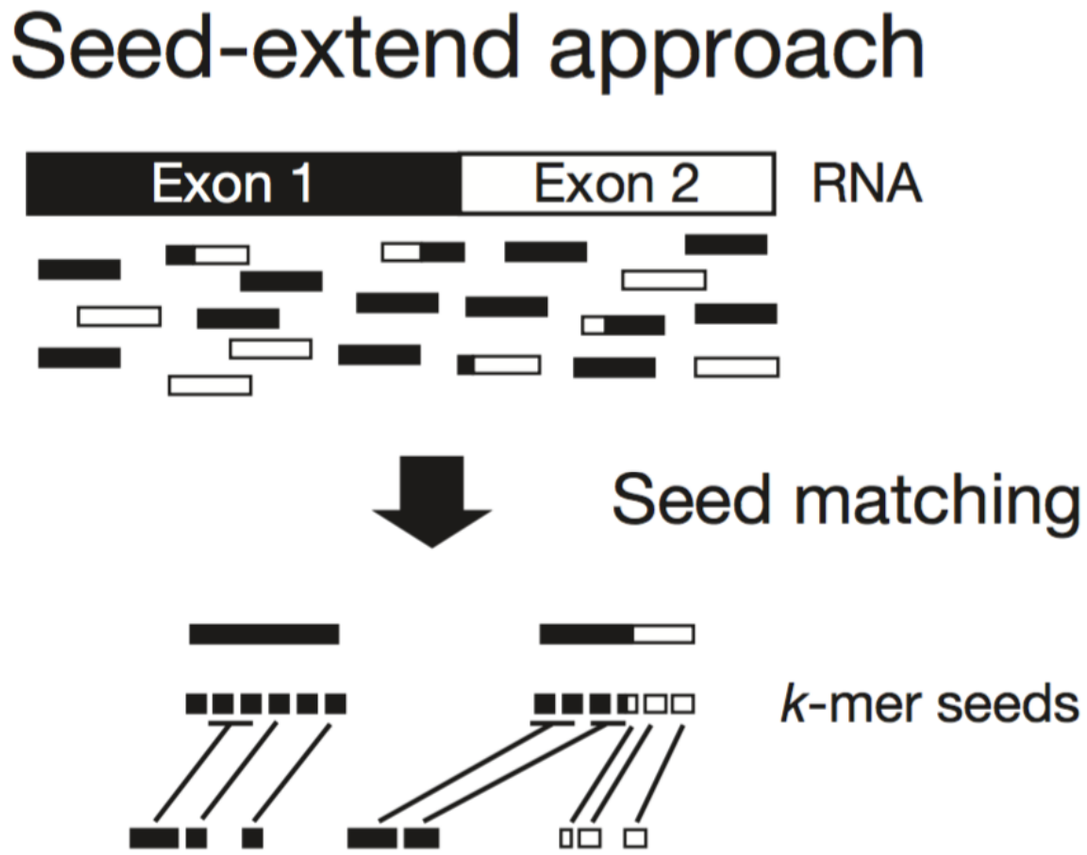
Rna-Seq Alignment



store a map of all small words (k-mers) of similar size in the genome in an efficient lookup data structure

Garber et al., Nature Methods 8, 469–477 (2011

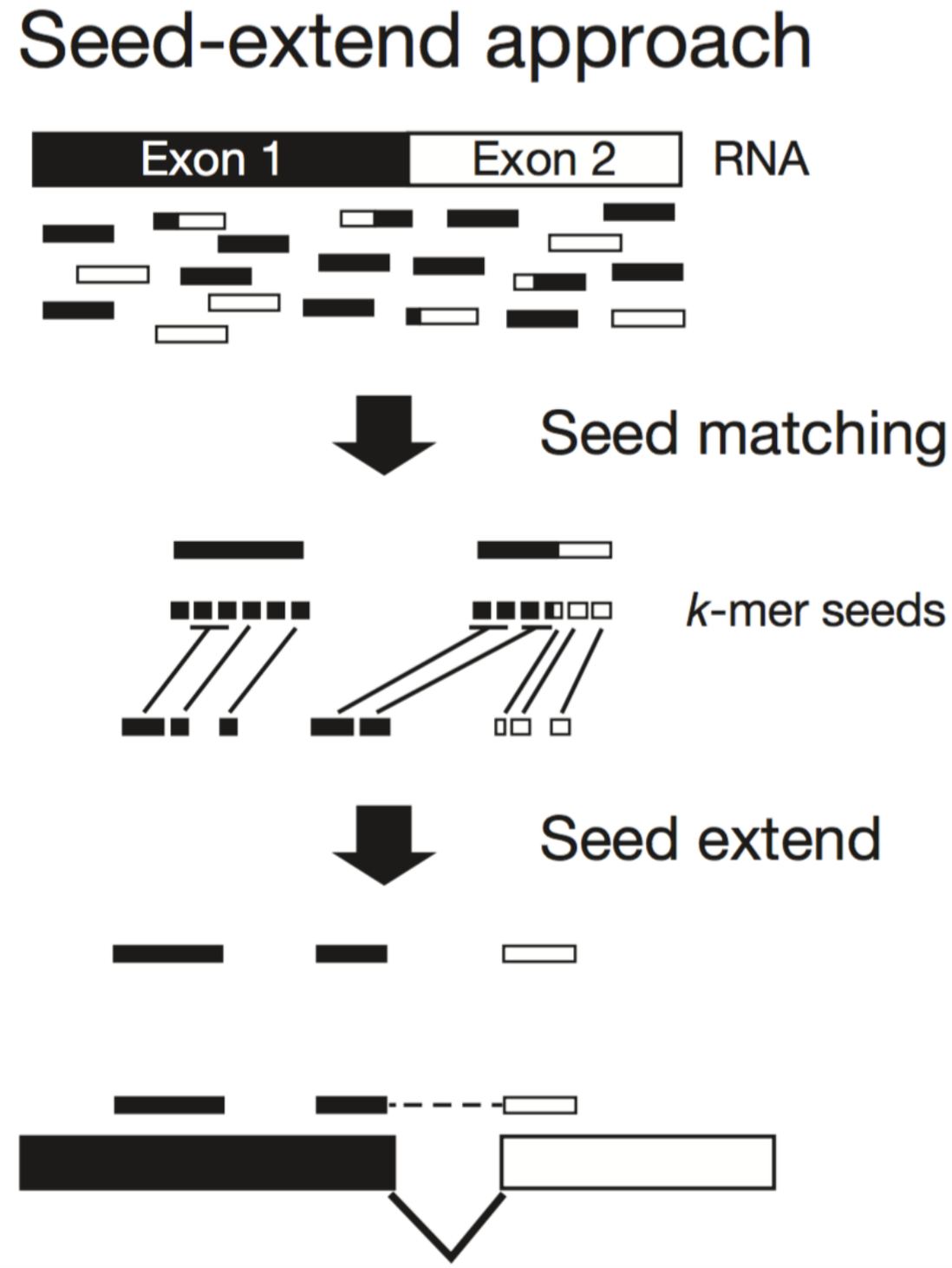
Rna-Seq Alignment



- each read is divided into k-mers
- k-mers mapped to the genome via the lookup structure.

Garber et al., Nature Methods 8, 469–477 (2011)

Rna-Seq Alignment

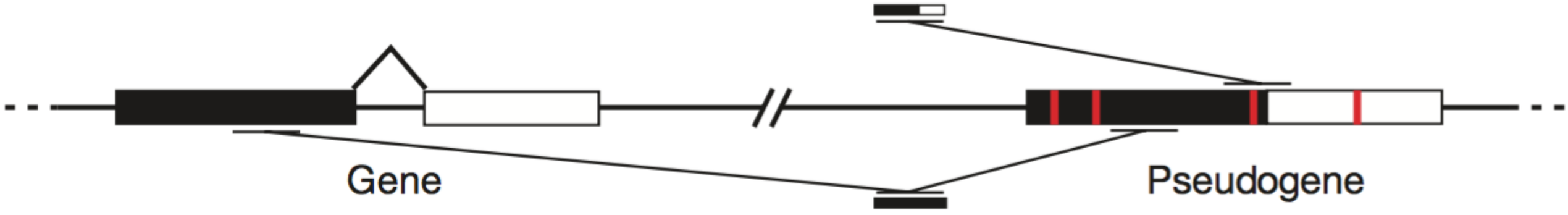


- Mapped k-mers extended into larger alignments
- May include gaps flanked by splice sites.

Garber et al., Nature Methods 8, 469–477 (201

Rna-Seq Alignment

Potential limitations of exon-first approaches



Example : gene & associated retrotransposed pseudogene.

Pseudogenes:

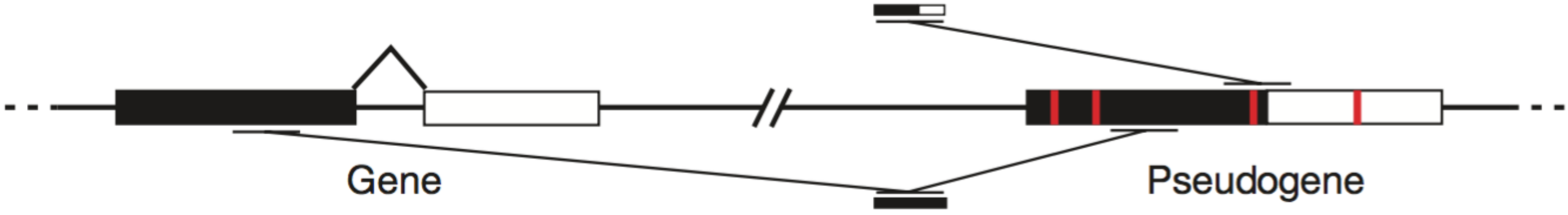
defunct genomic loci with sequence similarity to functional genes

but *lacking coding potential* due to the presence of disruptive mutations such as frame shifts and premature stop codons.

Pei et al., Genome Biology 13:R51 (2012).
Garber et al., Nature Methods 8, 469–477 (2011).

Rna-Seq Alignment

Potential limitations of exon-first approaches



Example : gene & associated retrotransposed pseudogene.

Categories of Pseudogenes

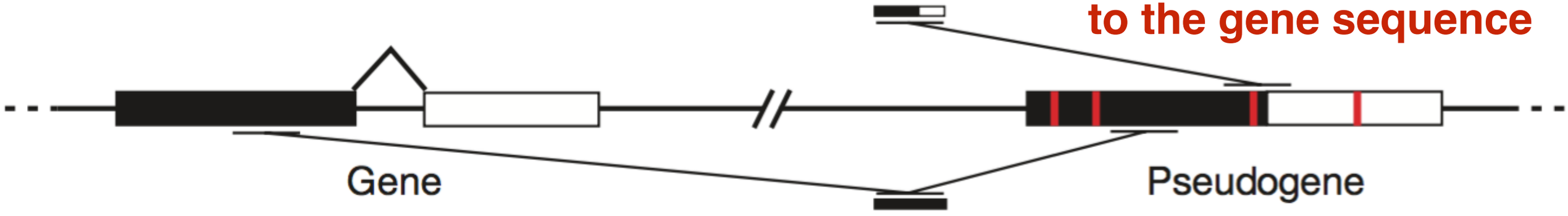
1. **Processed pseudogenes** Created by retrotransposition of mRNA from functional protein-coding loci back into the genome
2. **Duplicated (also referred to as unprocessed) pseudogenes** Derived from duplication of functional genes
3. **Unitary pseudogenes** Arise through in situ mutations in previously functional protein-coding genes

Pei et al., Genome Biology 13:R51 (2012).
Garber et al., Nature Methods 8, 469–477 (2011).

Rna-Seq Alignment

Potential limitations of exon-first approaches

Mismatches compared to the gene sequence



Example : gene & associated retrotransposed pseudogene.

- Exonic reads will map to both the gene and its pseudogene
- Gene placement preferred owing to lack of mutations
- Spliced read could be incorrectly assigned to the pseudogene as it appears to be exonic, preventing higher-scoring spliced alignments from being pursued.

Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

Reads originating from two different isoforms of the same genes colored **black** and **gray**.

Garber et al., Nature Methods 8, 469–477 (2011).

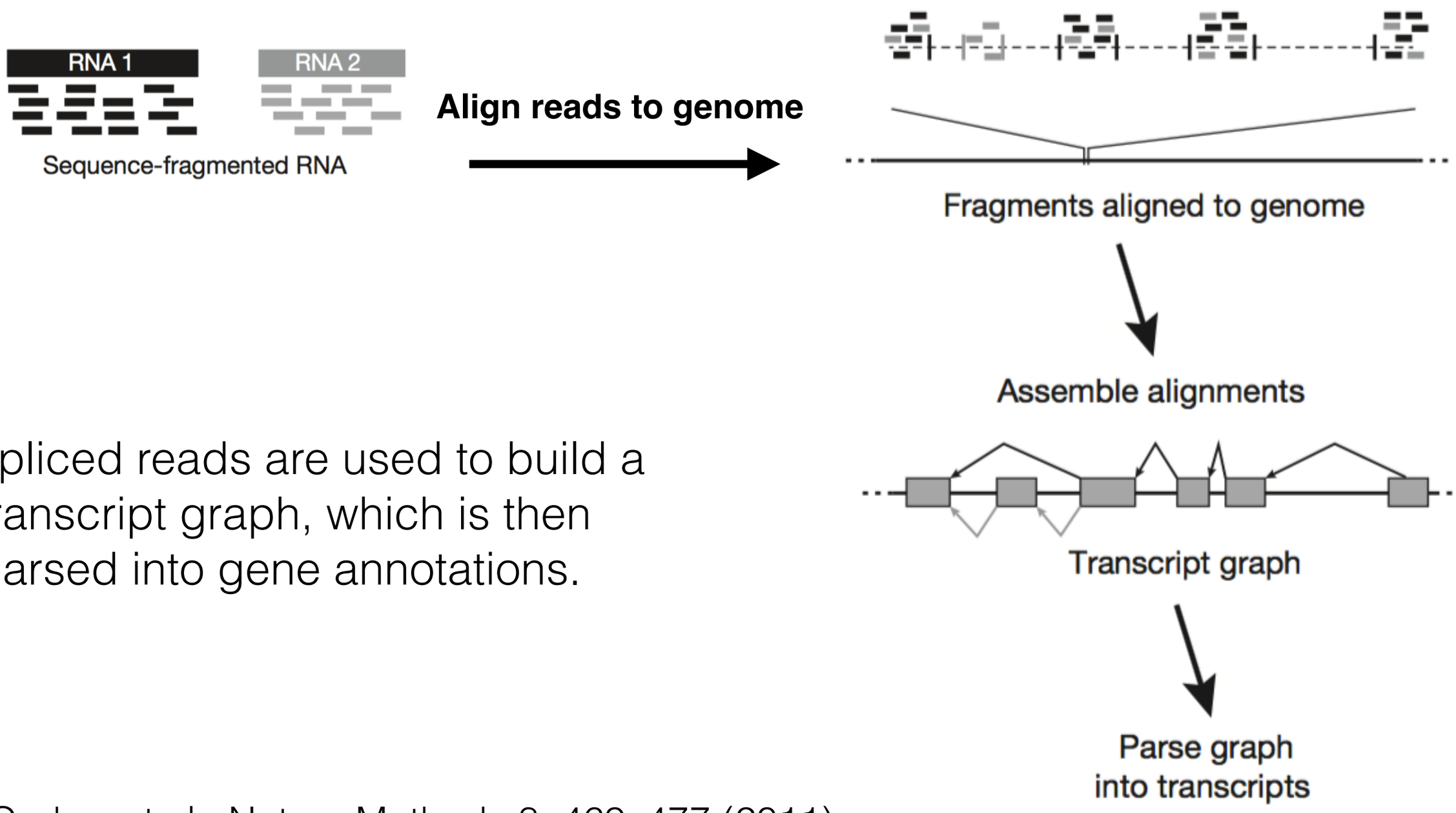
Transcriptome Reconstruction

Genome Guided Approach

Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

Genome Guided Approach

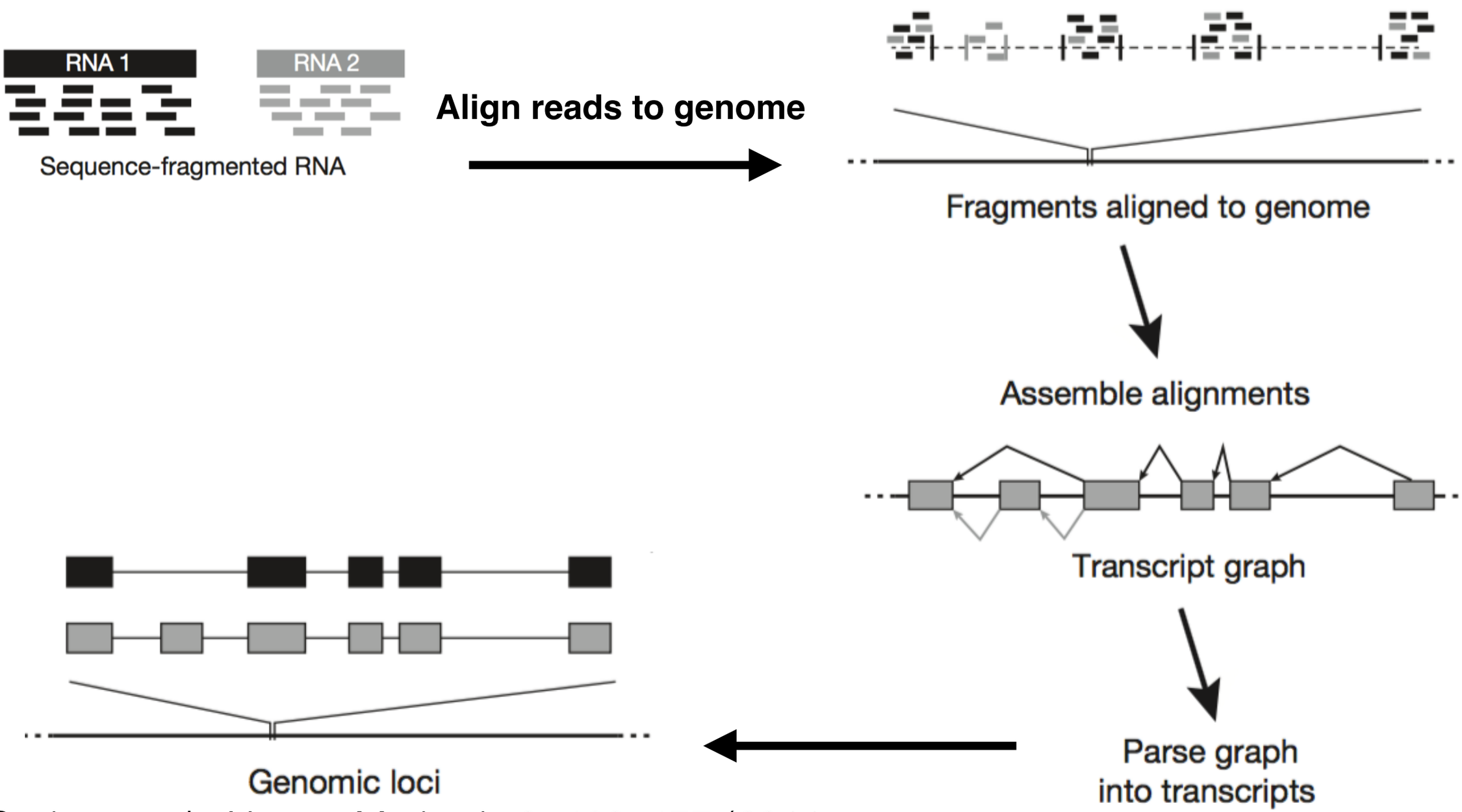


spliced reads are used to build a transcript graph, which is then parsed into gene annotations.

Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

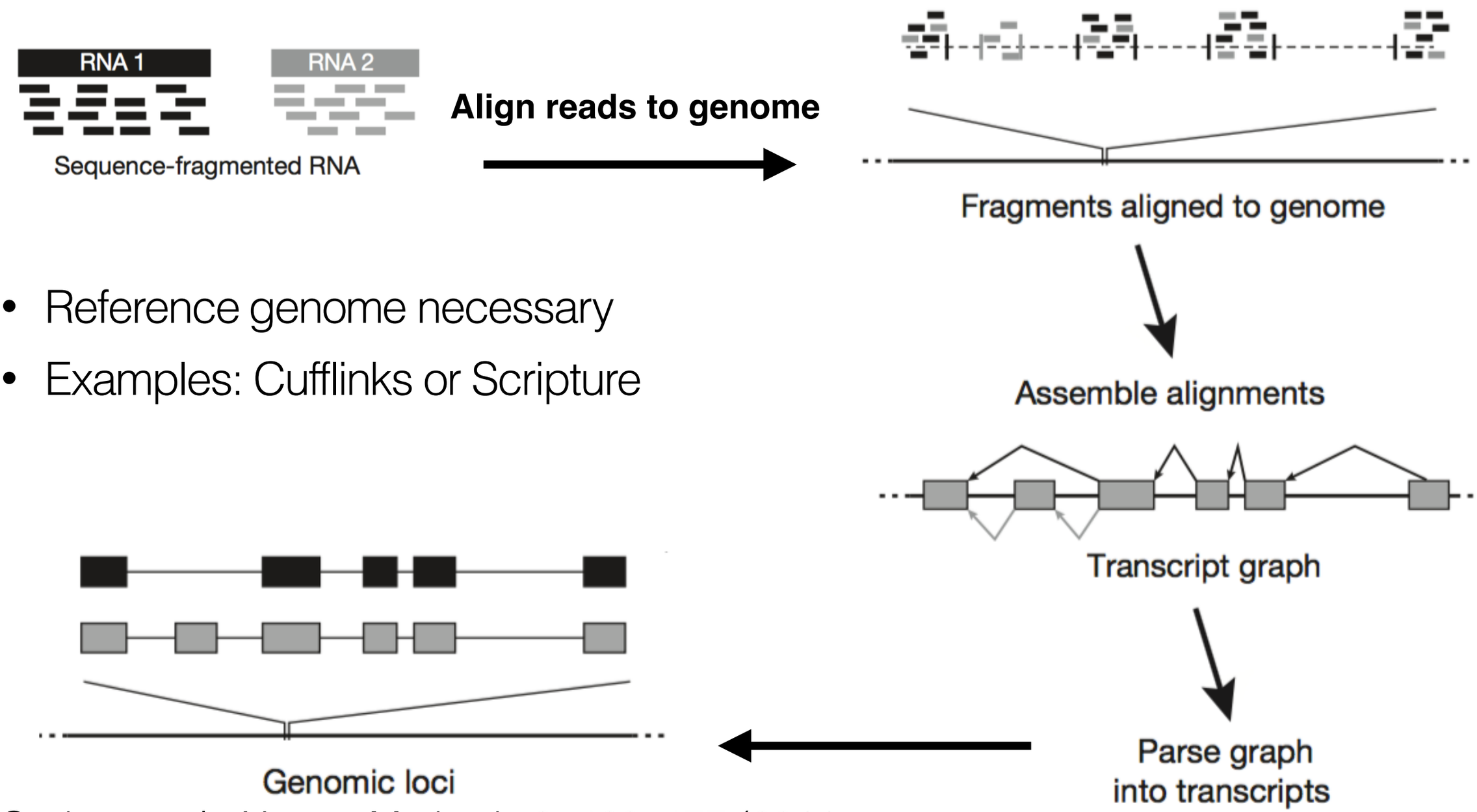
Genome Guided Approach



Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

Genome Guided Approach

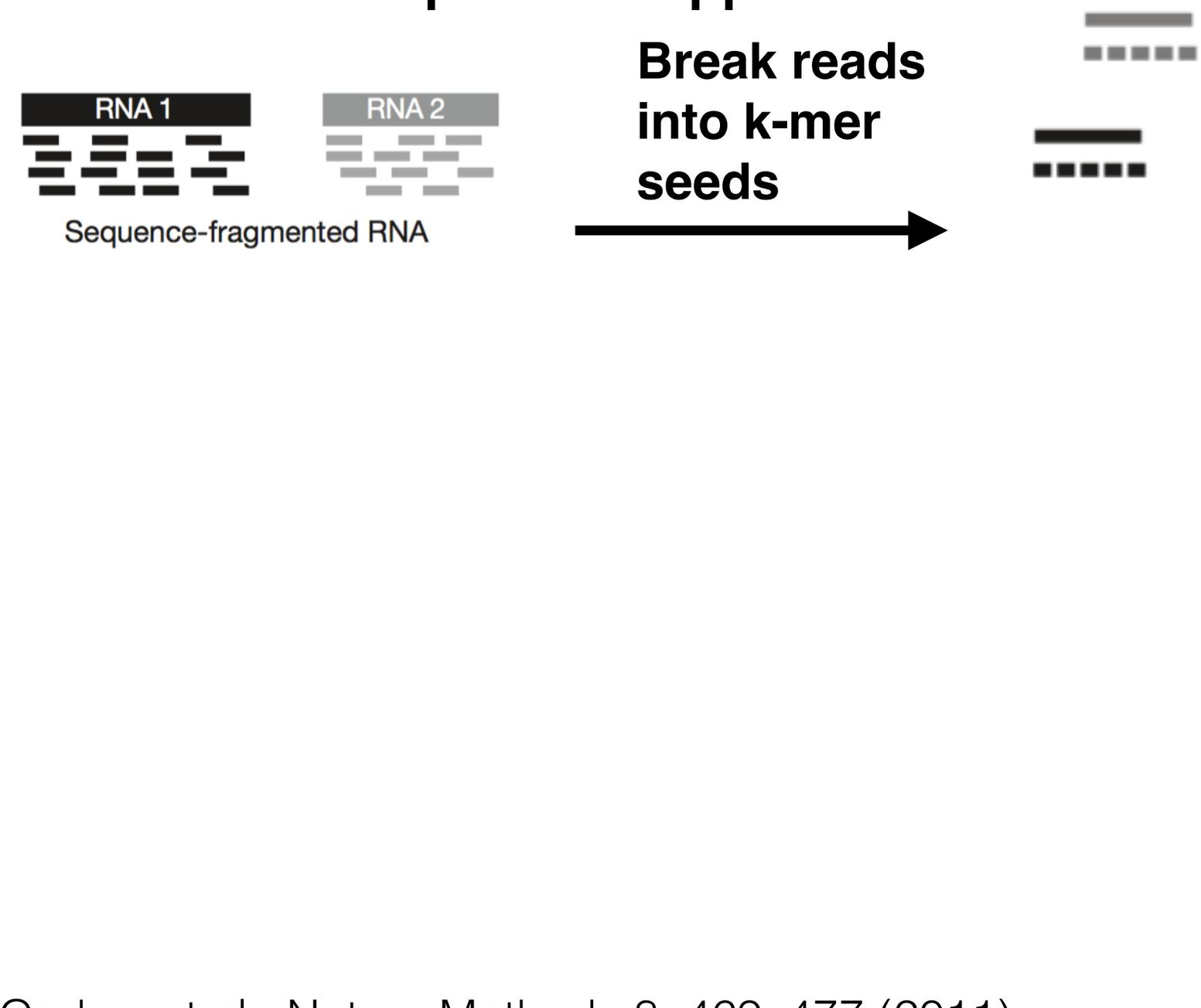


- Reference genome necessary
- Examples: Cufflinks or Scripture

Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

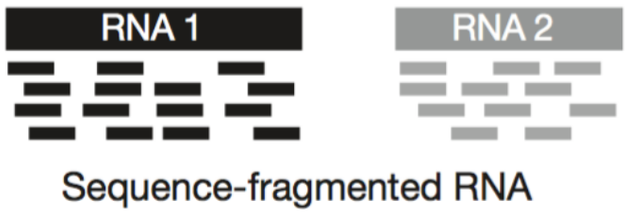
Genome-Independent Approach



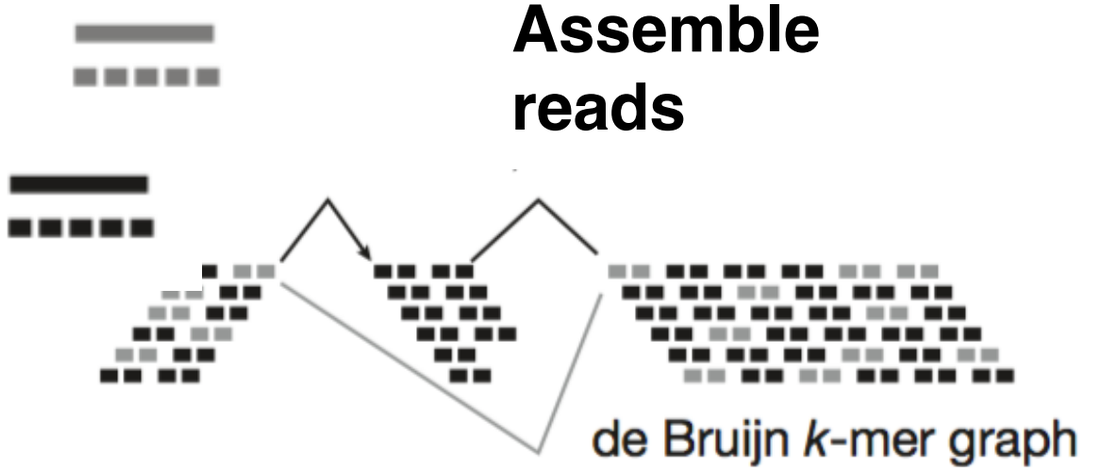
Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

Genome-Independent Approach



Break reads into k-mer seeds

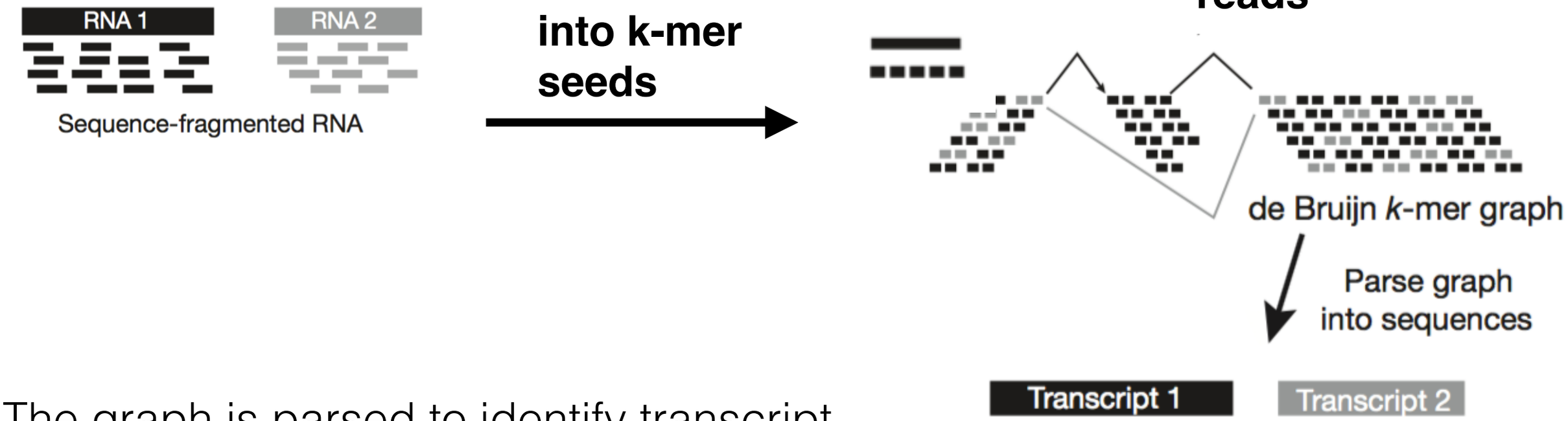


K-mer arranged into a de Bruijn graph structure.

Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

Genome-Independent Approach

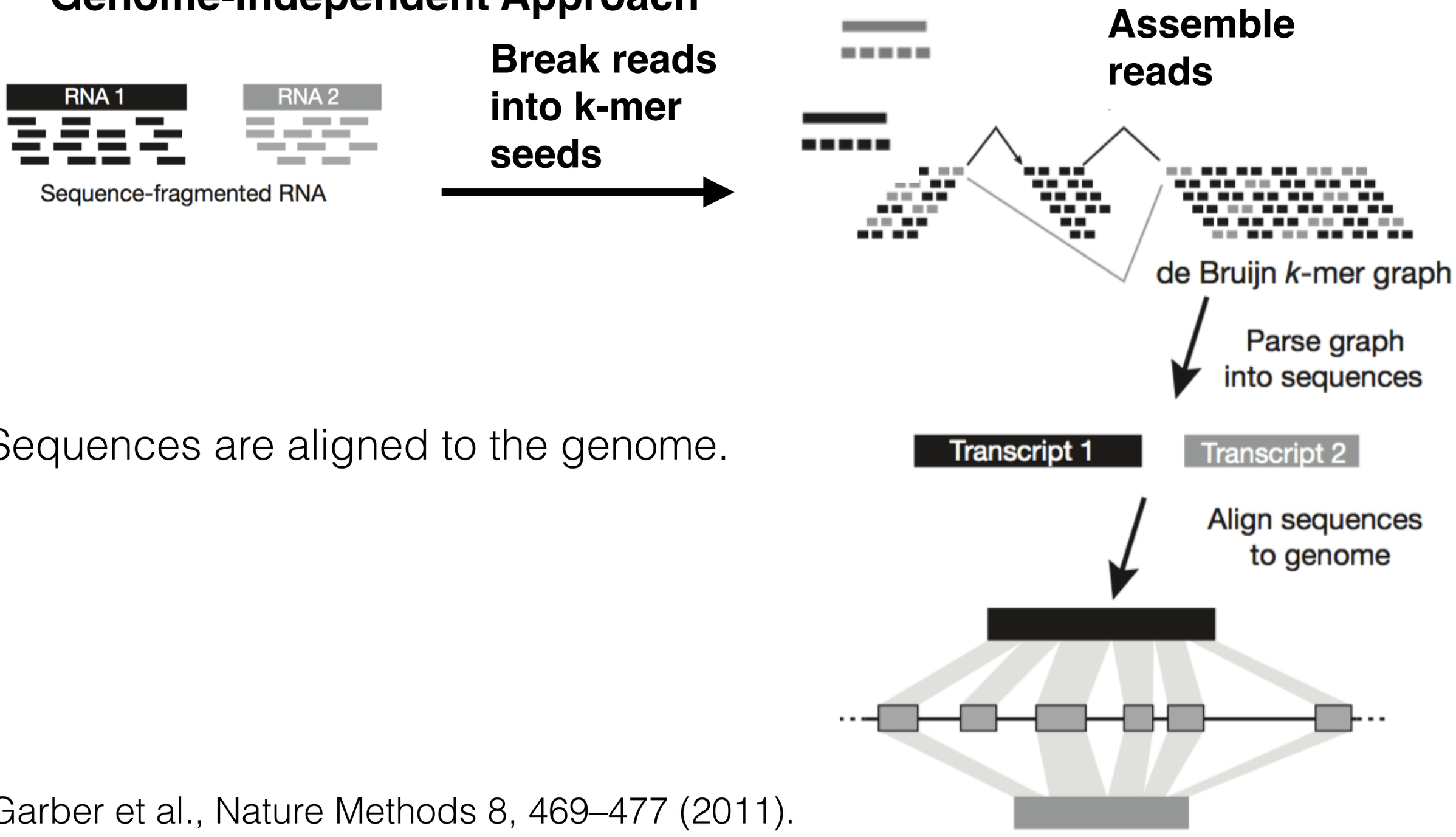


The graph is parsed to identify transcript sequences

Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

Genome-Independent Approach

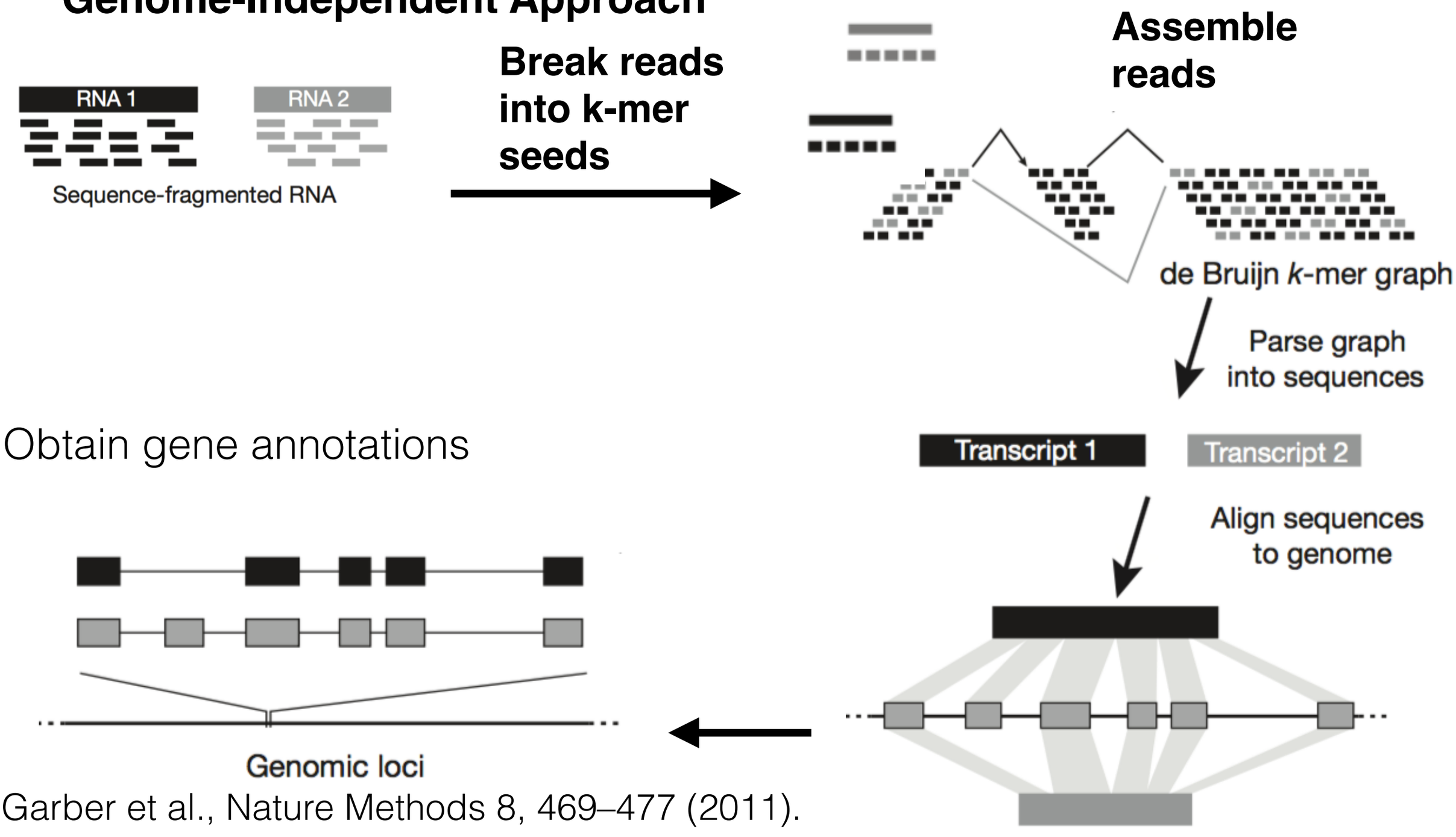


Sequences are aligned to the genome.

Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

Genome-Independent Approach



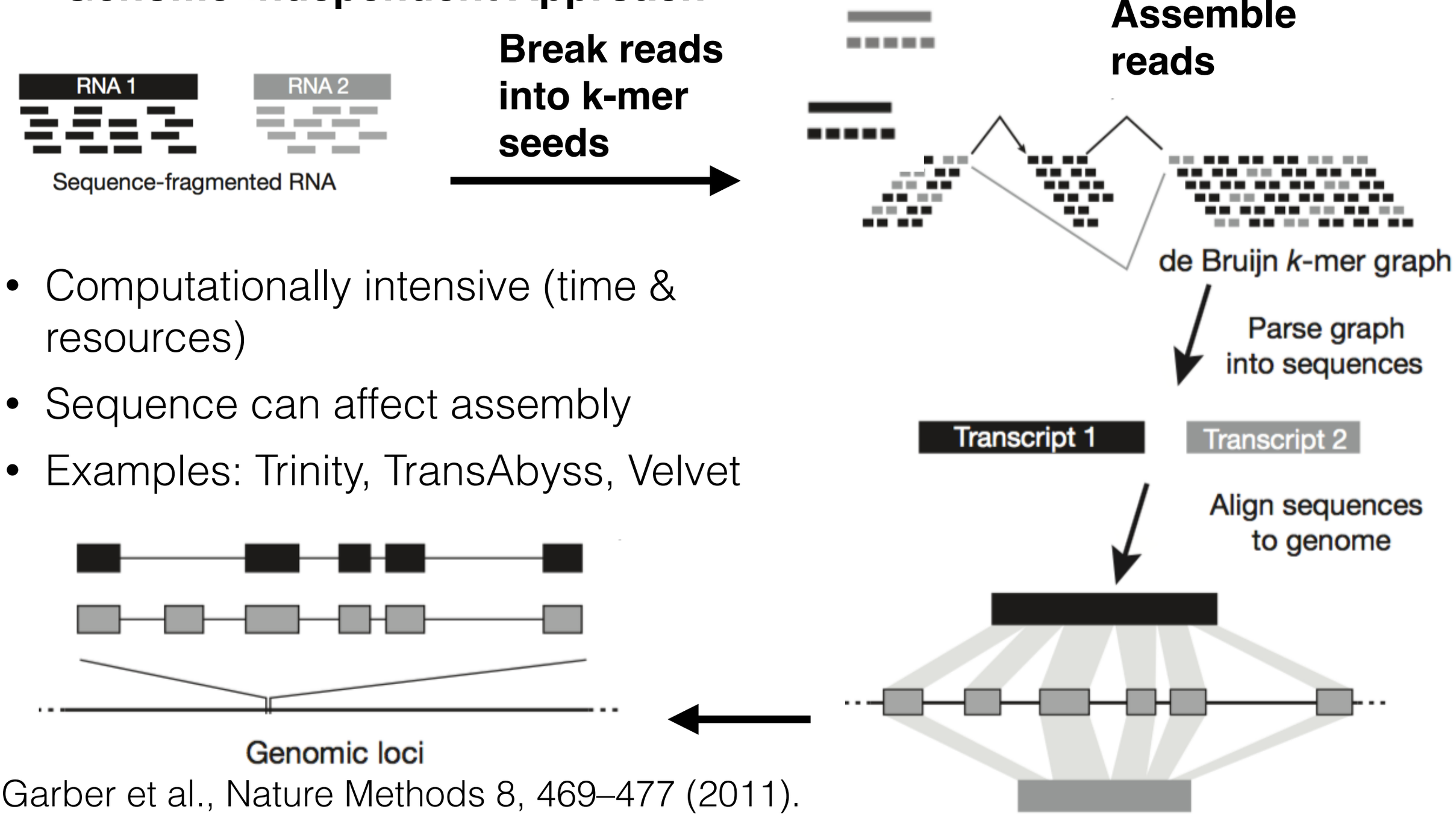
Obtain gene annotations

Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

Genome-Independent Approach

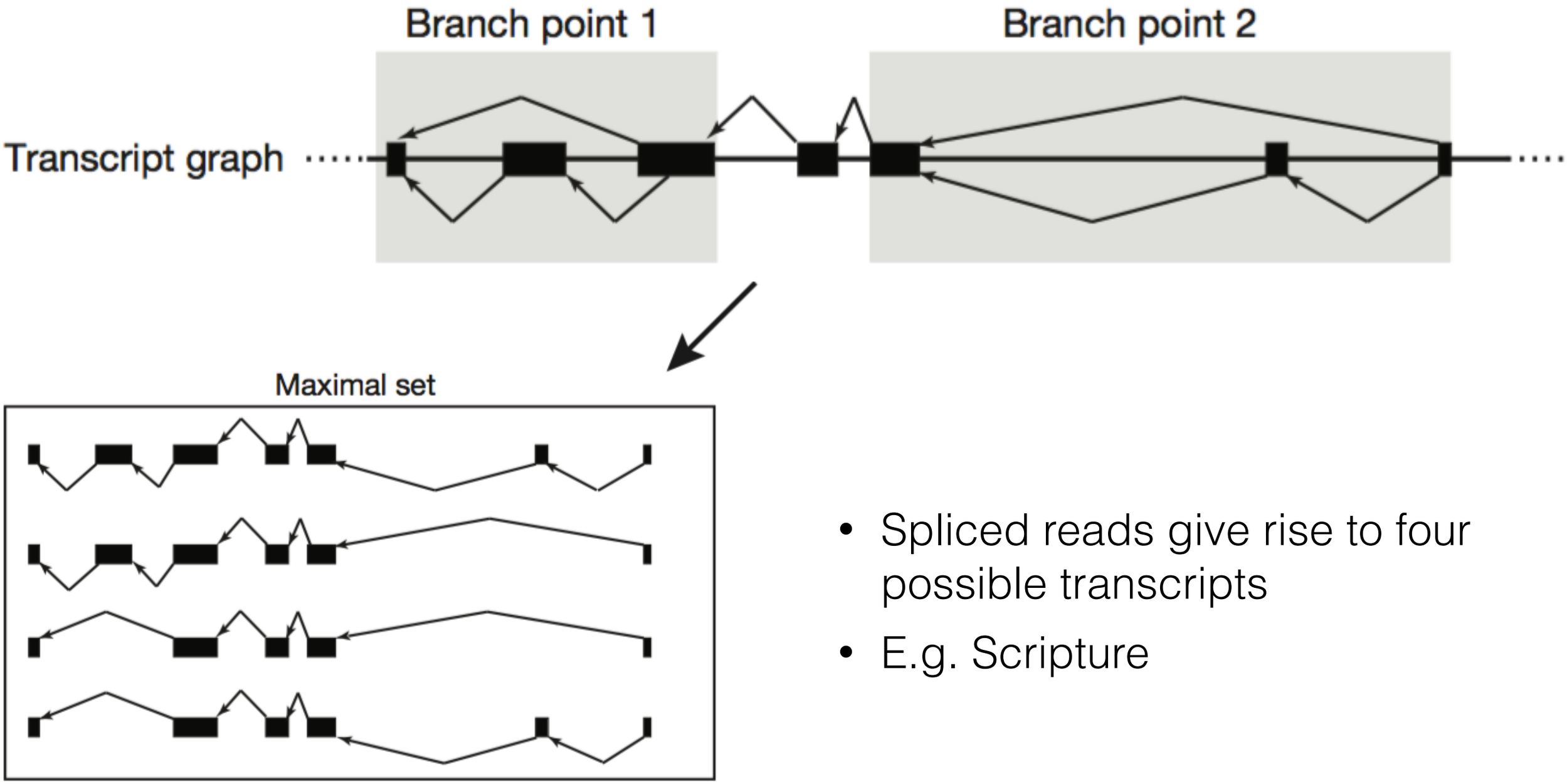
- Computationally intensive (time & resources)
- Sequence can affect assembly
- Examples: Trinity, TransAbyss, Velvet



Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

Branch points and transcript possibilities

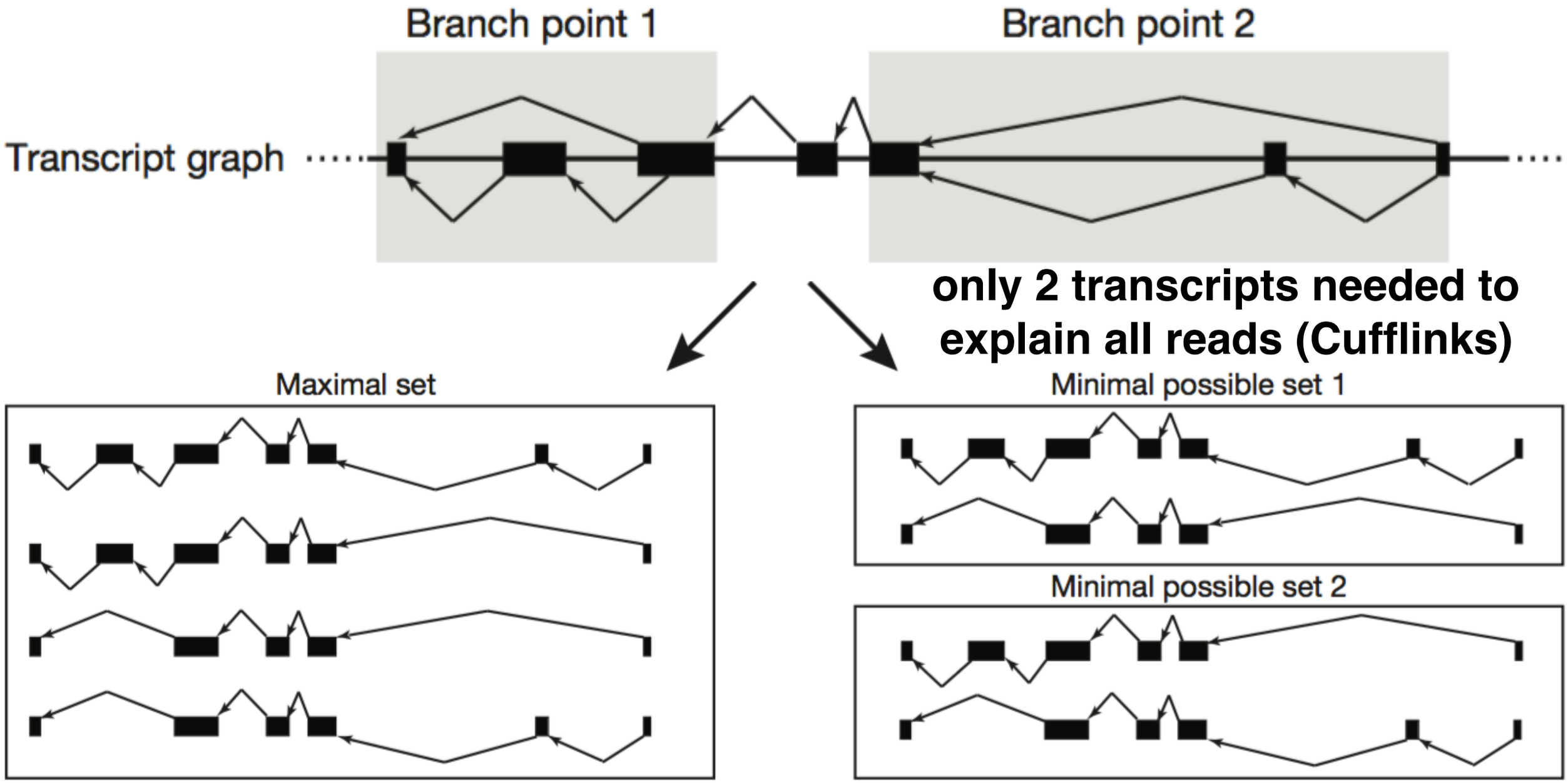


- Spliced reads give rise to four possible transcripts
- E.g. Scripture

Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

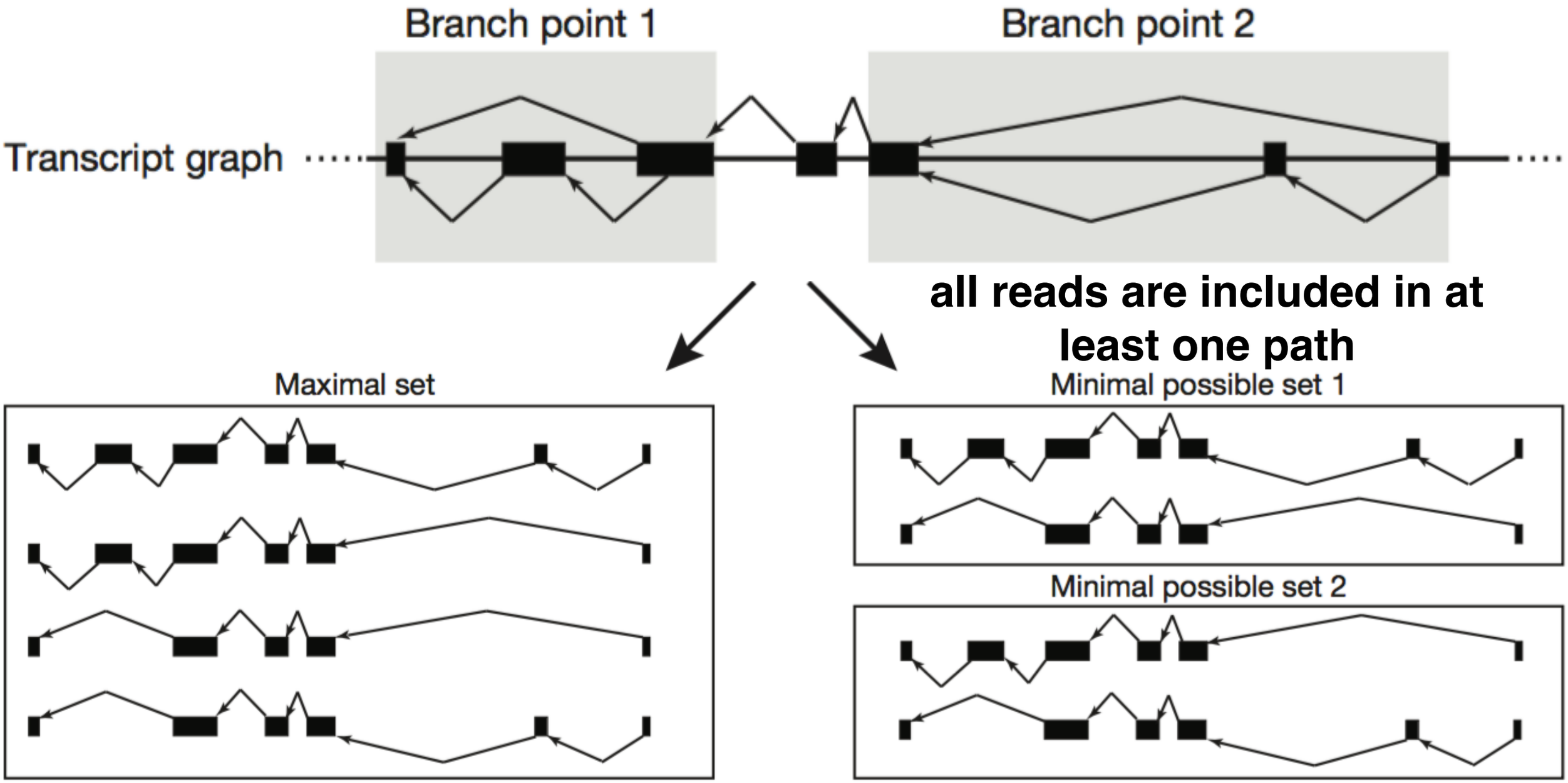
Branch points and transcript possibilities



Garber et al., Nature Methods 8, 469–477 (2011).

Transcriptome Reconstruction

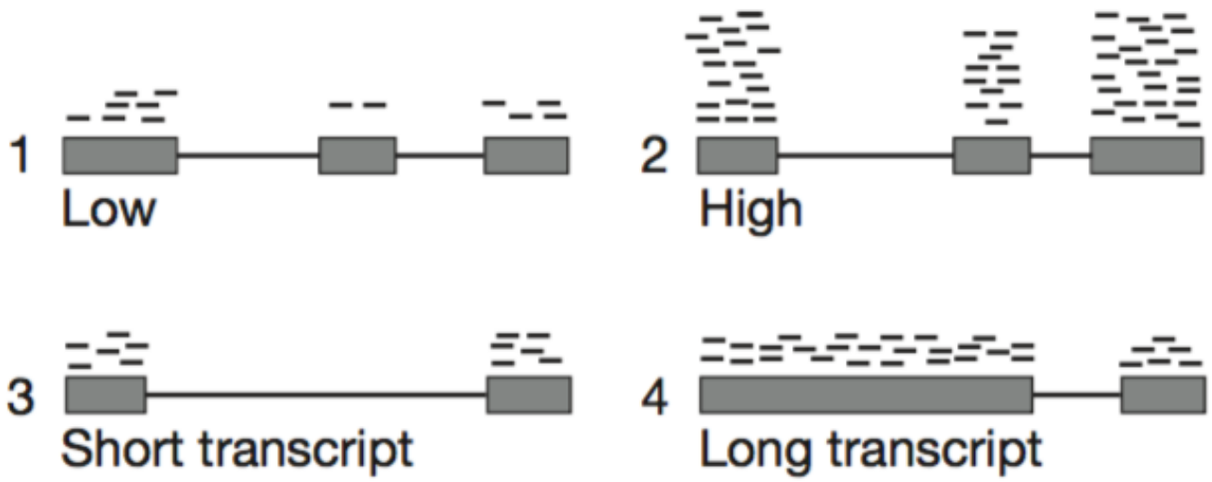
Branch points and transcript possibilities



Garber et al., Nature Methods 8, 469–477 (2011).

Select path based on coverage

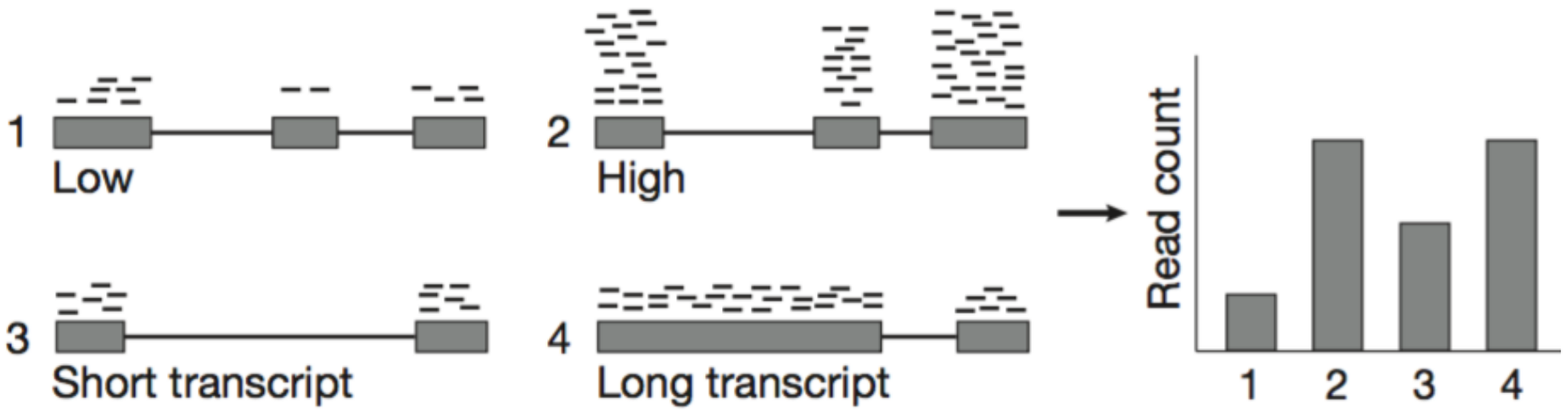
Expression Quantification



- Transcripts of different lengths & different read coverage levels

Garber et al., Nature Methods 8, 469–477 (2011).

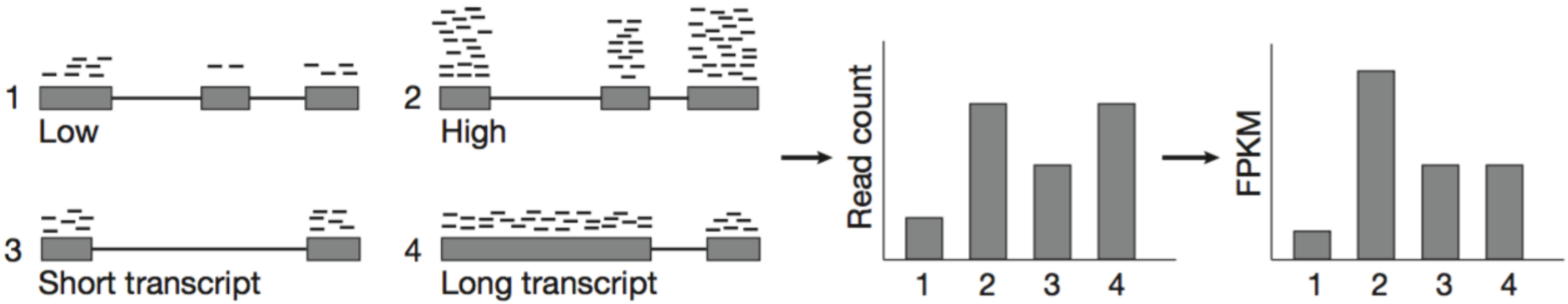
Expression Quantification



- Transcripts of different lengths & different read coverage levels
- Different total read counts observed for each transcript

Garber et al., Nature Methods 8, 469–477 (2011).

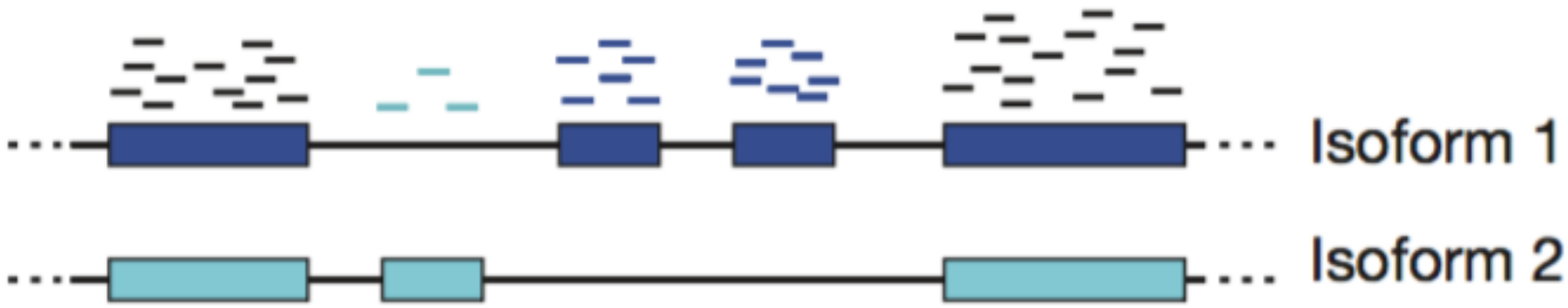
Expression Quantification



- Transcripts of different lengths & different read coverage levels
- Different total read counts observed for each transcript
- FPKM-normalized read counts

Garber et al., Nature Methods 8, 469–477 (2011).

Expression Quantification



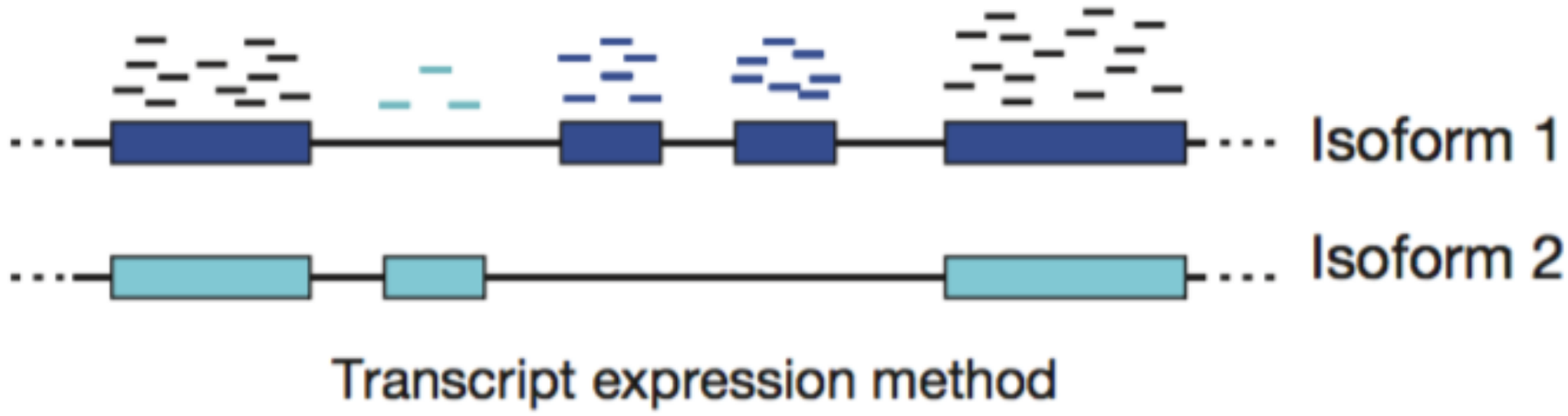
Transcript expression method

Reads from alternatively spliced genes may be attributable to a single isoform or more than one isoform.

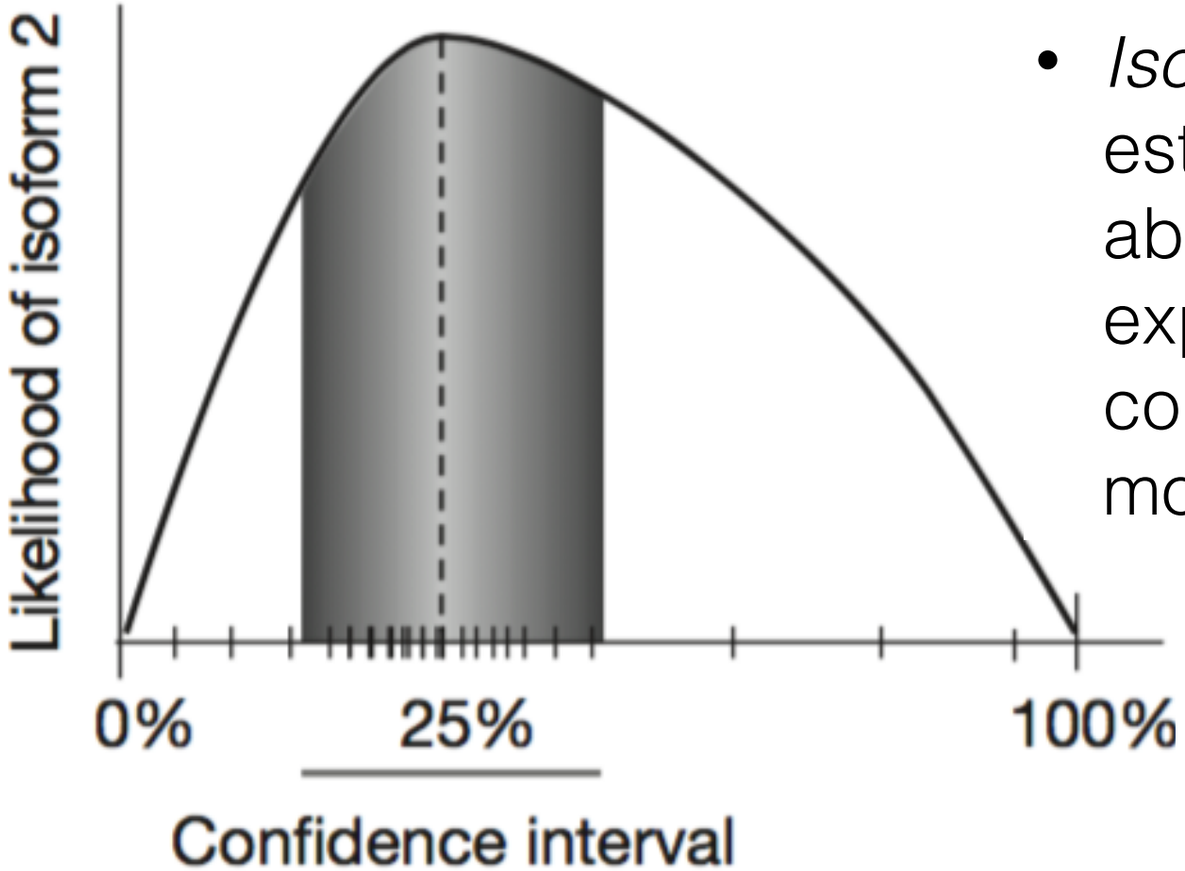
- Reads are color-coded when their isoform of origin is clear.
- Black reads indicate reads with uncertain origin.

Garber et al., Nature Methods 8, 469–477 (2011).

Expression Quantification



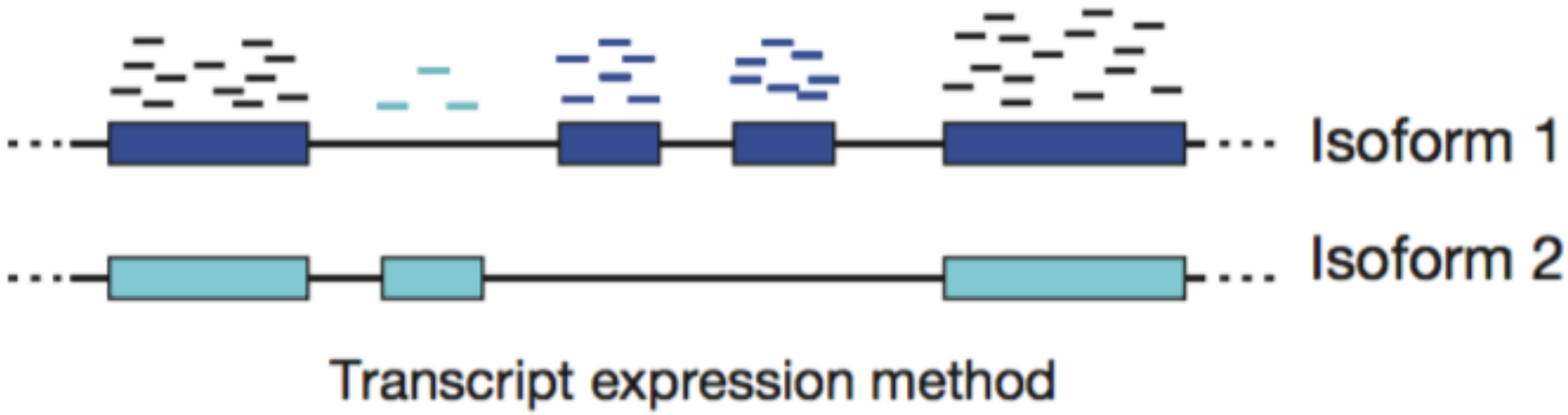
Transcript expression method



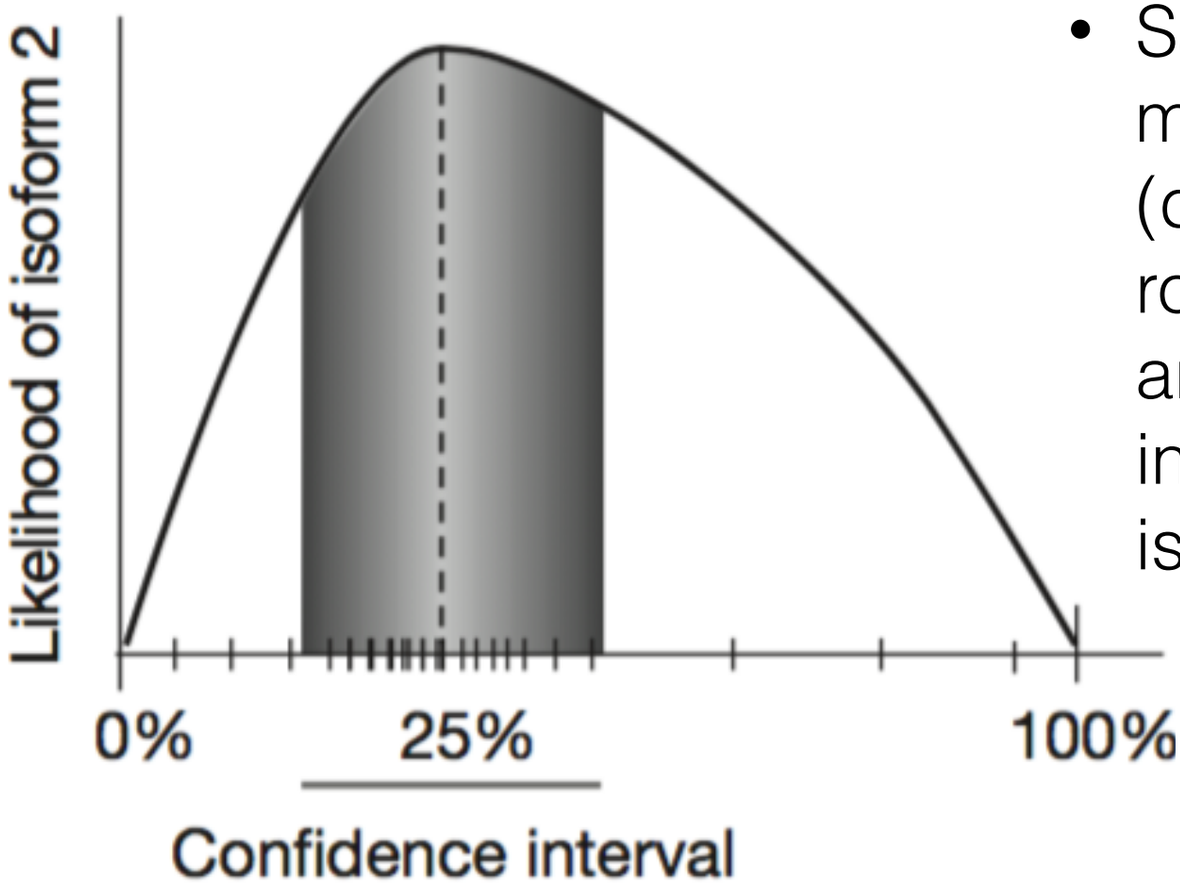
- *Isoform expression methods* estimate isoform abundances that best explain the observed read counts under a generative model.

Garber et al., Nature Methods 8, 469–477 (2011).

Expression Quantification



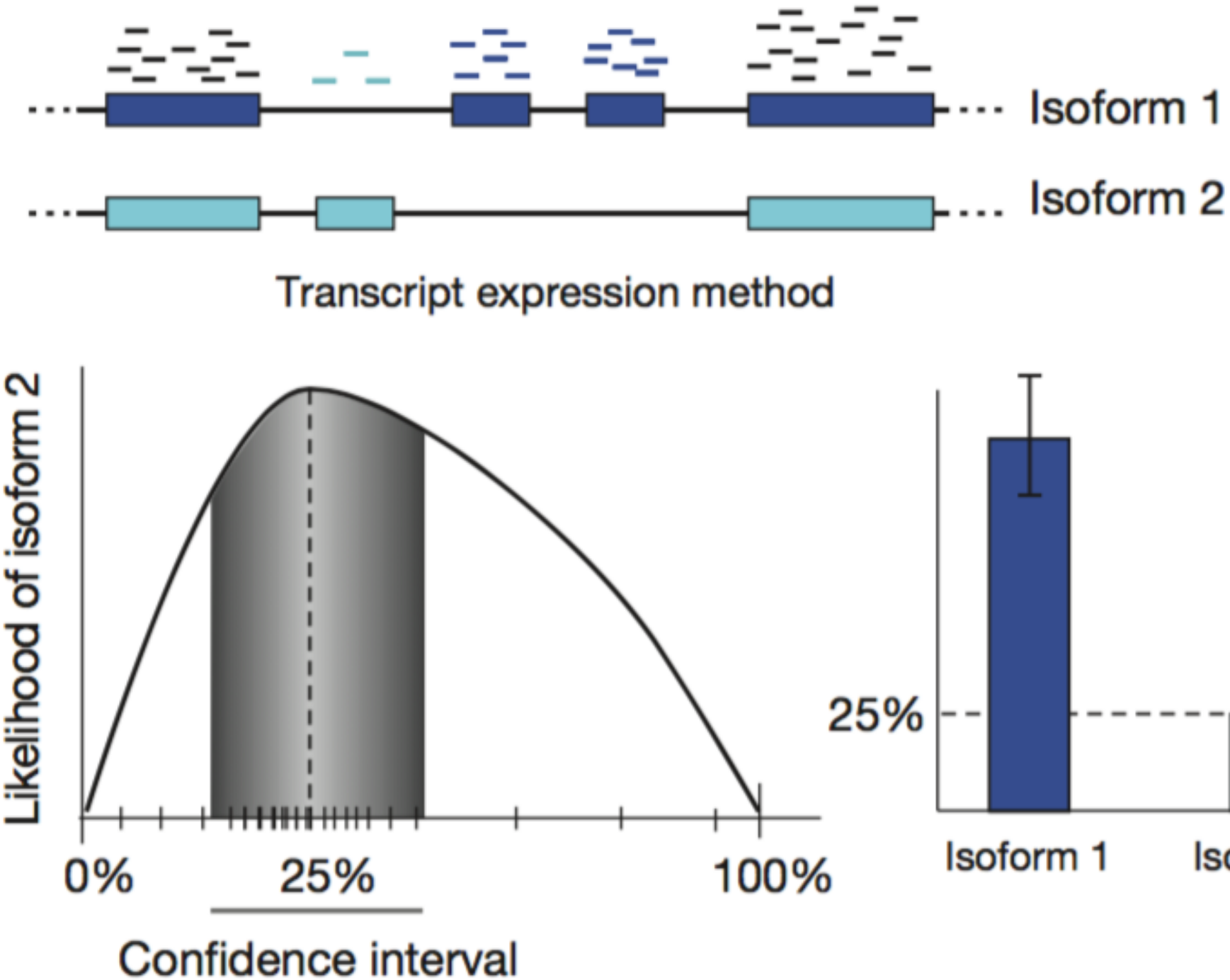
Transcript expression method



- 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.

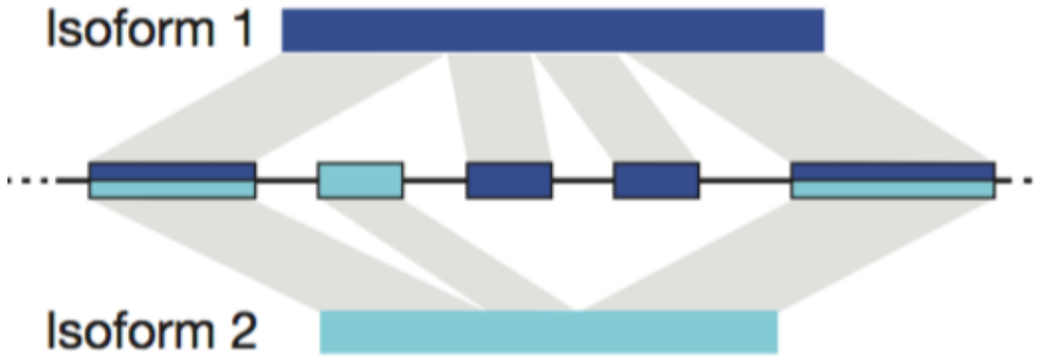
Garber et al., Nature Methods 8, 469–477 (2011).

Expression Quantification



Garber et al., Nature Methods 8, 469–477 (2011).

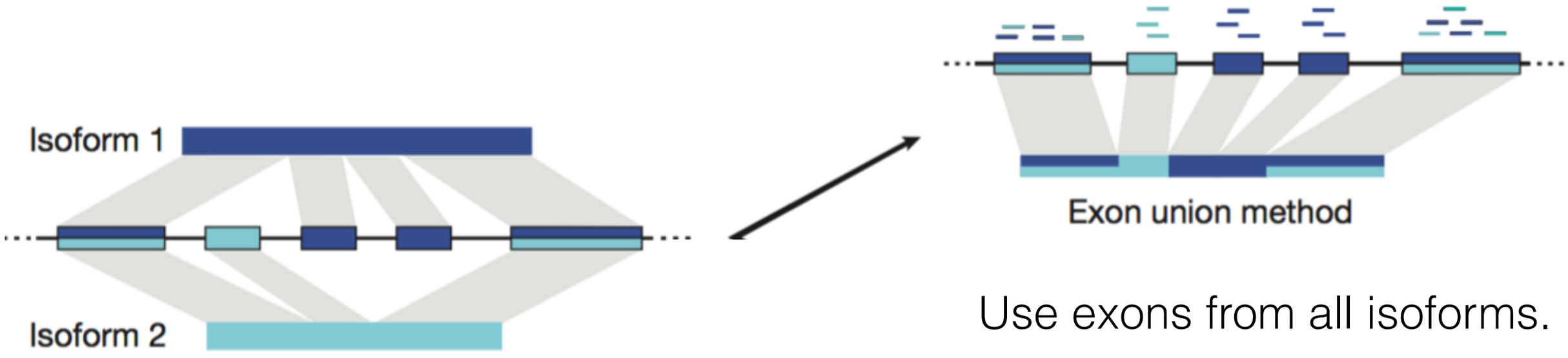
Expression Quantification



- E.g. gene with 2 expressed isoforms
- Exons colored according to the isoform of origin to all gene isoforms.

Garber et al., Nature Methods 8, 469–477 (2011).

Expression Quantification



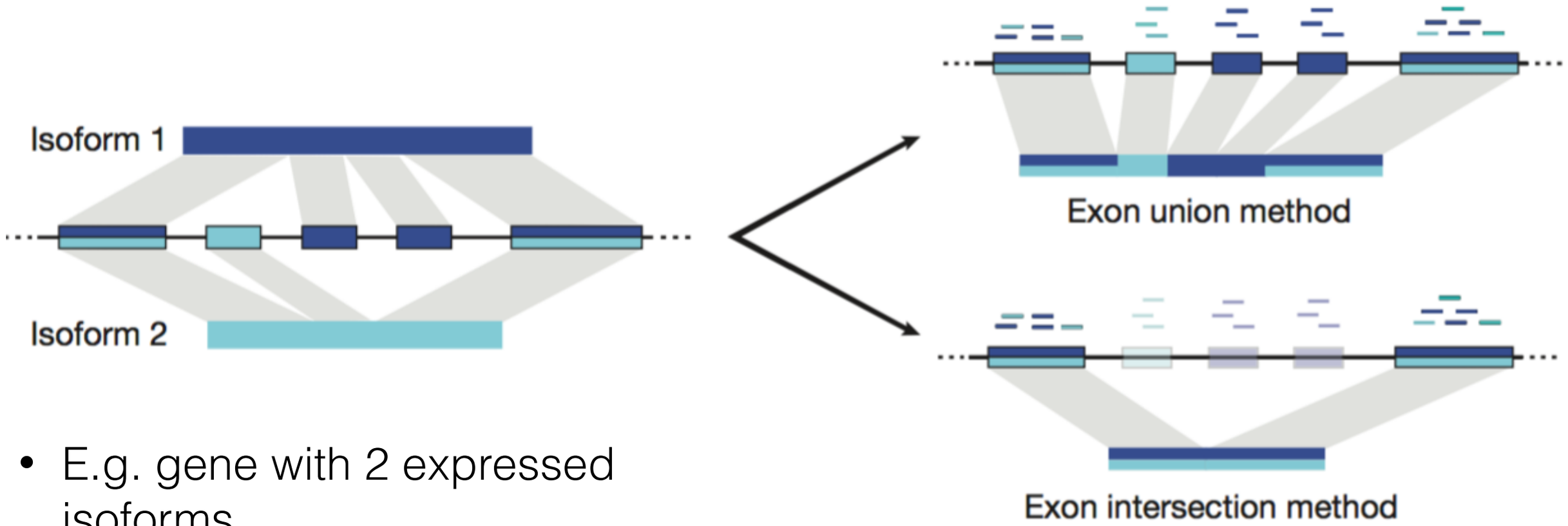
Use exons from all isoforms.

- E.g. gene with 2 expressed isoforms
- Exons colored according to the isoform of origin to all gene isoforms.

Gene models used for quantification

Garber et al., Nature Methods 8, 469–477 (2011).

Expression Quantification



- E.g. gene with 2 expressed isoforms
- Exons colored according to the isoform of origin to all gene isoforms.

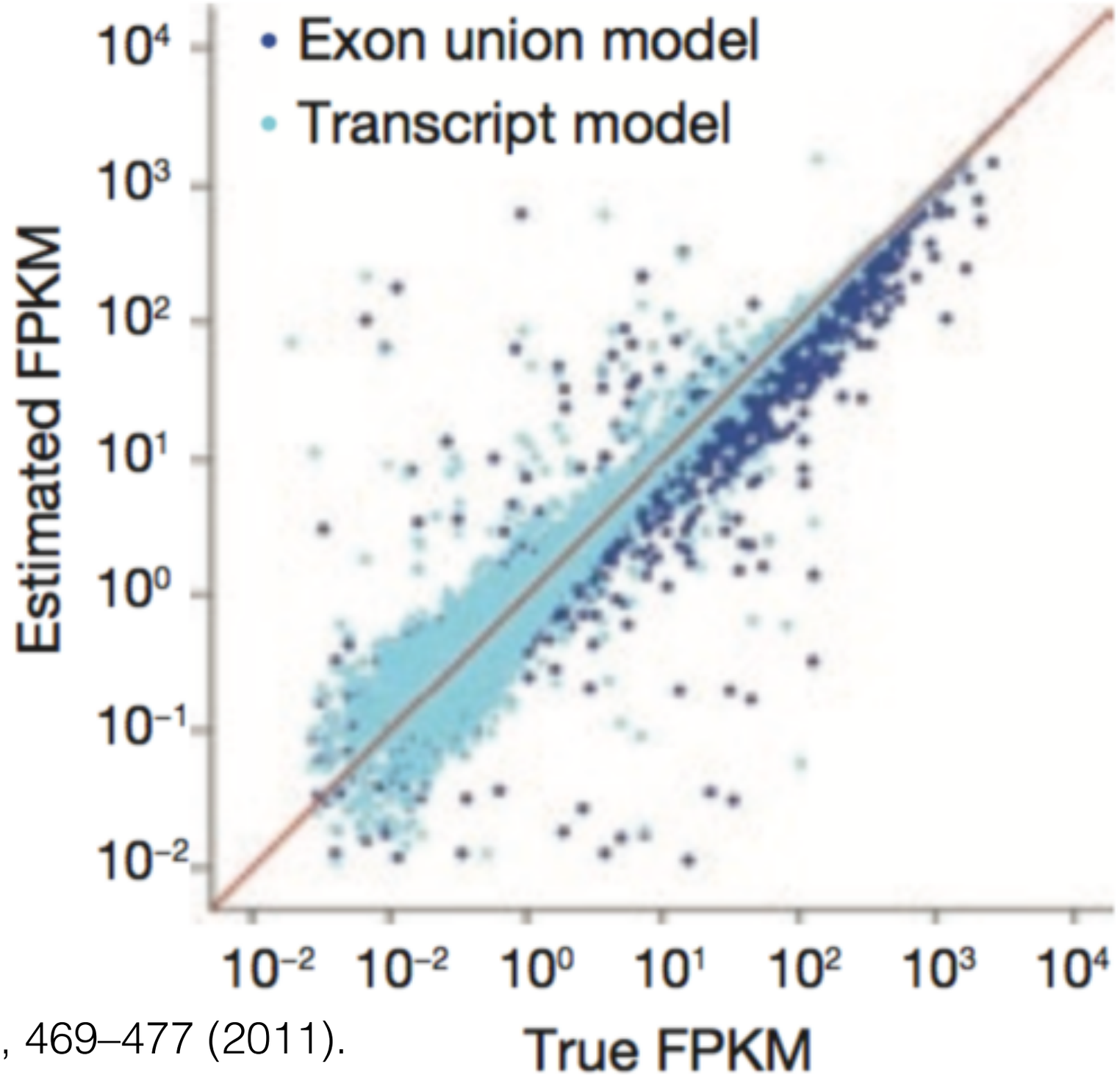
Uses only exons common to all gene isoforms

Gene models used for quantification

Garber et al., Nature Methods 8, 469–477 (2011).

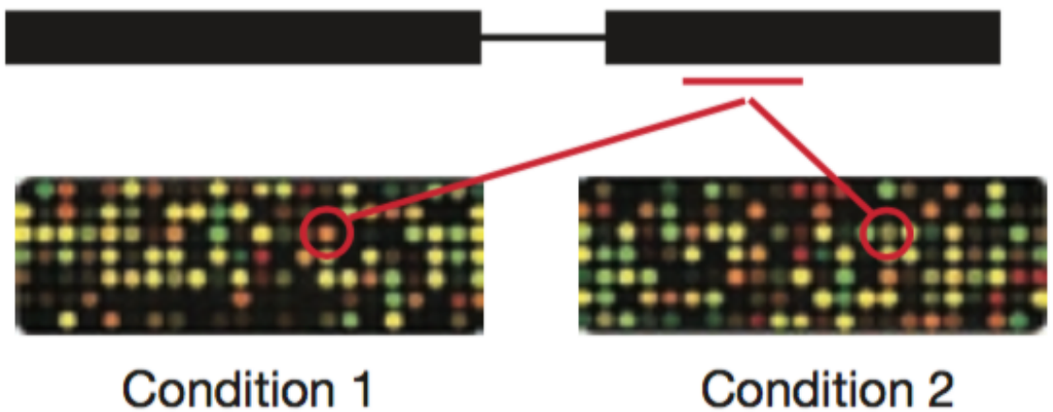
Expression Quantification

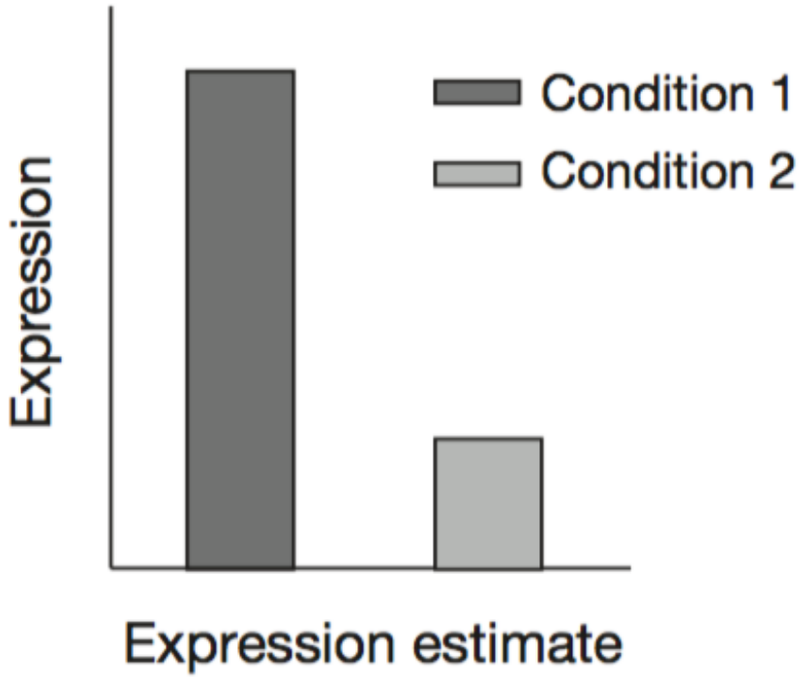
Comparison of true versus estimated FPKM values in *simulated* RNA-seq data.



Garber et al., Nature Methods 8, 469–477 (2011).

Differential expression analysis



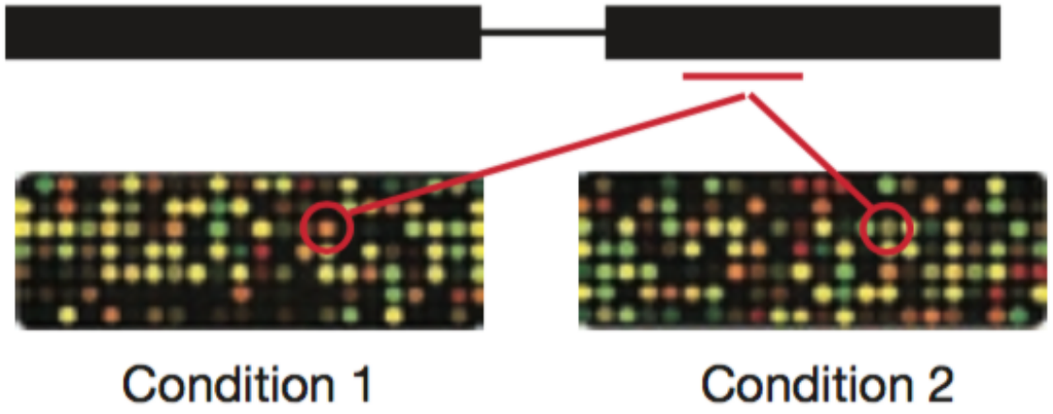


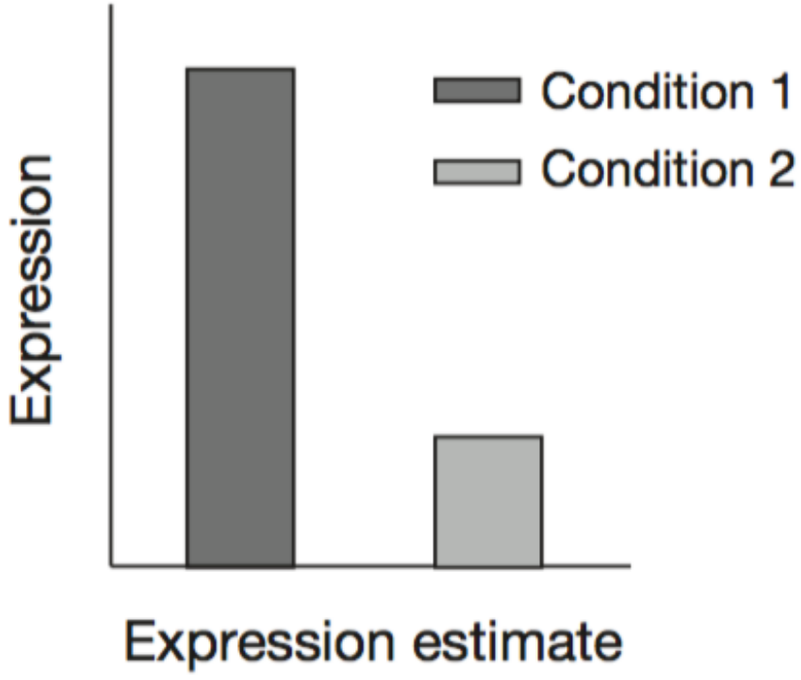
Expression Microarrays

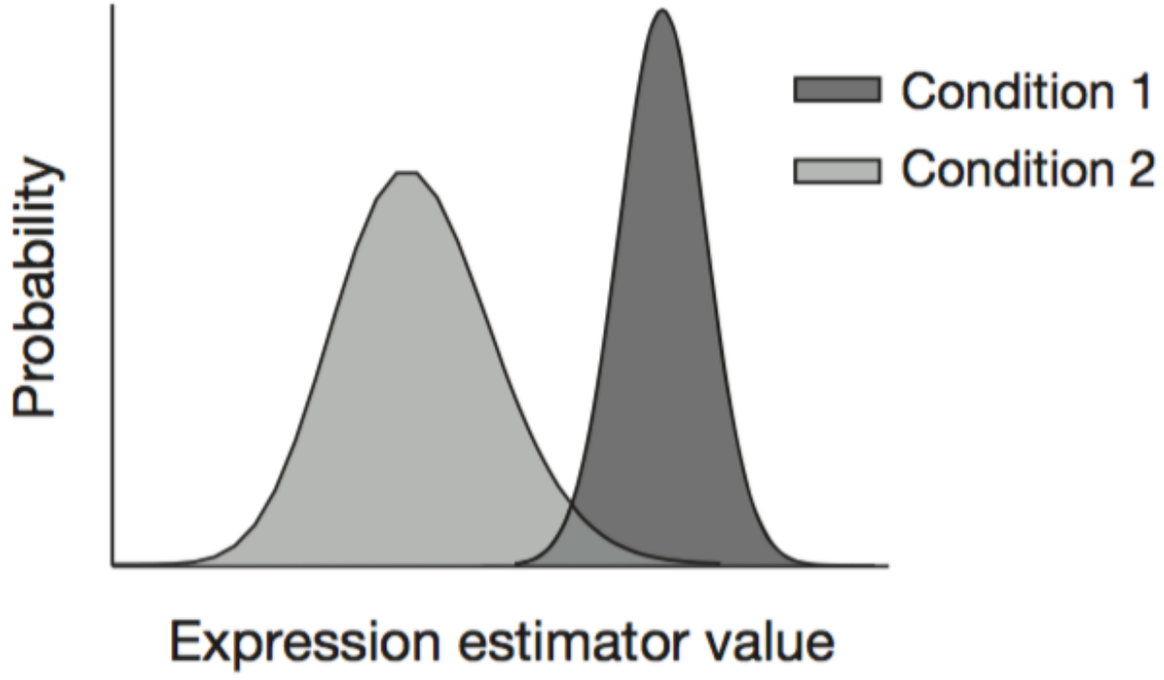
Fluorescence based comparison of intensity

Garber et al., Nature Methods 8, 469–477 (2011).

Differential expression analysis



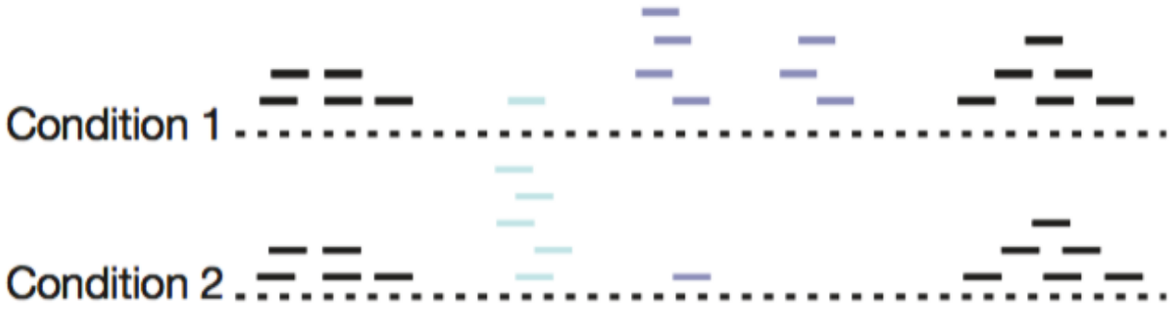




Rna-Seq: Fraction of aligned reads to a given gene

Garber et al., Nature Methods 8, 469–477 (2011).

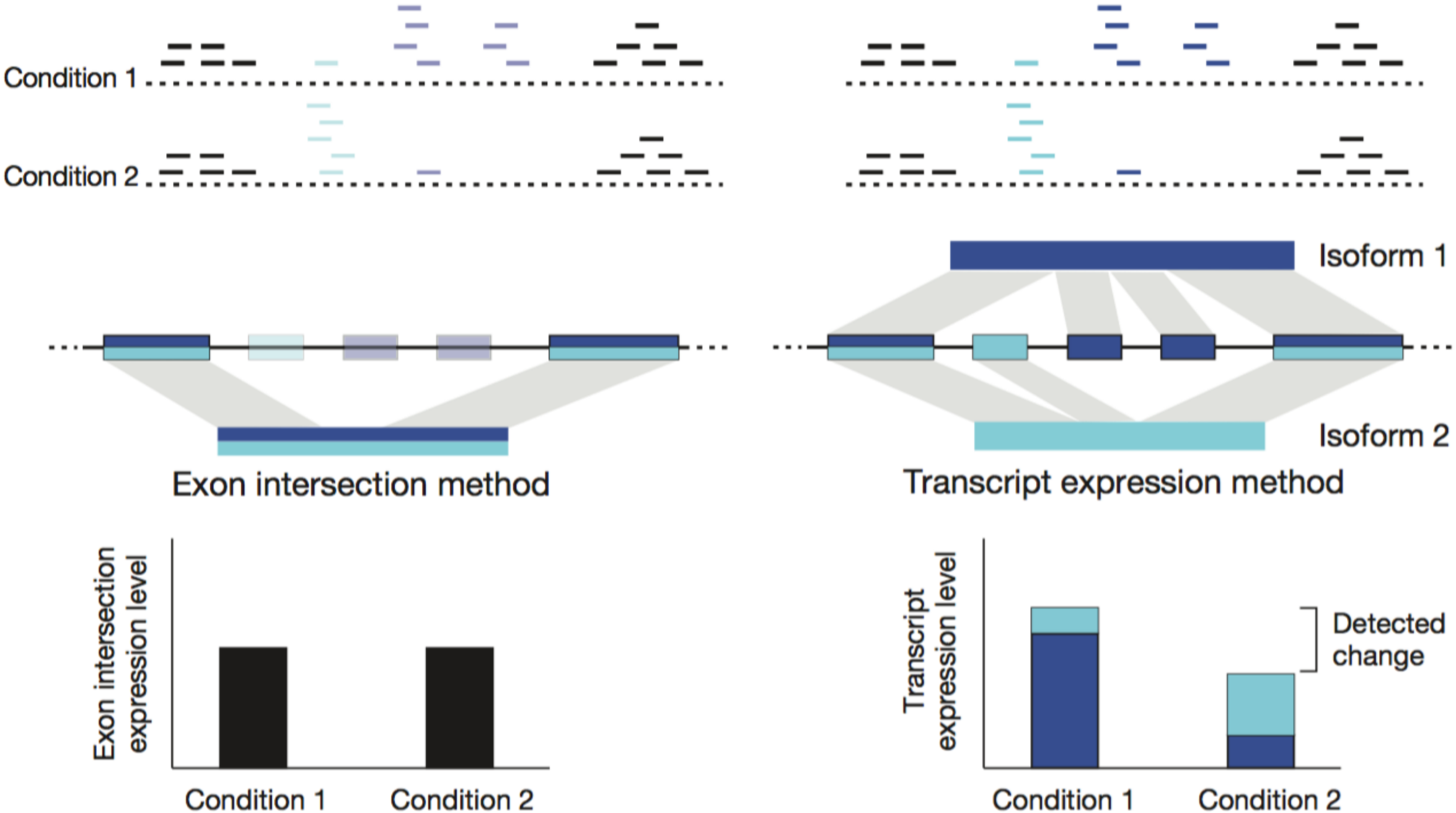
Differential expression analysis



- Example Gene with 2 isoforms with isoform switch in 2 conditions.
- Similar no. of reads
- Different distribution across isoforms

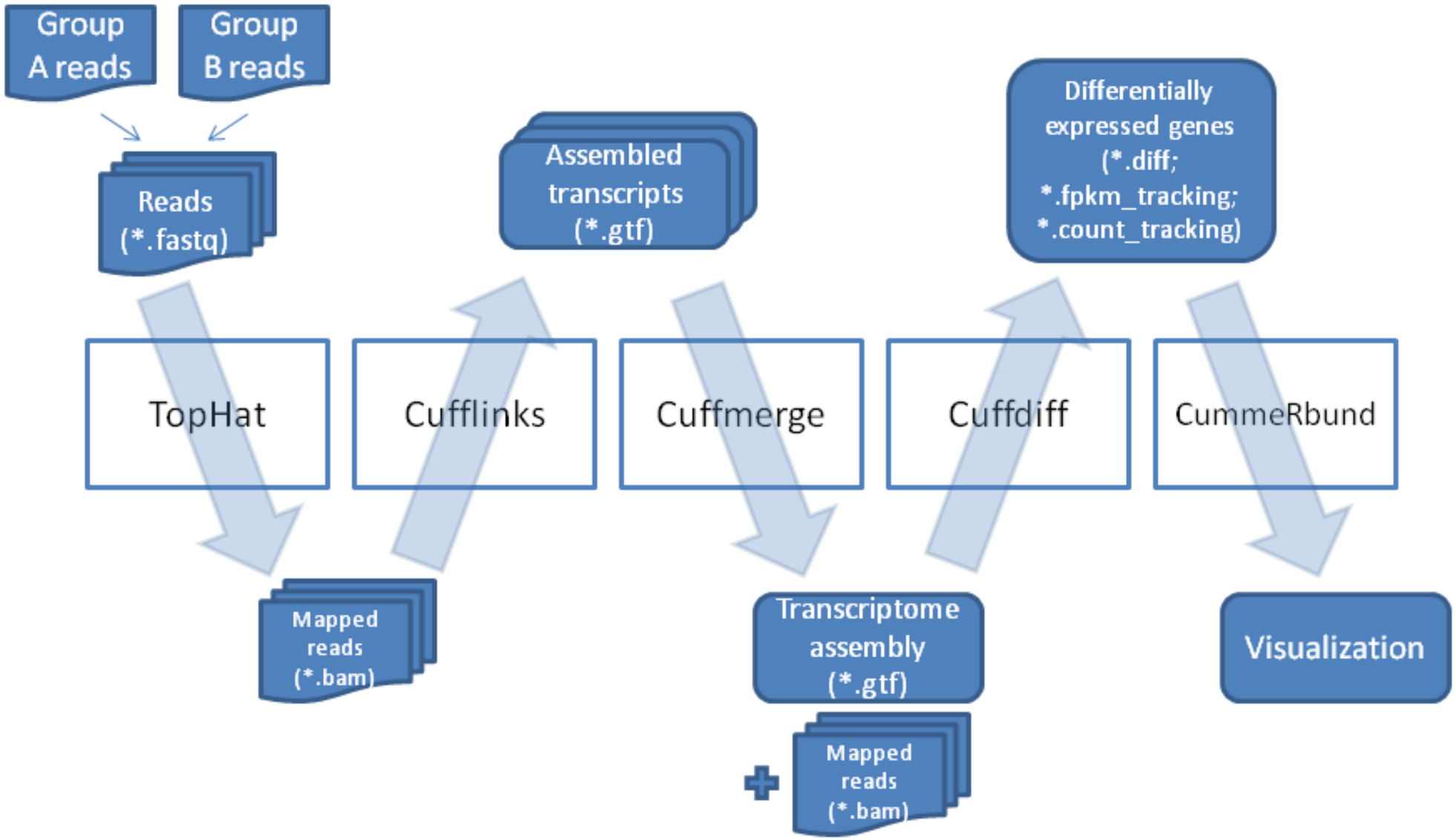
Garber et al., Nature Methods 8, 469–477 (2011).

Differential expression analysis



Garber et al., Nature Methods 8, 469–477 (2011).

Tuxedo Suite

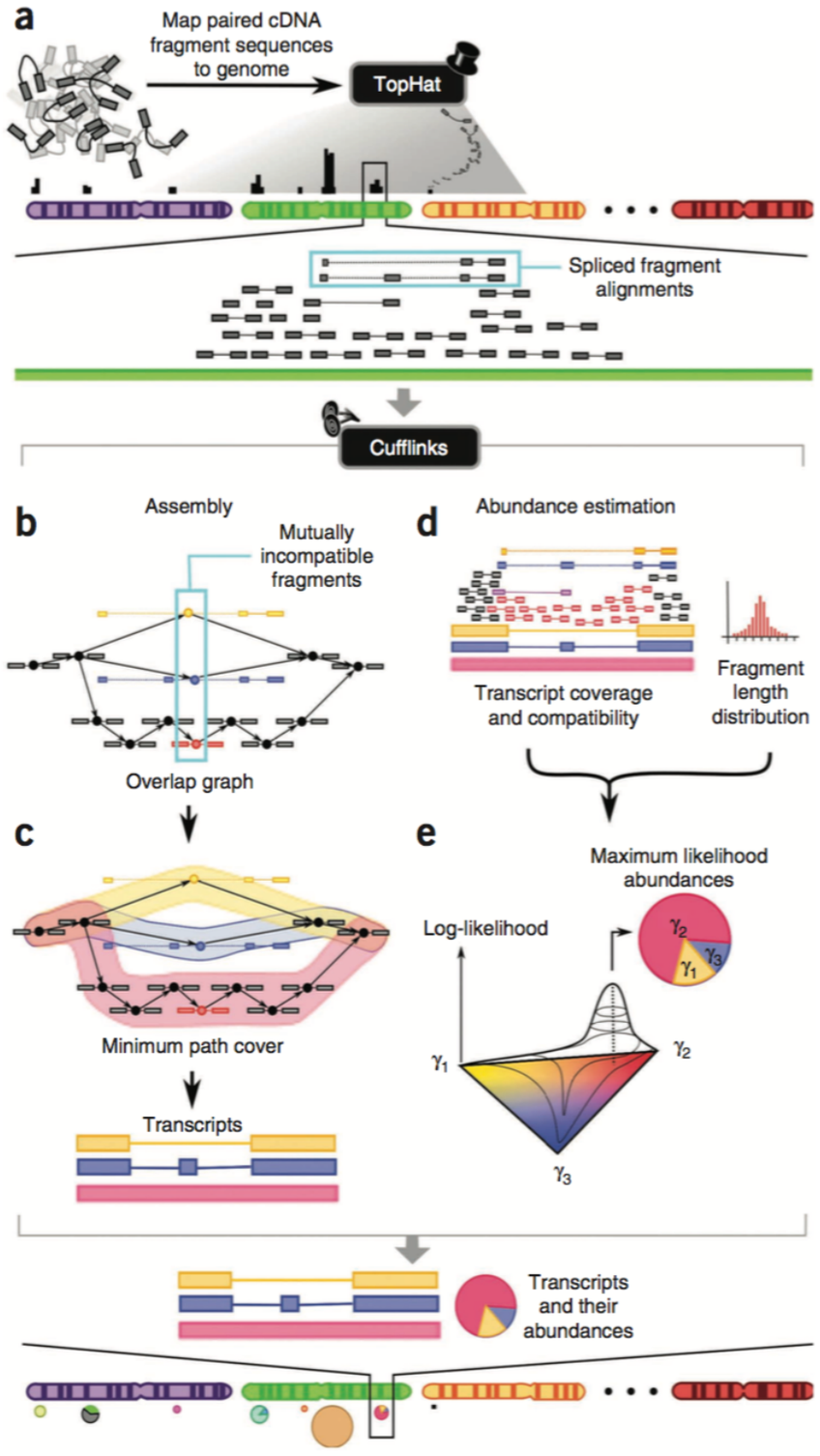


http://www.labome.com/method/RNA-seq-Using-Next-Generation-Sequencing.html
Corney (2013) Mater Methods 3:203

Tuxedo Suite

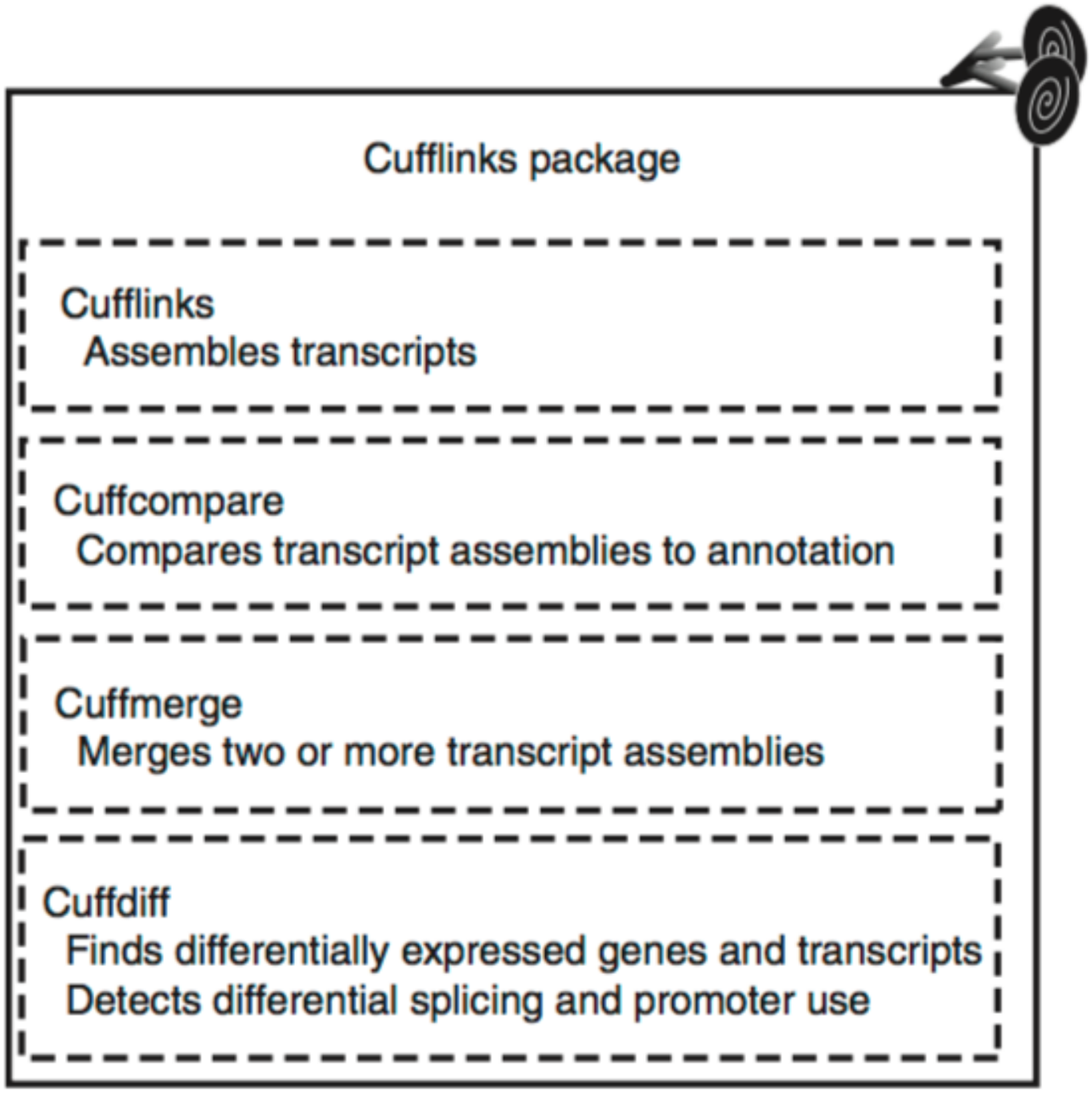
Bowtie

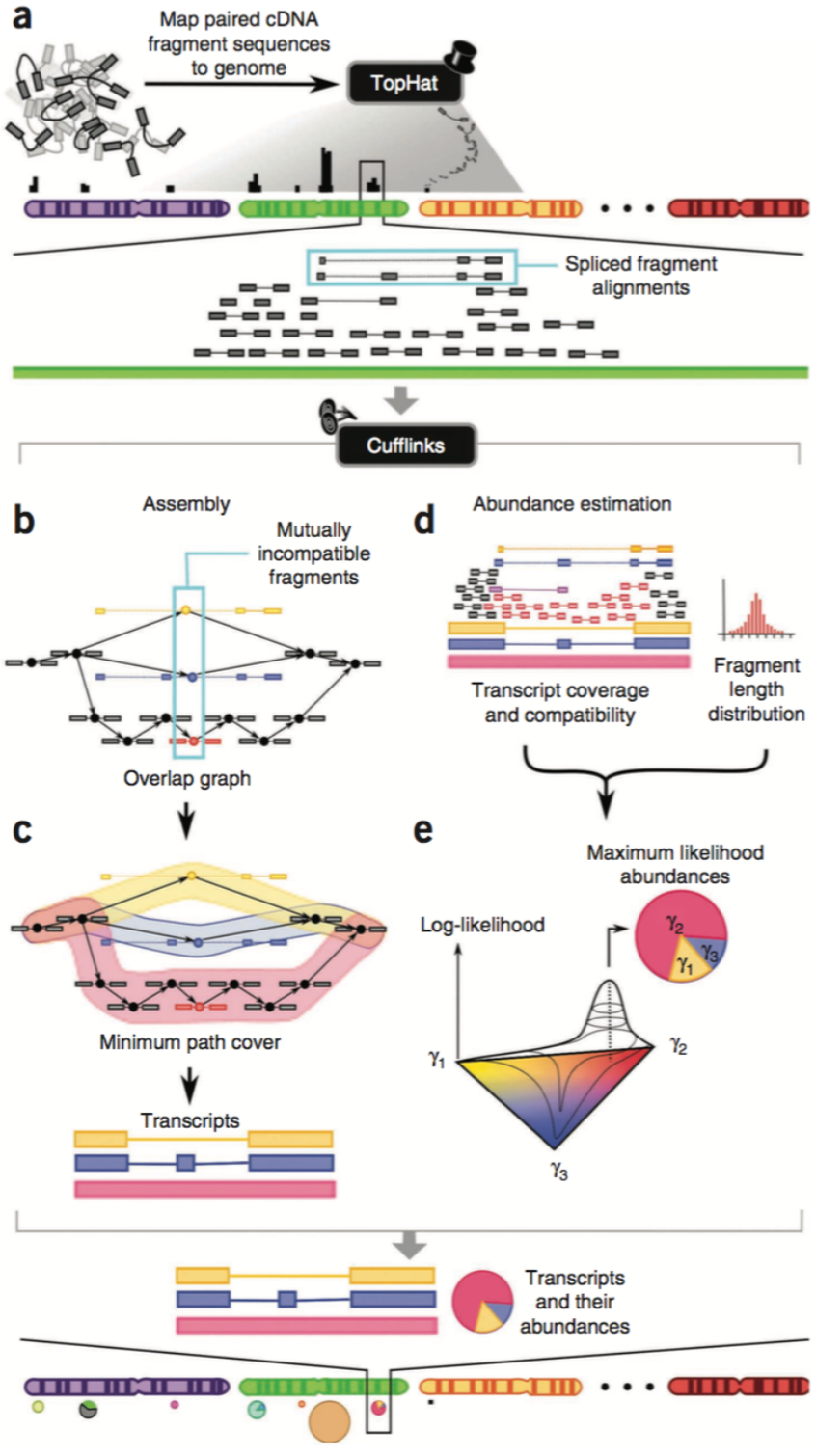
Extremely fast, general purpose short read aligner



Trapnell et al., Nature Biotechnology 28(5) p511(2010)
Trapnell et al., Nature Protocols 7, 562–578 (2012)

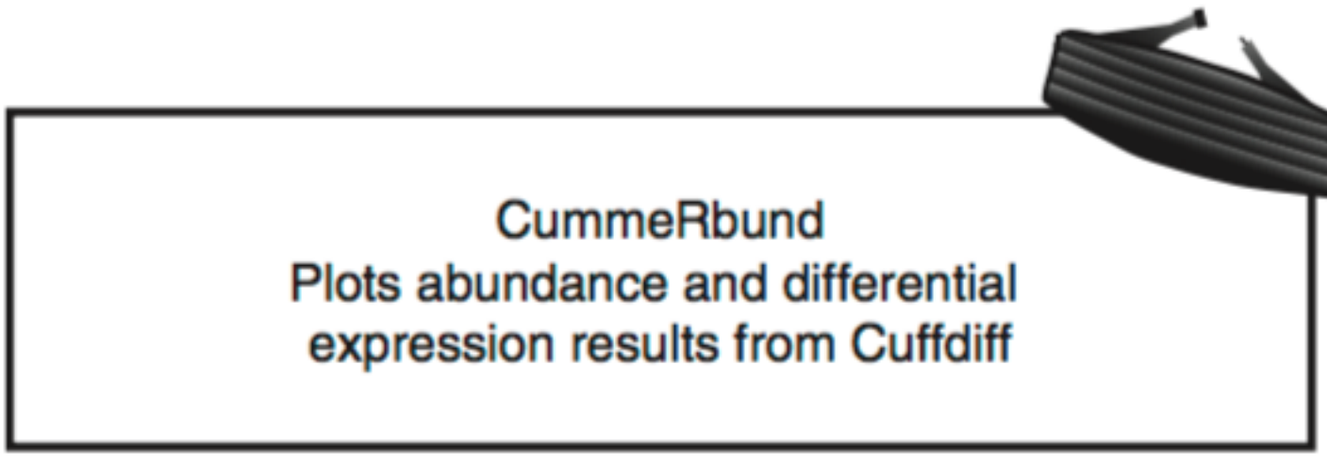
Tuxedo Suite

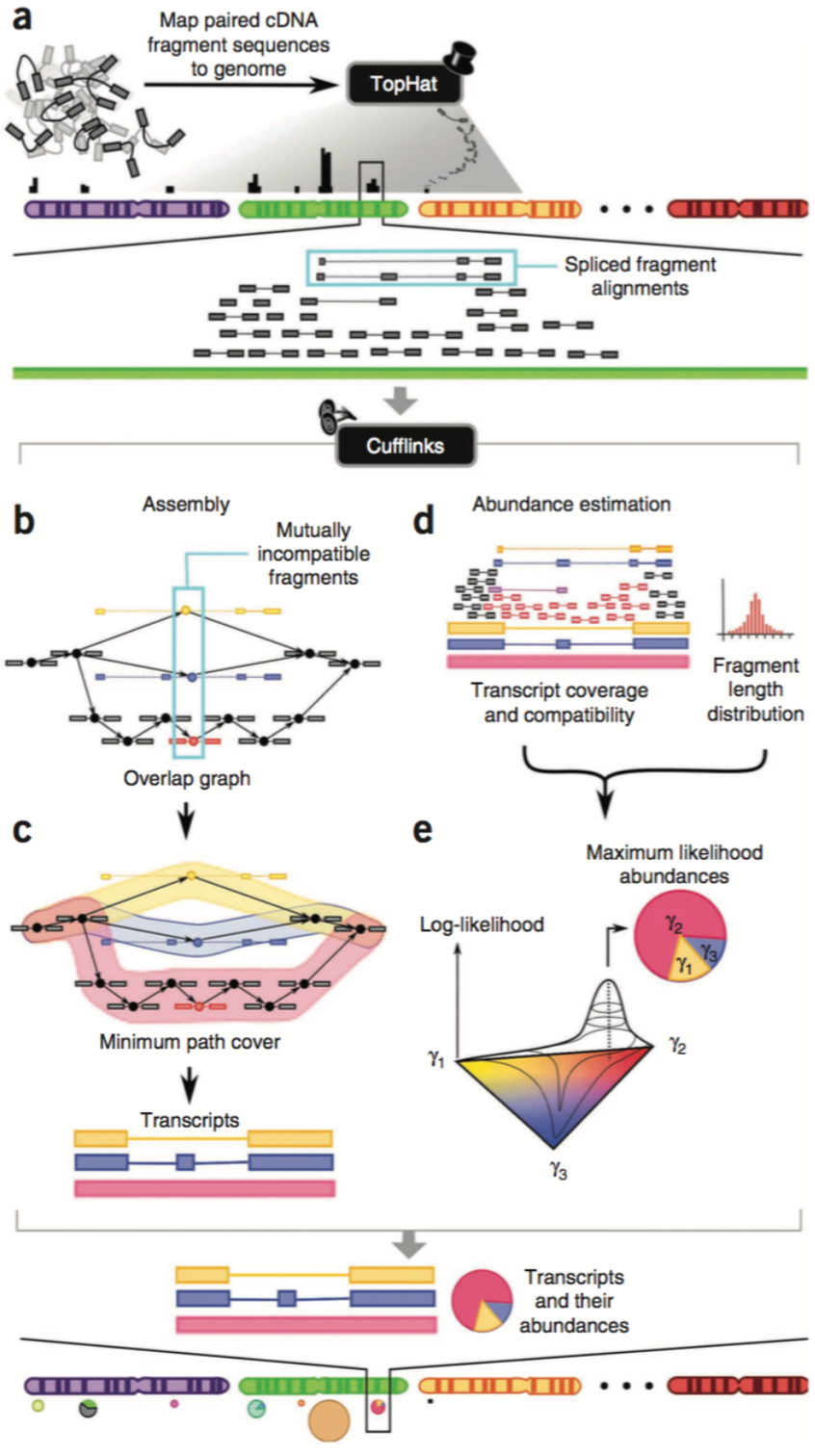




Trapnell et al., Nature Biotechnology 28(5) p511(2010)
Trapnell et al., Nature Protocols 7, 562–578 (2012)

Tuxedo Suite

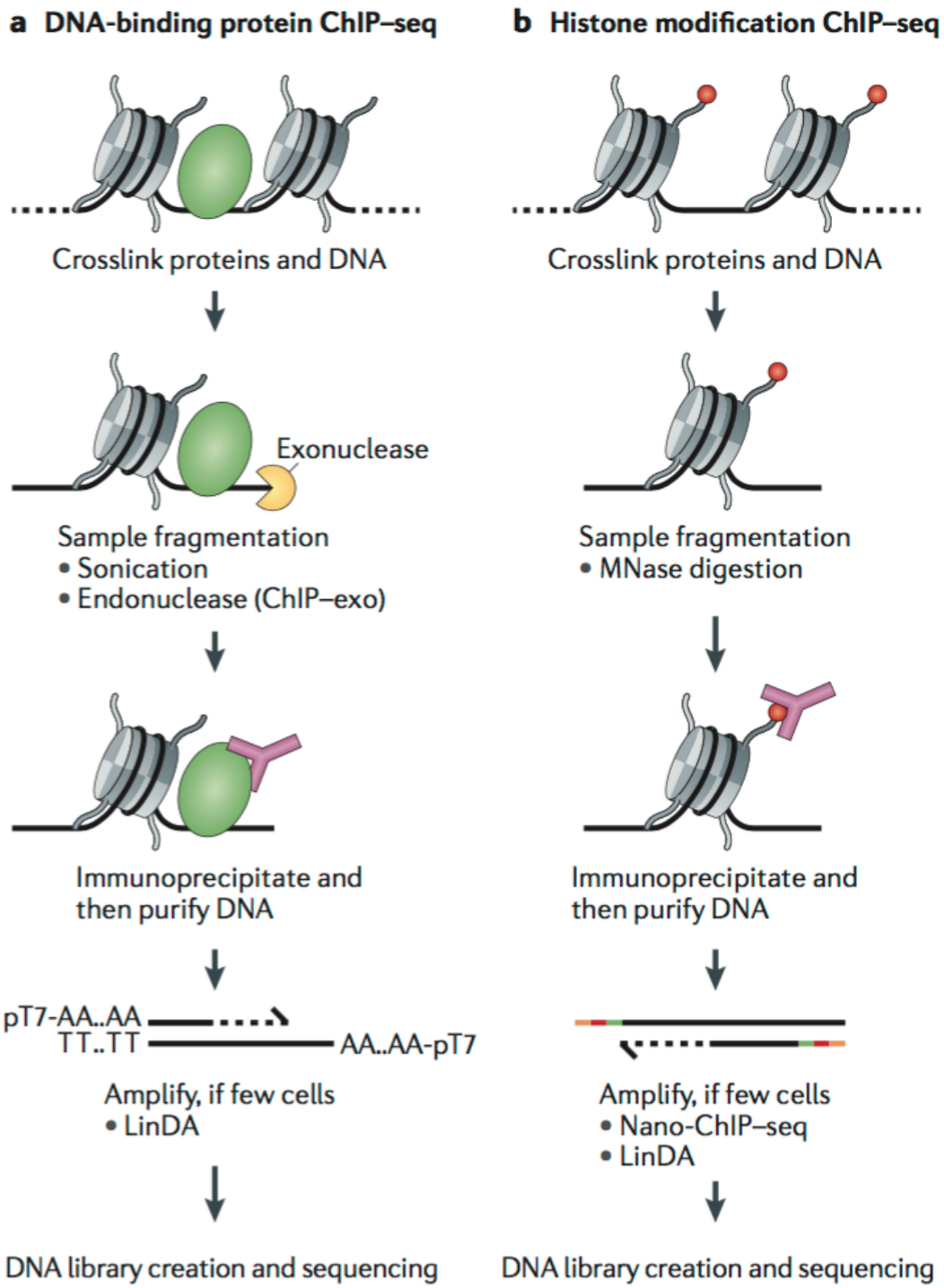




Trapnell et al., Nature Biotechnology 28(5) p511(2010)
Trapnell et al., Nature Protocols 7, 562–578 (2012)

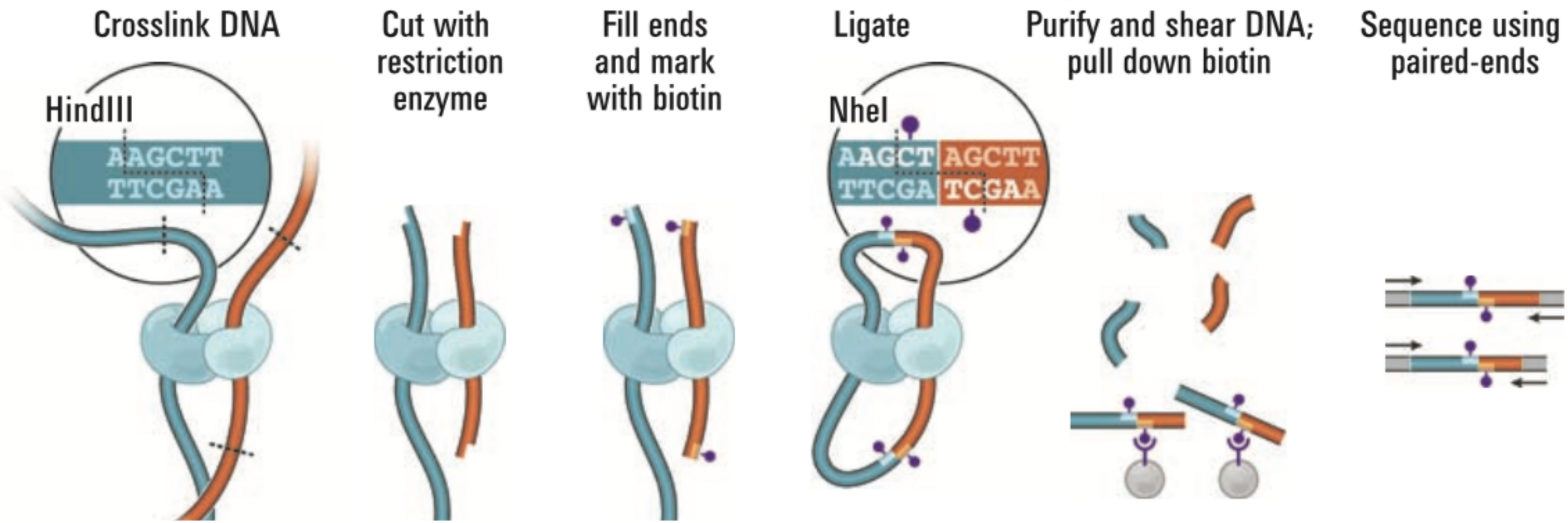
ChIP-seq

- Chromatin ImmunoPrecipitation followed by sequencing (ChIP–seq)
- Detect protein–DNA binding
- Detect chemical modifications of histone proteins



Furey, Nature Reviews Genetics 13, 840-852 (2012)

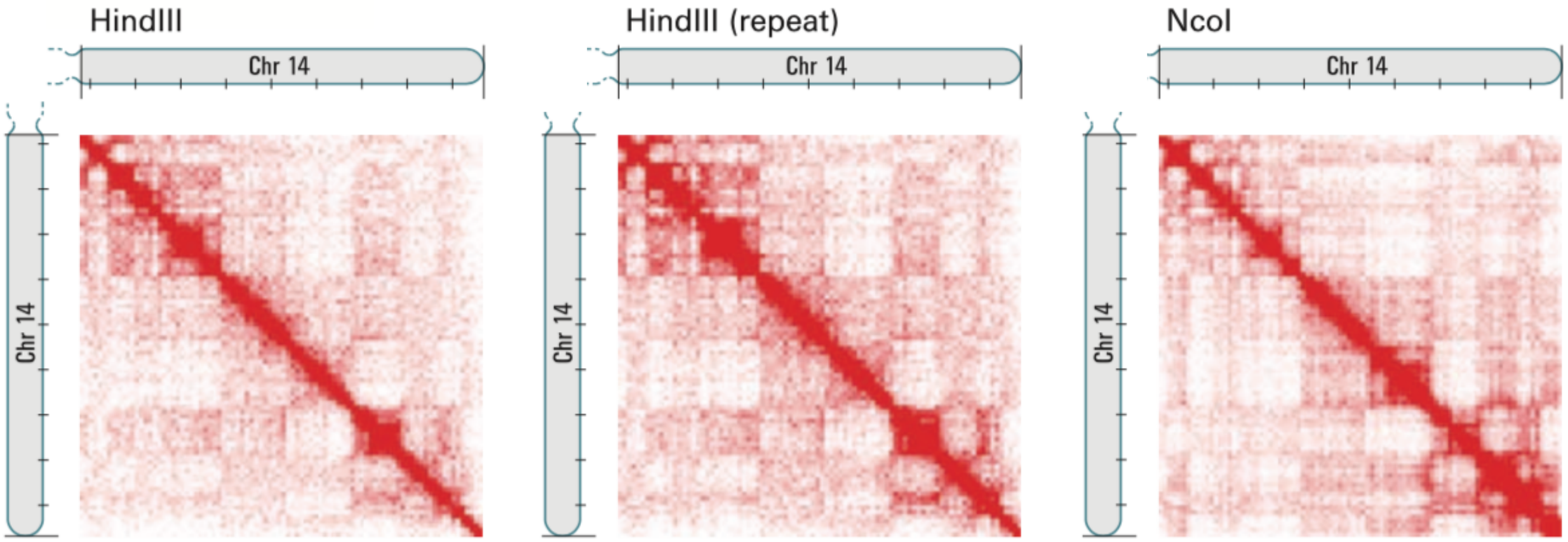
HI-C & DNA folding



- Cells cross-linked with formaldehyde; covalent links between spatially adjacent chromatin segments
- Chromatin digested with a restriction enzyme (here, HindIII)
- Resulting stickyends are filled in with nucleotides - one biotinylated (purple dot)
- DNA is purified and sheared.
- Biotinylated junctions are isolated with streptavidin beads and identified by paired-end sequencing.

Erez Lieberman-Aiden et al. Science 326, 289 (2009)

HI-C & DNA folding



- Hi-C produces genome-wide contact matrix
- Each pixel represents all interactions between 1-Mb loci
- Intensity as total number of reads

Erez Lieberman-Aiden et al. Science 326, 289 (2009)